



Bruker **Daltonics**



profileanalysis 2.0

- Statistical Processing of LC-MS Data –
Identification of Possible Biomarkers in Metabolomics

Statistics Made Easy in ProfileAnalysis

Nowadays, profiling complex samples using high-resolution chromatography and mass spectrometry is a routine task in many metabolomics laboratories.

An integral part of the process is the application of statistical methods to quickly pinpoint relevant information and generate knowledge. The full benefits of high-resolution separation techniques such as CE, GC or UHPLC can only be exploited in combination with the uncompromised performance of ultra-fast time-of-flight mass analyzers such as micrOTOF-Q II and maXis.

ProfileAnalysis enables smooth and easy analysis using fully unsupervised methods like Principal Component Analysis (PCA) and supervised methods such as student's t-test. It provides a complete set of tools for data pre-processing, in-depth statistical analysis, identification and feedback experiments.

ProfileAnalysis has been designed to give beginners a head start in metabolomics and also to meet the needs of expert users.

It's all in your data ...

It's not only statistics that play an important role in metabolomics applications. High-resolution, accurate mass and isotopic pattern data enable direct identification of possible biomarkers via their molecular formula using Bruker's unique SmartFormula algorithm.

Quick and easy searches of web-based databases with the CompoundCrawler facilitate structural assignments.

The scheduled precursor list (SPL)

provides intelligent design of an MS/MS experiment based on statistical analyses. The SPL can be created automatically from the t-test results using filtering options such as specific regulation ratios or p-values.

The hypothetical structure can be quickly verified by means of targeted MS/MS accurate mass measurements. Naturally, interesting loadings from PCA calculations can also serve as input for an SPL.



● Data Preparation

Data pre-processing for statistical analysis

The starting point for each statistical analysis is a series of LC-MS runs. Metabolic profiling data sets are typically too large to manually extract possible biomarkers. Find Molecular Features (FMF) is a peak finding algorithm for quantitatively pinpointing all relevant information from LC-MS analyses. Data reduction is achieved by efficiently differentiating real signals from noise. Retention-time alignment is performed by the pairwise comparison of an automatically assigned master run using a shifting vector algorithm that also takes non-linear retention-time shifts into account.

The alignment significantly improves data quality, and can be particularly important for long LC-MS runs.

Bucketing is the process of generating the data table for statistical analysis itself. ProfileAnalysis easily creates bucket tables of LC-MS data based on the extracted FMF compounds from raw or netCDF data. Different normalization and scaling options – such as Pareto scaling – complete the set of data pre-processing tools. All data pre-processing steps applied in the calculation of the bucket table form the basis for the subsequent statistical analysis and all parameters are easily defined in the ProfileAnalysis method.

Series of
LC-MS experiments

Find Molecular Features

Retention Time
Alignment

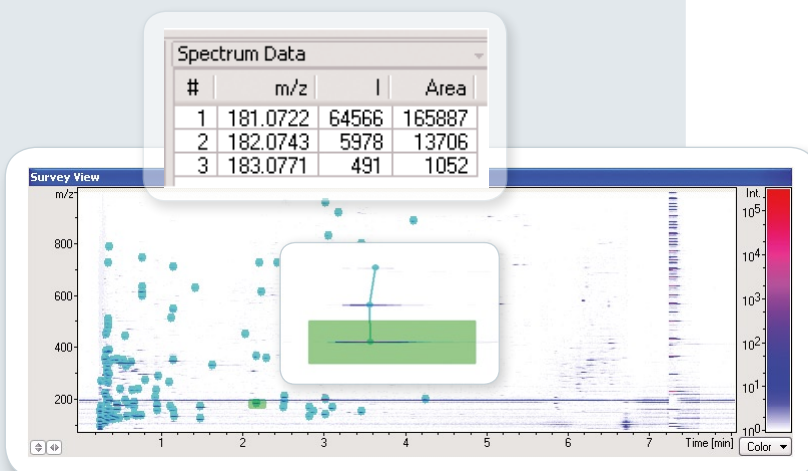
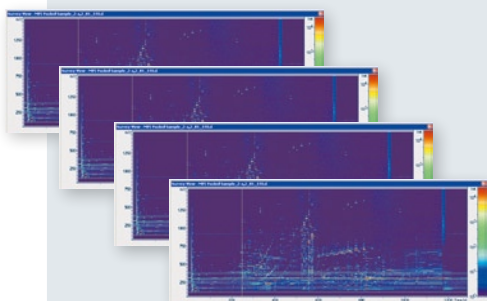
Bucketing

Scaling & Normalization

Statistical Analysis

Spectrum Data

#	m/z	I	Area
1	181.0722	64566	165887
2	182.0743	5978	13706
3	183.0771	491	1052



● Unsupervised Statistics

Principal Component Analysis (PCA)

As a non-supervised statistical method, PCA aims to find and rank variances in the data set and assists in the visualization of statistical patterns. ProfileAnalysis provides all the necessary tools for PCA calculation and visualization.

Direct access to individual compound information via the bucket table enables the fast identification of elemental compositions using SmartFormula. The structural assignment of possible biomarkers is enabled through public database queries based on correct molecular formula via the CompoundCrawler.

1. Project setup

TeaSampleTable.prof - ProfileAnalysis

Open Groups Calculate... Save Print... Options... Help

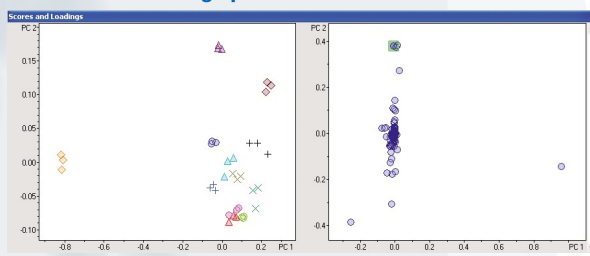
Sample Table	
	File Name
25	Tee_3_1-A_3_0
26	Tee_3_1-A_3_0
27	Tee_3_1-A_3_0
28	Tee_4_1-A_4_0
29	Tee_4_1-A_4_0
30	Tee_4_1-A_4_0
31	Tee_5_1-A_5_0
32	Tee_5_1-A_5_0
33	Tee_5_1-A_5_0
34	Tee_6_1-A_6_0
35	Tee_6_1-A_6_0
36	Tee_6_1-A_6_0

2. Bucket table

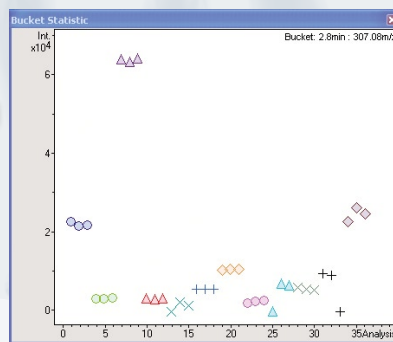
MS-PCA: 1

Bucket Table				
	Bucket	Include	TeeMix_1-B_7_01_100.d	TeeMi
224	2.7min: 741.20m/z	<input checked="" type="checkbox"/>	749.45796	
225	2.7min: 453.14m/z	<input checked="" type="checkbox"/>	760.30003	
226	2.8min: 778.16m/z	<input checked="" type="checkbox"/>	752.16848	
227	2.8min: 761.13m/z	<input checked="" type="checkbox"/>	1142.48293	
228	2.8min: 597.14m/z	<input checked="" type="checkbox"/>	3961.42065	
229	2.8min: 333.06m/z	<input checked="" type="checkbox"/>	4746.11533	
230	2.8min: 613.15m/z	<input checked="" type="checkbox"/>	2874.50332	
231	2.8min: 307.08m/z	<input checked="" type="checkbox"/>	22837.46130	
232	2.8min: 649.21m/z	<input checked="" type="checkbox"/>	1563.96833	
233	2.8min: 360.17m/z	<input checked="" type="checkbox"/>	832.12873	

3. Scores & loadings plot



4. Bucket statistics plot



6. Structural assignment

CompoundCrawler: DRUG-CC-981207, Round 6 Group

Search for compounds by formula (M) CH16O15 or (M+2) CH16O15, respectively, in my reference table (v. 3.1.3.3)

#	Compound	Formula	Weight	Charge
1	OH	C16H16O15	448.0	0
2	OH	C16H15O15	447.0	0
3	OH	C16H14O15	446.0	0
4	OH	C16H13O15	445.0	0
5	OH	C16H12O15	444.0	0
6	OH	C16H11O15	443.0	0
7	OH	C16H10O15	442.0	0
8	OH	C16H9O15	441.0	0
9	OH	C16H8O15	440.0	0
10	OH	C16H7O15	439.0	0
11	OH	C16H6O15	438.0	0
12	OH	C16H5O15	437.0	0
13	OH	C16H4O15	436.0	0
14	OH	C16H3O15	435.0	0
15	OH	C16H2O15	434.0	0
16	OH	C16HO15	433.0	0
17	OH	C16O15	432.0	0

5. Molecular formula generation

SmartFormula

Analysis: Tee_11_1-B_3_01_95.d

Bucket: 2.8min: 307.08m/z

Min: C₉ Max: C₉n

Generate Copy Save Results... Help

Note: for m < 2000 the elements C, H, N, and O are considered implicitly.

Compound: Compound 1

Measured m/z	Tolerance	mDa	Charge
307.08116695	1	mDa	1

#	Mol. Formula	m/z	err (mDa)	err (ppm)	mSigma	rb	N rule	e ⁻
1	C ₁₅ H ₁₅ O ₇	307.081	0.06	0.203	1.5	8.5	ok	even

Automatically locate monoisotopic peak. Maximum number of formulae: 200

Check rings plus double bonds. Minimum: 0 Maximum: 0

Electron configuration: even

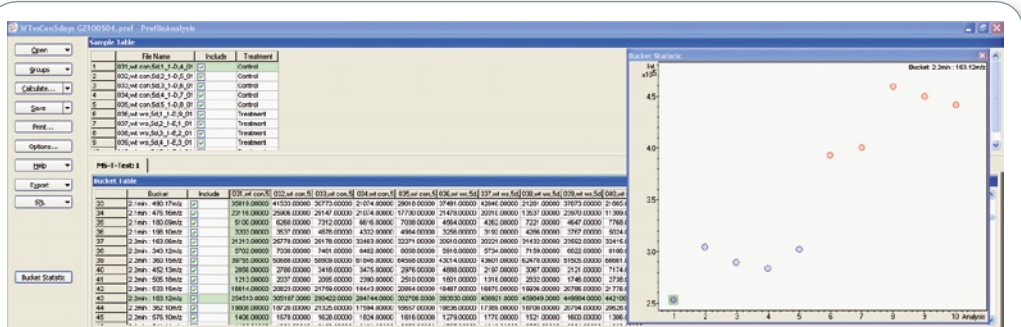
Filter H/C element ratio. Minimum H/C: 0 Maximum H/C: 5

Estimate carbon number

The ProfileAnalysis workflow starts with the setup of a new project (1): The core of the statistical calculation is the bucket table (2) for which the PCA is calculated and visualized in scores and loadings plots (3). The intensity distribution of a bucket for all analyses is quickly inspected in the bucket statistics plot (4) before a molecular formula is calculated by SmartFormula (5). Structural assignments are provided by web-based searches using the CompoundCrawler (6).

Supervised Statistics

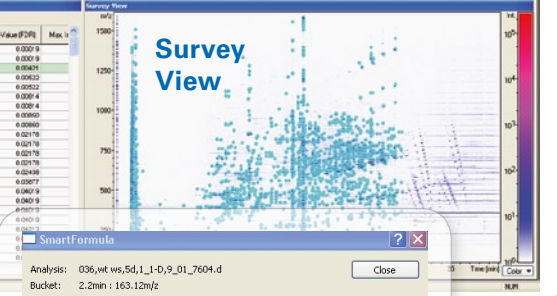
Sample table



Bucket table

Bucket	PValue	Average Ratio Control/Treatment	Fold Change Control/Treatment	PValue (FWER)	PValue (FDR)	Max. y
1	0.00000	0.23	-4.38	0.0026	0.0029	1500
2	0.00000	0.17	-5.97	0.0031	0.0039	1400
3	0.00005	0.17	-4.49	0.0136	0.0161	1300
4	0.00000	0.11	-9.24	0.0022	0.0027	1200
5	0.00011	0.30	-3.29	0.0265	0.0322	1100
6	0.00021	0.11	-9.24	0.0528	0.0644	1000
7	0.00023	0.32	-4.60	0.0701	0.0854	900
8	0.00028	0.41	-3.42	0.0822	0.1000	800
9	0.00028	1.87	1.87	0.0861	0.1049	700
10	0.00087	0.29	-3.51	0.21874	0.26178	600
11	0.00102	0.34	-2.87	0.25746	0.31278	500
12	0.00104	0.81	-1.65	0.38166	0.46278	400
13	0.00117	0.19	-15.14	0.39112	0.47878	300
14	0.00136	0.10	-10.48	0.4138	0.50488	200
15	0.00133	1.50	1.80	0.3838	0.46877	100
16	0.00139	0.16	-6.87	0.47746	0.58079	50
17	0.00142	0.31	-1.86	0.58886	0.71879	20
18	0.00159	0.88	-2.88	0.78871	0.95719	10
19	0.00160	0.48	-1.81	0.7886	0.95719	5
20	0.00164	0.29	-3.40	0.84351	1.03219	2
21	0.00166	0.27	-3.23	0.87221	1.06219	1
22	0.00168	0.24	-4.20	0.87728	1.06719	0.5
23	0.00182	1.74	1.74	0.98851	1.20819	0.2

t-test Result table



t-test & SmartFormula

Bucket	PValue	Average Ratio Control/Treatment	Fold Change Control/Treatment	PValue (FWER)	PValue (FDR)
1	0.00000	0.23	-4.38	0.0026	0.0029
2	0.00000	0.17	-5.97	0.0031	0.0039
3	0.00005	0.17	-4.49	0.0136	0.0161
4	0.00000	0.11	-9.24	0.0022	0.0027
5	0.00011	0.30	-3.29	0.0265	0.0322
6	0.00021	0.11	-9.24	0.0528	0.0644
7	0.00023	0.32	-4.60	0.0701	0.0854
8	0.00028	0.41	-3.42	0.0822	0.1000
9	0.00028	1.87	1.87	0.0861	0.1049
10	0.00087	0.29	-3.51	0.21874	0.26178
11	0.00102	0.34	-2.87	0.25746	0.31278
12	0.00104	0.81	-1.65	0.38166	0.46278
13	0.00117	0.19	-15.14	0.39112	0.47878
14	0.00136	0.10	-10.48	0.4138	0.50488
15	0.00133	1.50	1.80	0.3838	0.46877
16	0.00139	0.16	-6.87	0.47746	0.58079
17	0.00142	0.31	-1.86	0.58886	0.71879
18	0.00159	0.88	-2.88	0.78871	0.95719
19	0.00160	0.48	-1.81	0.7886	0.95719
20	0.00164	0.29	-3.40	0.84351	1.03219
21	0.00166	0.27	-3.23	0.87221	1.06219
22	0.00168	0.24	-4.20	0.87728	1.06719
23	0.00182	1.74	1.74	0.98851	1.20819

Analysis: 036.wt vs. 5d_1_1-1_0_9_01_7604.d
Bucket: 2.2min : 163.12m/z

Min: C₉
Max: C₅-n

Note: for m < 2000 the elements C, H, N, and O are considered implicitly.

Compound: Compound 1

Measured m/z: 163.123022791

#	Mol. Formula	m/z	err [mDa]	err [ppm]	nSigns	rdB	N rule	e*
1	C ₁₀ H ₁₅ N ₂	163.123	-0.05	-0.294	4.3	4.5	ok	even

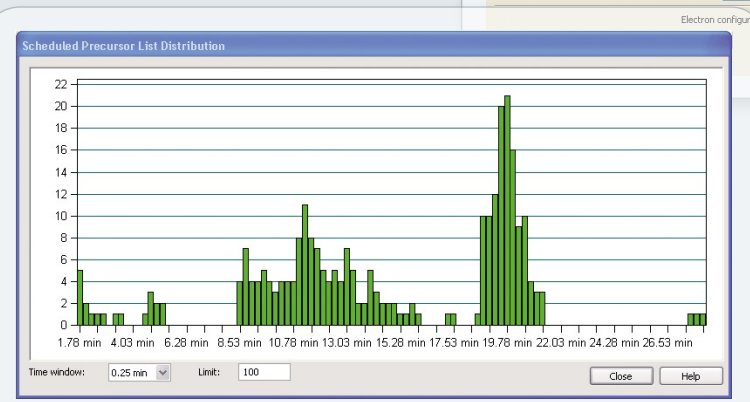
Automatically locate monoisotopic peak. Maximum number of formulae: 200

Check rings plus double bonds. Minimum: 0. Maximum: 80

Electron configuration: even

Maximum H/C: 3

SPL Distribution



The student's t-test or ANOVA (Analysis of Variance) is a standard hypothesis test to evaluate whether the mean intensity values of two or more groups of samples are statistically different from each other. The calculated t-test results comprise not only p-values, but also fold-change, average ratio or intensity information. These values allow for easy selection and quick identification of relevant differences using SmartFormula. Additionally, the t-test results can be filtered to create tailor-made scheduled precursor lists (SPLs) for dedicated MS/MS experiments. The SPL distribution plot visualizes the number of automatically selected precursor ions across the chromatographic time axis.

Technical Specifications

Data

- Processing of Bruker LC-MS data and netCDF-files
- Direct link to original data file
- Export of Bucket Table to ASCII and text format compatible to MatLab™, R™, SIMCA-P™
- Export of quantitative results to ProteinScape

Data processing

- Find Molecular Features (FMF) algorithm
- Processing of Find Molecular Features results calculated in DataAnalysis 4.0
- Recalibration of mass axis
- Retention Time Alignment
- Spectral Background Subtraction
- Rectangular and advanced bucketing
- Bucket filtering
- Various normalization algorithms
- Various scaling options available for Principal Component Analysis (PCA)
- Methods for data processing

Statistics

- Principle Component Analysis (PCA)
- Statistical plots for PCA review (e.g. Scores, Loadings, Influence, Hotelling's T²)
- Cross Validation, Test Set Validation
- Classification based on PCA models
- Student's t-test and ANOVA

Other features

- SmartFormula: Molecular Formula Generation
- Survey View: Graphical overview of data distribution
- Bucket Density view
- Automatic Scheduled Precursor List (SPL) generation from t-test results
- Installation Qualification
- Detailed Manual
- Tutorial Data

For research use only. Not for use in diagnostic procedures.



maXis or microTOF-Q II and Dionex UltiMate 3000 Rapid Separation LC-System (RSLC) are ideal for fast separation and high performance mass spectrometry

www.bdal.com

● **Bruker Daltonik GmbH**

Bremen · Germany
Phone +49 (0)421-2205-0
Fax +49 (0)421-2205-103
sales@bdal.de

Bruker Daltonics Inc.

Billerica, MA · USA
Phone +1 (978) 663-3660
Fax +1 (978) 667-5993
ms-sales@bdal.com