Using Omics skin in SIMCA®

Background

Skins are customized user interfaces that improve the usability for the dedicated application. The Omics skin is an add-on to SIMCA 14. Its user-friendly Analysis Wizard facilitates swift analysis of various types of omics-related data.

Data

The dataset used deals with the study of genetically modified poplar plants based on GC/MS measurements. Original literature reference is: Wiklund, et. al, *Visualization of GC/TOF-MS-Based Metabolomics Data for Identification of Biochemically Interesting Compounds Using OPLS Class Models, Analytical Chemistry, 2008, 80, 115-122*. In this paper, two genetic modifications are investigated, i.e., up-regulation and down-regulation within the PttPME1 gene. This gene is involved in the production of pectin methyl esterase, PME, which is an enzyme that de-esterifies methylated groups within pectin. The metabolic profile was of interest and both lines indicated several symptoms of oxidative stress response.

This tutorial illustrates the workflow in the Analysis Wizard when comparing the up-regulated group with the wild type group. The dataset contains N = 19 observations (plants) and K = 80 variables (resolved and integrated GC/MS profiles).

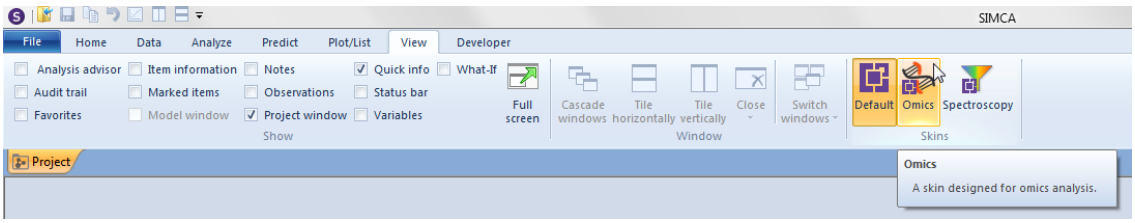
The observations (plants) are divided in two groups:

- Reference, wild-type plant, 10 plants
- Group 1, up-regulated poplar, 9 plants

Launching the Omics skin

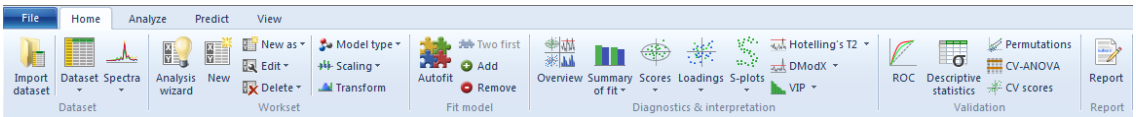
To use the Omics skin you must have SIMCA and the Omics skin installed.

Start SIMCA and launch the skin; on the View tab, in the Skins group, click Omics.



The omics skin is installed as part of the standard SIMCA installation. If you wish to use it and it are not shown in the view tab, you can enable it in **File | Options**, **SIMCA options**, **Skins** section. Set **Yes** on the skins you wish to enable.

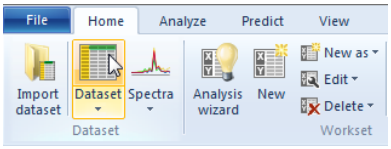
The Omics skin comprises five tabs (File, Home, Analyze, Predict and View). For modeling, the Home tab contains all functionality of interest to do a complete run-down of the information in your omics-data. (In order to change back to normal appearance of SIMCA, on the View tab, in the Skins group, click Default.)



Import of example dataset

To import the example dataset, click File | New and select the file "Poplar_Xylem_GCMS_two-class omics skin.xlsb" which is included in the tutorial package. Click Finish, give a name and storage location for the project, and the dataset will be imported into SIMCA.

To view the imported dataset, click Dataset on the Home tab:



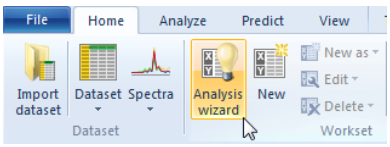
An excerpt of the dataset is shown below. The first row represents Primary ID for the variables. The first three columns represent Primary ID, Class ID and Secondary ID for the 19 observations.

Dataset - Poplar_Xylem_GCMS_two-class

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	Primary ID	$ClassID	Secondary ID	Dodecamı	HEPTANO	Dodecane	L-Valine (2	ETHANOL	Ethanolam	(2-Aminoe	ETHANOL	Butanoic a	2-Piperidir	L-PROLINt	1,2-Bis(trir	Glyceric ac	Not Assigr	L-Gluti
2	WT_10	WT	1	375833	543556	38276,9	69192,9	184317	1,04657e	260642	197055	300727	309177	6,21998e	882500	991475	113409	43045
3	WT_1	WT	2	398184	537285	34609,7	217668	193095	1,07511e	242647	416123	349135	310216	5,84856e	887659	774440	161768	1444
4	WT_2	WT	3	443126	695953	38394,2	314980	148169	763070	314753	502899	229606	253487	6,54093e	1,11863e	1,10805e	245764	2149
5	WT_3	WT	4	547311	627214	51519,6	205261	167343	792171	293795	423352	324328	245085	6,53183e	1,0511e+	1,23245e	166540	2415
6	WT_4	WT	5	417952	505763	30364	109736	166606	948087	226815	364643	291389	297992	5,70728e	756064	756439	109040	34363
7	WT_5	WT	6	269593	401569	26340	154390	130227	783297	171252	308511	261535	210111	5,60637e	644742	544435	141587	68027
8	WT_6	WT	7	268957	375125	26103,4	111417	132385	734289	164504	256760	222962	203051	6,03599e	635568	1,00474e	116548	59476
9	WT_7	WT	8	279926	394992	26409,3	141851	109391	535205	178013	303425	113054	175968	5,69597e	652005	684164	103411	1542
10	WT_8	WT	9	424989	585558	36533,1	166067	111414	526641	248937	416188	296331	179789	6,28718e	897827	936225	95369	3207
11	WT_9	WT	10	307961	299436	25833,7	80102,5	117767	777849	174648	254719	255771	172048	5,32846e	625137	665265	120449	457
12	2B_10	2B	11	447402	572535	47492	102955	174650	957931	225563	245923	198470	309649	6,51605e	951668	910418	93948,6	43609
13	2B_1	2B	12	379836	481590	34871,4	164956	170290	900165	219947	363968	309223	238715	6,32357e	834956	936802	110187	95629
14	2B_2	2B	13	418452	577945	35005,5	213249	201696	1,03829e	237693	423304	307948	261031	6,19838e	926526	717418	145883	1393
15	2B_3	2B	14	409893	450488	30647,8	90750,8	150054	880176	197708	282827	305534	229553	5,50231e	808893	894217	116253	27312
16	2B_4	2B	15	494234	596287	38420,8	134018	156044	784148	264308	394712	227411	213565	6,31834e	954939	1,27173e	135706	98677
17	2B_5	2B	16	440927	616060	37094,5	203504	138155	756563	289188	360051	354267	200753	6,36666e	1,00639e	1,49285e	158561	1263
18	2B_7	2B	17	360690	579813	35616,1	115003	174554	1,02391e	248006	308144	222155	298088	6,1469e+	951451	816339	144168	78704
19	2B_8	2B	18	506128	659398	49216,6	174616	148102	717770	283588	340819	261514	171632	6,94707e	1,00365e	1,24221e	131290	1474
20	2B_9	2B	19	425301	617222	42853,4	187411	120327	492021	256024	352901	312914	178906	6,4711e+	987386	1,40312e	142344	2611

The Analysis Wizard

The Analysis Wizard helps the user to create the basic multivariate models needed to summarize, document and interpret the information in an omics dataset. It is accessed by clicking Analysis Wizard on the Home tab.



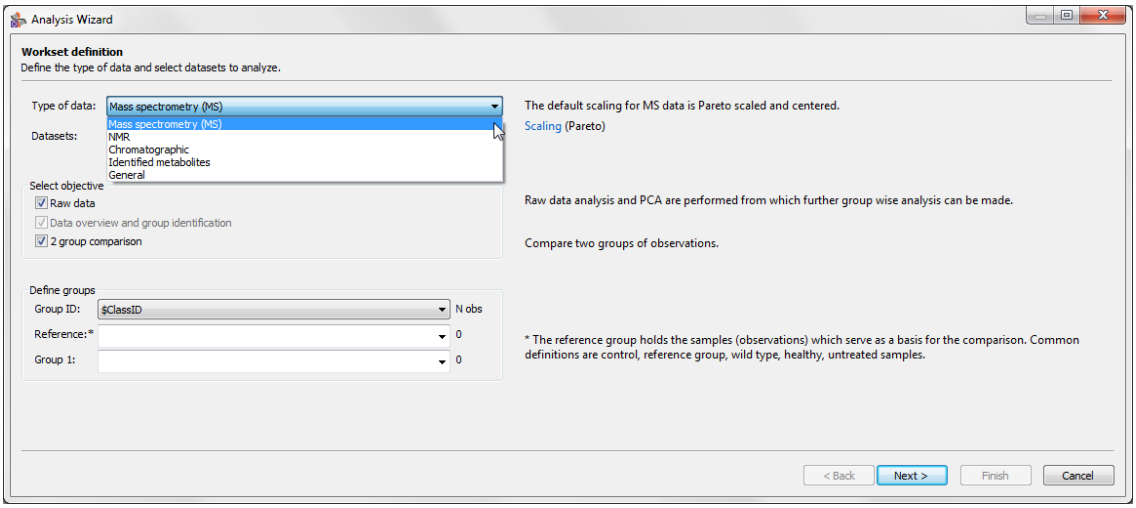
Workset definition

Prior to computing the multivariate models, the Workset must be defined. The Workset is the set of data being used for the modeling; it may contain data from multiple data sources. Additionally, the Workset specifies which groups of observations should be used and how the variables should be scaled.

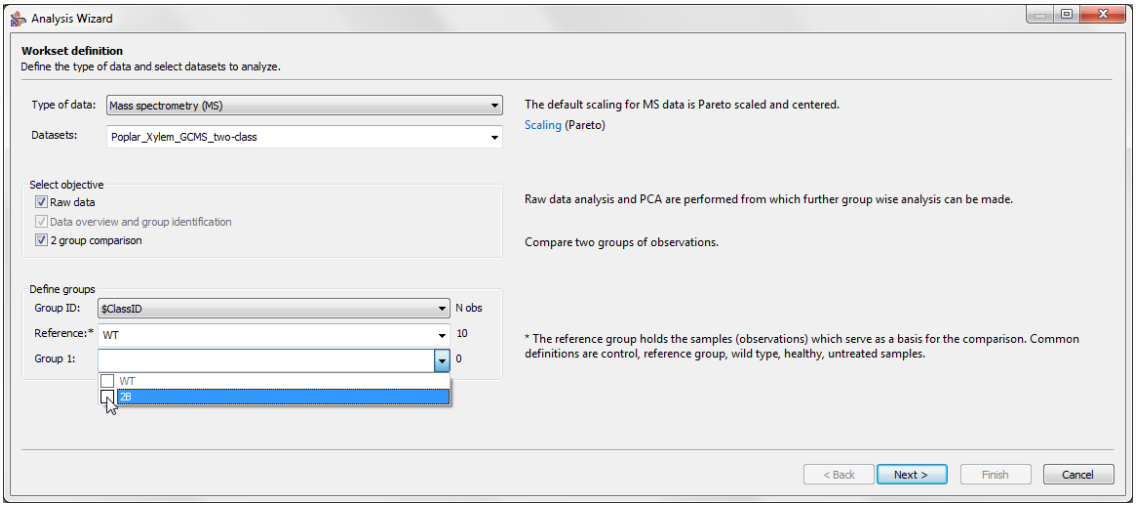
The Analysis Wizard recognizes five types of input data:

- Mass spectrometry (MS)
- NMR
- Chromatographic
- Identified metabolites
- General

For each type of data a default scaling procedure is pre-defined. The scaling can be changed manually. The data from the study of the genetically modified poplar plants are of the MS type. For MS data Pareto scaling and centering is the default option.

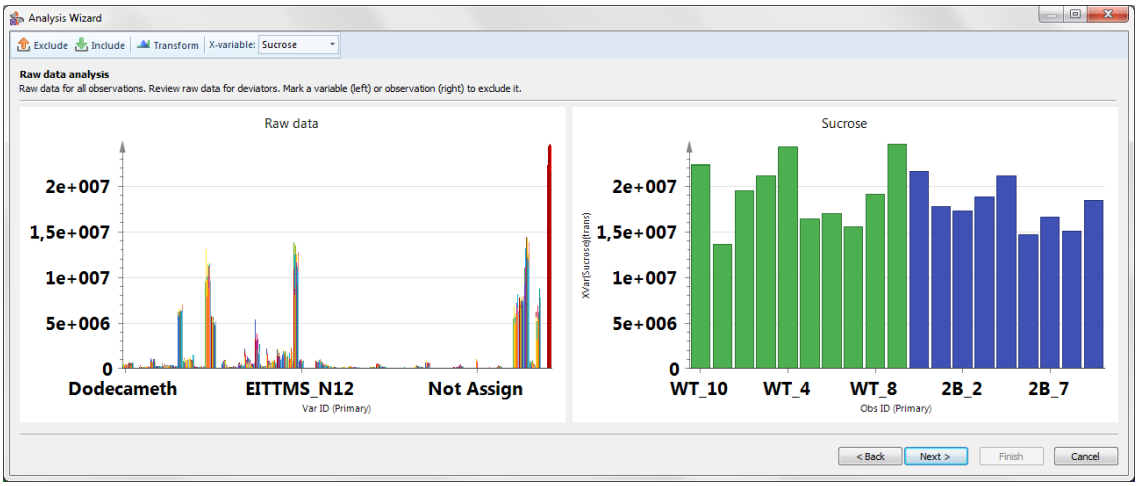


The Analysis Wizard must also be instructed which are the data analytical objectives and which groups of data are to be analyzed. The current objectives are to obtain an overview of the data (using principal components analysis, PCA) and the comparison of the two groups, i.e., wild type vs up-regulated plants (using orthogonal partial least squares discriminant analysis, OPLS-DA). As seen below, the wild type plants are assigned to the Reference group and the up-regulated plants to Group 1. Click Next to proceed.



Raw data analysis

The next stage in the Wizard involves evaluating properties of the raw data. Both variables and observations can be examined and, if needed, excluded prior to the PCA and OPLS-DA modeling. Variables in need of transformations can be identified and transformed. Click Next to proceed.

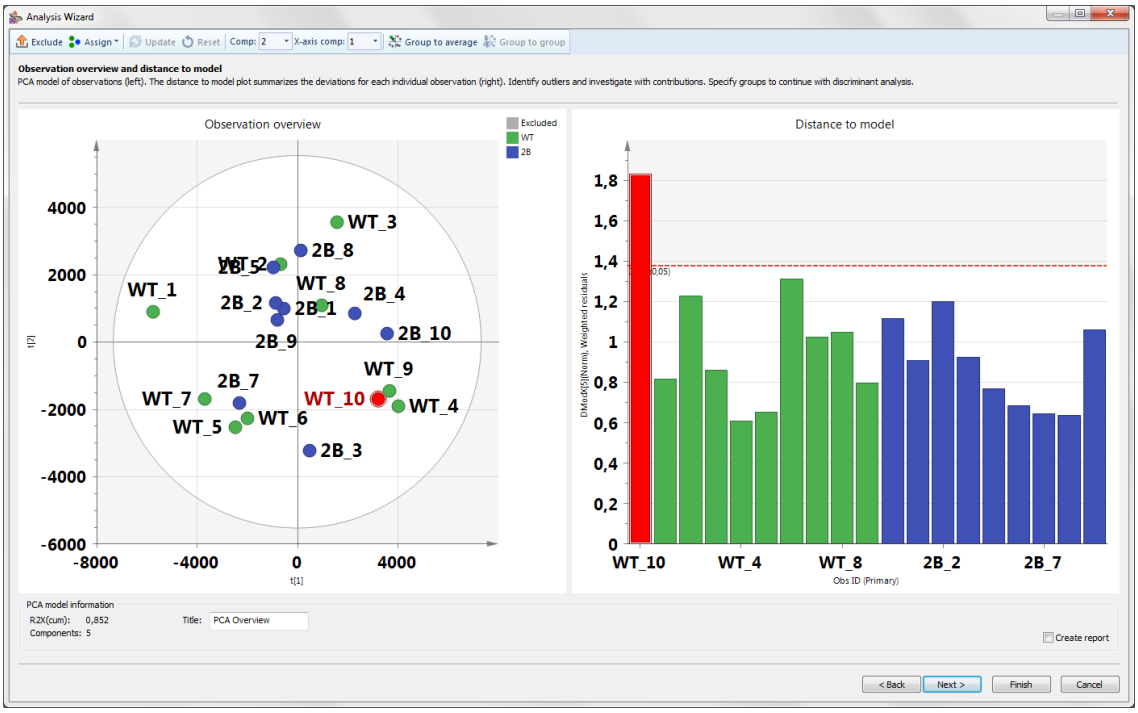


Overview of data using PCA

The first model is a PCA model. Its main purpose is to provide a quick overview of the information in the omics data. Of special interest is to investigate whether there are groups, trends, jumps and outliers in the data. The Observation overview (score plot) and the Distance to model (DModX plot) are color-coded according to the known information on the current group assignments.

Outliers can be detected both in the score plot (strong outliers) and the Distance to model plot (weaker outliers). However, if the raw data analysis (foregoing step in the wizard) did not indicate any strange observations, it is unlikely there will remain outliers in the score plot. Outliers may still show up in the Distance to model plot, however, and when their DModX values are 2-3 times higher than the critical distance (the red dashed line) they should be carefully examined using contribution plots (see below).

In addition to identifying possible outliers and interpreting them, this phase also allows specifying new (or modifying existing) groups for the ensuing discriminant analysis step.

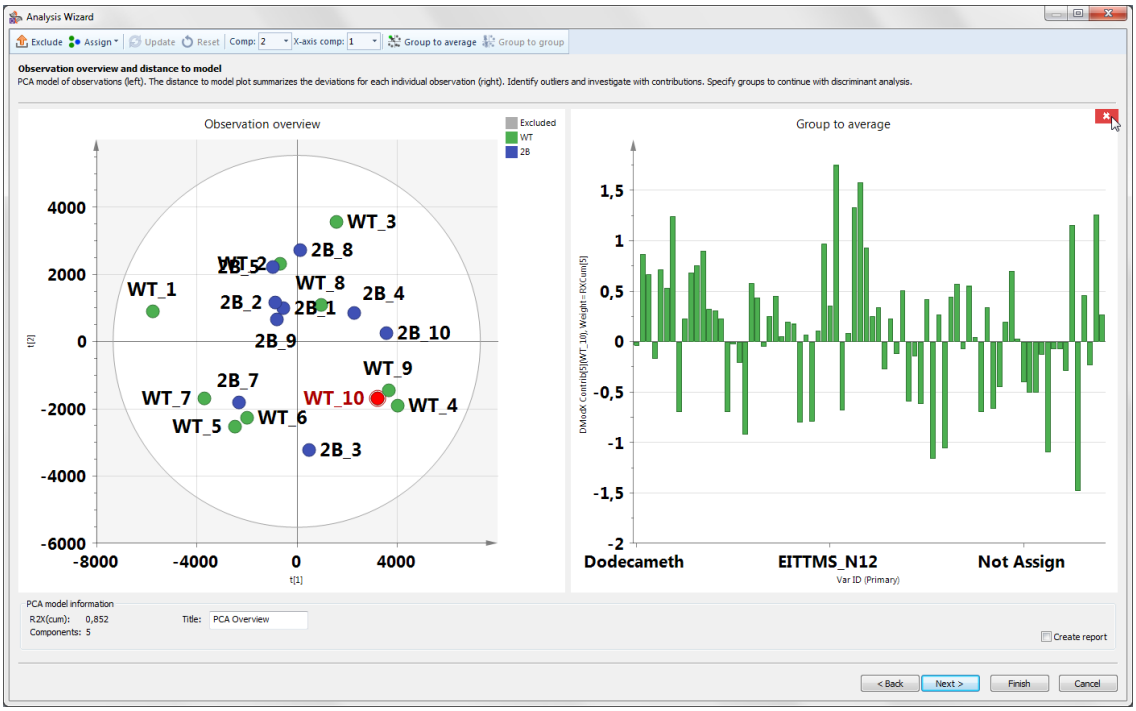


The Distance to model plot suggests a deviation from the PCA model for one of the samples of the Reference group (WT_10). When this sample is highlighted in the Distance to model plot its corresponding position in the Observation overview is marked. The highlighted wild type sample is inside the model in the Observation overview, but shows a slightly

different pattern in the residuals. This deviation is not large or critical. One can expect to have one of 20 observations outside the 95 % limit.

To understand why a sample deviates in the Distance to model plot a contribution plot is used. By double-clicking that sample in the Distance to model plot a contribution plot will open (see next screenshot).

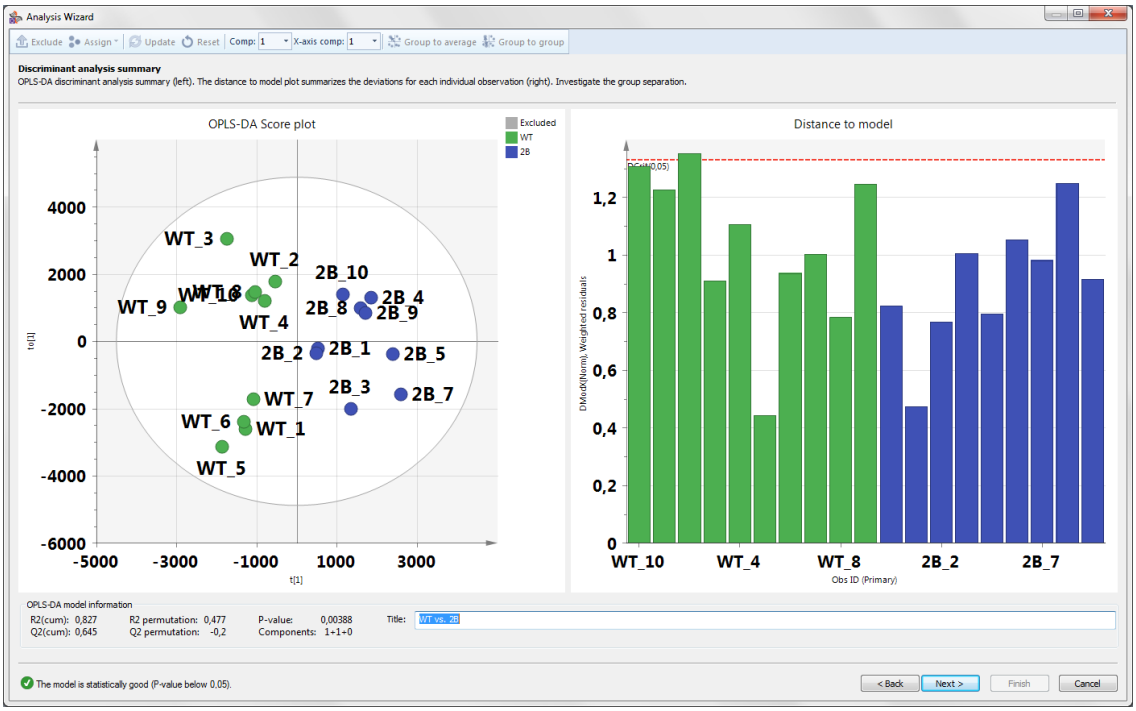
The Distance to model contribution plot (below, right) displays which variables contribute to the deviation of the WT_10 sample. For a variable where the bar is close to zero there is little deviation. Conversely, for a variable where the bar is large (in absolute value) the deviation from the model is more pronounced. As a rule of thumb contributions with absolute value above 3 should be investigated. The highest value for WT_10 is 1.75 why the conclusion is that the observation has no variable with critical deviation from the rest of the dataset. Click Next to proceed.



Group comparison using OPLS-DA

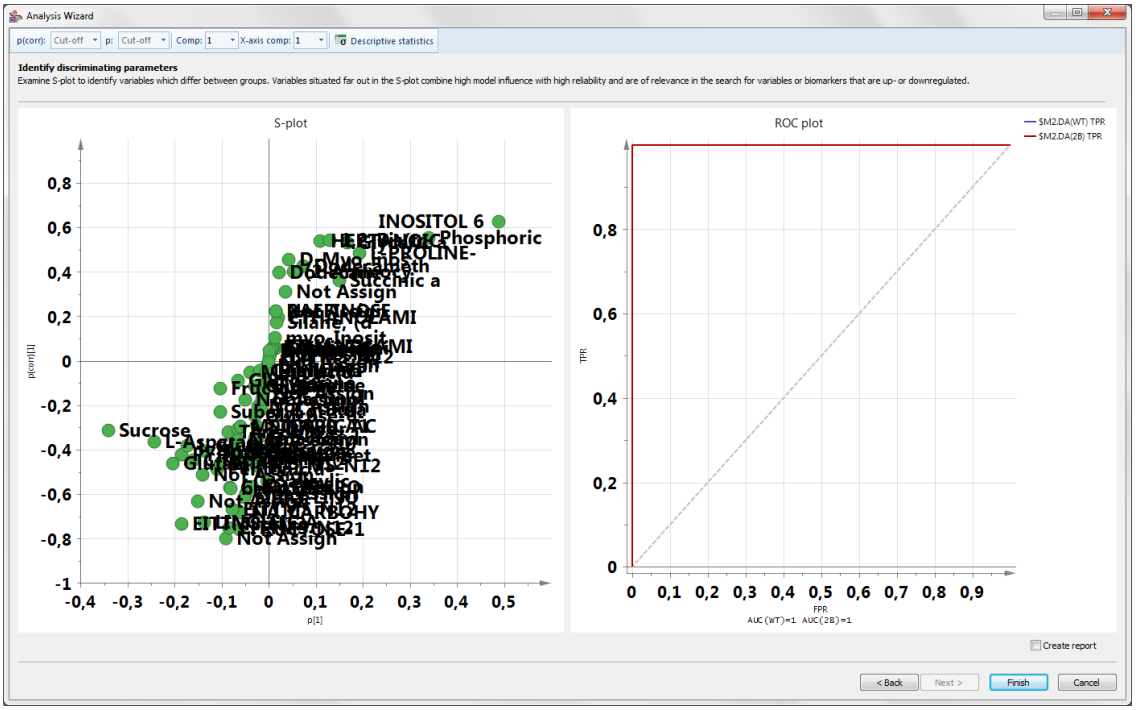
Provided that two groups of observations have been defined, the last stages in the Analysis Wizard correspond to computing and interpreting an OPLS-DA model comparing the two groups. Similarly to the previous PCA model, the first two plots in the wizard are the OPLS-DA score and Observation overview plots. As shown by the score plot, the two classes are completely resolved from one another. Moreover, the Observation overview plot demonstrates that all samples fit the OPLS-DA model well, i.e., there are no outliers.

In the lower left-hand corner of the Analysis Wizard window performance indicators of the OPLS-DA model are given. The obtained model is strongly significant. Additional information on how to interpret these performance indicators are given in the Appendix of this tutorial. Click Next to proceed.

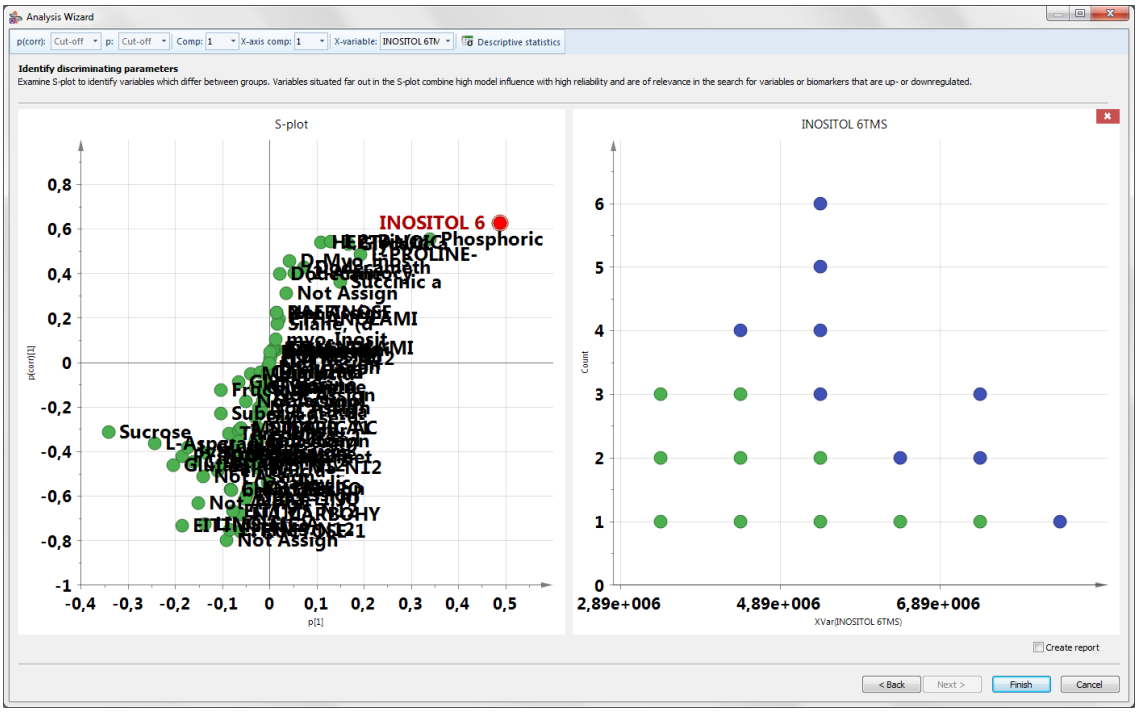


In order to identify which variables are the best discriminators between the two groups we consult the S-plot. Variables situated far out on the "wings" of the S-plot are the best discriminators. They combine high model impact with high reliability.

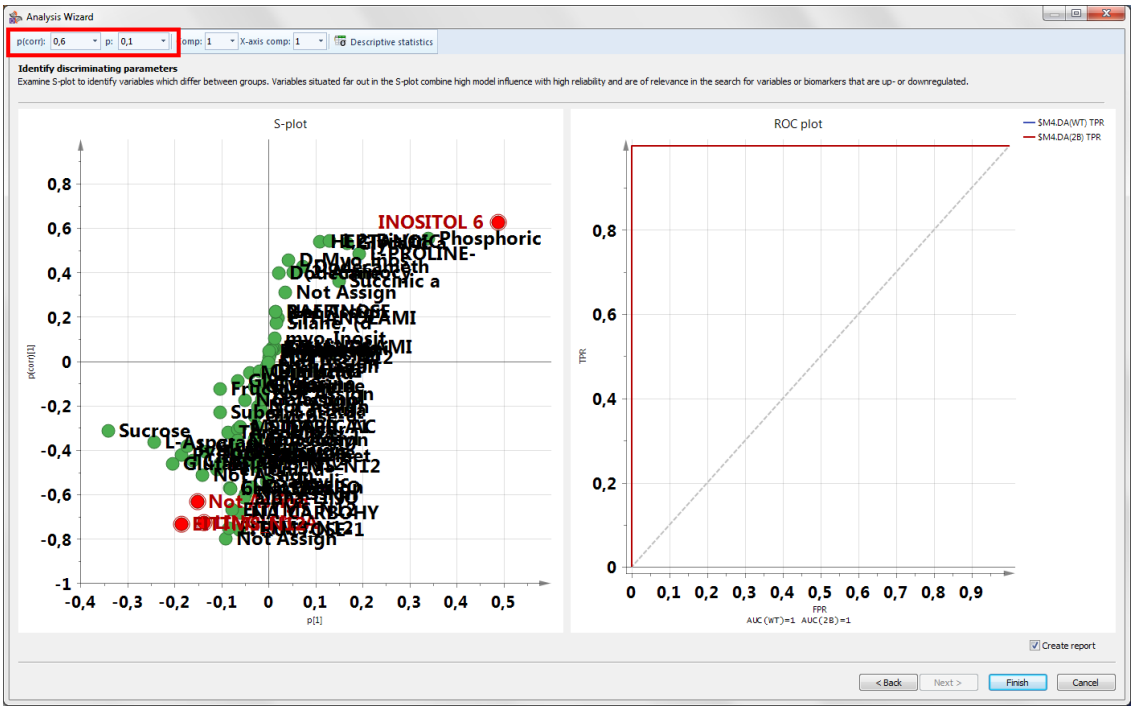
Alongside the S-plot a ROC (receiver operating characteristic) plot is provided. This plot is a graphical summary of the performance of a binary classifier (i.e., the OPLS-DA model). A classifier with a perfect discrimination has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Conversely, a ROC curve close to the 1:1 diagonal represents a very poor classifier. In the current example, perfect discrimination between the two groups is obtained.



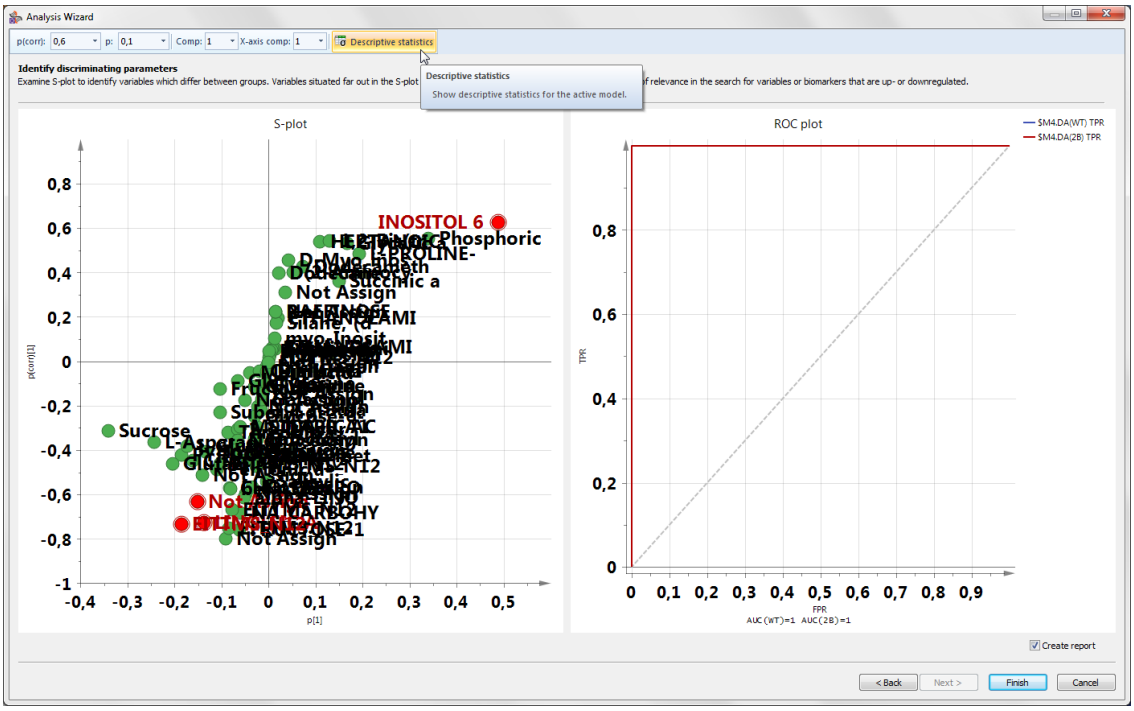
We can see that the S-plot highlights the influence of *INOSITOL 6TMS*. By clicking this point, or any other interesting variable, a dot plot of the raw data is created, which is color-coded according to the groups. The dot plot is an alternative to the histogram plot and shows the distribution of a vector, and also indicates group separation.



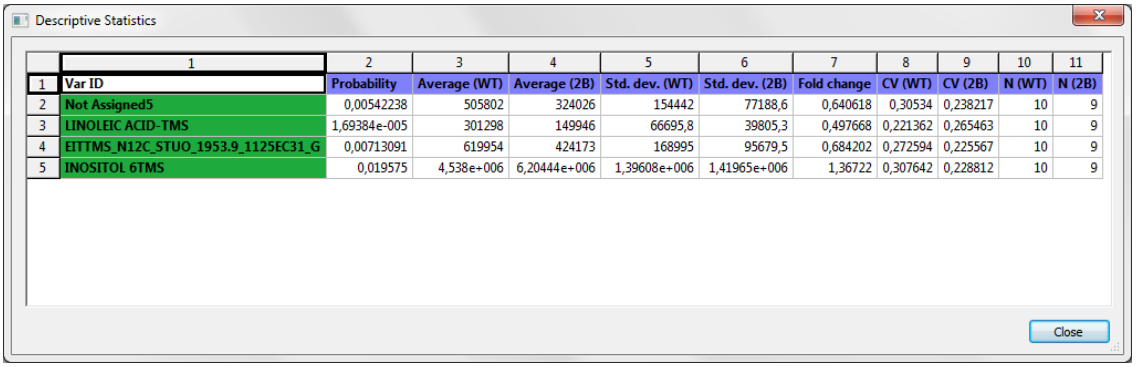
Moreover, instead of manual marking in the S-plot, automatic highlighting can be accomplished by specifying cut-off limits in the p(corr) and p fields. Use the arrows to set the desired cut-offs.



The Descriptive statistics is another tool that can be used to investigate raw data and the relevance of the discriminating variables uncovered by the S-plot.

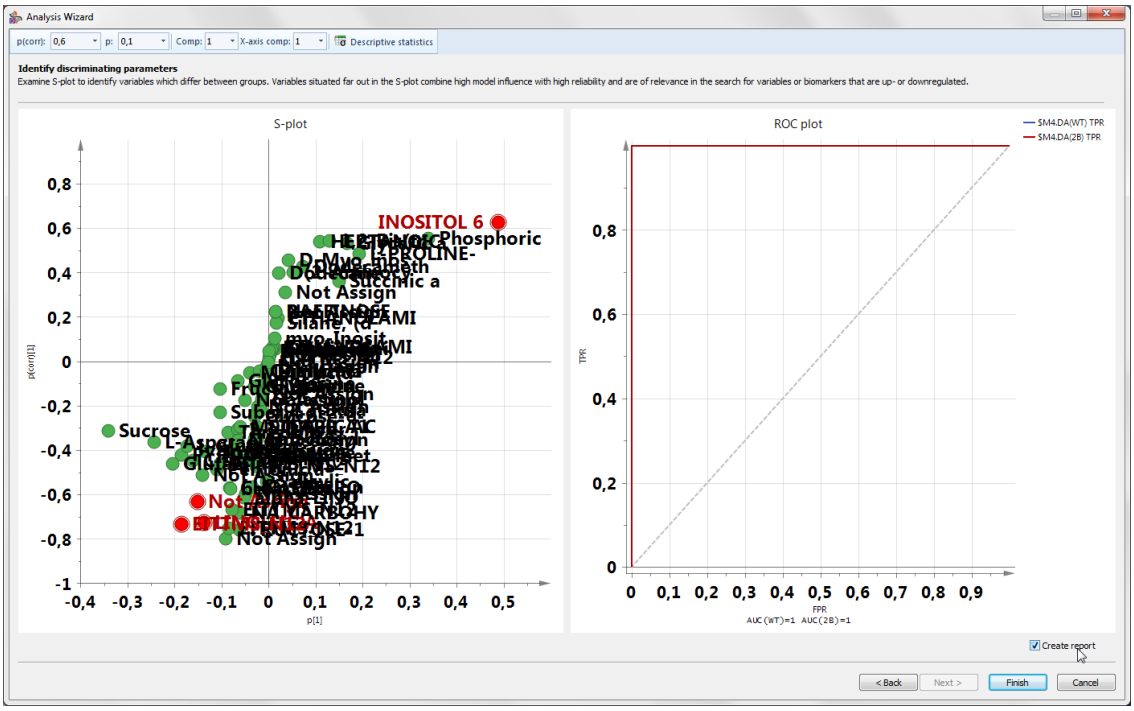


The contents of the Descriptive Statistics table is configurable. It may report up to 10 univariate statistics for all or a subset of the variables in the OPLS-DA model. Whether all or a subset of the variables are displayed in this table depends on the marking in the S-plot; with no marking all variables are displayed.

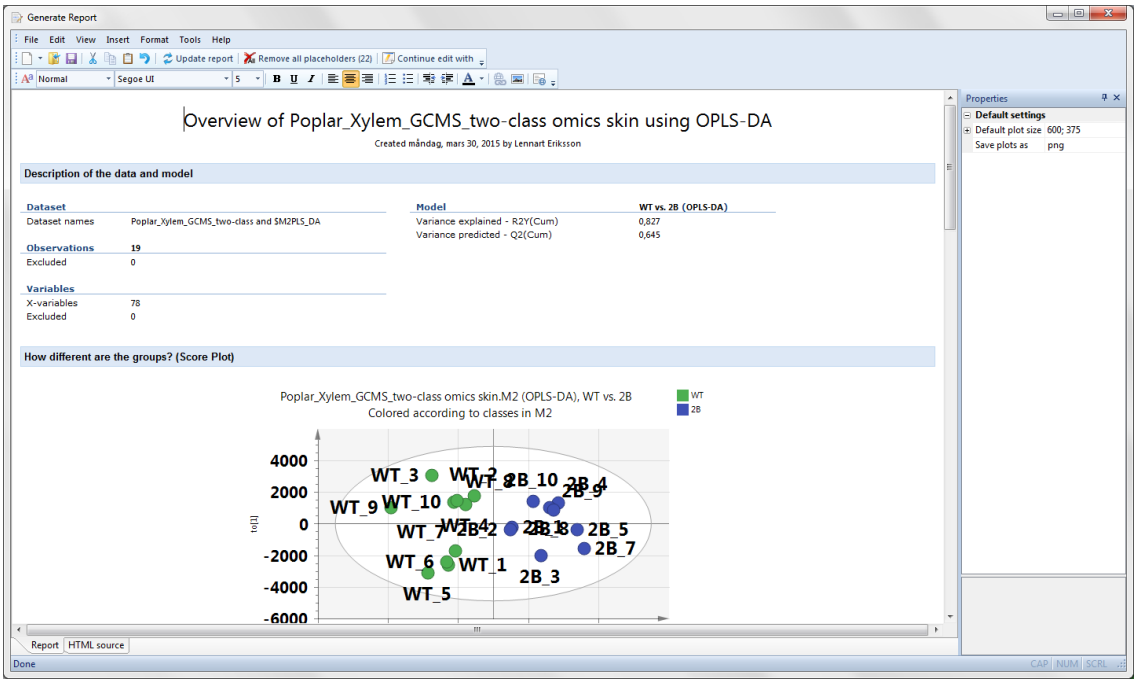


	1	2	3	4	5	6	7	8	9	10	11
1	Var ID	Probability	Average (WT)	Average (2B)	Std. dev. (WT)	Std. dev. (2B)	Fold change	CV (WT)	CV (2B)	N (WT)	N (2B)
2	Not Assigned5	0,00542238	505802	324026	154442	77188,6	0,640618	0,30534	0,238217	10	9
3	LINOLEIC ACID-TMS	1,69384e-005	301298	149946	66695,8	39805,3	0,497668	0,221362	0,265463	10	9
4	EITTMS_N12C_STUO_1953.9_1125EC31_G	0,00713091	619954	424173	168995	95679,5	0,684202	0,272594	0,225567	10	9
5	INOSITOL 6TMS	0,019575	4,538e+006	6,20444e+006	1,39608e+006	1,41965e+006	1,36722	0,307642	0,228812	10	9

As a final step before closing the Analysis Wizard a report can be automatically created by leaving the Create report check box selected. This will launch a separate window where the report is visible and editable. Click Finish to proceed.



The report that is created summarizes the main findings from the analysis of the data. This report is an HTML document that can be opened in regular word processors for further editing.

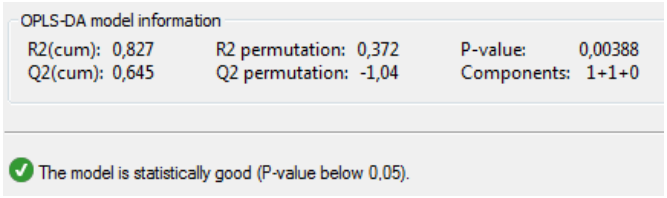


Conclusions

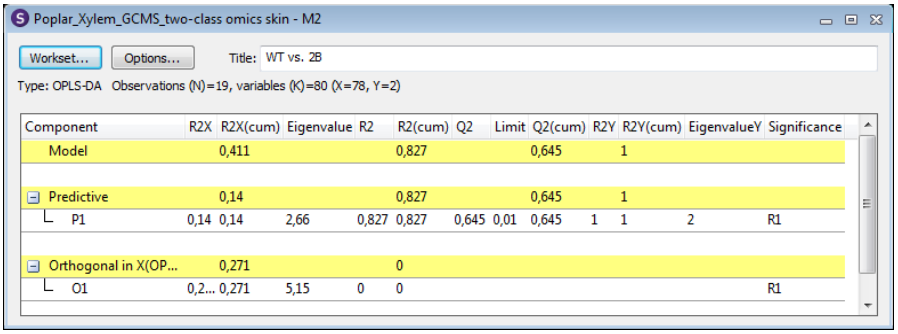
This tutorial shows how to easily create overview and group discrimination models from omics data using the Omics skin in SIMCA. The skin, in terms of its Analysis Wizard, ensures that the default scaling and plots in SIMCA are suited for omics data and discovery of putative biomarkers/discriminating variables. The report tool with its omics specific templates facilitates documentation of both PCA and OPLS-DA models.

Appendix – Performance indicators of the OPLS-DA model

The Analysis Wizard provides some quantitative performance indicators of the reliability of the obtained OPLS-DA model. These indicators are briefly explained here.

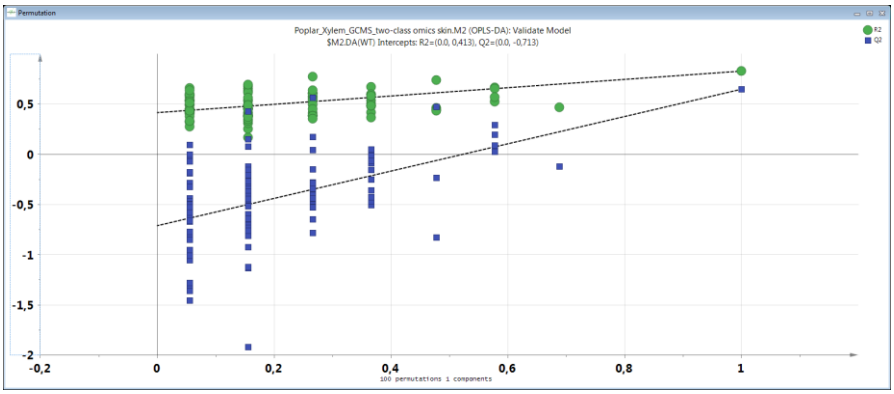


The R2(cum) and Q2(cum) values quantify explained and predicted variances of Y and are taken from the model window. The notation 1+1+0 components means that one predictive and one orthogonal component make up the current OPLS-DA model.



Component	R2X	R2X(cum)	Eigenvalue	R2	R2(cum)	Q2	Limit	Q2(cum)	R2Y	R2Y(cum)	EigenvalueY	Significance
Model		0,411			0,827			0,645		1		
Predictive		0,14			0,827			0,645		1		
P1	0,14	0,14	2,66	0,827	0,827	0,645	0,01	0,645	1	1	2	R1
Orthogonal in X(OP...		0,271			0							
O1	0,2...	0,271	5,15	0	0							R1

The R2 permutation and Q2 permutation numbers are obtained from the response permutation test (Home | Permutations). They suggest the level of R2 and Q2 that can be accomplished for the current example based on fitting the omics data (X) to random class assignment data (Y). The values that are reported are the intercepts of the two regression lines. As a rule-of-thumb, for an OPLS-DA model to pass this test, the Q2 intercept should be < 0.05.



Finally, the p-value that is reported is derived from an analysis of variance (ANOVA) of the cross-validation (CV) residuals (Home | CV-ANOVA). For an OPLS-DA model to be judged significant, the p-value should be smaller than 0.05.

CV-ANOVA [M2]

	1	2	3	4	5	6	7
1	M2(WT vs. 2B)	SS	DF	MS	F	p	SD
2	Total corr.	18	18	1			1
3	Regression	11,6185	4	2,90463	6,37234	0,00388457	1,7043
4	Residual	6,38147	14	0,455819			0,675144