BioSignature– Discoverer

User manual

Table of Contents

Introduction

The *BioSignature*– Discoverer plugin identifies biomarker signatures in biological data in a statistically robust, computationally efficient, and user-friendly way.

> *We consider a biomarker signature for an outcome of interest a minimal-size set of biomarkers whose values, when considered in combination, best determine (predict, diagnose) the most probable value of the outcome*

A typical example is the identification of a set of genes whose expression values discriminate between two different medical conditions, e.g., lupus erythematosus vs. healthy subjects. Upon such a set of genes is then possible to build a statistical, machine learning, or data-mining model that determines whether a new subject is affected by the disease.

BioSignature– Discoverer is designed to offer the following characteristics:

- **Automation**, requiring minimal input from the user and no data-analysis expertise
- **Quality of results**, employing state-of-the-art methods and analysis protocols that shield against methodological errors and are competitive against customized code by analysis experts
- **Efficiency of computations**, algorithmically optimizing the methods used
- **Understanding of output**, helping the user with the interpretation and visualization of results

BioSignature– Discoverer is able to find signatures within several types of biological data, such as (but not limited to):

- Transcriptomics data
- micro-RNA (miRNA) expression levels
- Methylation profiles
- Protein/Metabolite profiles

The plugin is able to find signatures and models for classification tasks with group-membership outcomes (e.g., diagnosing among four different cancer subtypes), regression tasks with continuous outcomes (e.g., predicting the level of expression of a specific gene), and time-to-event outcomes (e.g., time to death, disease relapse, occurrence of a complication, survival analysis).

These functionalities allow our plugin to solve problems related to extremely different research areas. Three case studies are introduced in order to illustrate the versatility of the plugin; each case study successfully analyzes a publicly available set of Next Generation Sequencing (NGS) or microarray data:

1. Identification of miRNA biomarkers for the early diagnosis of Alzheimer
2. Analysis of potato (solanum tuberosum) metabolic profiles for identifying early biomarkers of black spot bruising susceptibility.
3. Identification of a gene expression signature for estimating the survival probability of mantle cell lymphoma patients

What's in a signature

In principle, any subset of the input quantities could be an optimal signature. When the number of input quantities ranges above the hundreds, the number of probable signatures to consider becomes astronomical. *BioSignature* – Discoverer employs state-of-the-art machine learning and statistical methods to solve the problem both efficiently and with high-quality results. The signatures output by the tool have the following characteristics:

- **Minimality:** Smaller signatures are easier to interpret biologically, verify experimentally, and less costly to measure. While certain quantities may carry information regarding the output when examined in isolation, they may be superfluous given the selected signatures. The tool tries to identify and remove such quantities from the output. Thus, a gene expression that is correlated with low p-value with an outcome may actually not be part of a signature.
- **Collective Optimality:** The tool attempts to identify the set of quantities that can optimally determine the most likely outcome through a statistical model collectively as a group. Thus, a gene expression that is not correlated (high p-value) with the outcome when considered in isolation may actually become predictive given the other selected quantities and included in a signature.
- **Multiplicity of Signatures:** The tool attempts to identify as many signatures as possible that are statistically indistinguishable in terms of predictive capabilities. Any such signature could be employed to best determine the outcome value; the user can thus choose the signature that is more feasible or cost-effective to measure. Furthermore, contrasting the signatures against each other can provide additional biological insights into the biological mechanism generating the data.
- **Non-Monotonicity:** given more samples / measurements for training, the tool may include more or fewer quantities in a given signature. It may include additional quantities if the extra samples allow it to establish statistically significantly that they carry non-superfluous predictive information. It may decide to remove quantities if the extra samples allow it to determine statistically significantly that they are actually superfluous given the rest of the signature quantities.

What's in a model

In order to determine how well a given signature predicts, discriminates, or classifies the outcome *BioSignature* – Discoverer tries several standard and state-of-the-art machine learning, data mining, and statistical algorithms. This takes place transparently to the user. Models are also employed to explain the multi-variate correlations between the signature quantities and the outcome and to produce visualizations and explanations of the results.

Plugin installation

The BioSignature – Discoverer plugin can be installed as any other CLC bio plugin. In the CLC bio Workbench, click the "Help" tab, "Plugins and resources…", and then click on "Install from File". Select the CPA file that fits your version of CLC bio Workbench and press "Install".

The plugin is available for the CLC Genomics, Biomedical Genomics and Main Workbenches.

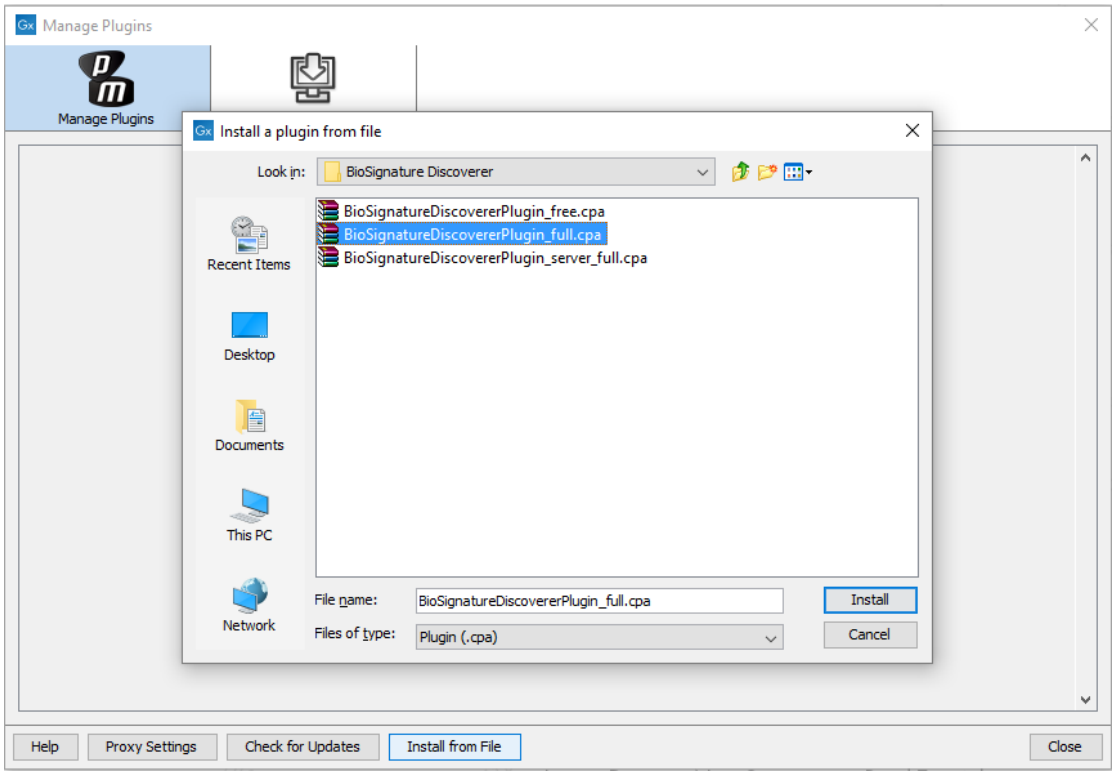

Figure 1: installing the BioSignature Discoverer plugin

BioSignature– Discoverer functionalities

The functionalities of the plugin are straightforward to use. Similarly to other CLC bio plugin, the user is required to specify the data to analyze and to configure the analysis to run. Once the computations are completed the results are presented to the user within detailed reports.

Analysis specification

1. Select data

When you first invoke *BioSignature*– Discoverer you are requested to specify the training samples and their outcome. There are two ways to specify the training samples, either as a list of individual samples or as an Experiment object.

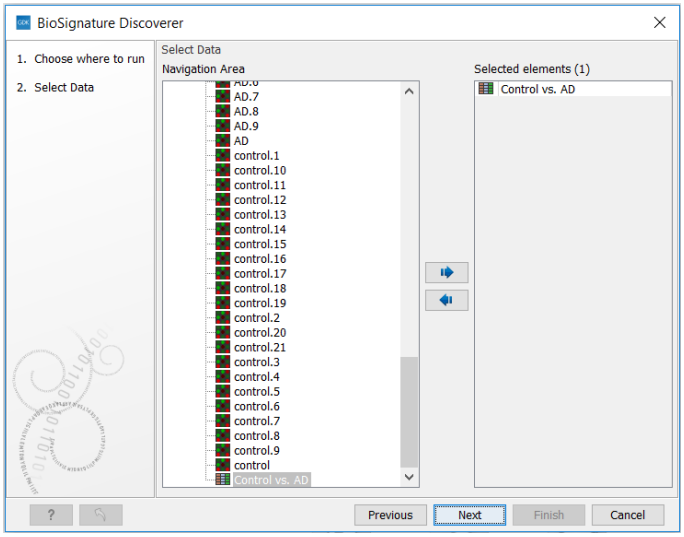

Figure 2: selecting an Experiment object as input for the BioSignature Discoverer plugin

> You can create an Experiment object with the standard CLC bio Workbench toolbox for "Expression Analysis" and the "Set up Experiment" option. In step 2 of the process, when you define the experiment type, choose "Unpaired"; the current version of the plugin is not designed for the analysis of paired samples. During the set-up of the experiment, samples can be assigned to 2 or more groups. In addition, using the toolbox "Transformation and Normalization" you can preprocess the samples in the Experiment with various transformation and normalization methods. The normalized and/or transformed values of the samples become associated with the Experiment object. See the relevant CLC bio tutorial for further information on how to create an Experiment.

Data can also be input as a list of samples you would like to include in the analysis. Notice that you cannot specify both an Experiment object and a list of samples at the same time. If an Experiment is already selected for analysis, then samples cannot be added to the selection and vice versa. The advantage of grouping your samples in an Experiment object is that *BioSignature*– Discoverer can make use of the group assignments and the preprocessing you have applied.

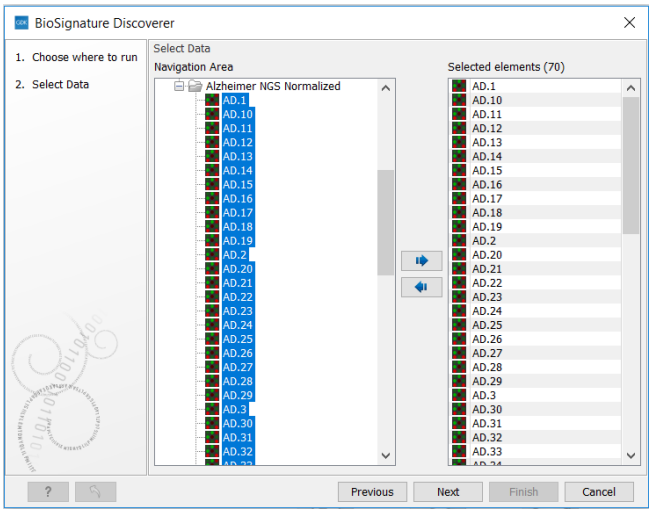

Figure 3: selecting a set of samples as input for the BioSignature Discoverer plugin

2. Specify Analysis Type and Outcome

There are three ways to specify the outcome in the data:

1. Use the already defined Experiment groups. *If you have selected an Experiment object to analyze, this step is omitted* and the analysis type is assumed to be Classification for the groups specified in the Experiment.
2. Use an existing feature (quantity, variable) that is measured in your samples. You can select this variable from the drop-down menu labeled “From the input features:”. Notice that if you select Classification as your type of analysis each different value of the feature will be considered as a different class / group.
3. Use a file to assign outcome values to your samples. The file must be in Comma Separated Values (.csv) format. Each row should contain a sample name and its outcome. In case of Survival Analysis there are two outcomes: the time-to-event (if known) and the status (censored or not) (see Section “Identification of a gene expression signature for estimating the survival probability of mantle cell lymphoma patients” below). Notice that this is the only available option for Survival Analysis.

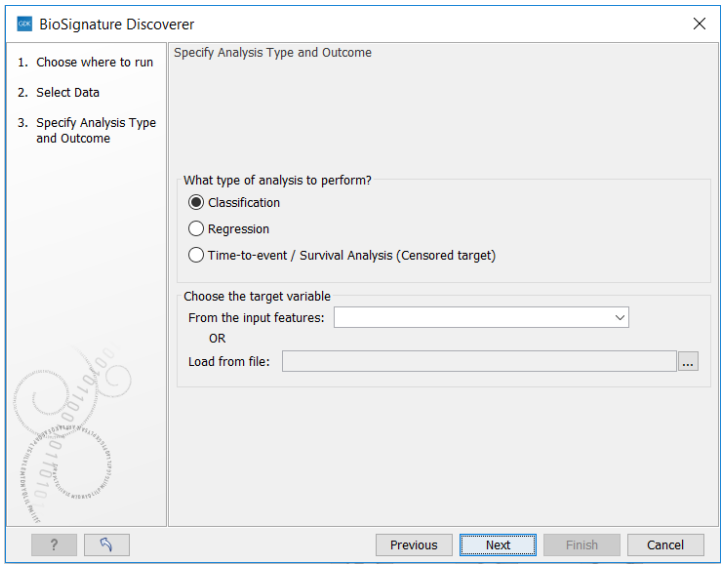

Figure 4: selecting the appropriate type of analysis and outcome

3. Specify Analysis Options

In this form you specify options that guide the analysis.

Which expression values to analyze? When an Experiment has been selected for the analysis, you have the option to analyze either the original values, or the values transformed or normalized with the "Transformation and Normalization" toolbox. Otherwise, these options are not selectable. Independently of the choice made at this step, the plugin internally scales the data in order to have zero mean and unit variance.

Choose the level of tuning effort for your analysis. The statistical and machine learning algorithms employed by the plugin require tuning the values of several options, called hyper-parameters, just as a TV receiver needs to be tuned to show a clear picture. Tuning the algorithms typically requires searching for the best hyper-parameter combination. Optimizing the analysis may return better performing models and different signatures, but of course requires more computation time. The plugin automatically searches for the best configuration of hyper-parameters in a transparent way to the user. The user is only required to specify how extensive the search should be. The plugin offers three possible choices: Quick, Normal and Extensive – which correspond to increasing levels of optimization.

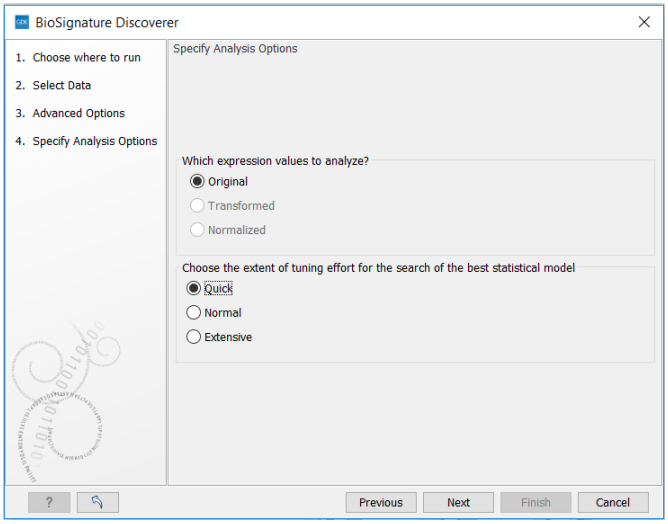

Figure 5: windows for specifying the analysis options

For a typical data analysis task (10 to 100 samples, 10,000 to 100,000 expression levels), a quick search should run for few minutes, while an extensive one may take hours. A good strategy in order to choose the most appropriate level of tuning is to perform a quick or normal search first, and then to estimate the time for a more thorough analysis with the help of the coefficients shown in Table 1. Note that those coefficients are only approximate and vary for different datasets.

Table 1: required computational time with respect to the Quick search. The left column reports the available tuning effort options, while the right column reports the required computational time. Times are scaled with respect to the Quick search; for example, if the Quick search runs for a minute, the user should expect the Normal search to run for 2-10 minutes and the Extensive one for 10-50 minutes.

Level of tuning effort	Computational time (Quick search = 1)
Quick Search	1
Normal Search	2 - 10
Extensive Search	10 - 50

4. Specify advanced options

The next window allows the user to specify a set of advanced options. The default values of these options are tailored to suit most of the usual analyses on omics data, so that users can safely forgo adjusting these parameters. More complicate or demanding data-analysis tasks may require a finer tuning.

Choose the percentage of CPU utilization. The plugin is able to exploit the full power of modern CPUs by parallelizing computations over multiple cores. This slider allows the user to specify the percentage of the total number of cores to use. By default, the plugin will use only one core (e.g., 12.5% of CPU power on a machine with 8 cores, as shown in Figure 6). The higher the percentage, the faster the computation; however, specifying 100% CPU power may significantly slow down other applications running on the same machine.

Choose the number of repetitions. The statistical analysis performed by the plugin involves splitting the data in a number of folders for better estimating the predictive performances of the identified signatures. For low sample sizes, the estimation of the performances can slightly depend on the random split used. In these cases the analyses can be repeated several times, each time using a different random split. The final results are then combined across repetitions and are not anymore dependent on a single partition of the data.

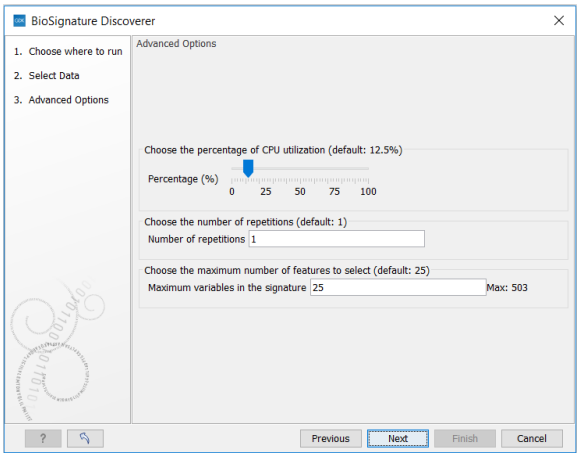

Figure 6: windows for specifying advanced options

Choose the maximum number of features to select. Signatures containing an excessive number of features can be hard to interpret or impractical to measure. This options allows the user to specify how many predictors can be included *at most* in each signature.

5. Result handling

In this form you specify whether you prefer the output open in a new tab in the main CLC Workbench window or saved in a new file.

This is it! Click Finish and find the molecular signatures.

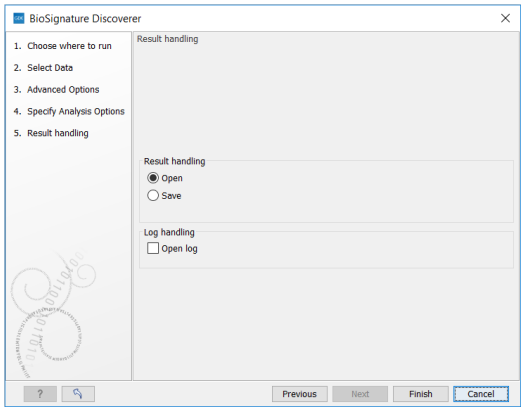

Figure 7: result handling options

Result reports

The results of the *BioSignature*– Discoverer computations are provided to the user in two different reports, the *Summary Report* and the *Detailed Report*. The first one contains the main findings of the analysis, while the latter shows detailed information about the retrieved signatures and their predictive performances.

Summary Report

The Summary Report is composed by three different types of information: (a) a description of the identified signatures, (b) performance estimation metrics, and (c) diagnostic plots.

1 Reference Signature

Feature 1 (Stability = 100,0%)	Feature 2 (Stability = 100,0%)	Feature 3 (Stability = 80,0%)	Feature 4 (Stability = 100,0%)
hsa-mir-148b:hsa-miR-148b-5p	hsa-mir-151a:hsa-miR-151a-3p	brain-mir-112:brain-mir-112	hsa-mir-98:hsa-miR-98

2 Lists of Equivalent Features

Feature 1	Feature 2	Feature 3	Feature 4
hsa-mir-148b:hsa-miR-148b-5p	hsa-mir-151a:hsa-miR-151a-3p	brain-mir-112:brain-mir-112	hsa-mir-98:hsa-miR-98

There is only 1 signature

3 Effect sizes

	hsa-mir-148b:hsa-miR-148b-5p (std = 0,758)	hsa-mir-151a:hsa-miR-151a-3p (std = 105,423)	brain-mir-112:brain-mir-112 (std = 0,439)	hsa-mir-98:hsa-miR-98 (std = 5,595)
Class Control vs Class AD	0,85	1,91	1,06	-0,67

Effect sizes are reported as log (base 10) odds ratio change per one standard deviation increase of the feature.

Figure 8: description of the retrieved signatures for an example classification analysis. From top to bottom the reference signature, the list of equivalent features and the effect sizes are reported. This specific example led to the discovery of only one signature.

Description of the identified signatures: the first information provided to the user is the *Reference Signature*, which represents the first molecular signature found by the algorithm. The *Stability* value reported for each quantity indicates the probability of selecting the same feature if the analyses were repeated on an independent set of samples.

The set of *Equivalent Signatures* is then reported. Each signature comprises of a quantity in the column named “Feature 1”, combined with one, and only one, quantity in column “Feature 2”, and so on. Any of such possible combinations is an equivalent signatures, and their total number is shown below the table. All signatures are statistically indistinguishable in terms of predictive capabilities, and the users can choose the one that best fits their needs (for example, the signature whose biomarkers are easier / more affordable to measure).

2 Lists of Equivalent Features

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
			Docosane-n-A220001		Analyte-A291005				
			Ethanolamine-A128002		Analyte-A293001				
			Pentadecane-n-A150001		Analyte-A294002				
			Weather						

There are 1 × 1 × 1 × 5 × 1 × 4 × 1 × 1 × 1 × 1 = 20 equivalent signatures.

Figure 9: list of equivalent signature from the analysis reported in Section "Reporting Multi-Class Classification Results". Note the presence of multiple equivalent predictors for Feature 4 and Feature 6.

Figure 9 shows the multiple signatures identified on an agriculture-related dataset (see page 21). Note that Feature 4 and Feature 6 list five and four equivalent biomarkers, respectively. The following is one of the signatures obtained by choosing one biomarker per Feature:

Feature1	**Feature2**	**Feature3**	**Feature4**	**Feature5**	**Feature6**	**Feature7**	**Feature8**	**Feature9**	**Feature10**
Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001

Which is equivalent to the signature obtained by replacing Feature4 (Analyte-A104001) with its equivalent biomarker Docosane-n-A220001:

Feature1	**Feature2**	**Feature3**	**Feature4**	**Feature5**	**Feature6**	**Feature7**	**Feature8**	**Feature9**	**Feature10**
Cultivar	Soil	Methionine-A142007	Docosane-n-A220001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001

Note again that exactly *one* biomarker per feature must be selected.

Finally, the *Effect size* of each element in the reference signature is provided. The effect size is a measure of the predictive strength of each element: the higher the effect size (in absolute value), the larger the expected variation in the outcome for a change in the value of the signature element. The way effect sizes are reported varies depending by type of outcome: $\log_{10}$ odds ratios for classification, linear regression standardized coefficients for regression, and log hazard ratios for survival analysis.

4 Performance Metrics

Metric	Average	95% Confidence Interval
Area Under the ROC Curve	0,946	[0,857, 1,000]
Accuracy	0,885	[0,783, 0,966]
Balanced Accuracy	0,859	[0,705, 0,976]
Average F1	0,860	[0,713, 0,961]
Precision for class AD	0,907	[0,786, 1,000]
Precision for class Control	0,841	[0,600, 1,000]
Recall for class AD	0,929	[0,800, 1,000]
Recall for class Control	0,789	[0,500, 1,000]
Sensitivity for class AD	0,929	[0,800, 1,000]
Sensitivity for class Control	0,789	[0,500, 1,000]
Specificity for class AD	0,789	[0,500, 1,000]
Specificity for class Control	0,929	[0,800, 1,000]

Figure 10: performance metrics for an example classification problem. For each metric the average (expected) value and the 95% confidence interval are presented.

Performance Estimation Metrics: these metrics provide a measure of the expected predictive performances of the selected signature(s) on an independent test set. The reported metrics vary depending on the type of outcome: for classification problems these are the Area Under the Curve (AUC), Accuracy, Balanced Accuracy, Average F1 score, along with Precision, Recall, Sensitivity and Specificity for each class. For regression tasks the out-of-sample R^2, the mean absolute and squared error, the relative absolute and squared error, and the correlation coefficient are displayed instead. The Concordance Index (CI) is reported for survival analysis. The 95% confidence interval for each metric is provided as well, as estimate through bootstrapping.

Furthermore, the contribution of each feature to the performance of the whole signature is assessed. Particularly, the impact of each feature is provided in terms of *individual* and *cumulative* contribution. The former is computed as the loss in performance when each element of the signature is removed in turn (see Figure 11).

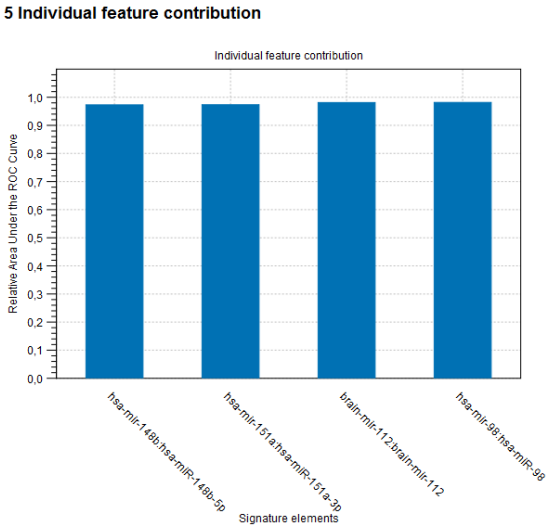

Figure 11: example of individual contribution graph. The graph reports the percentage of performance metric (in this case Area under the ROC Curve) achieved by the reference signature when each element is removed in turn. The most important feature causes the largest reduction in performance when removed.

The cumulative contribution shows the increase in predictive performance when the signature elements are added one after the other (following the order given by their individual contribution). Both the individual and cumulative contributions are reported as bar-graphs (Figure 11 and Figure 12, respectively).

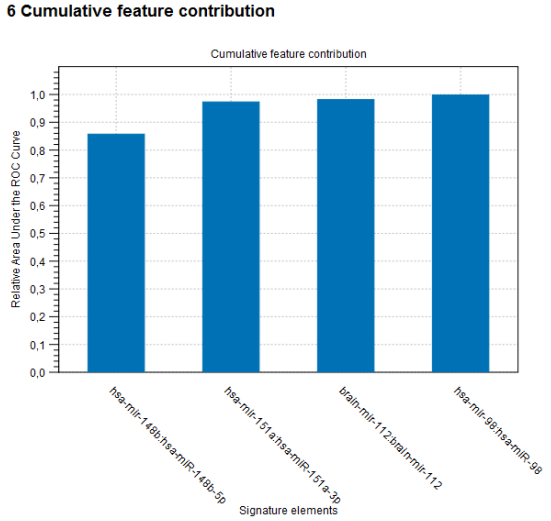

Figure 12: example of cumulative contribution graph. The graph reports the percentage of performance metric (in this case Area under the ROC Curve) achieved by adding to the reference signature each element in the order show by the X-axis (left to right).

Diagnostic plots: a set of diagnostic plots are provided in order to allow the user to identify possible anomalies in the data, for example outliers, unexpected trends and so on. The diagnostic plots to be shown depend by the problem at hand: for classification task, a Principal Component Analysis (PCA) plot of the data *using only the quantities in the Reference signature* is displayed along with the out-of-sample predicted probabilities of belonging to each class.

For regression task the diagnostic plots contrast the predicted values versus the residual and real values. For survival analyses the Deviance residual plot is reported instead. Such plots can reveal outlier samples that may be erroneously labeled, or hidden patters in the residuals that indicate bad fitting.

8 Principal Component Analysis (PCA)

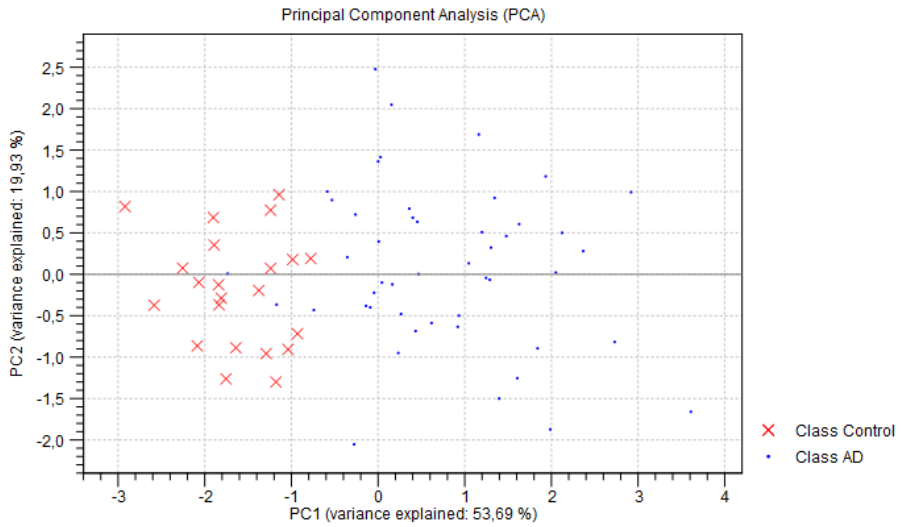

Figure 13: example of PCA diagnostic plot for a classification problem. The two axes correspond to the first two principal components (in order of explained variance) of the reference signature data.

Detailed Report

Two types of information are given in the Detailed Report: an extended list of equivalent signatures and the full list of out-of-sample predictions.

Extended list of equivalent signatures: for some specific problems the number of equivalent signatures can be quite high. For sake of clarity, the Summary Report shows only up to twenty equivalent signatures, while the remaining ones along with their respective effect sizes are reported in the Detailed Report.

Out-of-sample predictions: the predictions leading to the estimation of the signature(s) performances are reported in the Detailed Report, for the user's perusal.

3 Real values vs Predictions

Sample ID	Real Class	P(class = AD)	P(class = Control)
AD	AD	0,826	0,174
AD.1	AD	0,902	0,098
AD.10	AD	0,976	0,024
AD.11	AD	0,268	0,732
AD.12	AD	0,859	0,141
AD.13	AD	0,939	0,061
AD.14	AD	0,953	0,047
AD.15**	AD	0,531	0,469
AD.16	AD	0,989	0,011
AD.17	AD	1,000	~0
AD.18	AD	0,977	0,023
AD.19	AD	0,793	0,207
AD.2	AD	0,999	0,001
AD.20	AD	0,996	0,004
AD.21	AD	0,169	0,831
AD.22	AD	0,999	~0
AD.23	AD	0,995	0,005
AD.24	AD	0,990	0,010
AD.25	AD	0,952	0,048
AD.26	AD	0,974	0,026
AD.27	AD	0,965	0,035

Figure 14: example of out-of-sample predictions for a classification problem. For each sample the actual and predicted class are reported. Predictions are provided in terms of the probability of belonging to each class. The table shown in this example has been trimmed for representation purposes

Functionalities across plugin versions

The *BioSignature* – Discoverer plugin is released in two different versions. The "Full" version provides the whole set of *BioSignature* – Discoverer's functionalities. The "Free" version comes free of charge, and it allows analyzing binary classification problems, a common task in case-control studies. Table 2 details the functionalities available in each version. The server edition of the plugin is only available in the Full version.

Table 2: plugin functionalities across different versions

Functionality	Free	Full
Binary Classification	✓	✓
Extended hyper parameters optimization	✓	✓
Multiple Signatures	✓	✓
Multi-class Classification		✓
Regression		✓
Time-to-Event Analysis		✓

Case Studies

Identification of miRNA biomarkers for the early diagnosis of Alzheimer

In this case study we further elaborate the example partially shown in the previous sections. This study is a prototypical example of binary classification, where the aim is to find *NGS miRNA expression signatures for the early diagnosis of Alzheimer.* In this case the outcome is dichotomous (Alzheimer cases vs. healthy controls) and each sample belongs to one of the two groups. The identified signatures are the miRNA sets that provide the most accurate early diagnosis for Alzheimer when considered jointly.

Several studies have shown that non-coding micro RNAs can act as early diagnostic biomarkers for a number of diseases. The study reported in [1] identifies a 12 miRNA signature in blood samples able to nearly perfectly discriminate between Alzheimer and healthy subjects. The data for this study are publicly available on the Gene Expression Omnibus (GEO) website. The preprocessed CSV file ready to be imported in the CLC bio workbench can be downloaded directly from this link. The original count matrix was transformed as counts per kilo-base per million [2]; the original data (in excel format) can be also downloaded from this link.

Once you have saved/downloaded the GSE46579_NGS_miRNA_normalized.csv file, you can import it in the CLC bio Workbench with the Standard Import (Ctrl + i) utility. After selecting the CSV file, be sure of using the “Automatic import”.

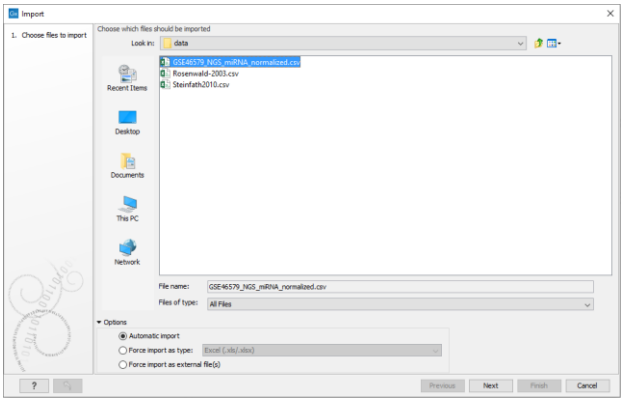

Figure 15: importing the miRNA data

When prompted for selecting the location where to save the files, create a new folder “NGS_miRNA”. Press “Finish” and wait for the file to be loaded.

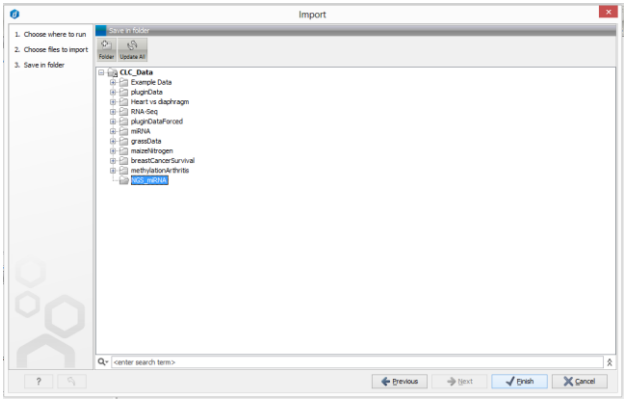

Figure 16: selecting the destination folder for the NGS miRNA data

Once the data have been imported, the NGS_miRNA folder will contain seventy expression profiles, whose names start either with “AD” (acronym for Alzheimer Disease) or “control” (healthy subject). We will now create an ‘Experiment” object that will contain and compactly represent these expression profiles. From the “Toolbox” panel, select “Expression Analysis” → “Set up experiment”. In the following dialog window select all the miRNA expression profiles, and click on “Next”.

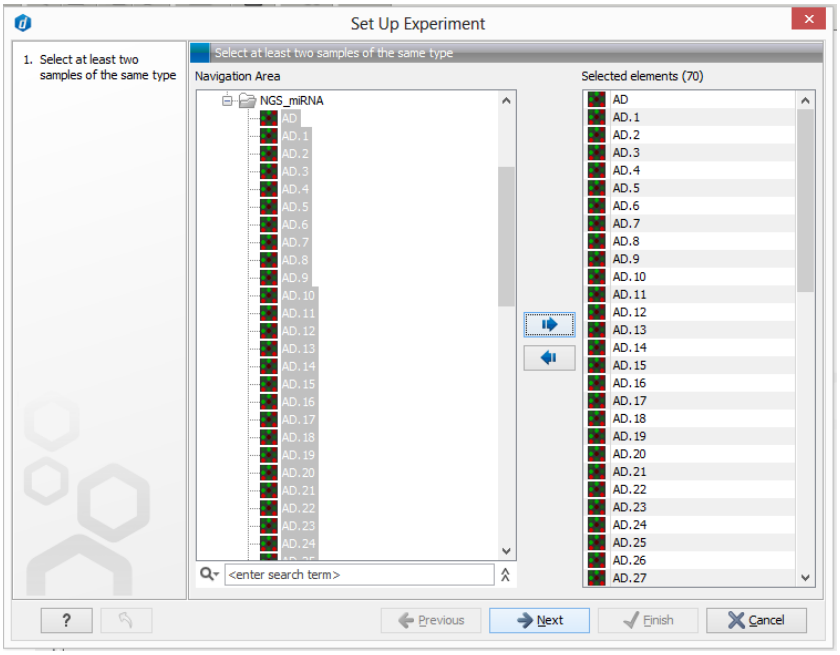

Figure 17: setting up the Control vs. AD experiment

In the next dialog window select “Two-group comparison”, “Unpaired” and proceed to the next window. Name the groups as in Figure 18 (Group 1: Control. Group 2: AD). Proceed to the next window, where you should assign each profile to its respective group. Finally, save the experiment in the NGS_miRNA folder.

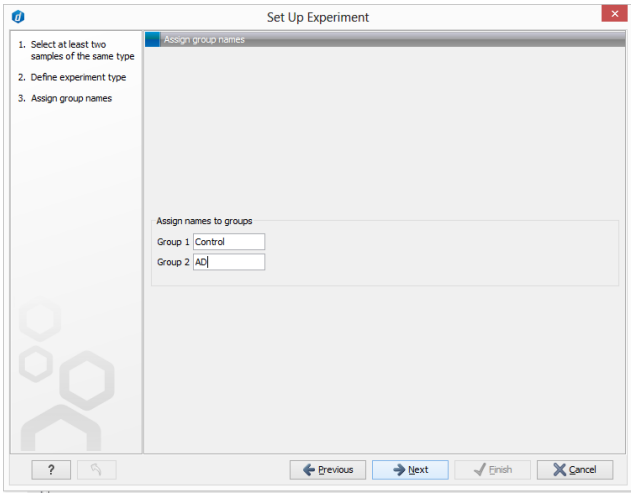

Figure 18: naming the groups for the Control vs. AD experiment

We are now ready for analyzing the “Control vs. AD” experiment with the BioSignature – Discoverer plugin. Start the plugin and select as input the “Control vs. AD” experiment.

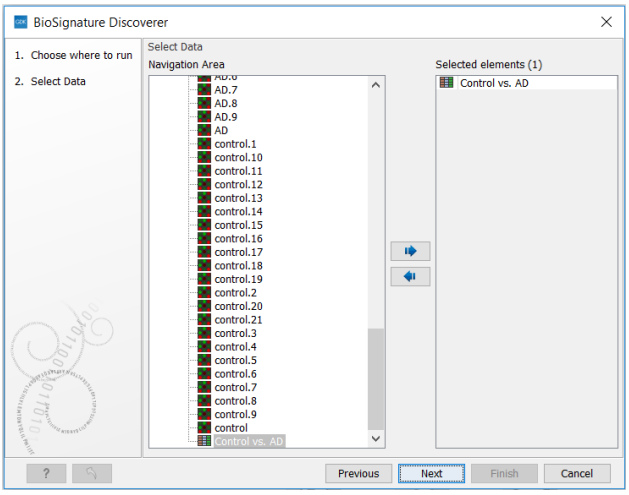

Figure 19: selecting the input for the BioSignature – Discoverer plugin

For the present case study let us set the plugin options as in Figure 20: original values and "Quick" as level of tuning. We leave the advanced options to their default values and we select "Open" in the Result handling options window. Let's click "Finish" for starting the plugin.

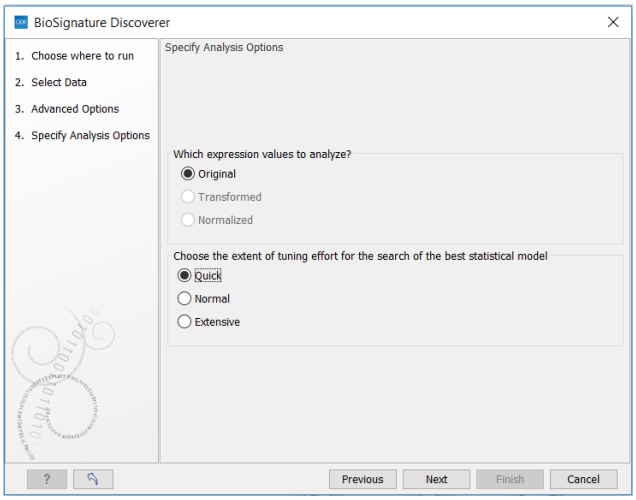

Figure 20: BioSignature – Discoverer plugin options

Reporting Binary Classification Results

At the end of the computations a new Summary Report will be generated, containing several pieces of useful information. The first piece is shown at the top of the Report and is the "Reference Signature":

1 Reference Signature

Feature 1 (Stability = 100,0%)	Feature 2 (Stability = 100,0%)	Feature 3 (Stability = 80,0%)	Feature 4 (Stability = 100,0%)
hsa-mir-148b:hsa-miR-148b-5p	hsa-mir-151a:hsa-miR-151a-3p	brain-mir-112:brain-mir-112	hsa-mir-98:hsa-miR-98

Figure 21: Reference Signature for the Control vs. AD experiment

The signature comprises of four different miRNA expression levels: 1) hsa-miR-148b-5p, 2) hsa-miR-151a-3p, 3) brain-mir-112, and 4) hsa-miR-98.

According to their stability levels, the two first and the last components should be certainly retrieved if the same study were to be performed on an independent sample, while the third component has less chances to do so (80% probability).

This is the only signature that has been identified, as reported in the subsequent "2 Lists of Equivalent Features" table.

2 Lists of Equivalent Features

Feature 1	Feature 2	Feature 3	Feature 4
hsa-mir-148b:hsa-miR-148b-5p	hsa-mir-151a:hsa-miR-151a-3p	brain-mir-112:brain-mir-112	hsa-mir-98:hsa-miR-98

There is only 1 signature

Figure 22: list of equivalent features. Only one signature was identified in this study

The next table of the Report gives an indication about the strength of the relationship between each element of the signature and the outcome.

3 Effect sizes

	hsa-mir-148b:hsa-miR-148b-5p (std = 0,758)	hsa-mir-151a:hsa-miR-151a-3p (std = 105,423)	brain-mir-112:brain-mir-112 (std = 0,439)	hsa-mir-98:hsa-miR-98 (std = 5,595)
Class Control vs Class AD	0,85	1,91	1,06	-0,67

Effect sizes are reported as log (base 10) odds ratio change per one standard deviation increase of the feature.

Figure 23: effect size of each element of the signature expressed as AD vs. Control log_{10} odds variation

In order to correctly interpret the percentages reported in table "3 Effect sizes", we must consider that:

1. the effect sizes are quantified through a logistic regression model. Logistic regression models redefine the outcome in terms of "log_{10} odds", i.e., the base-10 logarithm of the ratio between the probability of belonging to the first class ("Control") over the probability of belonging to the second class ("AD").
2. expression values have been standardized in order to have zero mean and unitary variance before fitting the logistic model.

Given these premises, the coefficients can be interpreted as follow: for the hsa-miR-148b-5p biomarker, an increment equal to its standard variation (std = 0.758) implies that a diagnosis of Alzheimer (i.e., belonging to the class AD) is $10^{0.85} \approx 7.1$ times more probable. On the other hand, an increment of 5.595 in the expression value of the miRNA hsa-miR-98 makes a diagnosis of Alzheimer approximately 5 times less probable ($10^{-0.67} \approx 1/5$).

The successive table of the Report, "3 Performance Metrics", provides the estimated predictive performances, along with their 95% Confidence Interval estimated through a boot-strapping approach.

4 Performance Metrics

Metric	Average	95% Confidence Interval
Area Under the ROC Curve	0,946	[0,857, 1,000]
Accuracy	0,885	[0,783, 0,966]
Balanced Accuracy	0,859	[0,705, 0,976]
Average F1	0,860	[0,713, 0,961]
Precision for class AD	0,907	[0,786, 1,000]
Precision for class Control	0,841	[0,600, 1,000]
Recall for class AD	0,929	[0,800, 1,000]
Recall for class Control	0,789	[0,500, 1,000]
Sensitivity for class AD	0,929	[0,800, 1,000]
Sensitivity for class Control	0,789	[0,500, 1,000]
Specificity for class AD	0,789	[0,500, 1,000]
Specificity for class Control	0,929	[0,800, 1,000]

Figure 24: performance metrics

The metrics reported in this table vary depending by the nature of the considered outcome. For a dichotomous outcome (AD class vs. Control class) the employed metrics are:

1. Area under the ROC Curve (AUC): it is a measure of the capability of the signature of correctly classifying the samples. A perfect classification would lead to an AUC equal to 1, while a random classification would produce an AUC equal to 0.5.
2. Accuracy: the fraction of correctly classified instances
3. Sensitivity for class AD: it is the fraction of correctly classified AD samples over the total number of AD samples. In other words, the probability that a sample belonging to the class AD is correctly classified as AD.
4. Specificity for class AD: it is the fraction of correctly classified Control samples over the total number of Control samples, i.e., the probability that a sample belonging to the class Control is classified correctly.
5. Precision for class AD: the fraction of correctly classified AD samples over the total number of samples classified as AD. In terms of probabilities, it is the probability that a sample classified as AD is actually belonging to the AD class.
6. Recall for class AD: same as Sensitivity for class AD.
7. Sensitivity/Specificity/Precision/Recall for class Control: as for the AD class.
8. Balanced Accuracy: similar to accuracy, but takes imbalanced classes into consideration. Specifically, it is the average recall over all classes. A random classification would result in a balanced accuracy of 1/C, where C is the number of classes.
9. Average F1: the F1 score for a class is defined as 2 * precision * recall / (precision + recall). The average F1 score is the average over all classes.

After the Performance metric table, the Reports includes two different graphics that quantify the impact of each signature element of the performances of the overall signature. The first graphic (Figure 25) represents the expected decrease in performance (AUC) caused by the elimination, in turn, of each element of the signature. The graphic shows that if any element is removed from the signature, it is possible to achieve up to ~97% of the original performance. This indicates that all elements contributes approximately equally to the performance of the whole signature.

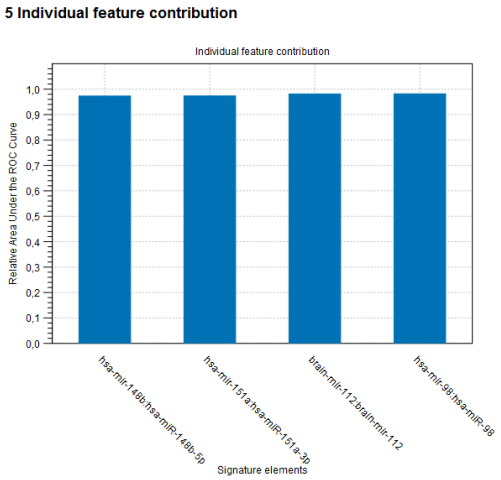

Figure 25: individual contribution of each element of the signature

The second bar chart (Figure 25) represents the percentage of performance that is achieved by adding one element at the time to the signature. Particularly, the graph shows that by considering only the first element, it is possible to arrive to ~85% of the predictive power of the whole signature. Considering the first and the second element, ~97% of the performance is reached. Adding the last two elements brings to the full predictive power (100%).

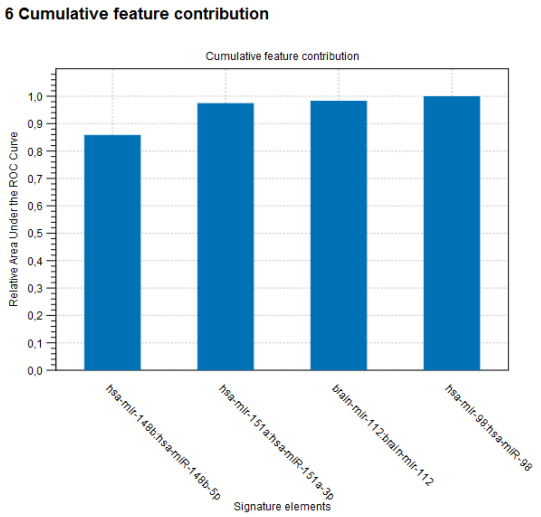

Figure 26: cumulative contribution of signature elements

The Report shows two further graphics, (a) the distribution of the predicted probability of belonging to class Control and (b) the distribution of the samples in the first two components of the PCA space built on top of the signature elements.

The first of the two graphics is shown in Figure 27.

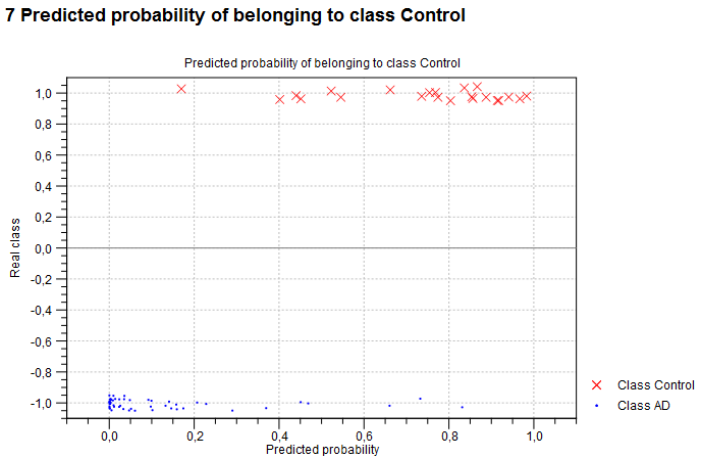

Figure 27: predicted probability of belonging to class Control

Each sample is represented as a dot in the graph. The dots have different shapes according to their class. The x-axis represents the predicted probability of belonging to the class Control; samples belonging to class Control are represented on the top, marked as "crosses", while samples belonging to class AD are on the bottom, represented as simple dots. The ideal behavior would be to observe the entire Control sample on the rightmost – top corner, while all the AD samples should be in the leftmost – bottom corner. Samples that do not obey to this rule are somewhat misclassified, and should be carefully investigated.

The last plot (Figure 28) represents the samples in the PCA space built on top of the signature elements. This plot provides a bi – dimensional graphical representation of the distribution of the samples in the space defined by the measurements included into the signature. Particularly, in this case it is evident that the two classes are almost perfectly separated by the first two components of the PCA space.

8 Principal Component Analysis (PCA)

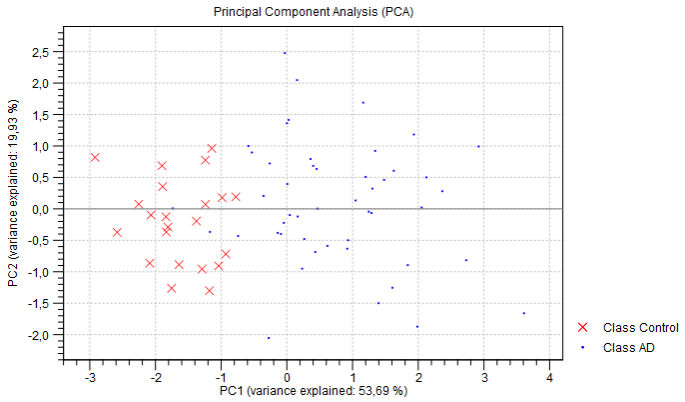

Figure 28: PCA plot built on the elements included in the reference signature

Analysis of potato (solanum tuberosum) metabolic profiles for identifying pre-harvest biomarkers of black spot bruising susceptibility.

Black spot bruising is the undesired formation of dark-blue to blackish melanin spots below the peel of potato tubers after being exposed to mechanical pressure [3]. Different harvests show different degree of susceptibility to this phenomenon, and black spots drastically reduce the commercial value of the tubers. A recent study [4] attempts to identify metabolic biomarkers able to discriminate, months ahead of the harvesting, potato crops highly susceptible to black spot bruising. The early identification of highly susceptible harvests allows the differentiation of the procedures for the collection and stock of the crops, in order to minimize both the deterioration of the tubers and the harvesting cost.

Tuber metabolic profiles employed in the study are publicly available on the journal website (link). For the present case study, the data have been formatted as Comma Separated Value (CSV file), in order to be easily imported in the CLC bio workbench. Please download the data file from this link. The data contain the metabolic profiles of a set of potato samples (growth in different soils and in different weather conditions) measured before the harvesting. For each profile an indication of the susceptibility to black spot bruising (as measured after the harvesting) is provided as well. Particularly, we consider three levels of susceptibility: 0, 1 and 2, corresponding to low, medium and high susceptibility, respectively[1]

Once you have saved/downloaded the CSV file, you can import it in the CLC bio Workbench with the "Automatic import" (Ctrl + i) utility.

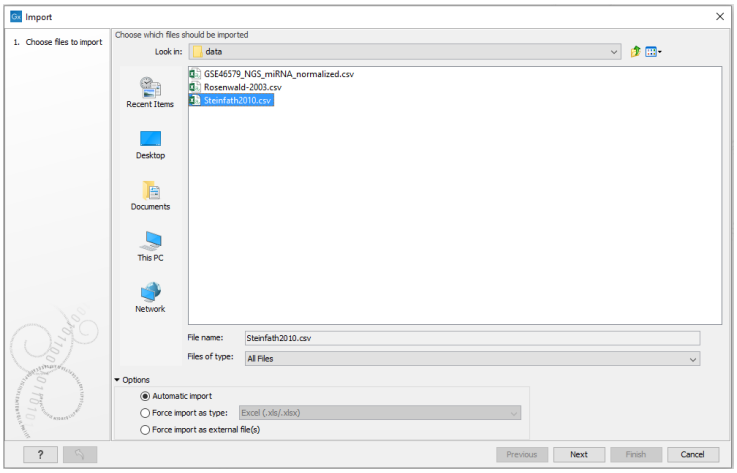

Figure 29: importing the metabolic profiles

When prompted for selecting the location where to save the files, create a new folder "Steinfath2010" in the CLC bio workspace. Press "Finish" and wait for the data to be loaded.

Once the data have been imported, the "Steinfath2010" folder will contain 478 metabolic profiles. We can now launch the *BioSignature* – Discoverer plugin for performing our analysis. In the "Select Data" panel, select all the metabolic profiles and click on "Next".

[1]We re-encode the nine levels (1 – 9) scale used in the original study as follows: 1 – 3 → 0, 4 – 6 → 1, 7 – 9 → 2.

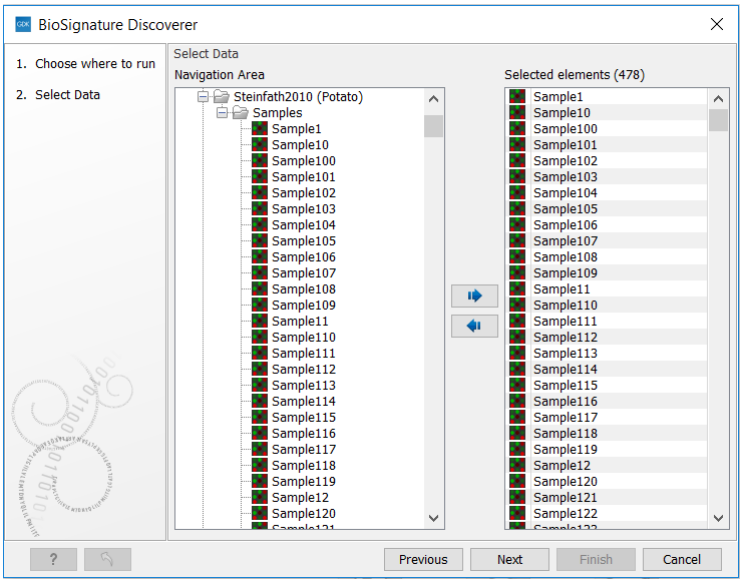

Figure 30: selecting the potato metabolic profiles

The successive window shows the "Specify Analysis Type and Outcome" options. In this study we want to classify the potato profiles according to their level of black spot bruising susceptibility. Thus, select "Classification" in the area named "What type of analysis to perform" and select "Blackspot Bruising" as target variable (see Figure 31). Click "Next".

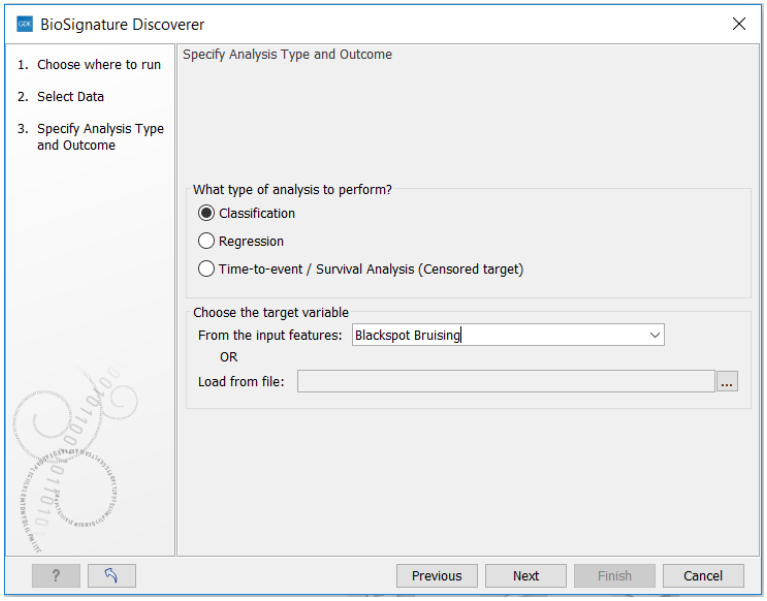

Figure 31: selecting the type of analysis and the target variable

For the present case study, let's set the plugin options as in Figure 32: original values and "Normal" as level of tuning. After clicking on "Next", the panel of advanced option will appear. We want to achieve a quite precise estimation of performances, so we set the number of repetitions to 3.

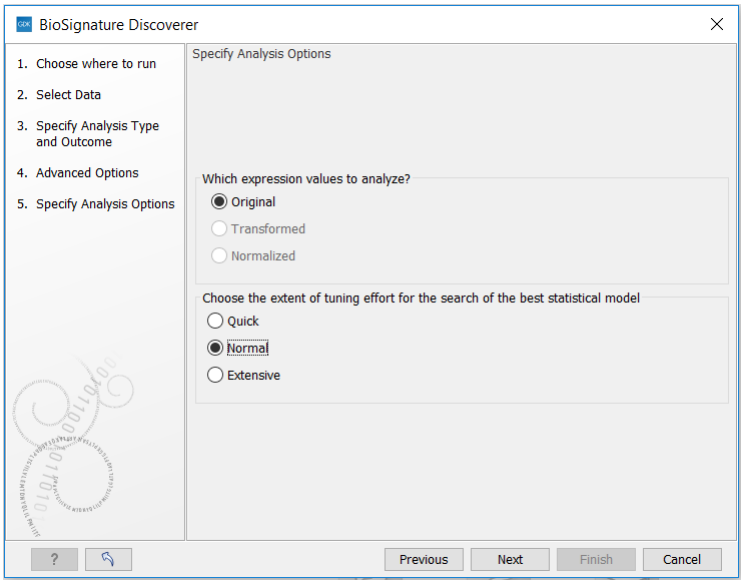

Figure 32: BioSignature – Discoverer plugin options

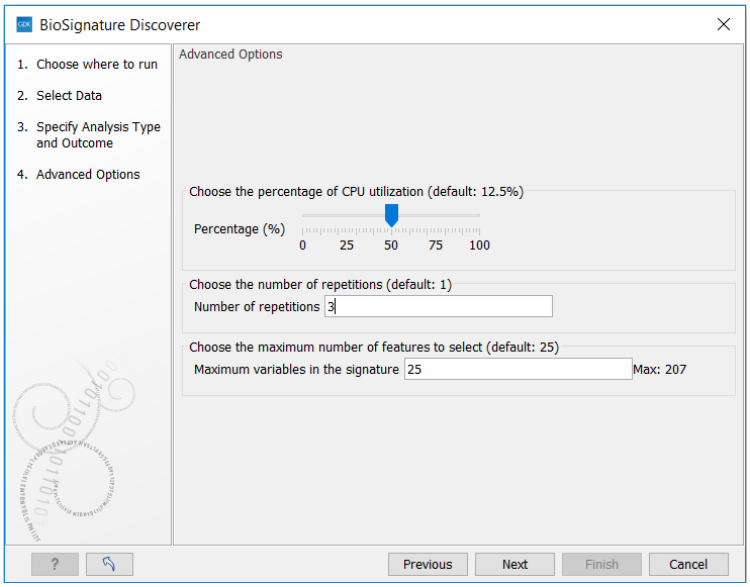

Figure 33: advanced settings panel

Finally, let's select "Open" in the Result handling options window, and then let's click "Finish" for starting the plugin.

Reporting Multi-Class Classification Results

The "Reference Signature" is reported right on the top of the Summary Report:

1 Reference Signature

Feature 1 (Stability = 100,0%)	Feature 2 (Stability = 96,7%)	Feature 3 (Stability = 100,0%)	Feature 4 (Stability = 76,7%)	Feature 5 (Stability = 100,0%)	Feature 6 (Stability = 100,0%)	Feature 7 (Stability = 63,3%)	Feature 8 (Stability = 73,3%)	Feature 9 (Stability = 53,3%)	Feature 10 (Stability = 96,7%)
Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A291005	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001

Figure 34: Reference Signature for potato black spot bruising

The signature is composed by ten different predictors. Notably, the type of cultivar and soil are included along with metabolic measurements (Feature 1 and 2). The stability values indicate that most of these predictors would have a high chance to be selected again if the analyses were repeated on a different, independent sample.

The table “2 Lists of Equivalent Features” indicates that some element of the reference signature can be substitute by other signatures that are equivalent in terms of predictive capabilities. For example, this means that if we substitute the fourth element of the Reference Signature, namely the “Analyte-A104001” variable, with the “Weather” variable, then we obtain a second signature that is equivalent to the reference one. In general, an equivalent signature can be built by picking one (and only one) element from each of the column of table “2 Lists of Equivalent Features” (see Figure 35).

2 Lists of Equivalent Features

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
			Docosane-n-A220001		Analyte-A291005				
			Ethanolamine-A128002		Analyte-A293001				
			Pentadecane-n-A150001		Analyte-A294002				
			Weather						

There are 1 × 1 × 1 × 5 × 1 × 4 × 1 × 1 × 1 × 1 = 20 equivalent signatures.

Figure 35: list of equivalent features. Eight different signatures can be constructed in this particular case

Consequently, in this case a total of 20 equivalent signatures can be built, as reported in the table “1 Lists of equivalent signatures” reported in the Detailed Report (see Figure 36).

1 List of equivalent signatures

Signatures	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
Reference Signature	Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A291005	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 1	Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 2	Cultivar	Soil	Methionine-A142007	Docosane-n-A220001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 3	Cultivar	Soil	Methionine-A142007	Ethanolamine-A128002	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 4	Cultivar	Soil	Methionine-A142007	Pentadecane-n-A150001	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 5	Cultivar	Soil	Methionine-A142007	Weather	Analyte-A281001	Analyte-A279001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 6	Cultivar	Soil	Methionine-A142007	Docosane-n-A220001	Analyte-A281001	Analyte-A291005	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 7	Cultivar	Soil	Methionine-A142007	Ethanolamine-A128002	Analyte-A281001	Analyte-A291005	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 8	Cultivar	Soil	Methionine-A142007	Pentadecane-n-A150001	Analyte-A281001	Analyte-A291005	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 9	Cultivar	Soil	Methionine-A142007	Weather	Analyte-A281001	Analyte-A291005	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 10	Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A293001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 11	Cultivar	Soil	Methionine-A142007	Docosane-n-A220001	Analyte-A281001	Analyte-A293001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 12	Cultivar	Soil	Methionine-A142007	Ethanolamine-A128002	Analyte-A281001	Analyte-A293001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 13	Cultivar	Soil	Methionine-A142007	Pentadecane-n-A150001	Analyte-A281001	Analyte-A293001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 14	Cultivar	Soil	Methionine-A142007	Weather	Analyte-A281001	Analyte-A293001	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 15	Cultivar	Soil	Methionine-A142007	Analyte-A104001	Analyte-A281001	Analyte-A294002	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 16	Cultivar	Soil	Methionine-A142007	Docosane-n-A220001	Analyte-A281001	Analyte-A294002	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 17	Cultivar	Soil	Methionine-A142007	Ethanolamine-A128002	Analyte-A281001	Analyte-A294002	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 18	Cultivar	Soil	Methionine-A142007	Pentadecane-n-A150001	Analyte-A281001	Analyte-A294002	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001
Equivalent Signature 19	Cultivar	Soil	Methionine-A142007	Weather	Analyte-A281001	Analyte-A294002	Analyte-A191007	Galactaric-acid-A204001	Analyte-A142003	Threonine-A140001

Figure 36: list of equivalent signatures (Detailed Report)

Table “3 Effect sizes” in the Summary Report gives an indication about the strength of the relationship between each element of the signature and the outcome.

3 Effect sizes

	Cultivar (std = 5,780)	Soil (std = 0,501)	Methionine-A142007 (std = 0,007)	Analyte-A104001 (std = 0,014)	Analyte-A281001 (std = 0,023)	Analyte-A291005 (std = 0,041)	Analyte-A191007 (std = 0,129)	Galactaric-acid-A204001 (std = 0,003)	Analyte-A142003 (std = 0,014)	Threonine-A140001 (std = 0,009)
Class 0.0 vs Class 2.0	-1,18	0,35	1,07	0,13	-0,47	-0,51	-0,21	-0,46	0,58	0,48
Class 1.0 vs Class 2.0	-0,12	0,29	0,79	-0,22	0,02	-0,01	-0,14	-0,39	0,54	0,41

Effect sizes are reported as log (base 10) odds ratio change per one standard deviation increase of the feature.

Figure 37: effect size of each element of the Reference Signature expressed $\log_{10}$ odds variation. Class 2 (corresponding to high susceptibility) is taken as reference

The coefficients are reported as $\log_{10}$ odds ratio, as explained in Section “Reporting Binary Classification Results”. Moreover, please note that:

- for outcomes comprising multiple classes, the Logistic Regression algorithm chooses one of the classes as baseline. In this case, class 2 (“high susceptibility”) acts as baseline

- all other classes (class 1 and class 0, in this case) are contrasted against the baseline

Let's focus on the fifth feature, "Analyte-A281001". According to the coefficients reported in the first column of table "3 Effect sizes", an increment of 0.023 (equal to its standard deviation) in the value of Analyte-A281001 corresponds to (a) an increment of the probability of being assigned to class 1 (with respect to the probability of being assigned to class 2) of $10^{0.02}$ = 1.05 times and (b) to a decrement of the probability of being assigned to class 0 (with respect to the probability of being assigned to class 2) of $10^{-0.47}$ = 0.39 times. In other words, the higher the value of Analyte-A281001, the most likely is for the potato sample to belong to class 1 (i.e., average susceptibility to black spot bruising).

4 Performance Metrics

Metric	Average	95% Confidence Interval
Area Under the ROC Curve	0,950	[0,930, 0,968]
Accuracy	0,837	[0,797, 0,875]
Balanced Accuracy	0,802	[0,754, 0,845]
Average F1	0,807	[0,761, 0,851]
Precision for class 0.0	0,898	[0,850, 0,945]
Precision for class 1.0	0,779	[0,698, 0,858]
Precision for class 2.0	0,779	[0,671, 0,882]
Recall for class 0.0	0,973	[0,946, 1,000]
Recall for class 1.0	0,762	[0,686, 0,840]
Recall for class 2.0	0,670	[0,562, 0,770]
Sensitivity for class 0.0	0,973	[0,946, 1,000]
Sensitivity for class 1.0	0,762	[0,686, 0,840]
Sensitivity for class 2.0	0,670	[0,562, 0,770]
Specificity for class 0.0	0,910	[0,866, 0,951]
Specificity for class 1.0	0,890	[0,850, 0,928]
Specificity for class 2.0	0,949	[0,922, 0,975]

Figure 38: performance metrics

The successive table of the Summary Report, namely "4 Performance Metrics", reports the estimated predictive performances, along with their 95% Confidence Interval estimated through a boot-strapping approach. For multi-class outcome the employed metrics are the same of the binary outcome (see Section "Reporting Binary Classification Results"). The AUC metric is defined as the average of all the possible 2-class AUCs.

After the Performance metric table, the Reports includes two different graphics that quantify the impact of each signature element on the performances of the overall signature. The first graphic (Figure 39) represents the expected decrease in performance (Accuracy) caused by the elimination, in turn, of each element of the signature, while the second bar chart (Figure 40) represents the percentage of performance that is achieved by adding one element at the time to the signature.

5 Individual feature contribution

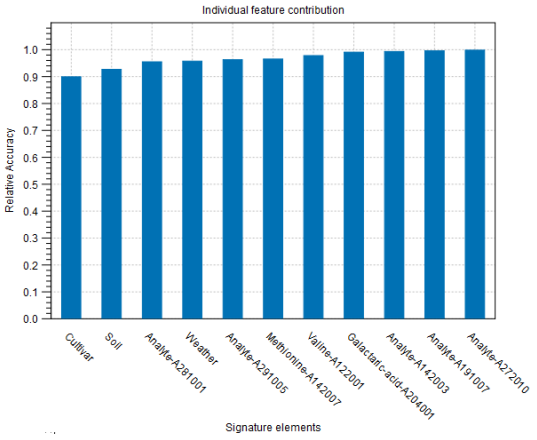

Figure 39: individual contribution of each element of the signature

The individual contribution graph shows that the first variables have an important predictive role, while the remaining features have less impact. The cumulative contribution graph strengthens this interpretation, showing that the first four variables are enough in order to achieve more than 95% of the performances of the whole signature (consequently, the remaining predictors are not shown for sake of clarity).

6 Cumulative feature contribution

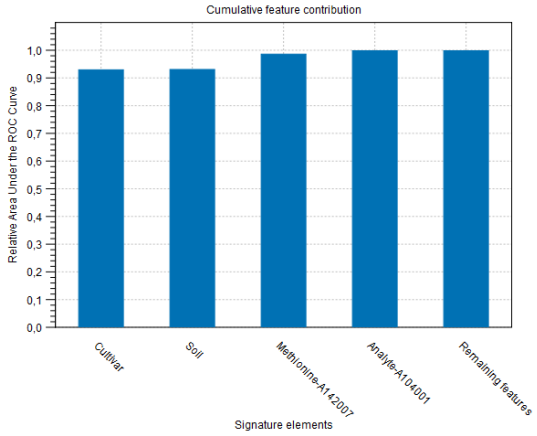

Figure 40: cumulative contribution of signature elements

The Report shows two further types of graphics, for checking the correctness of the classification model: (a) the distribution of the predicted probability of belonging to class 0, 1 or 2 and (b) the distribution of the samples in the first two components of the PCA space built on top of the signature elements.

Figure 41 shows the first type of graphics. Each plot shows the probability of belonging to class 0, 1 or 2 (red crosses, left to right, respectively) against the probability of belonging to any other class (blue dots in each plot). Each sample is represented as a mark in the graph. The ideal behavior would be to observe all the marks clustered in two groups, one on the rightmost – top corner and one on the leftmost – bottom corner. Samples that do not obey to this rule are somewhat misclassified, and should be carefully investigated.

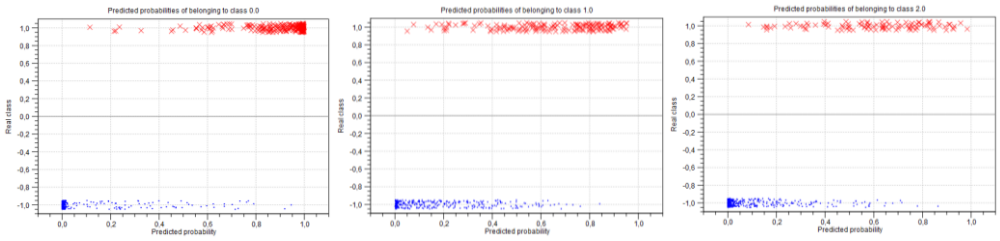

Figure 41: predicted probability of belonging to class 0, 1, and 2 (from left to right, respectively).

The last plot (Figure 42) represents the samples in the PCA space built on top of the signature elements. This plot provides a bi – dimensional graphical representation of the distribution of the samples in the space defined by the elements included in the signature. Particularly, in this case it is evident that the two classes are not perfectly separated by the first two components of the PCA space. This partly explains why the estimated accuracy is not optimal (accuracy = 0.837, see Figure 38).

10 Principal Component Analysis (PCA)

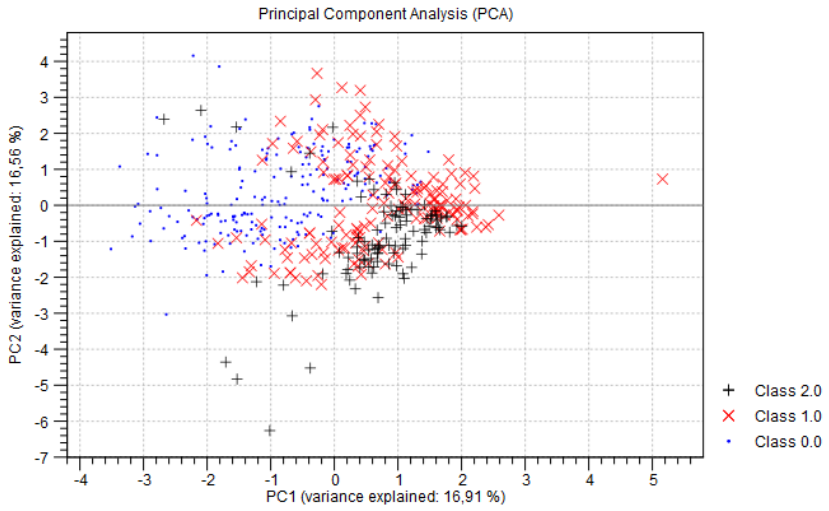

Figure 42: PCA plot

Identification of a gene expression signature for estimating the survival probability of mantle cell lymphoma patients

Predicting the survival time of breast cancer patients is a difficult task: multiple factors influence the mortality of cancer patients, and most of these factors may well be unknown or unmeasured. Moreover, the analysis of survival data presents an inherent technical difficulty, namely the presence of censored data. Censored observations appear when the exact time to event is unknown. For example, in a longitudinal study aimed at analyzing the survival of a cohort of cancer patients, it often happens that some of the subjects drop in advance from the study. The exact survival time for these patients is unknown; all that is known is that they have survived up to the moment when they left. Excluding these subjects from the analysis can produce biased results, since these are the patients that survive the longest. However, classical regression algorithms are not devised for dealing with censored data. Thus, specialized statistical methods must be employed for survival analysis [5].

In a pioneering study [6], Rosenwald et al. analyzed the survival of a cohort of 92 mantle lymphoma patients. Particularly, the authors investigated the possibility of predicting the time to death of the patients on the basis of their genome – wide transcriptome profiles and clinical information. The data from this study are available at this link. Download the CSV file for this study and load it into the CLC bio workbench with the Standard Import (Ctrl + i) utility. Save the data in the "Rosenwald" folder (Figure 43).

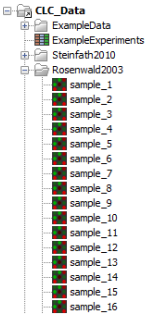

Figure 43: the mantle lymphoma expression profiles

Let's employ the *BioSignature* – Discoverer in order to identify the gene expression signature that best predicts the survival time. Start the plugin, and select all the expression profiles as input (Figure 44).

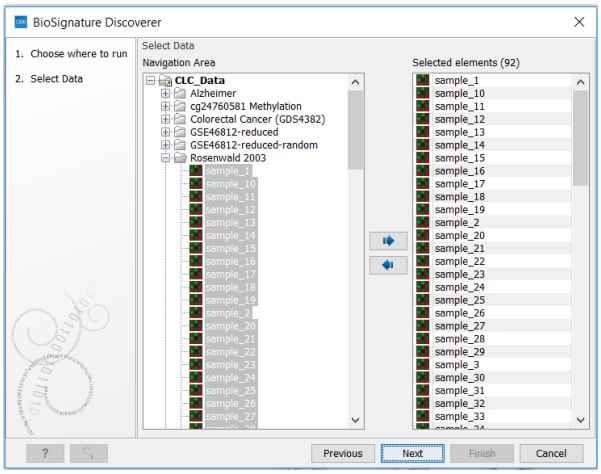

Figure 44: selecting gene expression profiles

In the next dialog window, let's select "Survival Analysis" for the type of analysis to perform (Figure 45). We are now required to "Choose the target variable". This means that we should indicate the survival time of each subject, which is the target variable that we want to predict.

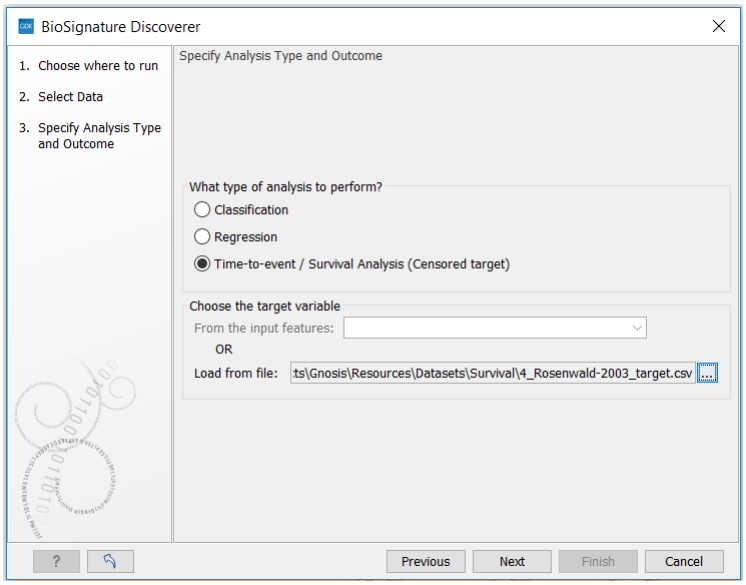

Figure 45: setting up the survival analysis

Survival times must be specified with a Comma Separated Value (CSV) file. An example file is shown in Figure 46. Each row of the file reports the survival information for a single subject, and it is formatted as <*sample_name*, *time_to_event*, *event_status*>, where:

- *sample_name* is the name of the expression profile the row refers to
- *time_to_event* is the time elapsed until the event or the censorship occurred
- *event_status* assumes value "1" if the time to event is known and "0" otherwise

```
sample_1,0.75291,0
sample_2,3.2772,0
sample_3,2.1218,0
sample_4,14.0534,0
sample_5,3.2361,0
sample_6,4.4873,0
sample_7,0.7778,1
sample_8,0.42984,0
sample_9,1.0568,0
sample_10,3.2882,0
sample_11,6.8966,1
```

Figure 46: survival time example file

For example, the patient corresponding to the expression profile sample_7 survived for 0.7778 years after the histological exam (*time_to_event* = 0.7778, *event_status* = 1). Conversely, the patient corresponding to sample_2 was still alive 3.2772 years after she underwent the histological exam, but no information are available after then (*time_to_event* = 3.2772, *event_status* = 0).

The file "survivalOutcome.csv" with the survival information for the 251 expression profiles is available at this link. Press the "Load from file" button and locate the CSV file on your computer (Figure 47).

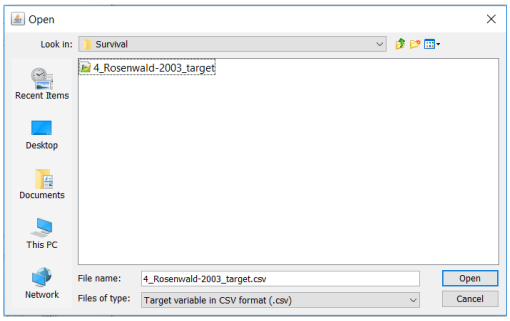

Figure 47: loading the survivalOutcome.csv file

Let's set up the options for the BioSignature – Discoverer plugin analyses with a "Extensive" level of tuning (Figure 48).

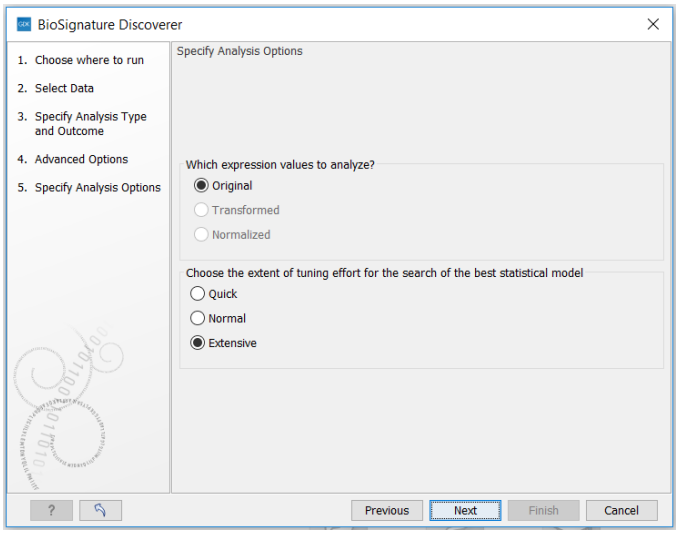

Figure 48: plugin configuration for the example survival analysis

Finally, let's have three whole repetition of the whole statistical pipeline for better estimating signatures' predictive performances (Figure 49).

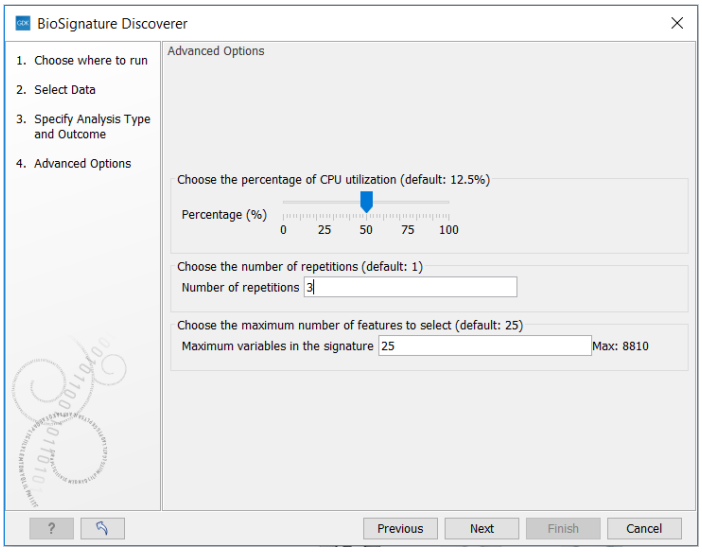

Figure 49: advanced options

Reporting Survival Analysis Results

At the end of the computation the Summary Report provides a Reference Signature with five different genes. Interestingly, some genes have a large number of equivalent features, leading to a total of 1890 possible equivalent signatures.

1 Reference Signature

Feature 1 (Stability = 70,0%)	Feature 2 (Stability = 40,0%)	Feature 3 (Stability = 43,3%)	Feature 4 (Stability = 36,7%)	Feature 5 (Stability = 43,3%)	Feature 6 (Stability = 20,0%)	Feature 7 (Stability = 50,0%)	Feature 8 (Stability = 6,7%)	Feature 9 (Stability = 56,7%)	Feature 10 (Stability = 43,3%)	Feature 11 (Stability = 0%)
M54992-1	NM_013242	AF230904-2	NM_003362-1	U03754-2	X76057	X77743	AF024636-1	AI361769	AA811800	AA262027

2 Lists of Equivalent Features (showing only first 25 features)

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11
M54992-1	NM_013242	AF230904-2	NM_003362-1	AA736568	X76057	AA279783	AF024636-1	AI361769	AA811800	AA262027
				AA804900		AA807503	AI400790		AF195765	
				NM_031300-1		AF009227	U68019-1		M54968-2	
				U03754-2		AF072934				
				U03754-5		AF106966				
				U90143		AI246382				
				X56841-5		AI281436				
						AK001360				
						AK024189				
						AL117563-1				
						AL832747				
						BC007568-1				
						D88208				
						L40636				
						M34065-2				
						NM_002913-1				
						NM_005583-1				
						NM_006297				
						S77154				
						S78187				
						U43195-1				
						U57650				
						U59321-1				
						U61167				
						UP36				
						Omitted 5 equivalent features				

There are 1 × 1 × 1 × 1 × 7 × 1 × 30 × 3 × 1 × 3 × 1 = 1890 equivalent signatures.

Figure 50: list of equivalent signature for the survival analysis task

The Effect Size table (Figure 51) reports how the risk of death for mantle lymphoma changes according to variations in the values of the signature's elements. Particularly, the effect sizes are reported as the natural logarithm of the hazard ratios, and all the predictors were standardized before the analysis. This means that a change in the value of "NM_003362-1" equal to 1.602 (i.e., equal to its standard deviation) implies an increase of the individual risk equal to $e^{0.862}$ = 2.367 times.

3 Effect sizes

M54992-1 (std = 5,187)	NM_013242 (std = 7,183)	AF230904-2 (std = 3,694)	NM_003362-1 (std = 1,602)	U03754-2 (std = 2,504)	X76057 (std = 1,356)	X77743 (std = 1,961)	AF024636-1 (std = 2,529)	AI361769 (std = 2,342)	AA811800 (std = 1,587)	AA262027 (std = 0,272)
-1,228	0,638	-1,366	0,862	-1,456	-1,551	-1,323	-0,820	1,342	1,536	0,336

Effect sizes are reported as the natural logarithm of the hazard ratio.

Figure 51: effect sizes for the survival analysis signatures

The Performance Metrics table shows only one metric, the Concordance Index (CI). This metric has an interpretation similar to the Area Under the ROC Curve, i.e., it represents the probability of correctly ranking, according to their respective risk, two randomly chosen subjects. Perfect predictions would grant a CI equal to 1, while a random ranking should achieve a 0.5 CI. In our case, CI is 0.636, indicating that the gene expressions carry some useful information in order to estimate the risk, but further information (e.g., clinical data) are necessary in order to provide better predictions.

4 Performance Metrics

Metric	Average	95% Confidence Interval
Concordance Index	0,636	[0,542, 0,718]

Figure 52: Performance Metrics for survival analysis

The contribution of each feature to the predictive performance of the signature is reported in the Individual and Cumulative Contribution graphs (Figure 53 and Figure 54, respectively). The first plots indicate that no variable has large predictive power when considered in isolation. The second plot points out that the first nine predictors are already enough to achieve nearly the 100% of the predictive power, while the remaining predictors provide a quite marginal contribution.

5 Individual feature contribution

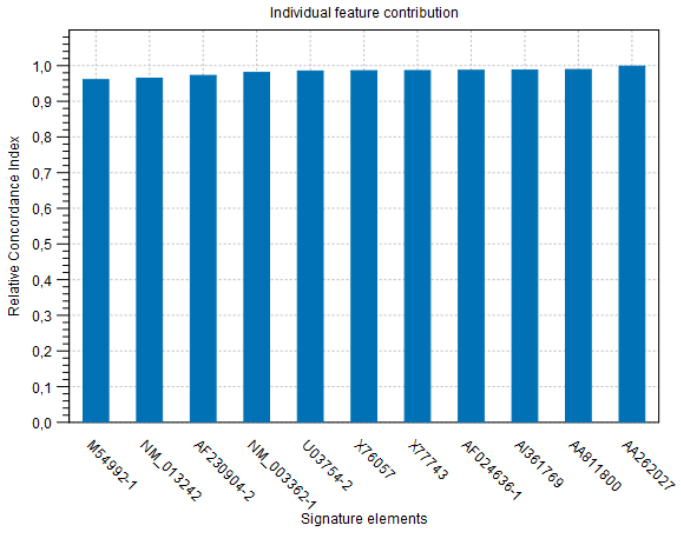

Figure 53: individual contribution plots

6 Cumulative feature contribution

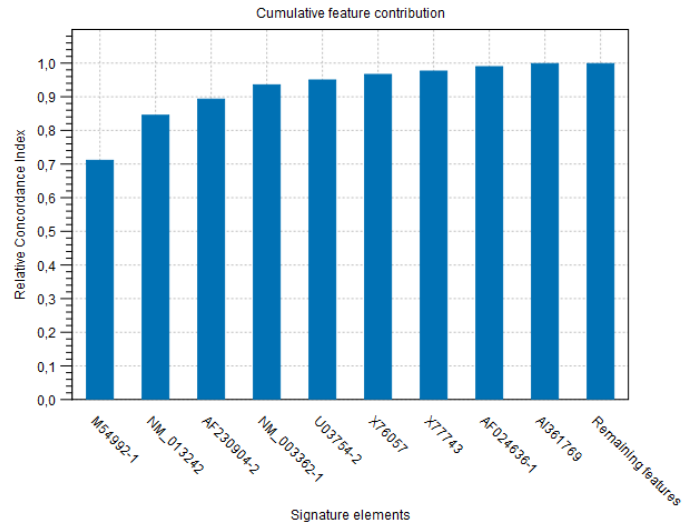

Figure 54: cumulative contribution plots

The deviance residuals' plot can be used for investigating the fit of the model. Deviance residuals indicate whether the model predictions depart from the real risk. They should ideally be randomly distributed around zero, without any identifiable pattern. This seem to be the case in Figure 55, with no samples largely departing from the zero line. The trimmed values are residuals that, for sake of clarity, are repositioned closer to the other ones on the X-axis.

7 Deviance residuals' plot

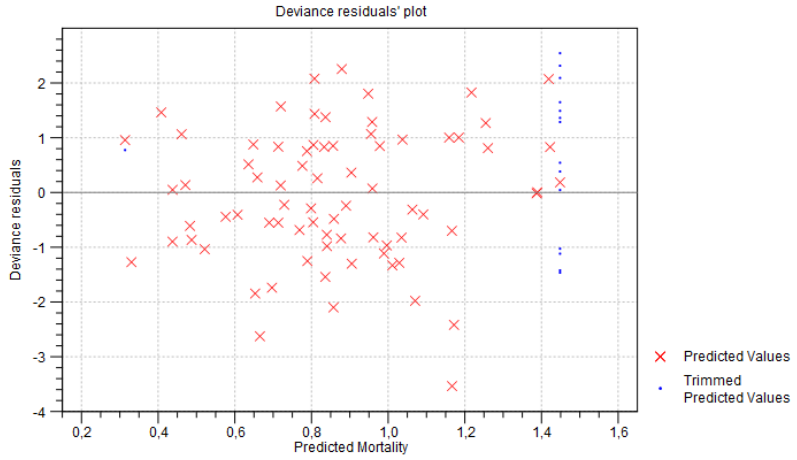

Figure 55: Deviance residuals' plot

Finally, the plugin provides Kaplan-Meier plots for better investigating whether the out-of-samples risk prediction allow to stratify the samples in different risk groups. Particularly, the samples are subdivided in *n* groups based on the out of samples predictions, where each group contains an equal number of samples and *n* is in turn set to 2, 3, 4 and 5. The corresponding Kaplan-Meier curves are then provided, along with a p-value assessing statistical differences among them (derived through a log-likelihood ratio test). In the breast cancer example, stratifying the patient in two level of mortality produces a first group at high risk (Figure 56, blue line) and a group with longer average survival (black line), which according to the log-likelihood ratio test are statistically different at level p = 0.012.

8 Kaplan-Meier Curves for 2 Levels of Mortality

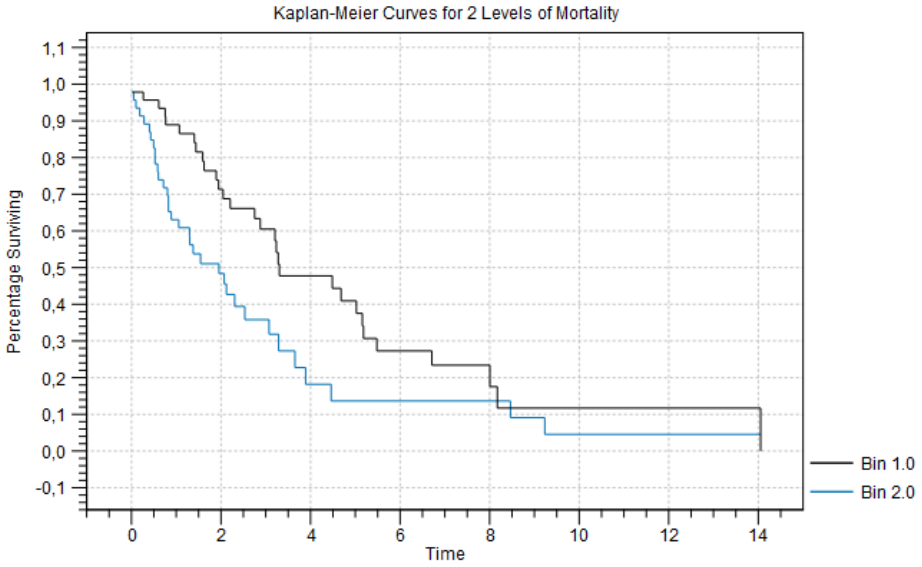

Figure 56: Kaplan-Meier curves for the breast cancer example

References

[1] P. Leidinger, C. Backes, S. Deutscher, K. Schmitt, S. C. Mueller, K. Frese, J. Haas, K. Ruprecht, F. Paul, C. Stähler, C. J. Lang, B. Meder, T. Bartfai, E. Meese, and A. Keller, "A blood based 12-miRNA signature of Alzheimer disease patients.," *Genome Biol.*, vol. 14, no. 7, p. R78, 2013.

[2] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: A Bioconductor package for differential expression analysis of digital gene expression data," *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2009.

[3] P. E. Lærke, J. Christiansen, and B. Veierskov, "Colour of blackspot bruises in potato tubers during growth and storage compared to their discolouration potential," *Postharvest Biol. Technol.*, vol. 26, pp. 99–111, 2002.

[4] M. Steinfath, N. Strehmel, R. Peters, N. Schauer, D. Groth, J. Hummel, M. Steup, J. Selbig, J. Kopka, P. Geigenberger, and J. T. Van Dongen, "Discovering plant metabolic biomarkers for phenotype prediction using an untargeted approach.," *Plant Biotechnol. J.*, vol. 8, no. 8, pp. 900–911, 2010.

[5] V. Lagani and I. Tsamardinos, "Structure-based variable selection for survival data," *Bioinformatics*, vol. 26, no. 15, pp. 1887–1894, 2010.

[6] A. Rosenwald, G. Wright, A. Wiestner, W. C. Chan, J. M. Connors, E. Campo, R. D. Gascoyne, T. M. Grogan, H. K. Muller-Hermelink, E. B. Smeland, M. Chiorazzi, J. M. Giltnane, E. M. Hurt, H. Zhao, L. Averett, S. Henrickson, L. Yang, J. Powell, W. H. Wilson, E. S. Jaffe, R. Simon, R. D. Klausner, E. Montserrat, F. Bosch, T. C. Greiner, D. D. Weisenburger, W. G. Sanger, B. J. Dave, J. C. Lynch, J. Vose, J. O. Armitage, R. I. Fisher, T. P. Miller, M. LeBlanc, G. Ott, S. Kvaloy, H. Holte, J. Delabie, and L. M. Staudt, "The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma," *Cancer Cell*, vol. 3, pp. 185–197, 2003.