



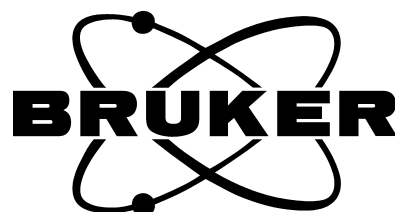
OPUS

Spectroscopy Software

Version 6

User Manual

IDENT



© 2006 BRUKER OPTIK GmbH, Rudolf-Plank-Straße 27, D-76275 Ettlingen, www.brukeroptics.com

All rights reserved. No part of this manual may be reproduced or transmitted in any form or by any means including printing, photocopying, microfilm, electronic systems etc. without our prior written permission. Brand names, registered trademarks etc. used in this manual, even if not explicitly marked as such, are not to be considered unprotected by trademarks law. They are the property of their respective owner.

The following publication has been worked out with utmost care. However, Bruker Optik GmbH does not accept any liability for the correctness of the information. Bruker Optik GmbH reserves the right to make changes to the products described in this manual without notice.

This manual is the original documentation for the OPUS spectroscopic software.

Table of Contents

- About OPUS IDENT1
- Introduction1
- 1 Setting Up an Identity Test Method3**
 - 1.1 Loading Existing Method 3
 - 1.1.1 Methods created by prior OPUS releases 6
 - 1.2 Creating New Method 6
 - 1.3 Setting Parameters 10
 - 1.4 Identity Test Limit 13
 - 1.5 Validating Library 14
 - 1.5.1 Validation Report 15
 - 1.6 Storing Method Files 16
- 2 Performing an IDENT Analysis19**
- 3 IDENT Report21**
 - 3.1 Identity Test Reports 21
 - 3.1.1 Standard Method 21
 - 3.1.2 Factorization Method 23
- 4 Cluster Analysis25**
 - 4.1 Theory 25
 - 4.1.1 Methods to Calculate Spectral Distances 28
 - 4.1.2 Cluster Algorithms 29
 - 4.2 Performing a Cluster Analysis 31
 - 4.3 3D Files/Filelist 37
- 5 Conformity Test39**
 - 5.1 Setting up Conformity Test 39
 - 5.2 Performing Conformity Test 48
- 6 IDENT Theory51**
 - 6.1 Algorithms 51
 - 6.1.1 Standard Method 51
 - 6.1.2 Factorization 53
 - 6.2 Factorization Theory 56
 - 6.2.1 Scaling to 1st Range and Normalize to Replevel 59
 - 6.3 Data Preprocessing 61
 - 6.3.1 Vector Normalization 61
 - 6.4 Determining Threshold Value for Identity Test 67

6.5	Identity Test	68
6.6	Class Test	70
6.7	Validation	74
7	Reference Section	77
7.1	Setup Identity Test Method - Load Method	77
7.2	Setup Identity Test Method - Reference Spectra	78
7.2.1	Sorting Reference Spectra	79
7.2.2	Missing Reference Spectra	79
7.2.3	Options	81
7.2.4	Set Sub Library	81
7.2.5	Assign Classes	83
7.3	Setup Identity Test Method - Parameters	84
7.3.1	Preprocessing	84
7.3.2	Regions	85
7.3.3	Interactive Region Selection	85
7.3.4	Clear Selected Regions	87
7.3.5	Method	87
7.3.6	Calculate Thresholds	88
7.4	Setup Identity Test Method – Threshold	90
7.4.1	Maximum Hit + X*SDev	90
7.4.2	Confidence Level	90
7.4.3	Set	91
7.4.4	Group Statistics	91
7.5	Setup Identity Test Method – Validate	93
7.5.1	Validation Report	94
7.5.2	Print	99
7.6	Setup Identity Test Method - Store Method	100
7.7	Identity Test	101
7.7.1	No Reference Defined	103
7.8	Cluster Analysis – Load Method	105
7.8.1	Load Method	105
7.8.2	General Information	106
7.9	Cluster Analysis – Reference Spectra	106
7.10	Cluster Analysis – Parameters	107
7.10.1	Preprocessing	107
7.10.2	Regions	107
7.10.3	Method	108
7.10.4	Calculate Distances	108
7.11	Cluster Analysis – Report	109
7.11.1	Score Plot	111
7.11.2	Options	112
7.12	3D File/Filelist	114
7.12.1	File List	119
7.13	Cluster Analysis – Store Method	120

About OPUS IDENT

This manual consists of two parts. The first part describes how to create a user-defined reference library, and generate an IDENT method. Apart from this the IDENT analysis and the resulting report files are explained in detail, as well as the theory of the IDENT software test routines. The cluster analysis and conformity test are described in a separate chapter.

The reference section refers to all IDENT functions and supports you if you have questions about the IDENT functionalities, or problems while using IDENT.

Introduction

OPUS IDENT (in the following referred to as IDENT) is a software package designed to identify substances by their IR spectra. An *unknown* spectrum (in the following called test spectrum) is directly compared to reference spectra of a library. IDENT identifies those reference spectra which are closest equivalent to the test spectrum, and determines the deviations between these spectra and the test spectrum. This allows IDENT to identify unknown substances, e.g. polymers, and to evaluate the conformity degree of a substance with a reference standard. The latter application is a typical task found in quality control.

To perform an identity test you first need to have a reference library which you compare the test spectrum with. If no suitable library exists, you have to measure a set of reference spectra, i.e. at least one spectrum per substance. However, it is recommended to measure several batches of the same substance to enable the program to get more information on possible allowed variations. Before, the samples required have to be classified as *identified* using common reference analytics.

If you have already measured a spectrum which you want to identify, the next step will be to generate an IDENT method. To perform an identity test you can select several algorithms and define identification parameters by using the IDENT software.

The results of an identity test are stored in a report which includes the analysis result, the method used as well as the parameters defined.

1 Setting Up an Identity Test Method

This chapter describes step by step how to set up a reference spectra library, and how to generate an IDENT method. You can perform an analysis using this IDENT method, and the analysis results will be displayed in the form of a report.

To perform an identity test the following steps are required:

- 1) Measuring at least 1 reference spectrum per substance
- 2) Incorporating the reference spectra into a library (the spectra need to belong to batches of one single substance which have been *identified* by means of conventional analytics)
- 3) Defining a suitable spectral range for identification
- 4) Selecting a data preprocessing method
- 5) Generating an IDENT method
- 6) Measuring test spectra
- 7) Analyzing test spectra

This chapter assumes theoretical basic knowledge, and therefore only briefly outlines the methods involved. The theory of the identity test method is explained in chapter 6.

If you generate an IDENT method, a main spectra library (IP1) will be created. All sub-libraries (e.g. IP2, IP3, IP4 etc.) which are related to this reference library will be stored in the same method file which uses the extension **FAA*.

The data structure of the complete library is displayed in a kind of browser window, and OPUS ensures a common (internal) validation of the entire library. For further details, see chapter 7.

1.1 Loading Existing Method

Start the IDENT software from the OPUS *Evaluate* menu. Select the *Setup Identity Test Method* command.

Setting Up an Identity Test Method

If you want to load an existing method, click on the *Load Method* button and select the respective method.

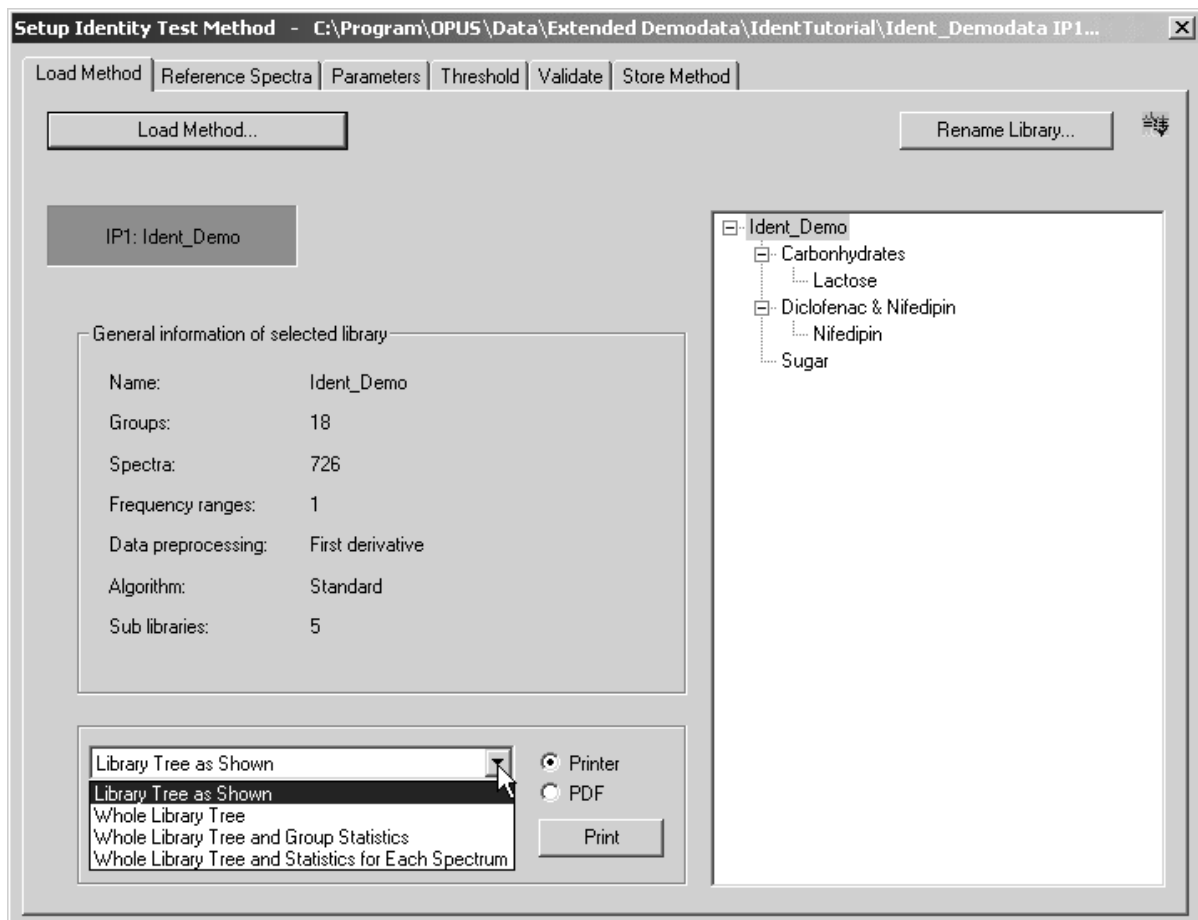


Figure 1: Setup Identity Test Method - Load Method tab (existing method)

Figure 1 exemplifies the exact description of the main library selected, including the data structure and the specific parameters in the *General Information of Selected Library* group field.

The name of the main library (*Ident_Demo*) is shown in the blue indicator field on the upper left side. If you click on one of the sub-libraries in the browser window, i.e. if you move to a different library level, the indicator field changes its color and description. In general, you can define as many sub-libraries as you like. Each library level will have a different color.

- IP1: Blue
- IP2: Green
- IP3: Yellow
- IP4: Orange
- IP5: Pink
- IP6: Magenta

If you defined IP7 as the next library level, the color cycle would start from the beginning, i.e. IP7 would be blue.

The library level you are currently working with is always displayed on the upper left side of each property tab, except for the *Store Method* tab. If you use a method which has not been properly calculated, an error message pops up and the library levels not yet re-calculated will be displayed in bold in the browser window.

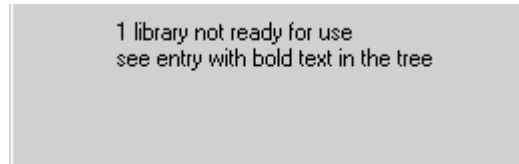


Figure 2: Error message on the Load Method tab

To rename the library description, click on the *Rename Library* button. The following menu pops up:

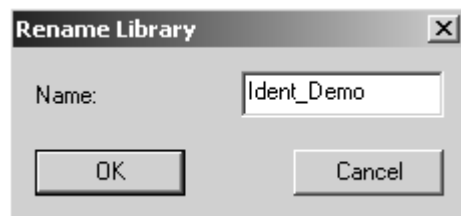


Figure 3: Rename reference library

Enter the new name of the current library and confirm it by clicking on the *OK* button. The new description is automatically displayed in the browser window on the right and in the *General information...* group field. The renaming procedure applies to each library level.

You can also delete libraries. Note that all sub-libraries assigned to a specific library level will be deleted as well. Select the particular library in the browser window and press the *DEL* key on the PC keyboard. A menu pops up and asks you whether you want to continue or cancel this deleting operation.

There are two possibilities to print the library structure displayed in the browser window. Activate the *Printer* option button to have the library structure printed on a connected printer. The drop-down list includes the following printing options:

- Library tree as shown
- Whole library tree
- Whole library tree and group statistics
- Whole library tree and statistics for each spectrum

Select one of the printing options and click on the *Print* button. If you activate the *PDF* option button, the library structure will be printed as a PDF file.

Clicking on the *Print* button opens the following dialog:

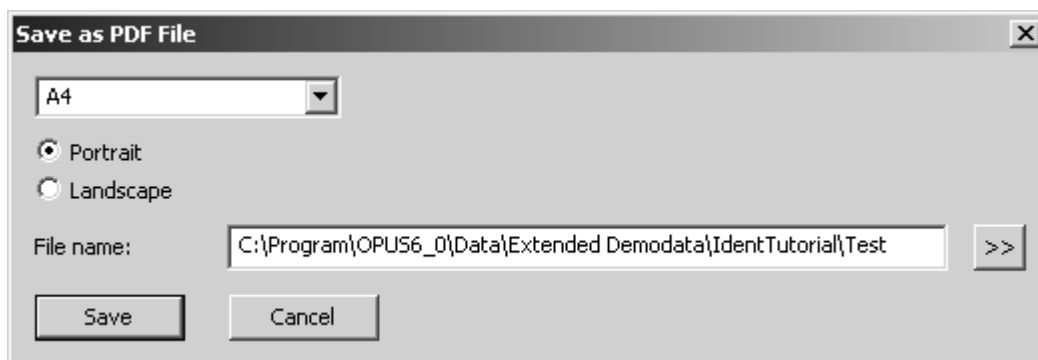



Figure 4: Printing as PDF file

Define the appropriate format, the file name and path of the PDF file. Click on the  button to open the *Define PDF File Name* dialog.

1.1.1 Methods created by prior OPUS releases

IDENT methods created by prior OPUS releases, and which also include sub-methods are automatically converted to OPUS 5.0 if you load them into the IDENT setup. If one of these sub-methods cannot be found in the path defined, OPUS tries to search in the path where the reference method (*.FAA) is stored. If the method cannot be found there either, OPUS writes a remark in the *.log ASCII-file.

Sub-methods containing spectra which are no member of the main library (IP1) will not be considered for the set-up of the converted OPUS 5.0 method. If spectra are assigned to more than one sub-method on the same library level, they will be considered only once. In both cases the library conversion log file includes an appropriate remark.

At the end of the conversion a dialog pops-up indicating that the conversion has been finished, but not completed. Confirm this dialog by clicking on the *OK* button.

1.2 Creating New Method

If you want to create a new IDENT method, click on the *Reference Spectra* tab. To create a reference library, define the spectra to be used for the library. Click on the *Add Spec. for New Group* button to add single spectra to a new product group. Click on *Add Spec. to Sel. Group* to add single spectra to a selected product group.

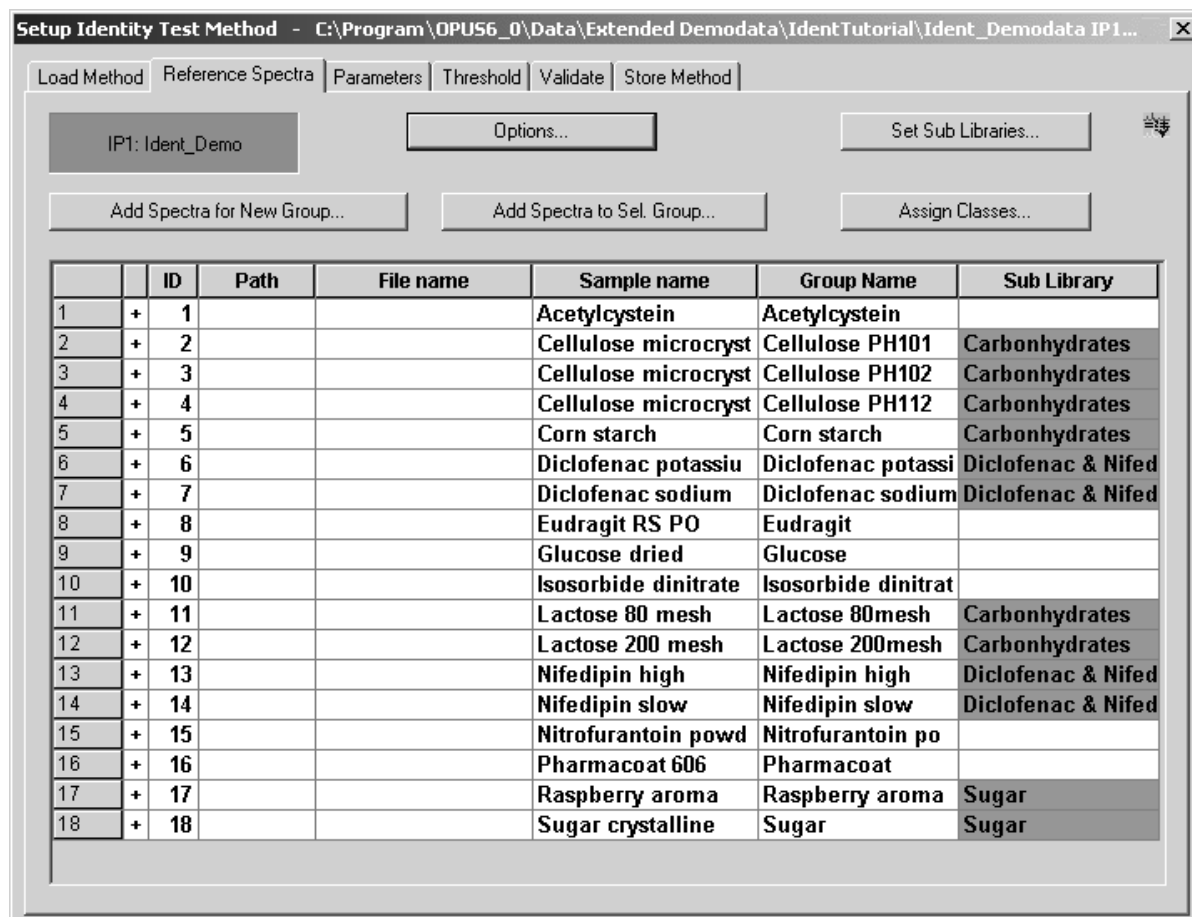


Figure 5: Setup Identity Test Method – Reference Spectra tab

In both cases, a dialog box opens where you can select the spectra to be used for the IDENT method. Having loaded spectra, they will be assigned to the respective group and are given a consecutive *ID* number. The *Sample Name* as well as the *Group Name* are automatically read from the file parameters.

To see the list of group-specific spectra, click on the sign in the first column (figure 5). Click on the sign to close the list. You can also remove entries from the file list. Select one or more entries by clicking on the left mouse button and holding down the *Shift* or *CTR* key. Use the *DEL* key on the PC keyboard to delete the entry(ies). Before you can delete a group or spectrum, a menu pops up and asks you whether you really want to delete the spectra selected.

If you want to create new sub-libraries, click on the *Set Sub Libraries* button. These sub-libraries are an integral part of the main IDENT method and are indicated in green in the *Sub Library* column (figure 5). The following dialog opens:

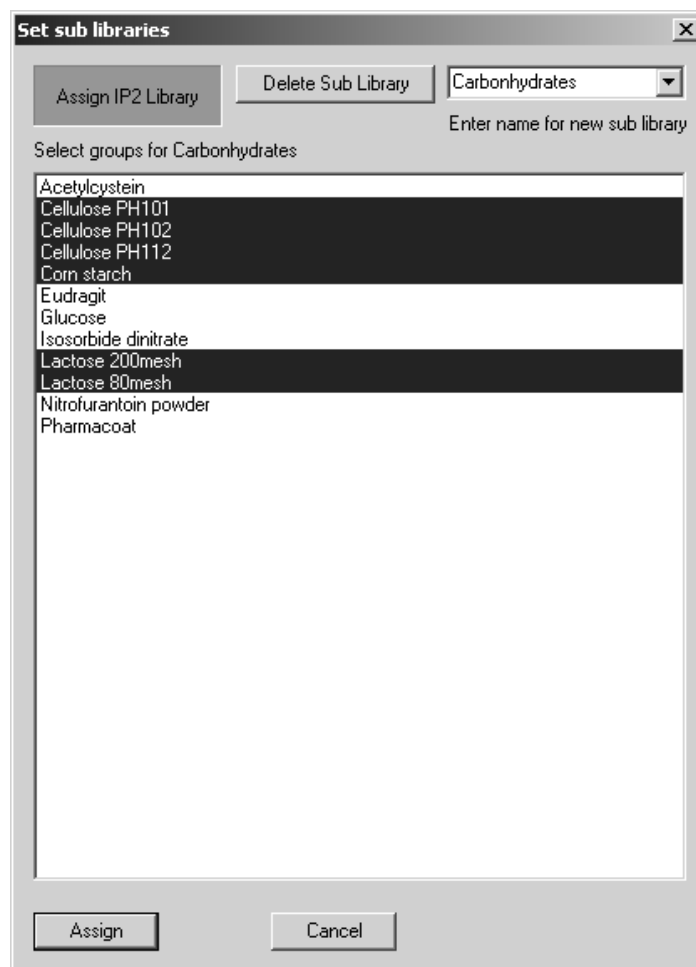


Figure 6: Set Sub Libraries

As the spectra of the *Carbonhydrates* group have already been assigned as a sub-library they are automatically highlighted (figure 6).

The indicator field shows the defined sub-library IP2 level in green. You can always set sub-libraries exactly one level below the current library. All available groups defined for the main method, and which have not yet been assigned to any sub-library on this level are displayed in the *Select groups for...* field. Select the *New* option from the upper right drop-down list and enter a unique name for the new sub-library. Then, click on the group(s) you want to assign. If you define a new sub-library name, OPUS automatically checks whether the new library name does already exist to avoid double naming. If a name already exists, a dialog pops up requiring unique naming. Click on the *Assign* button.

To delete sub-libraries, select them in the selection field and click on the *Delete Sub Library* button. A menu pops up and asks you whether you want to continue or cancel the deleting operation.

Groups which cannot be separated into further sub-libraries can, however, be assigned to common classes if *class identity* is sufficient as analysis result.

To assign groups to a class click on the *Assign Classes* button on the *Reference Spectra* tab. The following dialog opens:



Figure 7: Assign Classes

The *Select groups for...* selection field includes only groups which have not yet been a member of a sub-library or class on this level.

Select the *New* option from the upper right drop-down list and enter a unique name for the new class. Then, select the group(s) you want to assign and click on the *Assign* button. A menu pops up and asks you whether you want to continue or cancel the assigning operation.

Click on the *Options* button on the *Reference Spectra* tab if you want to change the path of the original and averaged spectra. The *Add selected spectra into one group* check box is checked by default. For further details, see chapter 7.2.3.

1.3 Setting Parameters

The quality of an IDENT analysis substantially depends on the data preprocessing method and spectral regions of each spectrum, which both have been selected for the samples and IDENT method.

For the main library (IP1) it is recommended to use the *Standard* method and define a large spectral region, as the spectral noise will be substantially smoothed. In case of sub-libraries (IP2, IP3 etc.) it may be better to use the *Factorization* method and limit the spectral region, which, of course, causes the spectral noise to be insufficiently smoothed, but the spectrum shows many significant details.

Note: If you update library data, new spectra have to be re-calculated on each library level.

Click on the *Parameters* tab. This tab defines the spectral regions which have to be considered for identification. You also have to select the preprocessing method from the drop-down list as well as the IDENT method (algorithm).

Vector normalization is frequently selected as data preprocessing method. Sometimes, however, you get even better results if you select *First* and *2nd Derivative*. For further details, refer to chapter 7. Select *Vector Normalization* from the *Preprocessing* drop-down list.

The *Always use lowest IP level* check box is only enabled on the first library level as this is a global setting for the entire library. If you activate the check box, the IDENT analysis will be performed on the lowest IP level available. This will also be done even if an IDENTICAL TO result has been achieved at any higher IP level.

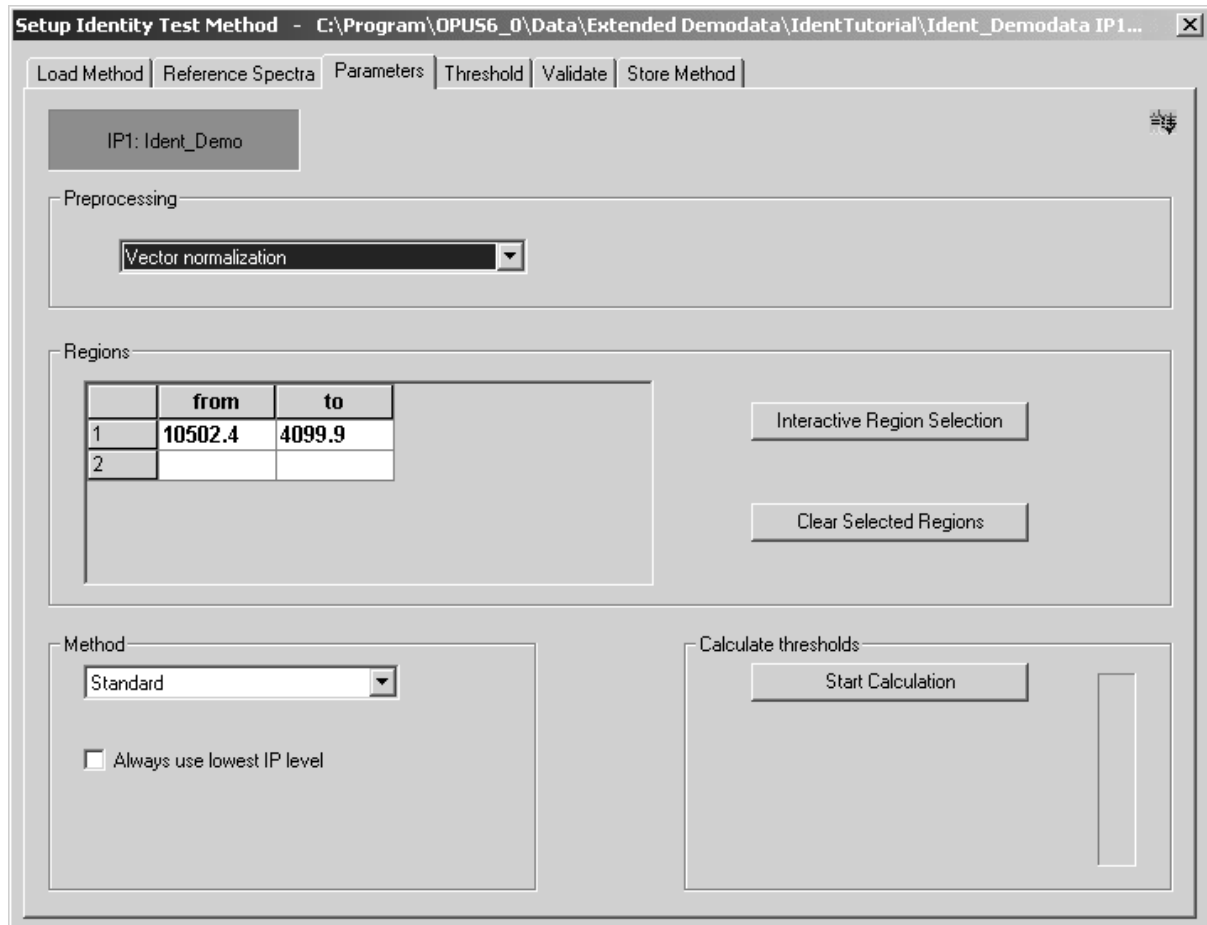


Figure 8: Setup Identity Test Method – Parameters tab

If you want to limit the IDENT analysis to certain frequency regions, enter them into the *Regions* table. Alternatively, you can also define these regions interactively.

Click on the *Interactive Region Selection* button. A separate window opens which displays mean spectra.

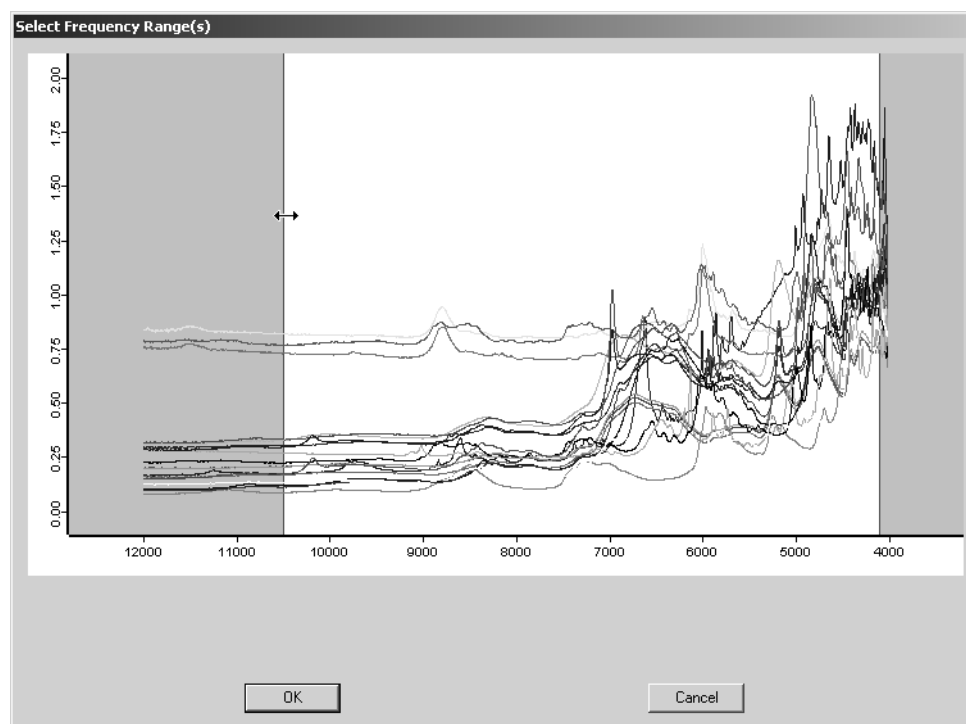


Figure 9: Interactive Frequency Range Selection

The gray areas indicate the selected frequency limits, and the white spectral range is the basis for the subsequent evaluation.

To move spectral regions place the cursor on the respective edge between the white and gray area, hold down the left mouse button and move the regions. To delete a spectral region, right click on the white area and select *Remove* from the pop-up menu. Click on the *OK* button to confirm the settings, and the *Parameters* tab will be displayed again.

You can also add a new frequency region by a right-click on the left or right window side. Select the *Add Region* option from the pop-up menu (figure 10).

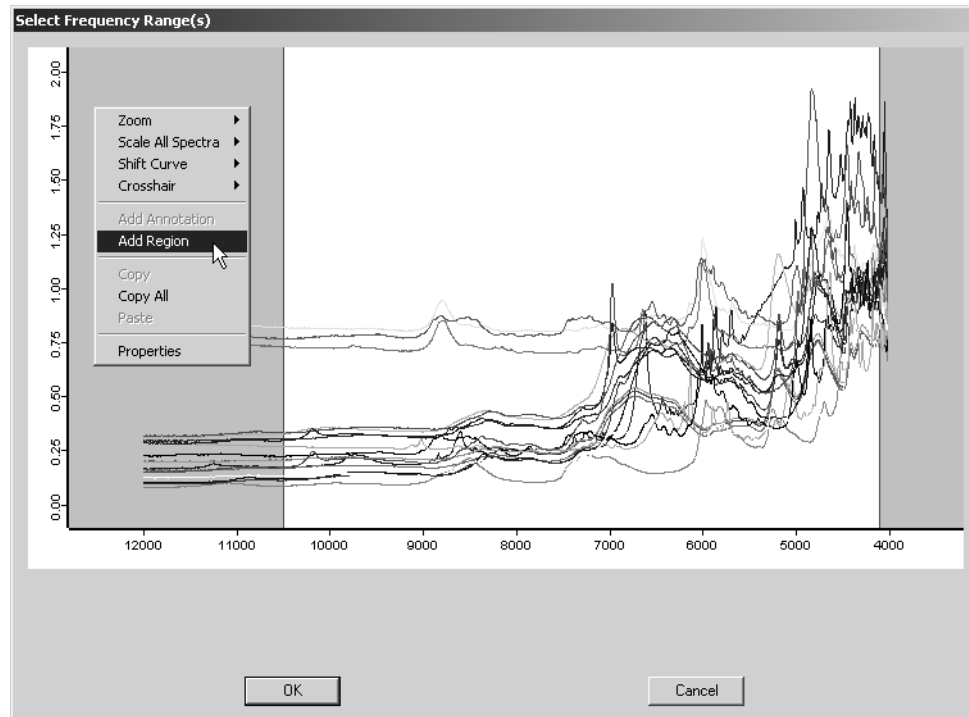


Figure 10: Interactive Frequency Range Selection with pop-up menu

Select *Standard* as method on the *Parameter* tab and click on the *Start Calculation* button to calculate the spectral distances.

1.4 Identity Test Limit

Click on the *Threshold* tab to have the identity test limits displayed. The threshold of a reference spectrum is the sum calculated from the maximum distance listed in the average report, plus the product resulting from the standard deviation (*SDev*) and any *x* factor. For each group the threshold values are listed in the *Threshold* column (figure 11).

You can enter any factor into the entry field, 0.25 is set by default. Click on the *Set* button to confirm your entry. The factor set is valid for all reference spectra, and the *Threshold* column is updated accordingly. You can also print the list. Click on the *Print List* button. For further details, see chapter 7.4.

In a similar way you can set the value for the *Confidence Level*. You can enter any factor between 95 and 99.9999% into the entry field, 99.99% is set by default. If you click on the *Set* button to confirm your entry, the *Threshold* column is updated accordingly. The *Outlier* column displays the number of spectra which are outside the threshold. If you, e.g., set the confidence level to 95%, then 5% of the total number of spectra will be identified as outliers.

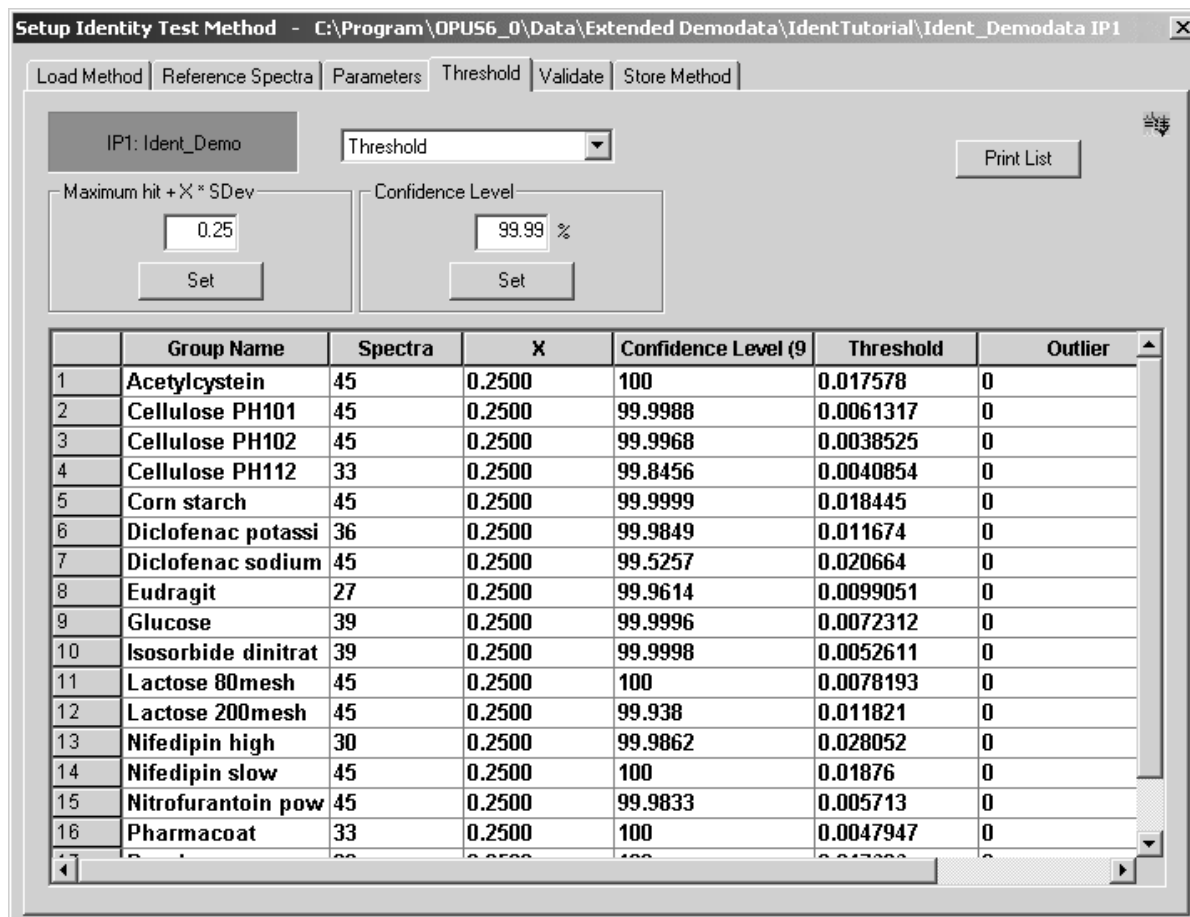


Figure 11: Setup Identity Test Method – Threshold tab

1.5 Validating Library

When creating a library, first check whether the IDENT parameters selected for the IDENT method are optimized for all reference spectra, and whether an unambiguous assignment of test spectra to a group can be ensured. This is done by a validation procedure which compares each original spectrum with the average spectra of all groups. To validate the library, click on the *Validate* tab. The following dialog opens:

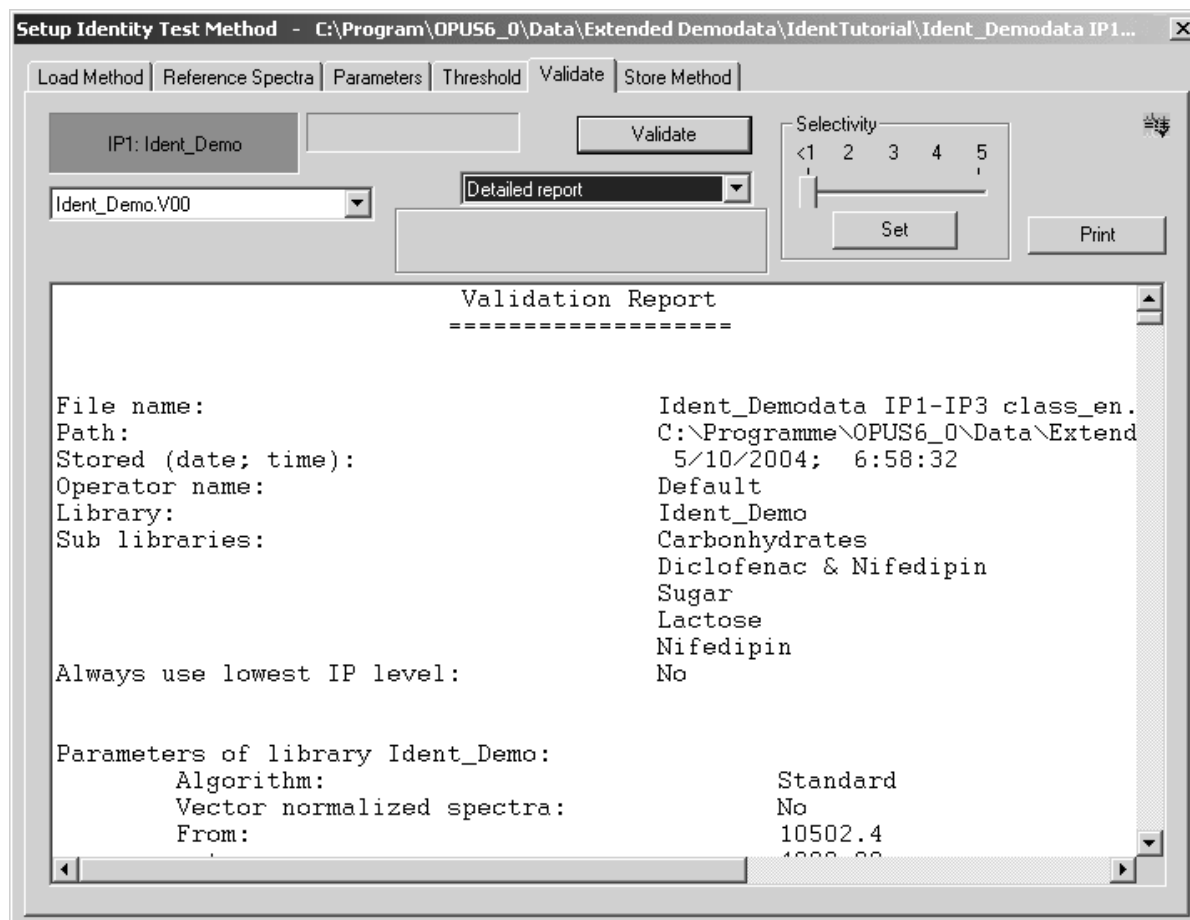


Figure 12: Setup Identity Test Method - Validate tab

Click on the *Validate* button to start validation. A dialog pops up and asks you whether to validate this library and all sub-libraries. For further details, see chapter 7.5.

1.5.1 Validation Report

The validation result is displayed in the form of a report and stored in a file which uses the extension **.VAL*. Validation reports can also be created for a single reference library, sub-library or even for the entire library data structure, beginning on the level from where the validation starts. For further details on the single reports, see section 7.5.1.

If you perform more than one validation, the result files will be consecutively numbered (**.v00*, **.v01*...). You can compare different validation results with each other by selecting the respective file from the drop-down list. You can also print the report by clicking on the *Print* button. This starts the *Windows Notepad* program which you can use to reformat the text, if desired. Use the *Notepad* print function to create a printout.

Note: It is recommended to select a small font in *Windows Notepad* to avoid extraordinary long reports. A proportional font may lead to a confusing display of the results. Therefore, it is advisable to use a monospace font, e.g. Courier New, 10.

The *Threshold* values listed on the *Threshold* tab are regarded as confidence region during validation. The results are classified in three categories: *uniquely identified*, *not identified* and *can be confused with*. In case of results belonging to the first category the spectral distance between the original and average spectrum is within the threshold value. The spectral distance is higher than the threshold value in case of results belonging to the second category. The *Can be confused* category indicates that the spectral distance of an original spectrum is smaller than the confidence level, compared to at least one different average spectrum. For further details, see section 6.7.

1.6 Storing Method Files

Click on the *Store Method* tab to store the method files you have created. The following dialog opens:

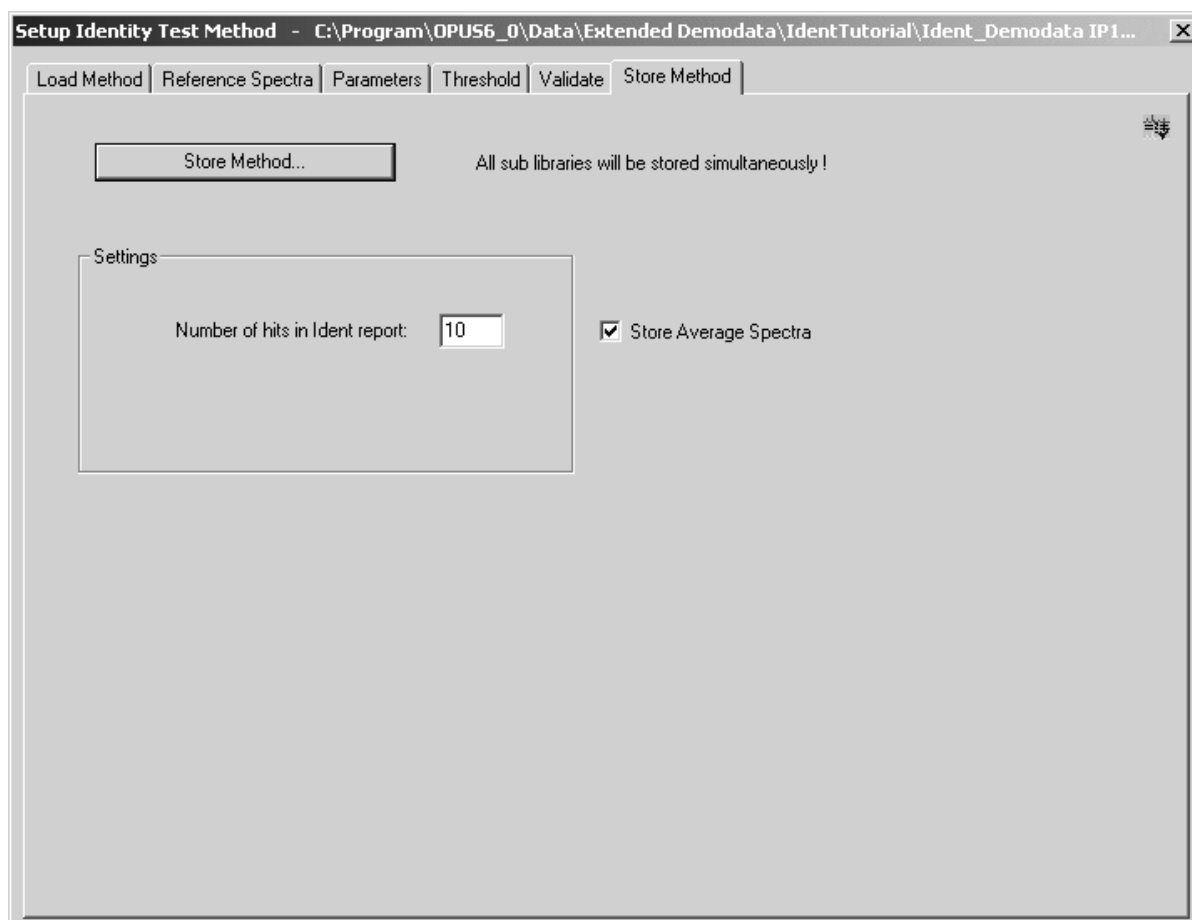


Figure 13: Setup Identity Test Method - Store Method tab

The parameter you can define in this tab is the *Number of Hits in Ident Report*, i.e. you enter the number of hits that have to be stored together with the IDENT report.

By default, the *Store Average Spectra* check box is activated. This means that you can store all average spectra of a library (IP1 level) in one separate directory which is a sub-directory of the directory in which the IDENT method has been stored.

The average spectra will be stored without data pre-processing, and have an AVERAGE (AVERAGE) data block appended. If you repeatedly store a particular IDENT method, the average spectra will NOT be overwritten. Instead, the file extension will be incremented.

Click on the *Store Method* button. The standard *Save File* dialog box opens to be used to save the method. The method file uses the extension *.FAA*, and all sub-libraries will be stored simultaneously.

For special details on method protection refer to the OPUS QUANT manual, chapter 9.

2

Performing an IDENT Analysis

Compared to the setup of an IDENT method the analysis of unknown samples is easy. Before you start the analysis load the spectra of your unknown samples into the OPUS browser window.

The analysis compares the test spectrum with all reference spectra. The result of a comparison between spectrum *A* and *B* results in the spectral distance *D*, which is also called *Hit Quality*. The better two spectra match, the smaller the spectral distance. The *Hit Quality* for identical spectra is 0 (i.e. a reference spectrum is compared with itself).

To start an IDENT analysis, select the *Identity Test* command from the *Evaluate* menu. The *Identity Test* dialog box opens.

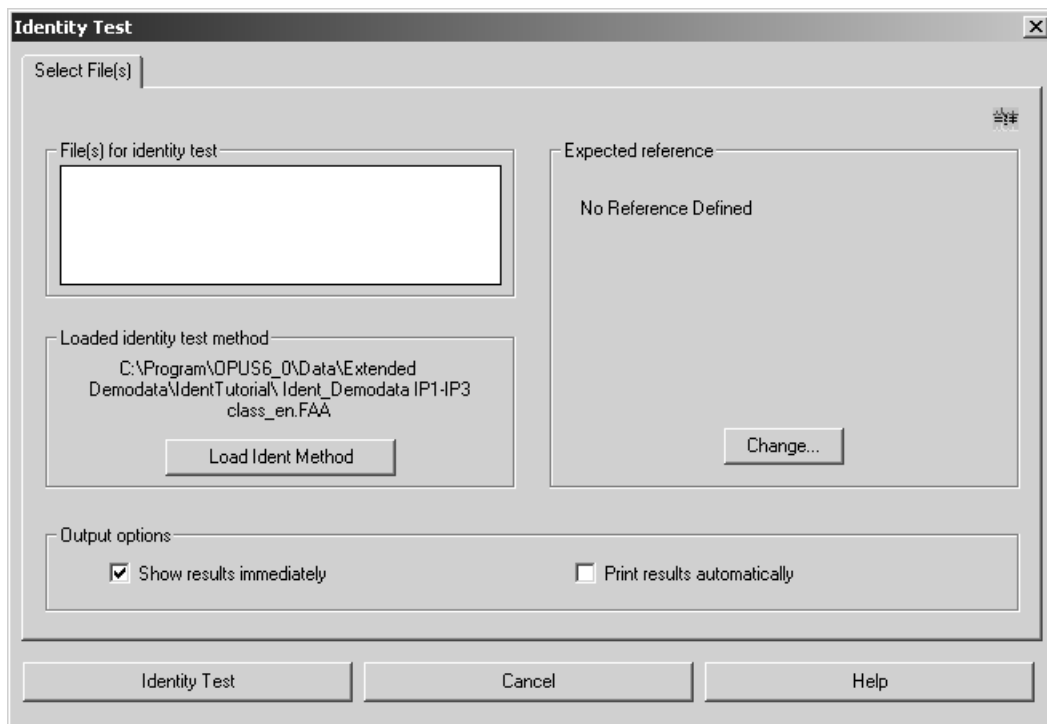


Figure 14: Identity Test - Select File(s)

Drag and drop the absorption block of the test spectra from the OPUS browser window into the *File(s) for Identity Test* field. If you release the left mouse button, the spectra are added to this entry field.

If an identity test method has already been loaded (e.g. if you have created a method prior to starting the analysis), the path and name of this method name will be indicated in the *Loaded Identity Test Method* field. To load or change an IDENT method, click on the *Load Ident Method* button and select the desired method from the dialog box that opens.

The analysis uses the *No Reference Defined* function for an IDENT method if you have not defined an expected reference. Click on the *Change* button to modify this default setting. The *Change* button will be disabled if the operator has no right to change parameters of this kind. For details on this dialog, see section 7.7.1.

Click on the *Identity Test* button to start the test. The result of the analysis is appended to the respective file in the form of an IDENT report block. If you click on this report block, a report window opens automatically and displays the results.

3 IDENT Report

Both the results of an identity test and the averaging of original spectra to generate reference files for an IDENT library are stored in report blocks. You can open these reports, like any other OPUS report, by double-clicking on the report block in the OPUS browser window.

3.1 Identity Test Reports

The results of the spectrum comparison between test spectrum and reference spectra are written into an IDENT report. This report is stored in the test spectrum file. The content of the report depends on the parameters and algorithms selected to run the identity test. For details, see section 6.4. If you click on the REPORT data block, a text window will open and show the results.

3.1.1 Standard Method

The IDENT report contains detailed information on the method and a list of spectral distances between the test spectrum and the reference spectra. This list includes distances in ascending order, i.e. Hit No. 1 is the reference spectrum which is most similar to the test spectrum. The number of listed distances can be defined when creating the IDENT method (see chapter 7.6).

Figure 15 shows an IDENT report using the *Standard* algorithm in combination with an identity test on the first reference spectrum (L-Leucin) of the library. An original (individual) spectrum of the reference library has been used as test spectrum. The threshold of the selected reference spectrum (*Threshold for expected reference*, see figure 15) is 0.023866. The spectral distance between the test spectrum and this reference spectrum is 0.010556 (*Hit quality with expected reference*), i.e. it is smaller than the threshold. Since no further hit can be found below this threshold (Hit No. 2 with 0.047006 is higher), the result is: *IDENTICAL TO* to the expected spectrum.

Compare Spectra		Values
Method file:		C:\OPUS\IDENT\examples\STANDARD.FAA
from (date):		16/06/00
(time):		10:15:01
Description:		IDENTICAL TO: 000001 Sample L-Leucin Av. of 11 : 1
Hit quality with expected reference:		0.010556
Threshold for expected reference:		0.023866
Threshold calculation:		- depends on each ref.-spectrum -
Algorithm:		Standard
Vector normalized spectra:		Yes
Order of Derivative:		0
Smoothing points:		1
No. of used factor sp.:		0
15 hits of 15		
X-Ranges:		1
Class Name:		BODO
Class Test OK:		1
Using residuals:		No
Order of Internal Derivative:		0
Smoothing Points for Internal Derivative:		1
Reduction Factor:		1

Hit No.	Hit Quality	Sample Name	Group	Threshold
1	0.010556	000001 Sample L-Leucin Av. of 11	000001	0.023866
2	0.047006	000003 Sample L-Isoleucin Av. of 11	000003	0.023866
3	0.095210	000002 Sample DL-Isoleucin Av. of 11	000002	0.033536
4	0.104591	000015 Sample L-Methionin Av. of 11	000015	0.061757
5	0.146645	000004 Sample DL-Alanin Av. of 11	000004	0.011690
6	0.158939	000005 Sample L-Alanin Av. of 11	000005	0.073910
7	0.172990	000014 Sample DL-Methionin Av. of 11	000014	0.072976
8	0.266797	000006 Sample DL-Tryptophan Av. of 11	000006	0.022966
9	0.288849	000007 Sample L-Tryptophan Av. of 11	000007	0.051395
10	0.668990	000011 Sample Xylit Av. of 11	000011	0.032523
11	0.671159	000010 Sample Fructose Av. of 11	000010	0.030262
12	0.690747	000012 Sample Sorbit Av. of 11	000012	0.029502
13	0.750100	000013 Sample Mannit Av. of 11	000013	0.017529
14	0.763355	000008 Sample Glucose H2Ofrei Av. of 11	000008	0.090445
15	0.848422	000009 Sample Glucose H2O Av. of 11	000009	0.026233

Figure 15: IDENT Report – Query spectrum identical

Figure 16 shows a test report for the same query spectrum. But this time DL-Methionin has been selected as expected reference. Now, the result is *NOT IDENTICAL* to the expected spectrum. The spectral distance to this reference spectrum is 0.172990 and Hit No. 7. This value exceeds the threshold of 0.072976, and therefore the test spectrum is classified as not being identical.

Compare Spectra		Values
Method file:	C:\OPUS\IDENT\examples\STANDARD.FAA	
from (date):	16/06/00	
(time):	10:15:01	
Description:	>> NOT IDENTICAL << to: 000014 Sample DL-Methionin Av. of 11 : -2	
Hit quality with expected reference:	0.172990	
Threshold for expected reference:	0.072976	
Threshold calculation:	- depends on each ref.-spectrum -	
Algorithm:	Standard	
Vector normalized spectra:	Yes	
Order of Derivative:	0	
Smoothing points:	1	
No. of used factor sp.:	0	
15 hits of 15		
X-Ranges:	1	
Class Name:		
Class Test NOT PERFORMED:	0	
Using residuals:	No	
Order of Internal Derivative:	0	
Smoothing Points for Internal Derivative:	1	
Reduction Factor:	1	

Hit No.	Hit Quality	Sample Name	Group	Threshold
1	0.010556	000001 Sample L-Leucin Av. of 11	000001	0.021304
2	0.047006	000003 Sample L-Isoleucin Av. of 11	000003	0.023866
3	0.095210	000002 Sample DL-Isoleucin Av. of 11	000002	0.033536
4	0.104591	000015 Sample L-Methionin Av. of 11	000015	0.061757
5	0.146645	000004 Sample DL-Alanin Av. of 11	000004	0.011690
6	0.158939	000005 Sample L-Alanin Av. of 11	000005	0.073910
7	0.172990	000014 Sample DL-Methionin Av. of 11	000014	0.072976
8	0.266797	000006 Sample DL-Tryptophan Av. of 11	000006	0.022966
9	0.288849	000007 Sample L-Tryptophan Av. of 11	000007	0.051395
10	0.668990	000011 Sample Xylit Av. of 11	000011	0.032523
11	0.671159	000010 Sample Fructose Av. of 11	000010	0.030262
12	0.690747	000012 Sample Sorbit Av. of 11	000012	0.029502
13	0.750100	000013 Sample Mannit Av. of 11	000013	0.017529
14	0.763355	000008 Sample Glucose H2Ofrei Av. of 11	000008	0.090445
15	0.848422	000009 Sample Glucose H2O Av. of 11	000009	0.026233

Figure 16: Ident Report – Query spectrum not identical

3.1.2 Factorization Method

If you use the *Factorization* algorithm (on the *Parameters* tab), *Eigen values* and *Eigen vectors* (see section 6.1.2) may also be interesting for you. The factor values can be retrieved from the report file of the IDENT method file. Load the method file (extension *.*FAA*) into the OPUS browser window and click on the *REPORT* data block to open the report. Open the *Identity Search Method* subdirectory as shown in figure 17. If you click on *Eigen Vectors* or *Eigen Values*, these values will be displayed in the report window. The *T* values (see section 6.1.2) are listed in the *Eigen vectors* sub-directory.

The screenshot shows the IDENT software interface. On the left is a tree view of the file structure: "C:\OPUS\IDENT\examples\FACTOR.FAA" 1, 1024, Reports, Identity Search Method (highlighted with an arrow), Eigen Values, Eigen Vectors, and x Ranges. The main window displays the following parameters:

Identity Search Method	Values
Algorithm:	Factorization
For Threshold Individual:	Yes
No. of used factor sp.:	6
For Threshold Info Entry:	1
No. of Hits to be Listed:	10
Vector normalized spectra (if standard):	No
No. of Ref. Spectra:	15
Delta x of x raster:	3.857202
F\X (or L\X) mod Delta x:	0.000000
min x of x raster:	3999.918628
max x of x raster:	10001.725171
Whole x range:	No
No. of x Points in all x Ranges:	1454
No. of x Ranges:	1
BlockID:	4111
Order of Derivative:	0
Description:	
For Threshold Info Entry:	
Constant conf. level [%%]:	99.000000
Path of origin reference spectra:	H:\ident\example2\
Subpath from name, first character:	1
Subpath from name, length:	8
Smoothing points:	1
Using residuals:	No

Below the parameters is a table with the following columns: File Name, Sample Name, Frequency of First Point, and Frequency.

File Name	Sample Name	Frequency of First Point	Frequency
H:\ident\example2\An000001.100	000001 Sample L-Leucin Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000002.100	000002 Sample DL-Isoleucin Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000003.100	000003 Sample L-Isoleucin Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000004.100	000004 Sample DL-Alanin Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000005.100	000005 Sample L-Alanin Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000006.100	000006 Sample DL-Tryptophan Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000007.100	000007 Sample L-Tryptophan Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000008.100	000008 Sample Glucose H2Ofrei Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000009.100	000009 Sample Glucose H2O Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000010.100	000010 Sample Fructose Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000011.100	000011 Sample Xylit Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000012.100	000012 Sample Sorbit Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000013.100	000013 Sample Mannit Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000014.100	000014 Sample DL-Methionin Av. of 11	10001.725171	3999.9186
H:\ident\example2\An000015.100	000015 Sample L-Methionin Av. of 11	10001.725171	3999.9186

Figure 17: Report file of a method file using factorization

4 Cluster Analysis

The cluster analysis tests FT-IR spectra for their similarity. In contrast to the identity test, no input information is required. The cluster analysis divide similar spectra into groups. These groups are called **classes** or **clusters**. The clustering can be displayed in a **dendrogram**. Figure 18 shows a simplified dendrogram including 5 spectra.

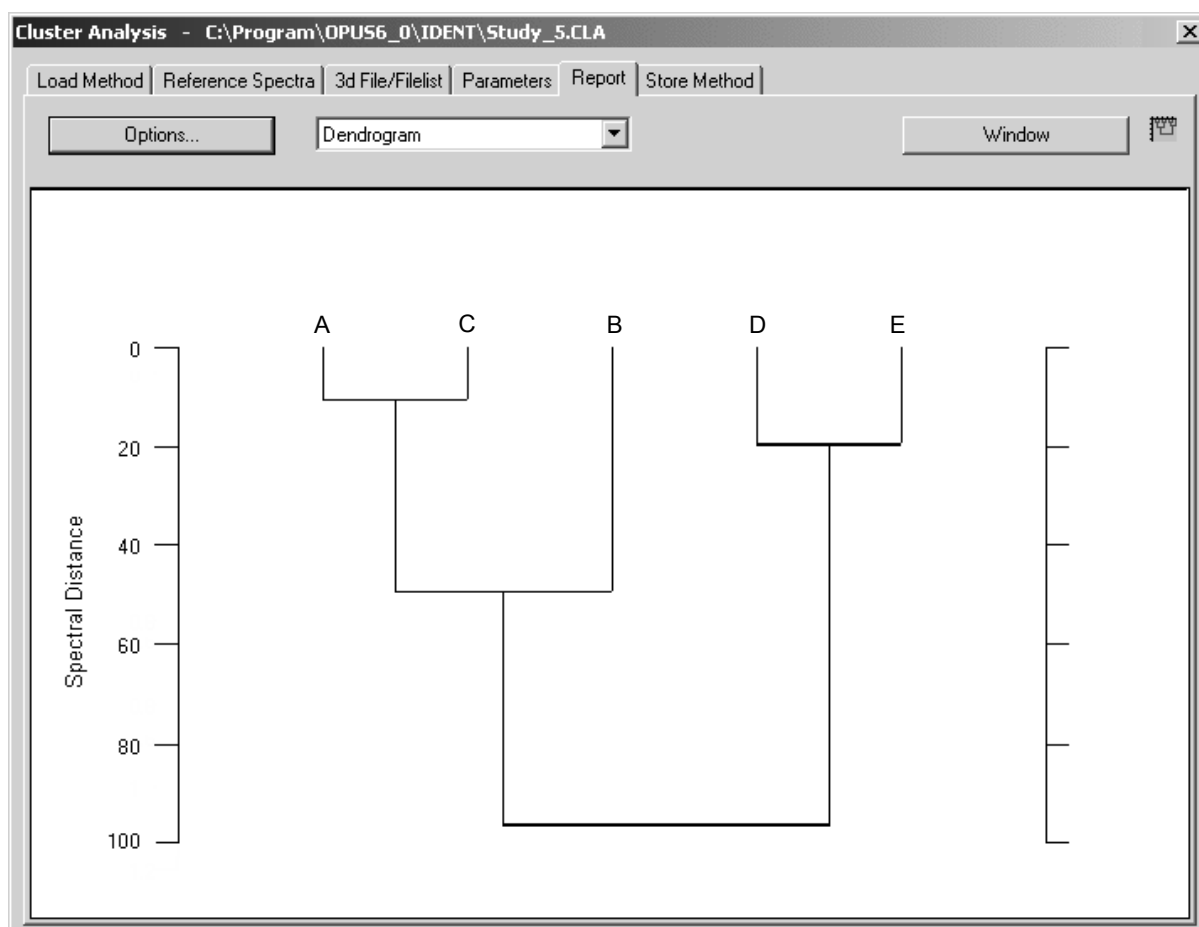


Figure 18: Cluster Analysis – Dendrogram

4.1 Theory

The spectral distance indicates the degree of spectral similarity. Two spectra with a spectral distance of 0 are entirely identical (within the frequency ranges tested). The higher the difference between two spectra, the higher the spectral distance.

The hierarchical cluster algorithms perform the following tasks:

- First, the spectral distances between all spectra are calculated.
- The two spectra with the highest similarity (i.e. spectra with the smallest spectral distance) are merged into a cluster.
- The distances between this cluster and all other spectra are calculated. Several methods (*Single Linkage*, *Complete Linkage*...) are available to calculate the distances.
- The two spectra (spectrum/spectrum or spectrum/cluster) with the smallest distance are merged again into a new cluster.
- The distances between this new cluster and all other spectra (spectra, cluster) are calculated.
- The two spectra (spectrum/spectrum or spectrum/cluster or cluster/cluster) are merged into a new cluster.
- ...

This procedure will be repeated until only one big cluster will be left.

Figure 19 shows this procedure in more detail. The spectral distances between any two spectra of a set of n spectra can be represented in a $n \times n$ matrix. This matrix is symmetrical and the main diagonal elements are 0. Subsequently, it is sufficient to test only a triangle of this matrix.

Five spectra A , B , C , D and E are used in the example. The triangular matrix which contains the spectral distances between these 5 spectra is shown in the upper part of figure 19. The A and C spectra are mostly similar to each other. The spectral distance is 11.0. Both spectra are merged into the AC cluster.

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	<i>E</i>
<i>A</i>	0				
<i>B</i>	44.0	0			
<i>C</i>	11.0	54.4	0		
<i>D</i>	101.6	68.1	97.4	0	
<i>E</i>	118.3	92.1	115.9	21.2	0

	<i>AC</i>	<i>B</i>	<i>D</i>	<i>E</i>
<i>AC</i>	0			
<i>B</i>	49.2	0		
<i>D</i>	99.5	68.1	0	
<i>E</i>	117.1	92.1	21.2	0

	<i>AC</i>	<i>B</i>	<i>DE</i>
<i>AC</i>	0		
<i>B</i>	49.2	0	
<i>DE</i>	108.3	80.1	0

	<i>ABC</i>	<i>DE</i>
<i>ABC</i>	0	
<i>DE</i>	94.2	0

Figure 19: Spectral distances calculated using a hierarchical cluster method

Now, the distances between *AC* and the other spectra have to be calculated. This time the *Average Linkage* algorithm has been used. The distance between *A* and *D* is 101.6 and the distance between *C* and *D* is 97.4. The distance between the *AC* cluster and *D* spectrum is the mean value of the two original distances: $(101.6 + 97.4) / 2 = 99.5$. The new distance values can be seen in the second matrix of figure 19.

As the smallest spectral distance is 21.2 in this matrix, the *D* and *E* spectra are merged into the *DE* cluster. Then, the distances between this new cluster and all other spectra will be calculated again.

Example: The spectral distance between *AC* and *D* is 99.5 and between *AC* and *E* is 117.1. Based on these values, the distance between *AC* and *DE* will be 108.3. The third matrix in figure 19 includes these new distance values.

In the next step, the *B* spectrum is merged with the *AC* cluster into a new *ABC* cluster. The distance between the *ABC* and *DE* cluster is 94.2. Finally, the *ABC* and *DE* clusters are merged into the *ABCDE* cluster.

The y axis of the dendrogram shows the spectral distances between different clusters. The horizontal lines indicate the fusion levels, which are the spectral distances of the different clusters and spectra prior to new clustering.

Table 1: Clustering Process

Number of Clusters	Clusters
5	A – B – C – D – E
4	AC – B – D – E
3	AC – B – DE
2	ABC – DE
1	ABCDE

You have to generate a cluster analysis method before a dendrogram can be graphically displayed. While you generate the method the spectral distances between the different spectra are calculated.

The clustering is repeated until all spectra are merged in one single cluster. Sometimes the intermediate states are of great interest for the user. If you use the *Make Diagnosis* function, you can get a cross section of the dendrogram. Simply enter the number of classes, and a list including the components of each single class will be generated. Additionally, the spectral distance for each cluster is indicated which has recently been merged.

As already mentioned, the spectral distances between different spectra can be represented by a symmetrical matrix. If you use the *Make Histogram* function, the whole matrix or part of it can be statistically tested. The results will be shown in the form of a histogram. However, this part of the program does not include clustering in general, but calculates the distances between the different spectra only.

4.1.1 Methods to Calculate Spectral Distances

Four different methods can be used to calculate spectral distances:

- Standard algorithm
- Factorization
- Scaling to 1st Range
- Normalize to Replevel

These methods are an integral part of the IDENT software. For details, see chapter 6 and 7. The *Standard* algorithm uses the Euclidian distance to determine spectral distances, while *Scaling to 1st Range* and *Normalize to Replevel* use the correlation coefficient.

Using the *Standard* or *Factorization* method the spectral distances calculated by the cluster analysis differ from those calculated by the identity test. Overlapping frequency ranges will not be merged when using cluster analysis. An artificial spectrum is derived from the selected spectral ranges of the measured spectrum. The artificial spectrum is used for the calculation of the spectral distances (and the data preprocessing) and includes numerous data points of the overlapping frequency regions.

The *Scaling to 1st Range* and *Normalize to Replevel* algorithms separately calculate the spectral distances for each frequency range. Then, an average value is calculated, and each frequency range can be weighted differently. If you use *Normalize to Replevel*, you can additionally specify a reproduction level for each frequency range. This level can be determined by the *Make Histogram* function.

The calculated spectrum-to-spectrum distances of the cluster analysis are equal to those calculated by the identity test if you use the *Normalize to Replevel* method. This, however, does not apply to *Scaling to 1st Range*. The identity test uses the spectral distances between the test spectrum and n reference spectra to determine extrema. This means that n distance values have to be taken into account per each frequency range. The cluster analysis, however, uses the distances between all reference spectra to determine extrema. These are $(n \cdot (n - 1))/2$ distances for n reference spectra as the spectral distance of a reference spectrum to itself is not considered.

To be able to compare the results achieved by the identity test and cluster analysis using the *Normalize to Replevel* and *Scaling to 1st Range* methods, the parameters required for the identity test and cluster analysis must be identical (same reference spectra, same frequency ranges etc.). Then, start the identity test. Use the first or last reference spectrum as test spectrum. Compare the *Hit Qualities* of this IDENT test report with the spectrum-to-spectrum distances of the cluster analysis.

4.1.2 Cluster Algorithms

There are 7 methods available to calculate spectral distances between a newly-created cluster and all the other spectra or clusters. The algorithms most frequently used are *Average Linkage* and *Ward's Algorithm*.

Single Linkage

The p and q clusters are merged to the new r cluster. $D(p,i)$ is the spectral distance between the p and i clusters, while $D(q,i)$ is the spectral distance between the q and i clusters. The $D(r,i)$ distance between the new r cluster and the i cluster is the smaller one of the two original distances:

$$D(r, i) = \min[D(p, i), D(q, i)] \quad (4-1)$$

This method can be used to create large clusters.

Complete Linkage

The new distance is the larger one of the two original distances.

$$D(r, i) = \max[D(p, i), D(q, i)] \quad (4-2)$$

This method prefers to create small groups.

Average Linkage

The arithmetic mean value is calculated as follows:

$$D(r, i) = \frac{D(p, i) + D(q, i)}{2} \quad (4-3)$$

Weighted Average Linkage

$n(p)$ is the number of spectra which are merged in the p cluster and $n(q)$ is the number of spectra which are merged in the q cluster. The spectral distance between the new r cluster and the i cluster is calculated as follows:

$$D(r, i) = \frac{n(p) \cdot D(p, i) + n(q) \cdot D(q, i)}{n(p) + n(q)} \quad (4-4)$$

This algorithm is a generalization of *Average Linkage*.

Median Algorithm

$D(p, q)$ is the spectral distance between p and q .

$$D(r, i) = \frac{D(p, i) + D(q, i)}{2} - \frac{D(p, q)}{4} \quad (4-5)$$

Centroid Algorithm

n is the total number of reference spectra. $D(r, i)$ is calculated according to the following equation:

$$D = \frac{n(p) \cdot D(p, i) + n(q) \cdot D(q, i)}{n} + \frac{n(p) + n(q)}{n^2} \cdot D(q, p) \quad (4-6)$$

Ward's Algorithm

The previous algorithms merge the two groups which are most similar. Ward's Algorithm, however, tries to find as homogeneous groups as possible. This means that only two groups are merged which show the smallest growth in heterogeneity factor H . Instead of determining the spectral distance, the Ward's Algorithm determines the growth of heterogeneity H . This method can

especially be used for the cluster analysis of bacteria spectra, i.e. these clusters correlate extremely well with microbiological affinity.

$n(i)$ is the number of spectra merged in the i cluster. $H(r,i)$ is calculated according to the following equation:

$$H(r, i) = D(r, i) = \frac{[n(p) + n(i)] \cdot D(p, i) + [n(i) + n(q)] \cdot D(q, i) - n(i) \cdot D(q, i)}{n + n(i)} \quad (4-7)$$

If you want to test different cluster algorithms, you do not need to calculate the spectrum-to-spectrum distance matrix again, because the clustering does not have any effect on this kind of matrix.

4.2 Performing a Cluster Analysis

To perform a cluster analysis the following steps are required:

- Measuring at least 1 spectrum per substance
- Incorporating the spectra into the list
- Defining a suitable spectral range for identification
- Selecting a data preprocessing method
- Calculating spectral distances
- Defining a cluster algorithm
- Generating a dendrogram, histogram or diagnosis

Several tabs of the *Cluster Analysis* dialog box are identical to the *Setup Identity Test Method* dialog and will only be briefly explained. For details, see the respective sections in chapter 1.

Click on the *Cluster Analysis* command from the *Evaluate* menu.

On the *Load Method* tab you can load an already existing method file generated for cluster analysis. This tab is identical to the *Load Method* tab of the *Setup Identity Test Method* command. As you will generate a new *Cluster Analysis* method, click on the *Reference Spectra* tab. The following dialog opens:

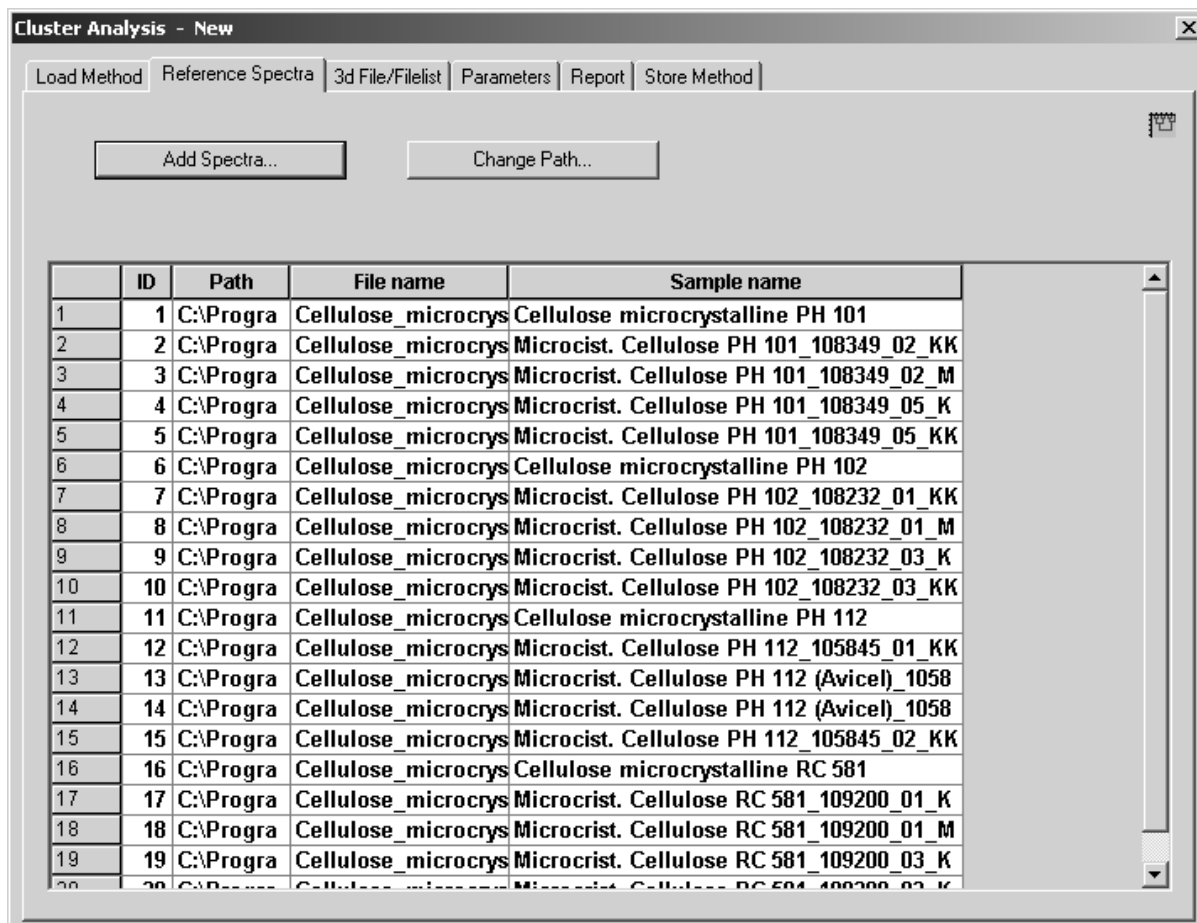


Figure 20: Cluster Analysis – Reference Spectra tab

To create a list, first define the spectra to be used on the *Reference Spectra* tab. Add these spectra as described in chapter 1.

Now, click on the *Parameters* tab. Define the spectral regions to be considered for cluster analysis and select a data preprocessing method as well as the cluster analysis algorithm.

In case of data preprocessing you can select between *Vector Normalization*, *First* and *2nd Derivative*, as well as combinations of both. *Vector Normalization* is set by default.

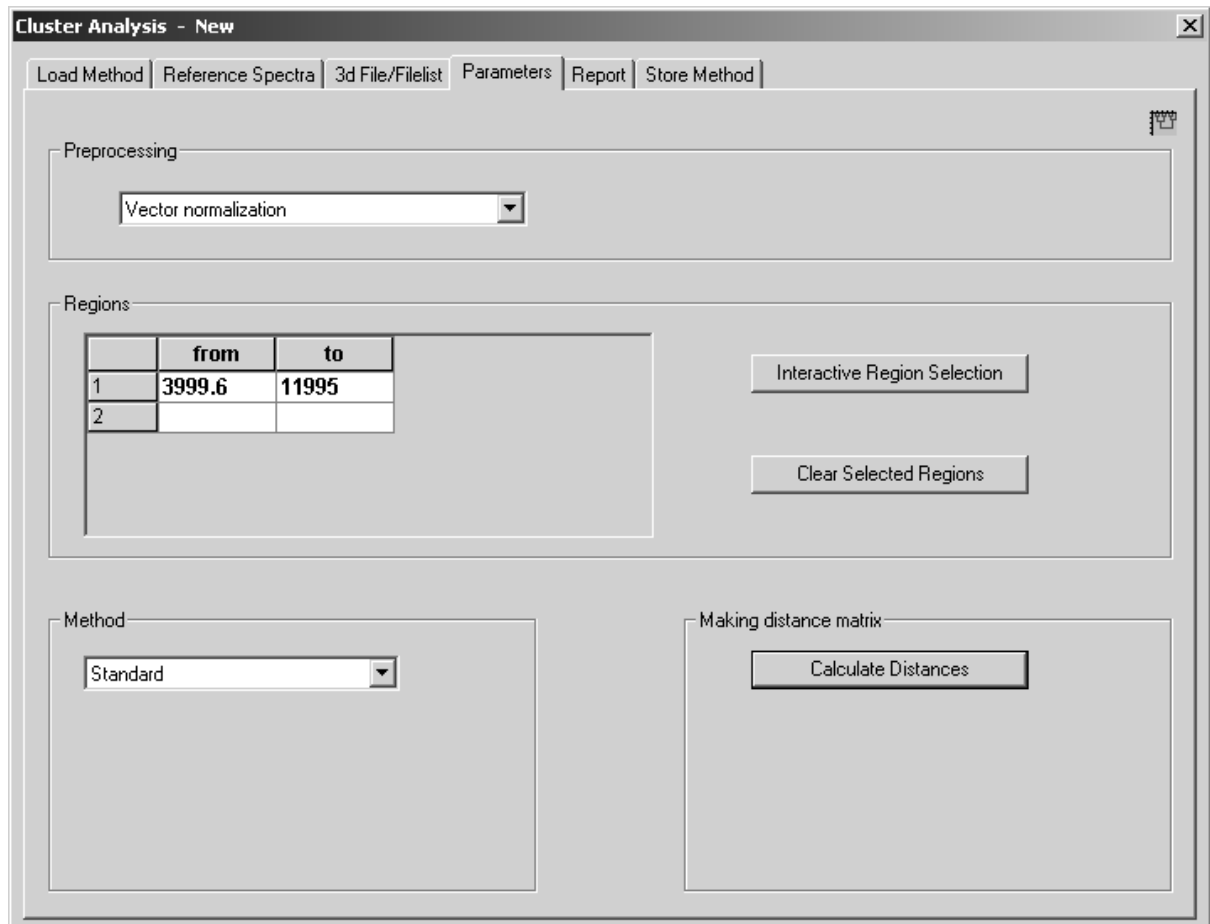


Figure 21: Cluster Analysis – Parameters tab

Define the frequency regions you want to use, see chapter 1. Select the *Standard* method to calculate the spectral distance.

Click on the *Calculate Distances* button. After the calculation of the distances you first have to save the method. Click on the *Store Method* tab and on the *Store Method* button to open the standard *Save File* dialog box. Enter a new file name for the method file and click on the *Save* button.

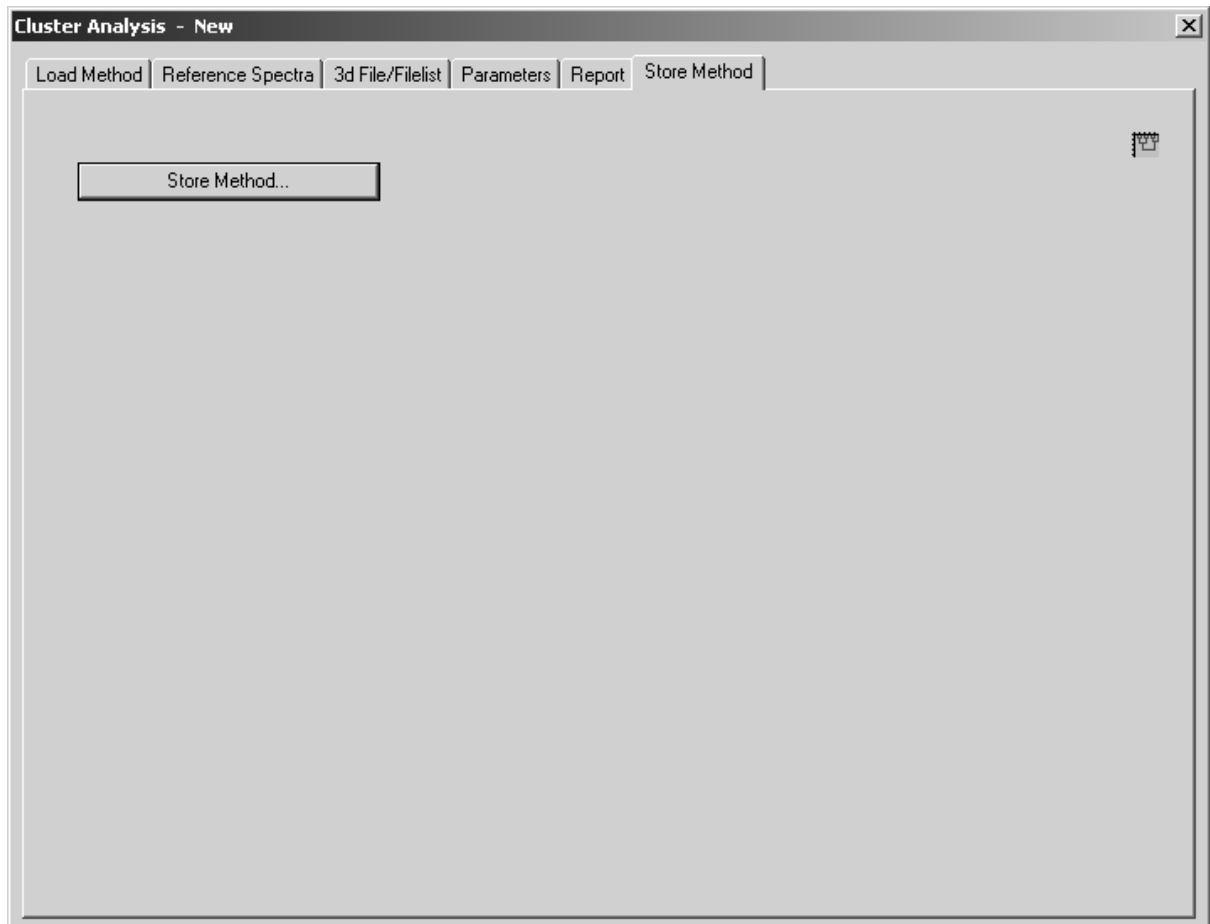


Figure 22: Cluster Analysis – Store Method tab

Now, click on the *Report* tab to have the results displayed. You can select between different display views.

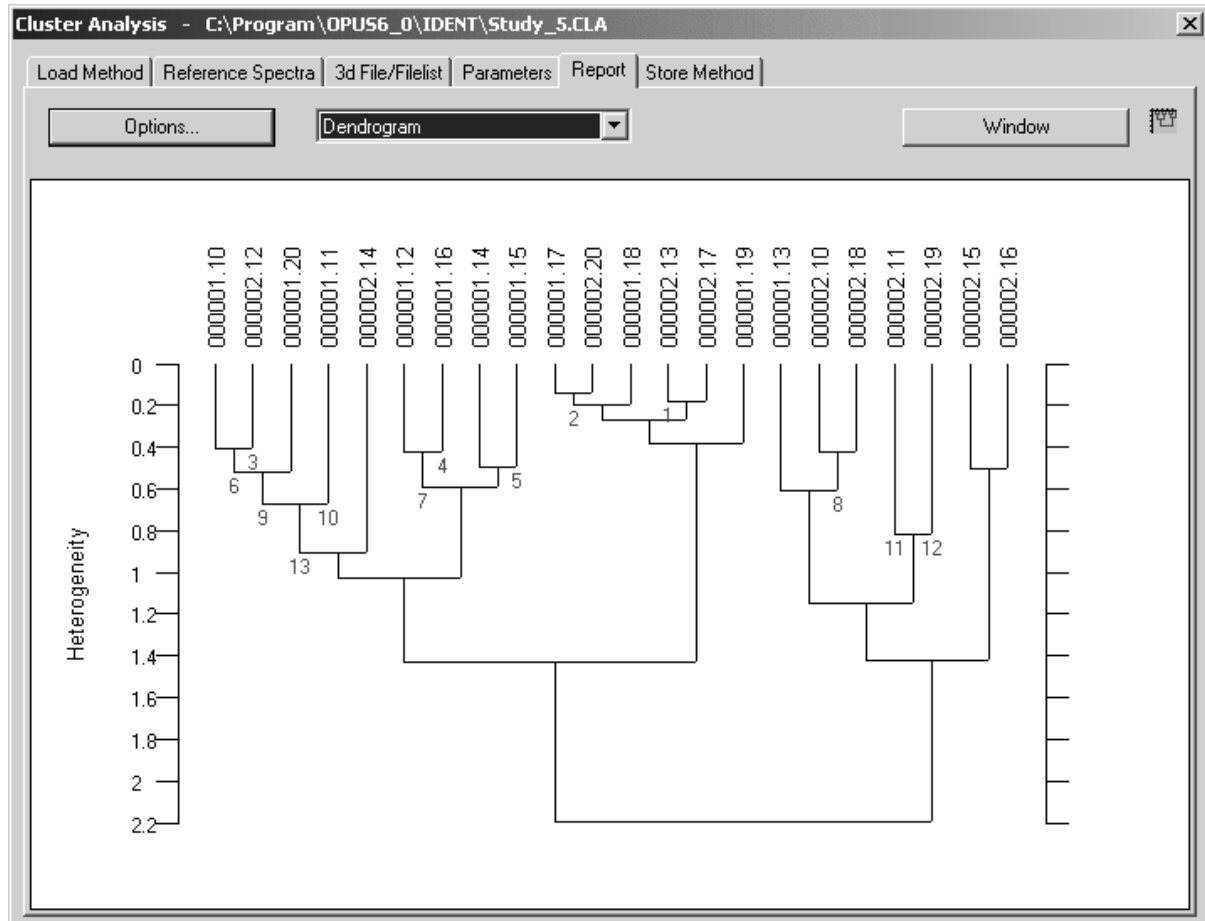


Figure 23: Cluster Analysis – Report tab with dendrogram

If you click on the *Window* button, you can modify the report display. A new dialog will open which allows to change the algorithm and the kind of labeling.

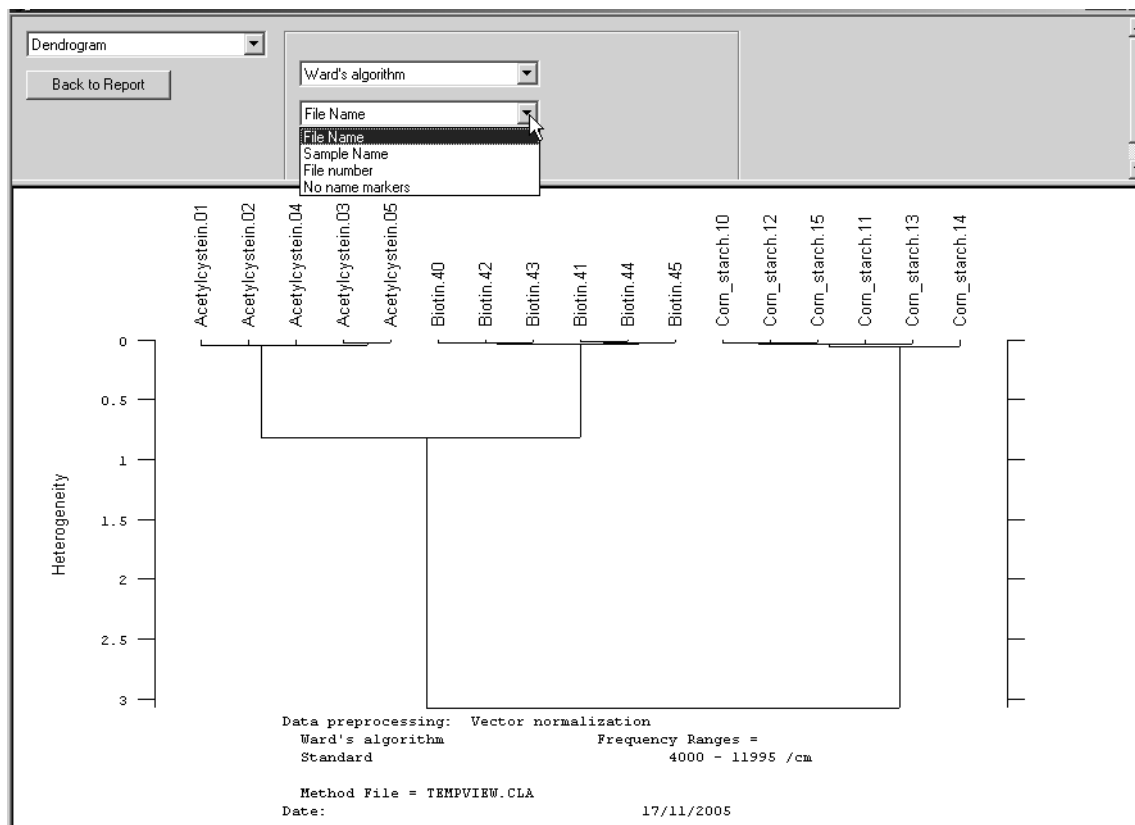


Figure 24: Cluster Analysis - Options to modify report display

To return to the report click on the *Back to Report* button.

The *Options* button on the *Report* tab also enables to modify the report display, and to define the algorithm and matrix parameter in more detail. You can select *Sample Name* from the *Dendrogram* drop-down list in the *Cluster Analysis* dialog to change the dendrogram labeling accordingly. You can also add file names, sample names or file numbers to the dendrogram or have the dendrogram displayed without any labeling.

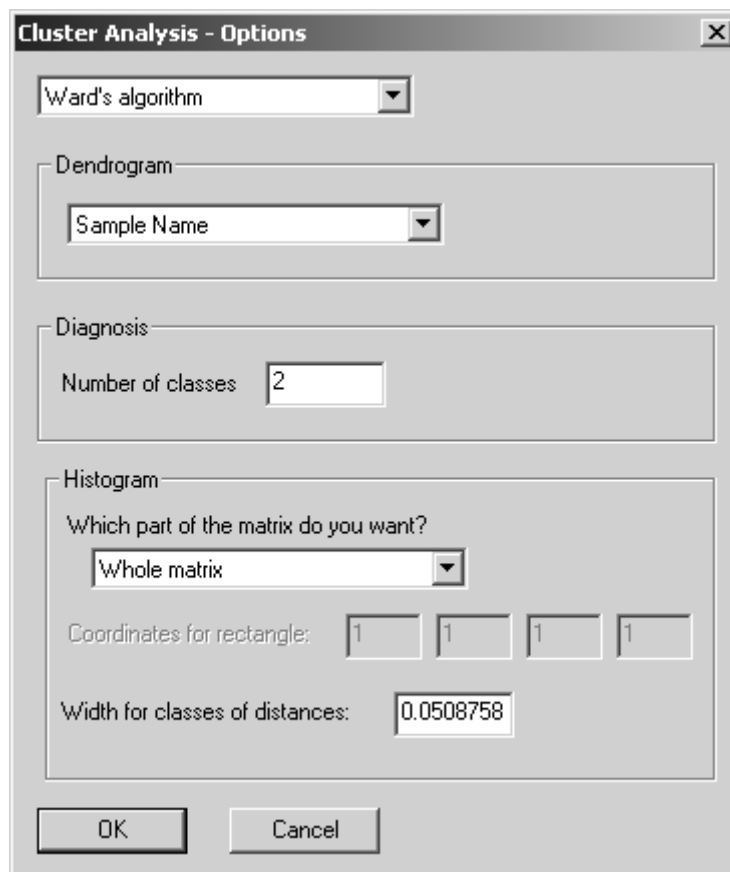


Figure 25: Cluster Analysis – Options

For further details on the *Cluster Analysis* dialog, see section 7.11.

4.3 3D Files/Filelist

Instead of single spectra it is also possible to use 3D files or file lists in connection with the cluster analysis. Note that not more than one 3d file or file list can be loaded to perform a cluster analysis. For further details on this subject refer to chapter 7.12.

5 Conformity Test

The conformity test is an easy method to test the deviations of measured NIR spectra within certain limits. To set these limits you need samples, which belong to at least one batch or one production cycle, of the final product as reference spectra. These reference spectra vary within the accepted range of specifications. The NIR spectra of these samples reflect the different sample variations and form a *confidence band* in the spectral range. To pass the conformity test, the spectrum of a new sample has to be within this *confidence band* on each wavelength.

The conformity test is mainly used for the quality control of defined products for which a quantitative calibration would be too time-consuming or even impossible.

First, you have to calculate the average and the standard deviation σ of the absorbance values for each wave length i . The mean value plus/minus the standard deviation determine the *confidence band* within the spectral range, and define which amount of variations on each spectral wavelength is acceptable for the particular product.

Second, you have to check whether the spectrum of a sample to be tested is within the defined *confidence band* in the spectral range. The difference between this sample and the average of the reference samples is calculated on each wave length i . This absolute deviation is now weighted by the corresponding standard deviation σ on the respective wavelength, which results in a relative deviation referred to as *Conformity Index (CI)*.

$$CI = (A_{\text{reference},i} - A_{\text{sample},i}) / \sigma_{\text{reference},i}$$

The maximum of all CI values is derived as test result.

5.1 Setting up Conformity Test

Select the *Setup Conformity Test* command in the *Evaluate* menu. On the *Load Method* tab you can load an already existing method file generated for a conformity test.

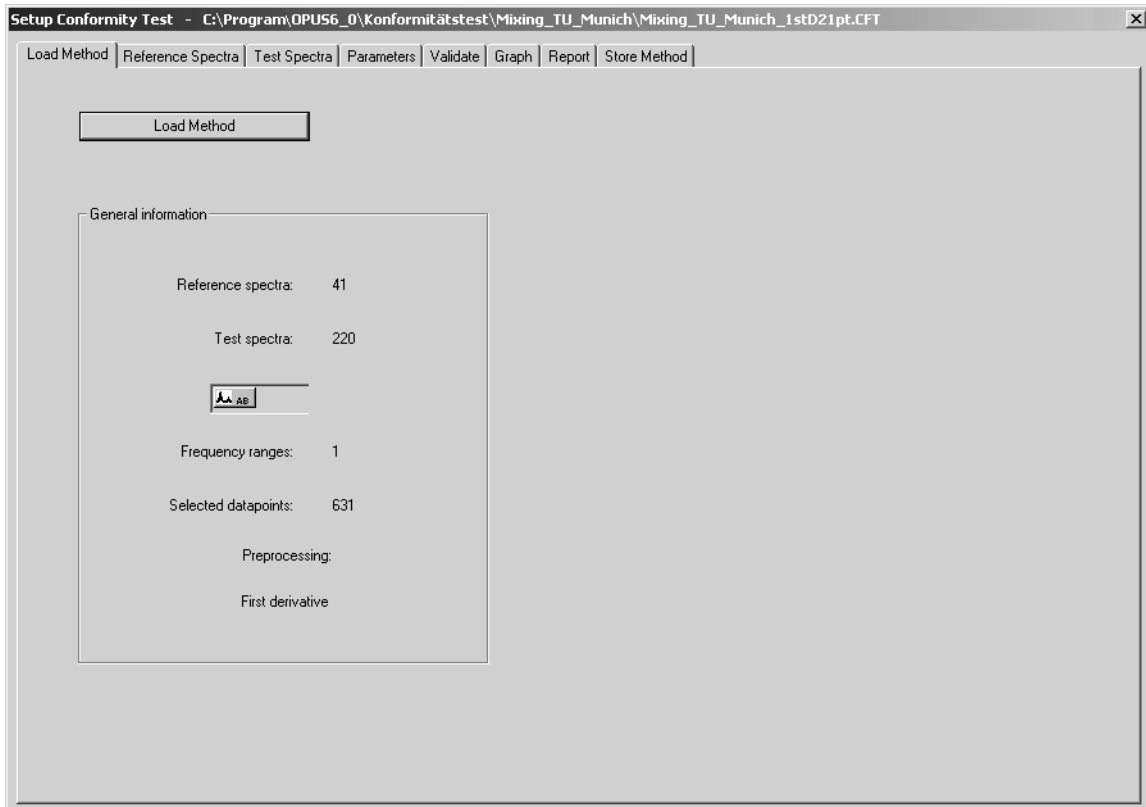


Figure 26: Setup Conformity Test - Load Method tab

The *General information* group field includes information on the kind of data block, data points selected as well as the data preprocessing type. It is distinguished between reference and test spectra. Reference spectra have been created by a specific method, whereas test spectra can be tested for their conformity with this specific method for validation purposes.

To create a new method click on the *Reference Spectra* tab and load the respective reference spectra by clicking on the *Add Reference Spectra* button. A dialog pops up and displays a browser window which you have to use to search for and select the spectra. Click on the *Open* button to load the spectra.

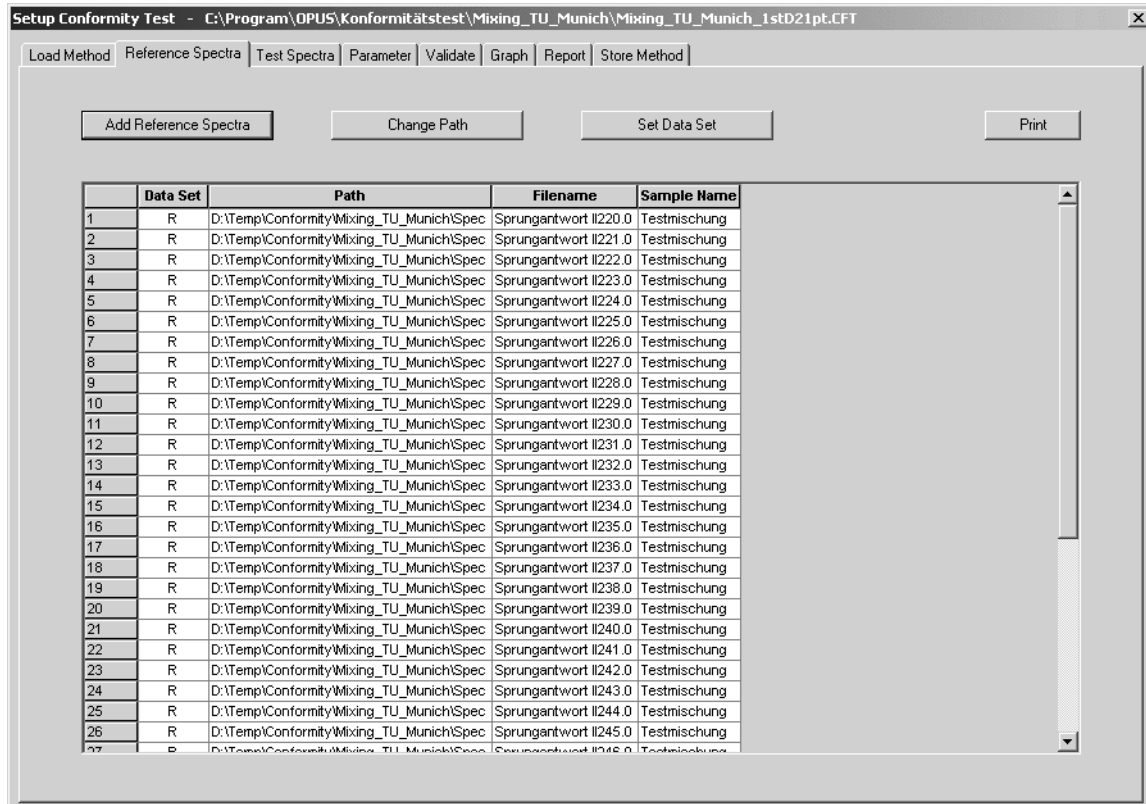


Figure 27: Setup Conformity Test - Reference Spectra tab

The *Reference Spectra* and *Test Spectra* tab are based on the same principle. Figure 27 exemplifies a method which consists of a large number of reference spectra. The *Data Set* column specifies the type of spectra, in this case *R* indicates reference, *T* test. Further spectra features are the path, file and sample name, which are the same for both tabs.

If you want to change the path for the conformity test spectra, click on the *Change Path* tab. A dialog opens which you use to define the new path.

It is also possible to modify the data set. Select the respective reference or test spectra and click on the *Set Data Set* button.

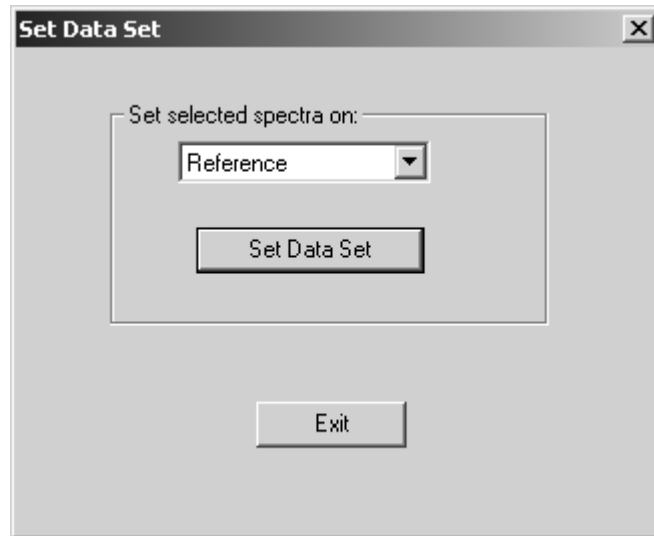


Figure 28: Setup Conformity Test - Set Data Set option

You can temporarily exclude some spectra or change them from reference to test spectra (or vice versa), or restrict the data set to a few spectra (*Excluded* option) only and create a new method. As the quality of a method depends on the reference and test spectra, make sure that the data sets are carefully assembled.

Select one of the options from the drop-down list and click on the *Set Data Set* button. To continue click on the *Exit* button.

The spectra excluded will be marked in gray, see figure 29.

29	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II248.0	Testmischung	
30	E	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II249.0	Testmischung	←
31	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II250.0	Testmischung	
32	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II251.0	Testmischung	
33	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II252.0	Testmischung	
34	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II253.0	Testmischung	
35	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II254.0	Testmischung	
36	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II255.0	Testmischung	
37	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II256.0	Testmischung	
38	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II257.0	Testmischung	
39	R	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II258.0	Testmischung	
40	E	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II259.0	Testmischung	←
41	E	D:\Temp\Conformity\Mixing_TU_Munich\Spec	Sprungantwort II260.0	Testmischung	

Figure 29: Setup Conformity Test - Spectra list with changed data set

On the *Parameters* tab you define the spectral regions to be used for the conformity test and select a data preprocessing method.

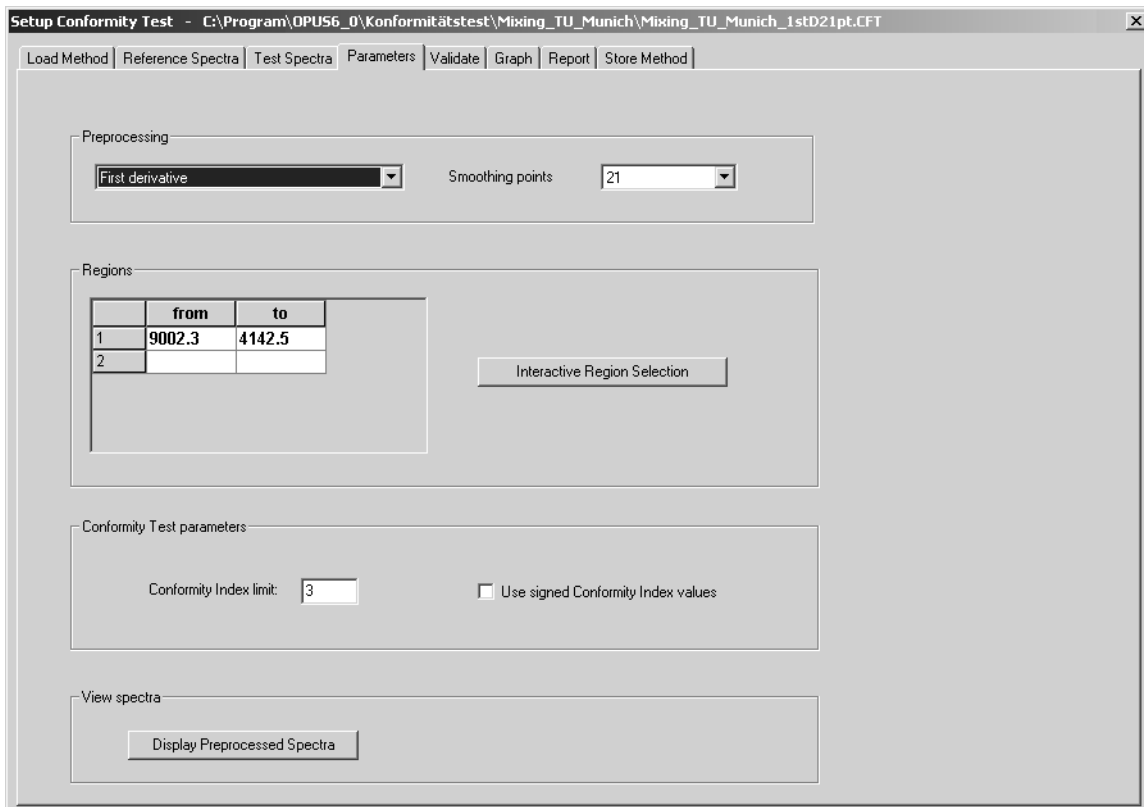
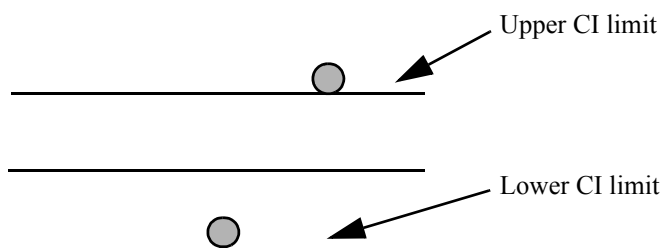


Figure 30: Setup Conformity Test - Parameter tab

Frequently, the *Vector normalization* is selected as preprocessing method. Sometimes better results can be obtained by using the *First* or *2nd derivative* method. In both cases you have to additionally define the amount of *Smoothing points*, you can select between 5 (9) and 25. The optimal number of smoothing points, however, has to be evaluated empirically.

The *Conformity Index Limit* parameter records spectra between an upper and lower limit (see graph below). The best possible scaling factor is between 3 and 4 which is, of course, not mandatory. The evaluation basis for the self-adapting conformity test is the reference spectra scaling, indicated by means of the standard deviation.



Activate the *Use signed Conformity Index Values* check box if you want to use non-absolute index values.

Click on the *Display Preprocessed Spectra* button to have the respective spectra and confidence band displayed. The confidence band is displayed in red and is calculated as follows:

$$\text{Average value} \pm \text{CI limit} \bullet \text{standard deviation}$$

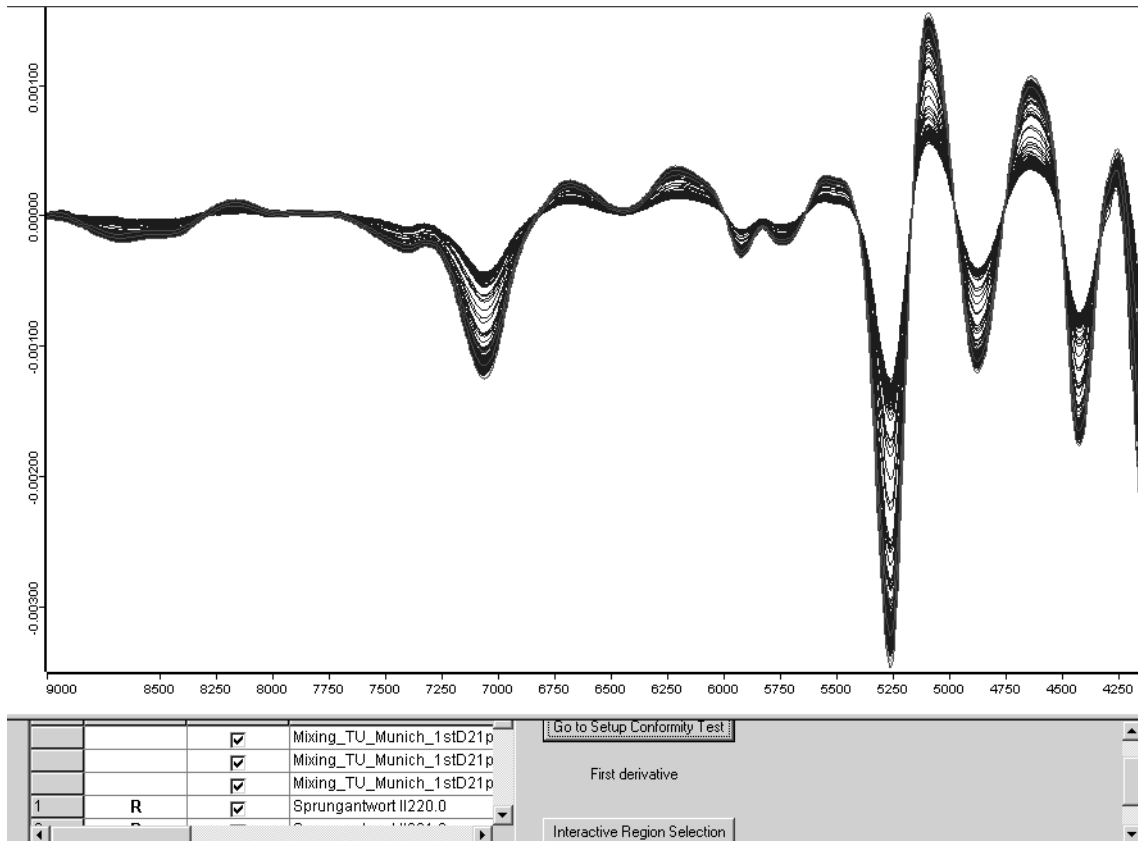


Figure 31: Preprocessed spectra plot

Use the selection box on the lower part of the dialog (figure 31) to have the spectra selectively displayed. Deactivate the *Show* check box of the spectra which you do not want to have displayed. If you click on the *Interactive Region Selection* button, the standard *Select Frequency Range(s)* dialog opens which you can use to interactively set the frequency range. To continue with the conformity test click on the *Go to Setup Conformity Test* button.

If you have defined all the parameters required, click on the *Validate* tab. This tab only includes the *Validate* button. Click on this button, and you will automatically be transferred to the *Graph* tab.

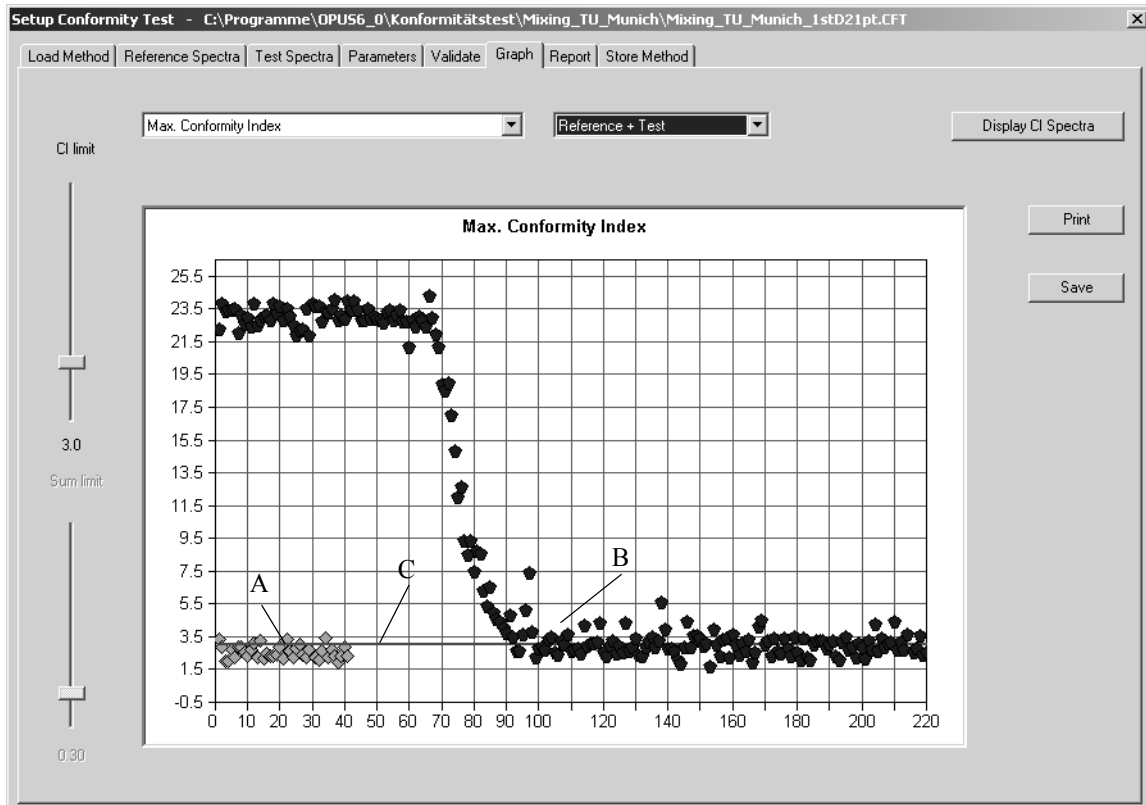


Figure 32: Setup Conformity Test - Graph tab

The green data points (A) represent the reference spectra, and the blue ones (B) represent the test spectra. The CI Limit is indicated by the red line (C) which can be moved by using the *CI limit* slider on the left side. To display the reference and test spectra separately, select either *Reference* or *Test* from the upper drop-down list. In most cases it is recommended to select the *Reference + Test* option, to be able to directly compare the scattering of reference vs test spectra.

If you position the cursor on one specific data point, a text frame pops up indicating the exact spectrum identity (see figure 33).

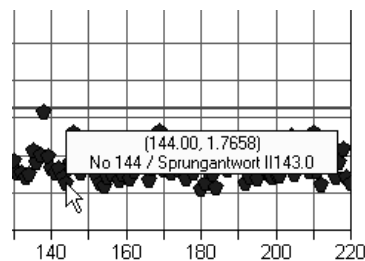


Figure 33: Spectra plot with spectrum description

You can also select single data points only. Move the cursor to the respective data point section, press the left mouse button and draw the mouse over this section. If you leave the mouse button, only the data points selected will be displayed in the plot. To undo this, just right click into the plot.

You can select between the following algorithms:

- **Max Conformity Index**
The maximum value will be calculated based on the frequency ranges selected.
- **Sum 1: Sum over CI > Limit (/N total)**
All y-values above the CI limit are added up and divided by the total number of data points within the frequency ranges selected.
- **Sum 2: Sum over CI > Limit (/N over Limit)**
All y-values above the CI limit are added up and divided by the number data points which are above the CI limit.

Depending on your specific quality control problem, you select either one of these algorithms. General recommendations cannot be made, you have to empirically find out which procedure would be the best for your specific requirements.

If you click on the *Display CI Spectra* button, the CI spectra and the CI limit will be displayed.

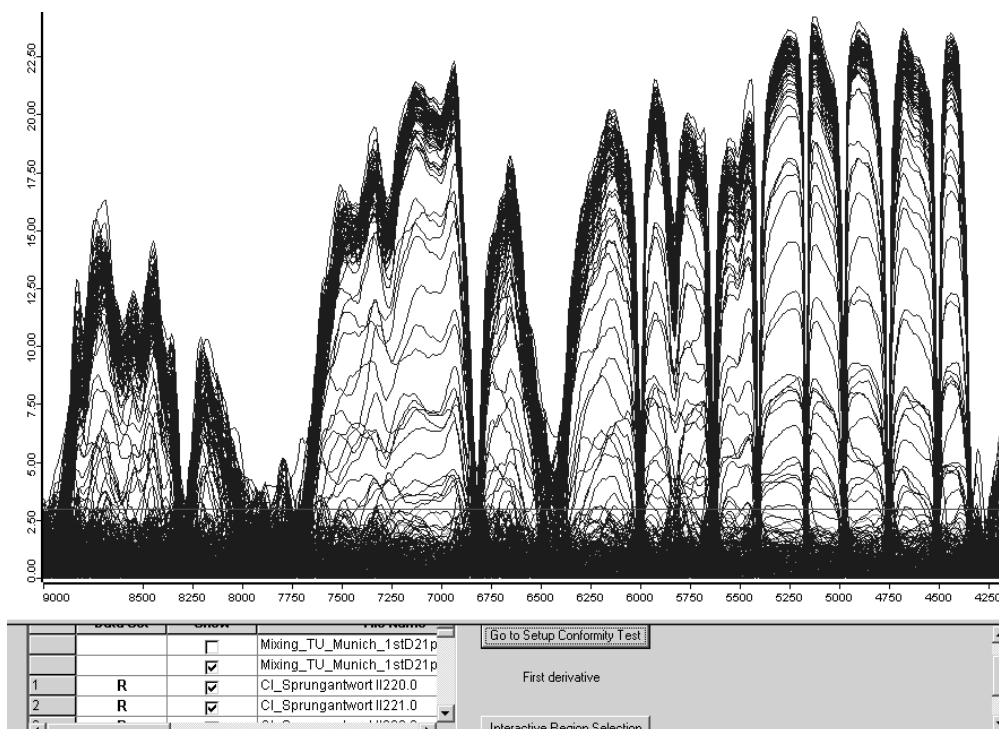


Figure 34: Setup Conformity Test - CI spectra

The reference spectra are displayed in blue, and the test spectra are displayed in green. To localize a specific spectrum and analyze the standard deviation, deactivate and re-activate the *Show* check box. Click on the *Go to Setup Conformity Test* button to return.

To display the spectra report click on the *Report* tab.

	D	File Name	Sample Name	Max. CI Value	at Freq.	Std. Dev.	Sum 1	Sum 2	Datapoints > CI Limit
1	R	Sprungantwort II220.0	Testmischung	3.37	4505.02	9.24E-006	0.000647	0.204	2
2	R	Sprungantwort II221.0	Testmischung	2.85	7868.36	6.28E-006	0	0	0
3	R	Sprungantwort II222.0	Testmischung	2.00	5160.72	1.07E-005	0	0	0
4	R	Sprungantwort II223.0	Testmischung	1.99	6448.97	5.71E-006	0	0	0
5	R	Sprungantwort II224.0	Testmischung	2.74	5407.57	3.98E-006	0	0	0
6	R	Sprungantwort II225.0	Testmischung	2.21	4227.31	6.92E-005	0	0	0
7	R	Sprungantwort II226.0	Testmischung	2.86	8956.04	9.75E-006	0	0	0
8	R	Sprungantwort II227.0	Testmischung	2.84	8886.61	8.37E-006	0	0	0
9	R	Sprungantwort II228.0	Testmischung	2.56	8014.92	5.74E-006	0	0	0
10	R	Sprungantwort II229.0	Testmischung	2.29	8354.34	6.48E-006	0	0	0
11	R	Sprungantwort II230.0	Testmischung	2.73	6387.25	6.53E-006	0	0	0
12	R	Sprungantwort II231.0	Testmischung	3.08	7621.51	5.57E-006	0.000168	0.0529	2
13	R	Sprungantwort II232.0	Testmischung	2.26	6587.82	8.54E-006	0	0	0
14	R	Sprungantwort II233.0	Testmischung	3.23	6904.10	9.64E-006	0.00119	0.15	5
15	R	Sprungantwort II234.0	Testmischung	2.19	8346.63	6.16E-006	0	0	0
16	R	Sprungantwort II235.0	Testmischung	2.46	4273.60	4.65E-005	0	0	0
17	R	Sprungantwort II236.0	Testmischung	2.35	8824.90	6.82E-006	0	0	0
18	R	Sprungantwort II237.0	Testmischung	2.35	6641.82	8.12E-006	0	0	0
19	R	Sprungantwort II238.0	Testmischung	2.38	7899.21	5.20E-006	0	0	0
20	R	Sprungantwort II239.0	Testmischung	2.87	7706.36	5.11E-006	0	0	0
21	R	Sprungantwort II240.0	Testmischung	2.20	4751.87	8.57E-006	0	0	0
22	R	Sprungantwort II241.0	Testmischung	3.37	8261.77	8.12E-006	0.00299	0.118	16
23	R	Sprungantwort II242.0	Testmischung	2.63	5407.57	3.98E-006	0	0	0
24	R	Sprungantwort II243.0	Testmischung	2.26	7783.50	4.88E-006	0	0	0
25	R	Sprungantwort II244.0	Testmischung	2.82	8161.49	8.53E-006	0	0	0
26	R	Sprungantwort II245.0	Testmischung	3.04	6780.67	6.35E-006	6.88E-005	0.0434	1
27	R	Sprungantwort II246.0	Testmischung	2.38	5777.84	5.68E-006	0	0	0

Figure 35: Setup Conformity Test - Report tab

Depending on the algorithm selected from the drop-down list, the corresponding outliers are marked in grey color.

The *Data Points > CI Limit* column refers to the data points above the CI limit, which is important to analyze the integral. The values displayed in the *Sum 1* column represent the integrals divided by all data points, and the values displayed in the *Sum 2* column represent the integrals divided by the data points which are above the CI limit. If you want to print the report, click on the *Print* button.

Before you store the method, you have to define the analysis method. Either activate the *Use CI Limit*, *Use Sum 1 Limit* or *Use Sum 2 Limit* option button on the *Store Method* tab.

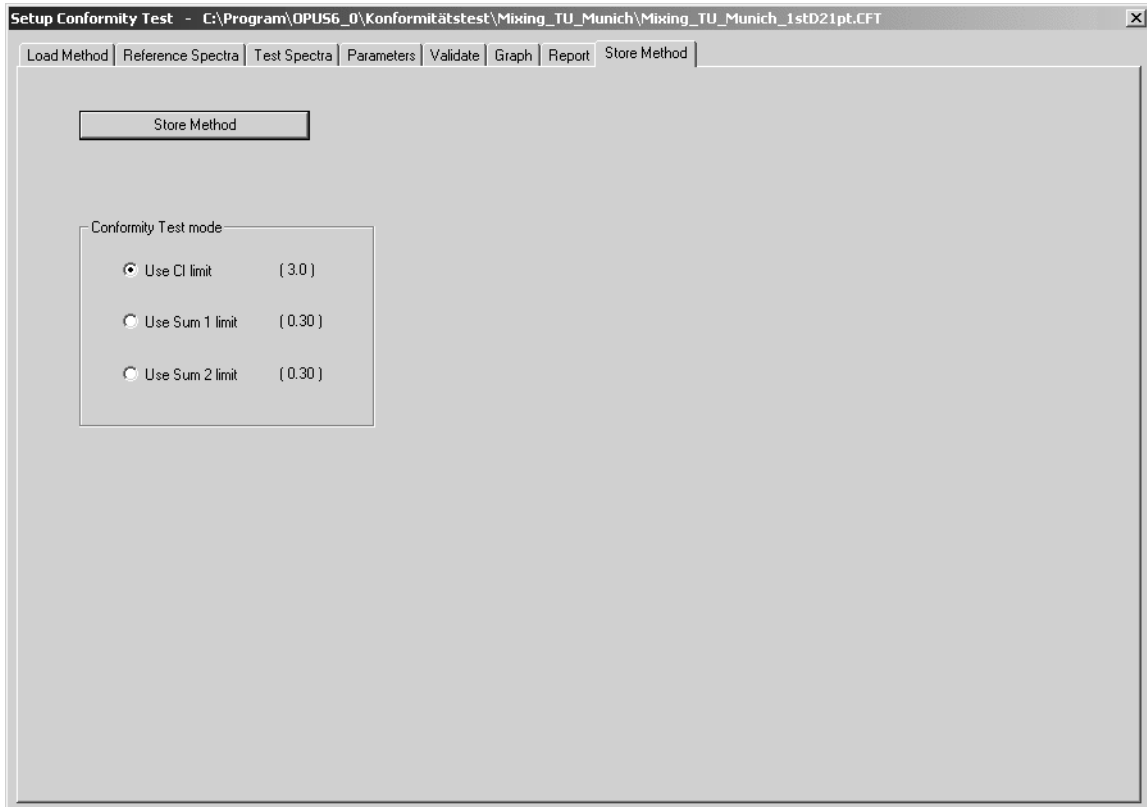


Figure 36: Setup Conformity Test - Store Method tab

5.2 Performing Conformity Test

Start the conformity test by clicking on the *Conformity Test* command. The following dialog opens:

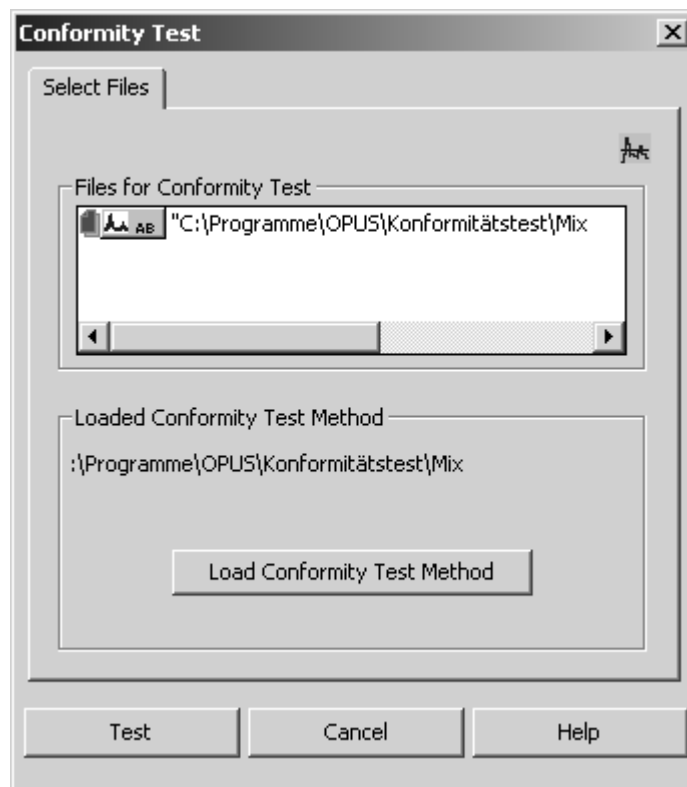



Figure 37: Conformity Test - Select Files tab

Drag & drop the file(s) to be evaluated from the OPUS browser window into the *File(s) for Conformity* selection field.

Click on the *Load Conformity Method* button and load the particular method which path is displayed above this button.

The *Conformity Test* results will be stored in a CONF data block () and displayed in a specific report view.

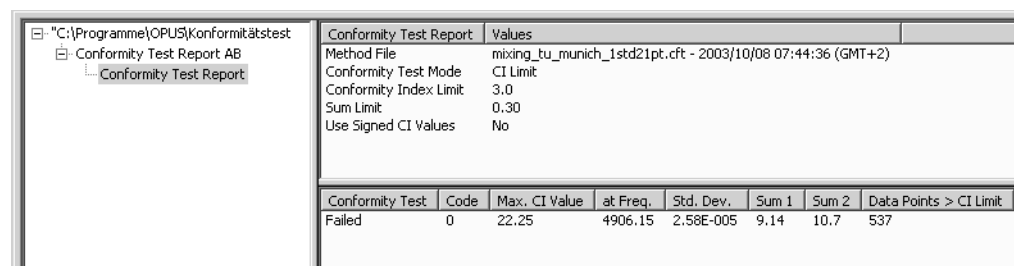


Figure 38: Conformity Test Report

6 IDENT Theory

The aim of an IDENT analysis is to determine the differences between a test spectrum and the reference spectra of a library. You have to define a method to test the similarity of spectra, and a threshold. This threshold determines whether a spectrum is only similar or even identical to the reference spectrum.

6.1 Algorithms

Basically, there are two algorithms to perform the IDENT analysis: the *Standard* and *Factorization* method. During the analysis both methods compare the test spectrum with all reference spectra. The result of a comparison between two spectra is the *Hit Quality*, also referred to as spectral distance D . The better two spectra match, the smaller the spectral distance. The *Hit Quality* for identical spectra is 0 (i.e. if a reference spectrum is compared with itself).

6.1.1 Standard Method

Figure 39 shows two spectra a and b , one test and one reference spectrum. The spectral distance D is proportional to the area between these two curves. The following formula for the so-called Euclidean distance is used in the *Standard* method:

$$D = \sqrt{\sum_k (a(k) - b(k))^2} \quad (6-2)$$

where $a(k)$ and $b(k)$ are the ordinate values of the a and b spectra. The sum incorporates all selected k data points.

In the current IDENT report (see figure 40) the smallest spectral distance which has been determined by spectrum comparison is 0.96. In this case one of the sample spectra previously used to generate an average spectrum for the reference library serves as a test sample, which explains the extremely small spectral distance. As this example shows, the first two hits are well separated from each other. The spectral distance for Hit No. 2 is 6.19, which is about 6 times the smallest spectral distance.

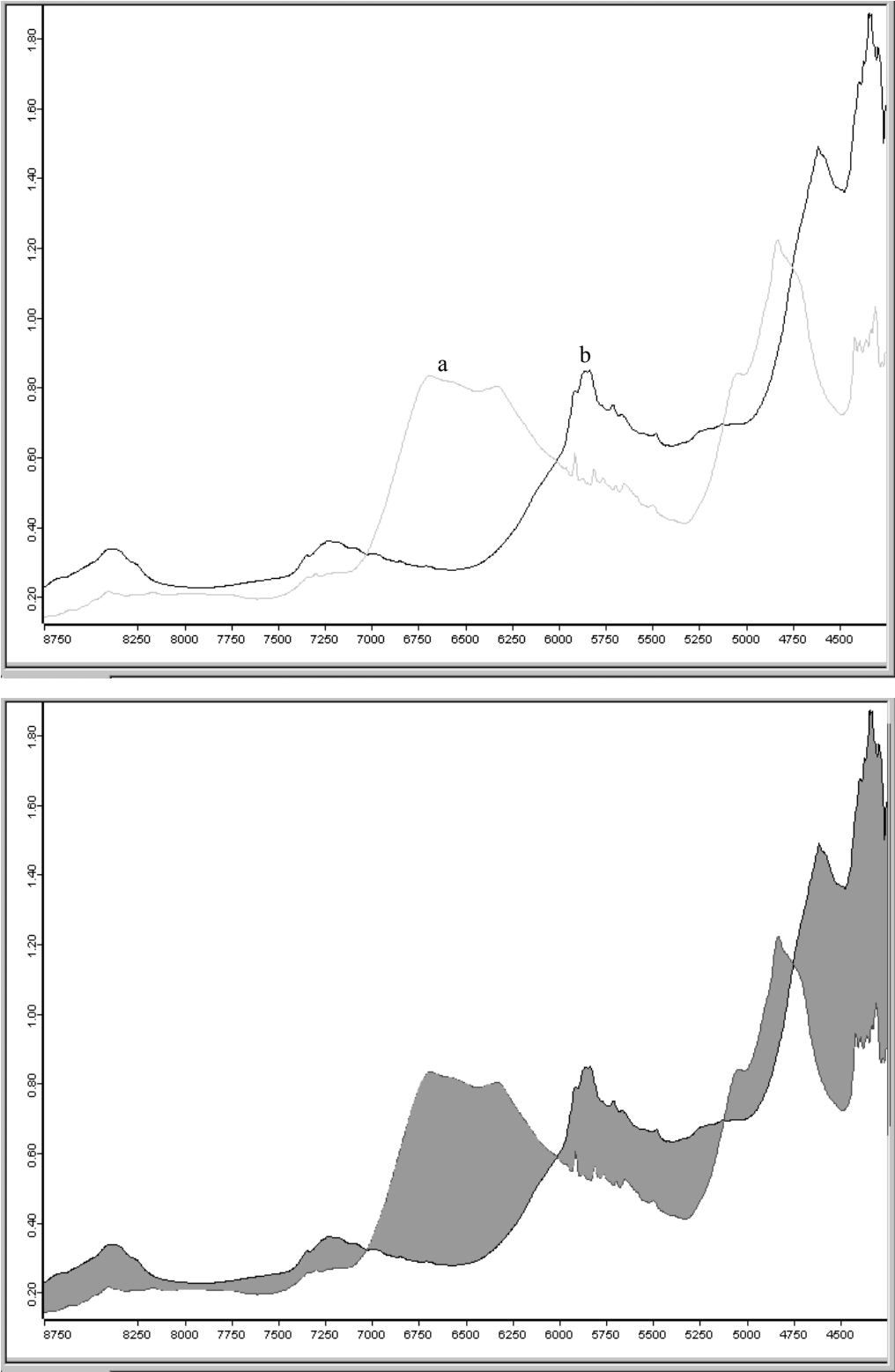


Figure 39: Two spectra and their spectral distance

Report of Correlation Search		Values
Method file:	\\WS\Ident\example2\New.FAA	
from [date]:	29/05/00	
(time):	12:11:17	
Description:		
IDENTITY NOT CHECKED:	0	
Hit quality with expected reference:	0.000000	
No Threshold avail:	0.000000	
Threshold calculation:	- depends on each ref.-spectrum -	
Algorithm:	Standard	
Vector normalized spectra:	No	
Order of Derivative:	0	
Smoothing points:	1	
No. of used factor sp.:	0	
10 hits of 15		
X-Ranges:	1	
Class Name:		
Class Test NOT PERFORMED:	0	
Using residuals:	No	
Order of Internal Derivative:	0	
Smoothing Points for Internal Derivative:	1	
Reduction Factor:	1	

Hit No.	Hit Quality	Sample Name	File Name	Threshold
1	0.964332	000001 Sample L-Leucin Av. of 11	An000001.100	15.463419
2	6.198979	000004 Sample DL-Alanin Av. of 11	An000004.100	17.56358
3	8.598950	000005 Sample L-Alanin Av. of 11	An000005.100	16.701758
4	8.967188	000015 Sample L-Methionin Av. of 11	An000015.100	17.573791
5	9.232257	000003 Sample L-Isoleucin Av. of 11	An000003.100	16.73790
6	9.693203	000014 Sample DL-Methionin Av. of 11	An000014.100	18.17575
7	11.084043	000002 Sample DL-Isoleucin Av. of 11	An000002.100	8.758205
8	11.231035	000006 Sample DL-Tryptophan Av. of 11	An000006.100	15.760675
9	11.261703	000007 Sample L-Tryptophan Av. of 11	An000007.100	21.082051
10	12.836960	000008 Sample Glucose H2Ofrei Av. of 11	An000008.100	18.293217

Figure 40: Ident Report

6.1.2 Factorization

The *Factorization* method represents spectra as linear combinations of so-called factor spectra (loadings):

$$a = T_{1a} \cdot f_1 + T_{2a} \cdot f_2 + T_{3a} \cdot f_3 + \dots \quad (6-3)$$

The a vector shows the a spectrum and the factor spectra are denoted f_1, f_2, f_3 etc. T indicates the coefficients (scores) required to reconstruct the original a spectrum.

To calculate the spectral distance D between the two spectra a and b , the T coefficients are used in the *Factorization* method:

$$D = \sqrt{\sum_i (T_{ia} - T_{ib})^2} \quad (6-4)$$

The summation is performed for a certain number of coefficients. These T coefficients are also called *scores*. The differences between the original and reconstructed spectrum are known as *spectral residuals* (figure 41).

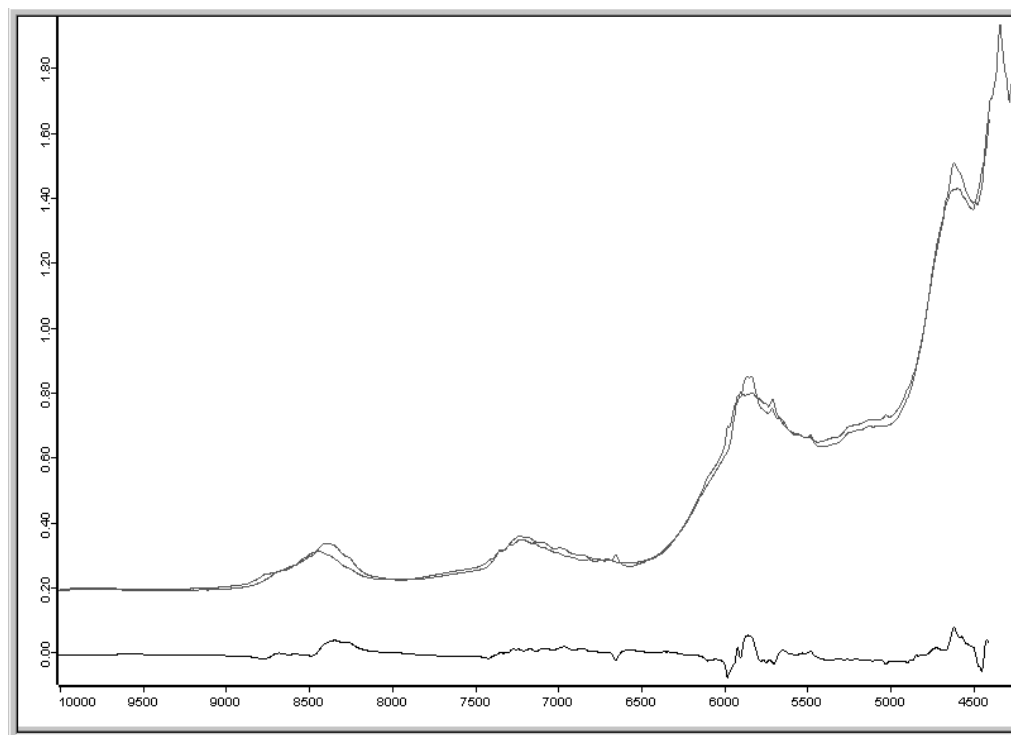


Figure 41: Reference spectrum, reconstructed reference spectrum and difference spectrum

The *Standard* method directly uses the spectral intensities to calculate the spectral distance, and the summation is performed for all data points within the specified frequency regions (which could be more than 1000 points). How many factor spectra or score coefficients have to be included in an IDENT library is a very important aspect and will be explained in the following.

When factorizing an IDENT library, s average spectra are transformed into s factor spectra. These factor spectra are orthogonal to each other. The effect a certain factor has on the reproduction of reference spectra is indicated by the respective *Eigen value*. The factor spectra are sorted according to these *Eigen values*. The first factor spectrum is the most important one and thus has the highest *Eigen value*.

The more the *Eigen value* decreases, the lower the spectral intensities (ordinate values) of the factor spectra, and the more intensive the noise. Factor spectra which mainly consist of noise must not be used for an IDENT method.

Factor spectra are stored in the IDENT method directory, using the OPUS file format. They have the same file name as the method. However, the spectra have an additional numeric file extension, starting with 0 for the first factor spectrum. These spectra can be loaded into OPUS like any other spectrum.

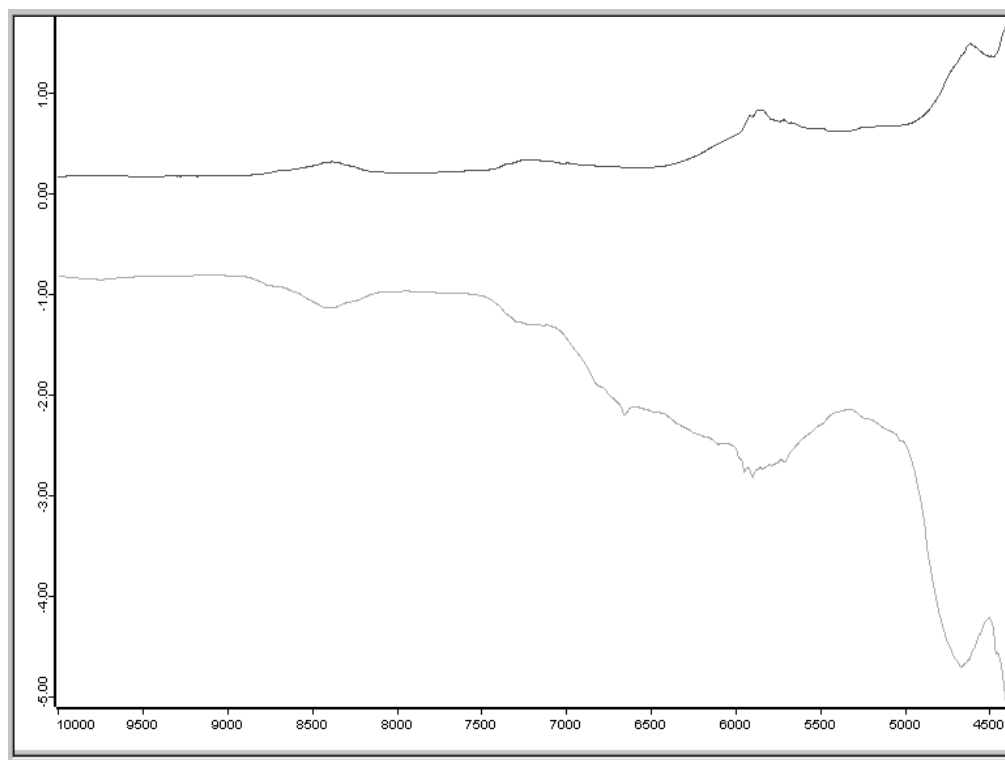


Figure 42: Reference spectrum and first factor spectrum

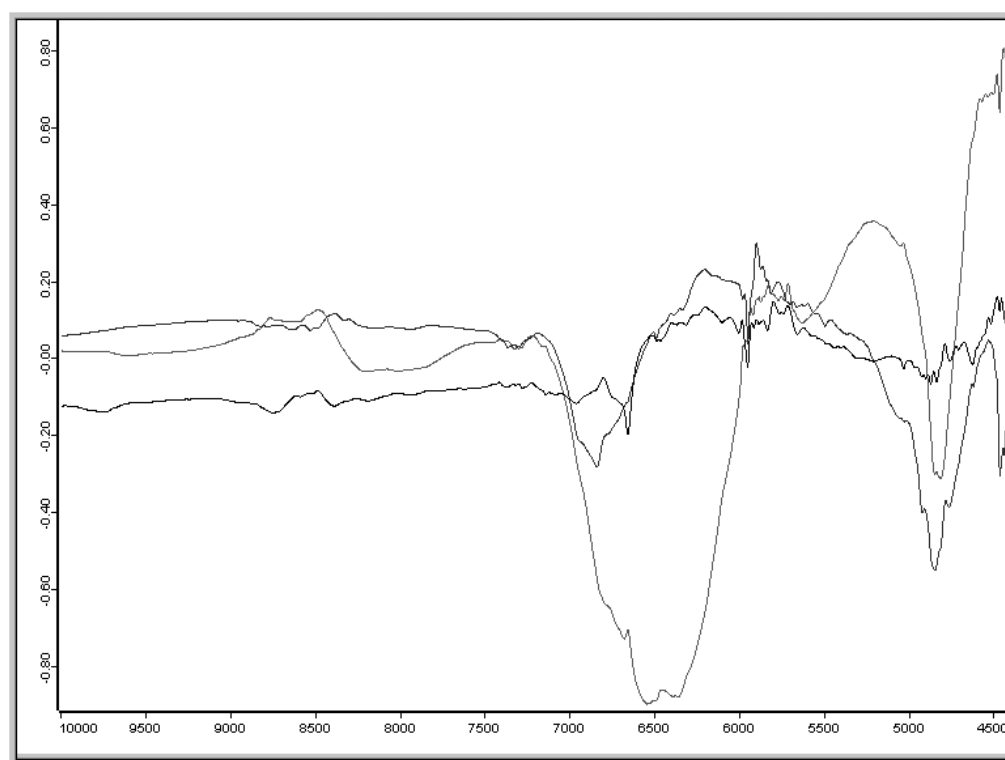


Figure 43: Second, third and fourth factor spectrum

The spectra (see figure 42 and 43) show the signal-to-noise ratio of a certain factor spectrum. The factor spectrum displayed in figure 44 mainly consists of noise. It is not recommended to use this spectrum to calculate spectral distances.

You can easily check the factor spectra orthogonality by multiplying two factor spectra using the OPUS *Spectrum Calculator*. This is followed by an integration across the whole frequency range of the result spectrum. The integration result will be 0 (or approximately 0 due to the finite computing accuracy).

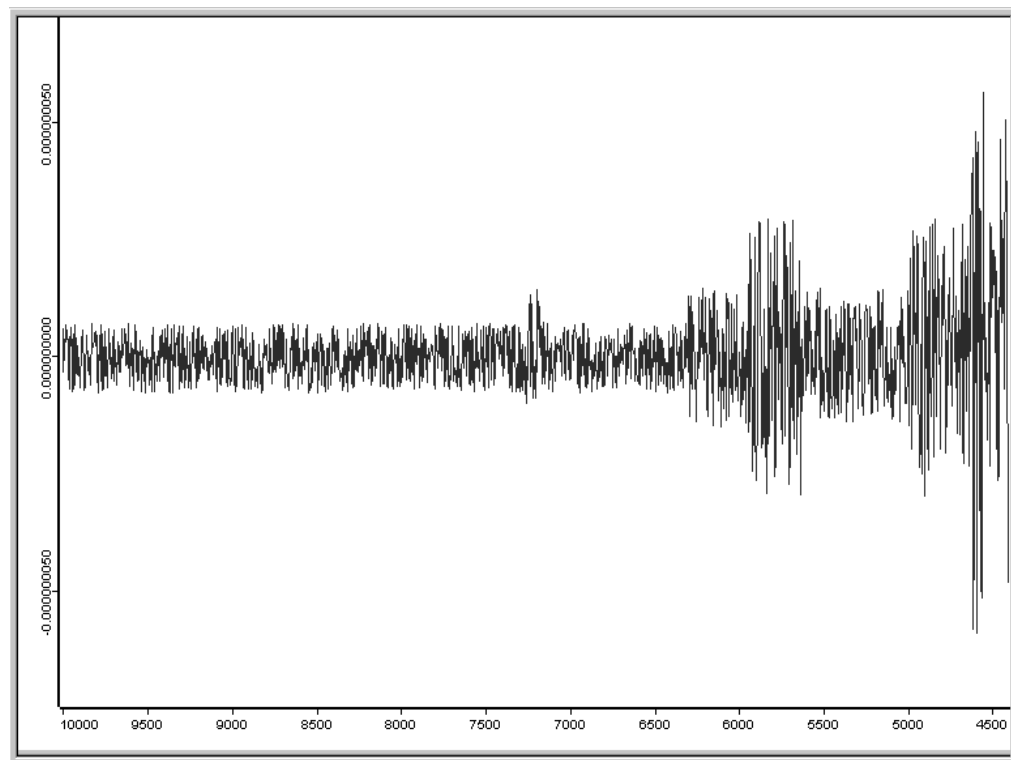


Figure 44: Factor spectrum with excessive noise

6.2 Factorization Theory

Assuming that s reference spectra consist of d data points each. The reference spectra are represented by $d_1, d_2, d_3 \dots$ column vectors which form D data matrix ($d \times s$ dimension):

$$D = [d_1 \ d_2 \ d_3 \ \dots \ d_s] \quad (6-5)$$

When exchanging the rows for the columns in this matrix, you obtain D^T transposed matrix of $s \times d$ dimension. Multiply D^T transposed matrix by D original matrix to obtain Z covariance matrix:

$$Z = D^T \cdot D \quad (6-6)$$

A diagonalization and orthogonal transformation of Z produce *Eigen vectors* and *Eigen values* of Z .

$$\Lambda = L^T \cdot Z \cdot L \quad (6-7)$$

The column vectors of L matrix ($s \times s$ dimension) are $l_1, l_2, l_3 \dots$ *Eigen vectors* of Z matrix. These *Eigen vectors* are orthonormal, i.e. the following equations are valid for the scalar product of two *Eigen vectors*:

$$l_i \cdot l_j = 0 \quad i \neq j \quad (6-8)$$

$$l_i \cdot l_i = 1 \quad (6-9)$$

Λ matrix ($s \times s$ dimension) contains $\lambda_1, \lambda_2, \lambda_3 \dots \lambda_s$ *Eigen values* of Z matrix as the main diagonal, all other matrix elements are 0. This means:

$$Z \cdot l_i = \lambda_i \cdot l_i \quad (6-10)$$

D data matrix is factorized by L Eigen vector matrix, using multiplication:

$$F = D \cdot L \quad (6-11)$$

F matrix has the same dimensions as D data matrix ($d \times s$) and includes the vectors of f_1, f_2, f_3, \dots factor spectra as columns. Multiplying F^T transposed matrix by F yields:

$$F^T \cdot F = (D \cdot L)^T \times (D \cdot L) = L^T \cdot D^T \cdot D \cdot L = L^T \cdot Z \cdot L = \Lambda \quad (6-12)$$

The elements of $F^T F$ quadratic matrix ($s \times s$ dimension) are the scalar products which can be created in pairs together with factor spectra. The 6-12 equation causes the factor spectra to be orthogonal to each other:

$$f_i \cdot f_j = 0 \quad i \neq j \quad (6-13)$$

$$f_i \cdot f_i = \lambda_i \quad (6-14)$$

The vector norm of a factor spectrum is equal to the square root of the corresponding *Eigen value*. Using L orthogonality, D data matrix can be as follows:

$$D = D \cdot 1 = D \cdot L \cdot L^T = F \cdot L^T \quad (6-15)$$

The reference spectra are represented as linear combinations of the factor spectra, and the coefficients are contained in the columns of L^T matrix. Based on the 6-15 equation the following applies to the first reference spectrum:

$$d_1 = L_{1,1}^T \cdot f_1 + L_{2,1}^T \cdot f_2 + L_{3,1}^T \cdot f_3 + \dots + L_{s,1}^T \cdot f_s \quad (6-16)$$

The score coefficients are the coordinates of the reference spectra in the factor spectra system.

Any u spectrum can be represented as linear combination of the factor spectra:

$$u = F \cdot k \cdot e \quad (6-17)$$

The unknown k column vector corresponds to the column elements of L^T matrix. E error spectrum is the difference between the u spectrum and reconstructed spectrum. The *Least Squares* solution for k which minimizes the error is as follows:

$$k = (F^T \cdot F)^{-1} \cdot F^T \cdot u = \Lambda^{-1} \cdot F^T \cdot u \quad (6-18)$$

If only the first r factor spectra are taken into account, D spectral distance between one u spectrum and one d_a reference spectrum is:

$$D = \sqrt{(k_1 - L_{1a}^T)^2 + (k_2 - L_{2a}^T)^2 + \dots + (k_r - L_{ra}^T)^2} \quad (6-19)$$

Instead of using the column vectors of L^T matrix you can use L row vectors. The equation for D spectral distance in r dimensional factor space is then:

$$D = \sqrt{(k_1 - L_{a1})^2 + (k_2 - L_{a2})^2 + \dots + (k_r - L_{ar})^2} \quad (6-20)$$

When using the *Standard* method you can specify a range of values for D distances. Select *Vector Normalization* preprocessing. The range of values reaches from 0 (identical spectra) to 2 (maximum spectral difference). This does not apply to D when using the *Factorization* method. If all factor spectra are used, the spectral distances between reference spectra are constant, i.e. $D = \sqrt{2}$. To calculate spectral distances, the elements of L Eigen vector matrix are used. This matrix consists of orthogonal unit vectors in s dimensional space. The distance between two orthogonal unit vectors always has to be $\sqrt{2}$. However, when using the *Factorization* method not all factor spectra are used, because the higher factor spectra mainly increase the noise level of the reconstructed spectrum.

The *Factorization* method also allows to calculate spectral distances by using residuals. For details see chapter 7.3. The spectral residual is calculated from the difference between the original and reconstructed spectrum. To calculate $SpecRes_u$ spectral residual (u being an arbitrary spectrum) the equation is as follows:

$$SpecRes_u = \sqrt{\sum_k (u(k) - (k_1 \cdot f_1(k) + k_2 \cdot f_2(k) + \dots + k_r \cdot f_r(k)))^2} \quad (6-21)$$

The summation is performed for all selected k data points.

D spectral distance between u spectrum and d_a reference spectrum is calculated as follows:

$$D = \sqrt{(k_1 - L_{a1})^2 + (k_2 - L_{a2})^2 + \dots + (k_r - L_{ar})^2 + (SpecRes_u - SpecRes_a)^2} \quad (6-22)$$

Figure 45 shows a scheme representing the *Factorization* method.

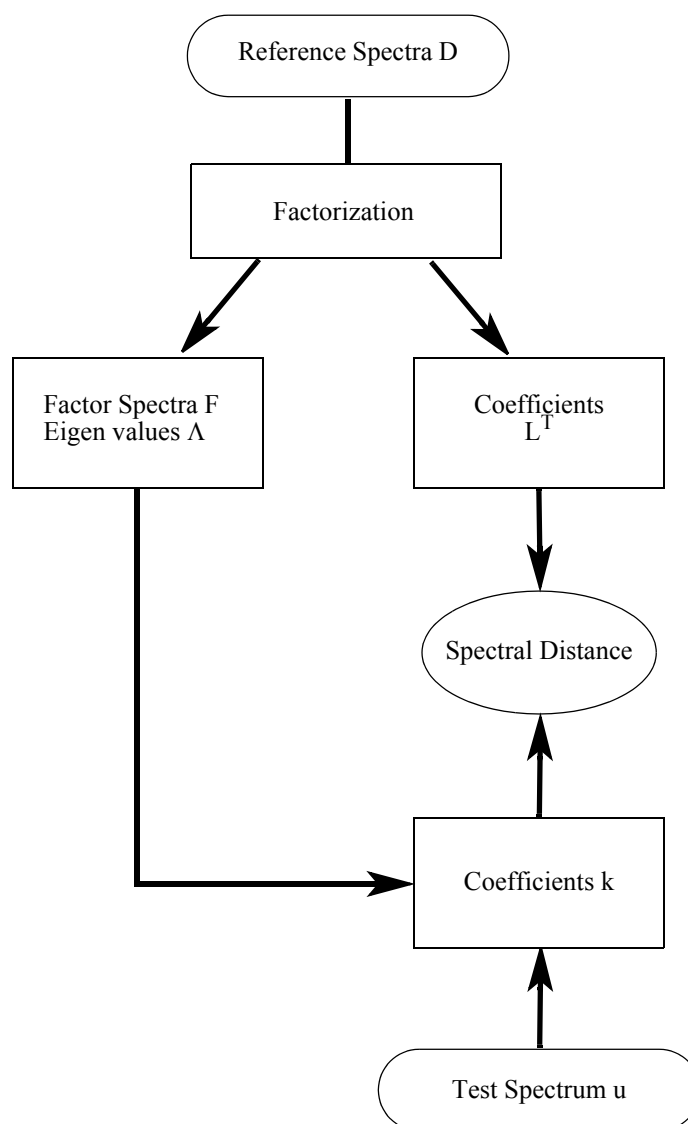


Figure 45: Factorization – Spectral distance calculation

6.2.1 Scaling to 1st Range and Normalize to Replevel

Scaling to 1st Range and *Normalize to Replevel* are algorithms that can be used to identify microorganisms. Contrary to the *Standard* and *Factorization* method overlapping spectral ranges are not merged. For example, if you set $1500\text{-}1200\text{cm}^{-1}$ as first frequency range and $1500\text{-}1400\text{cm}^{-1}$ as second frequency range, they will not be combined into one frequency range, and the data points in the $1500\text{-}1400\text{cm}^{-1}$ range will be considered twice to calculate spectral distances. The single spectral ranges can be weighted by different factors. These factors are defined in the *Weight* column (see chapter 7.3).

The *Vector Normalization* preprocessing is not available in combination with the *Scaling to 1st Range* and *Normalization to Replevel* algorithms, as a vector normalization will be automatically performed in this case. Contrary to the

Standard and *Factorization* algorithm the vector normalization is calculated separately for each spectral range. The resulting values are used to determine a mean distance. The vector normalization considers all data points selected, when using the *Standard* and *Factorization* algorithm.

Calculating spectral distances

- 1) To calculate spectral distances spectra have to be vector normalized first, for each frequency range.
- 2) Then, the r Pearson correlation coefficient is calculated. This kind of coefficient defines the correlation between a and b spectra:

$$r = \sum a_n(k) \cdot b_n(k) \quad (6-2)$$

This correlation is calculated separately for each frequency range. The summation covers all k data points of a frequency range: a_n and b_n are the normalized spectral intensities. The normalization yields:

$$r = \frac{\sum (a_n(k) - a_m) \cdot (b_n(k) - b_m)}{\sqrt{\sum (a_n(k) - a_m)^2} \cdot \sqrt{\sum (b_n(k) - b_m)^2}} \quad (6-3)$$

a_m and b_m are the mean spectral intensities within the spectral range, while $a(k)$ and $b(k)$ are the original spectral intensities.

The value range of r correlation coefficient reaches from -1 (inverse spectra) to +1 (identical spectra).

- 3) The correlation coefficient is transformed into D spectral distance by the following equation:

$$D = (1 - r) \cdot 1000 \quad (6-4)$$

D spectral distance can be between 0 (identical spectra) and 2000 (inverse spectra).

- 4a) The *Scaling to 1st Range* determines the minimum and maximum value of spectral distances for the first spectral range. Then, the distances of all the other spectral ranges are calculated and scaled to the same range of values, i.e. the same minima and maxima like the first spectral range.

Example: The spectral distances in the first spectral range are between 2 and 10, in the second between 6 and 22. The distances of the second spectral range are transformed as follows:

$$D \rightarrow 0.5 \cdot D - 1 \quad (6-5)$$

After this transformation the spectral distances of the second spectral range have the same values as the distances belonging to the first spectral range.

As the scaling has referred to the first spectral range, it does matter which spectral range is selected first. Make sure to select the correct spectral range as first range.

If spectral distances are sorted according to ascending values, this order directly results from the spectral ranges selected. For example, when comparing a test spectrum with a reference spectrum using an IDENT

test, the spectral distance may have the lowest value to the fourth reference spectrum (best *Hit Quality*) based on the first spectral range. If you consider the second spectral range, the spectral distance may have the lowest value to the eleventh reference spectrum.

- 4b) When using the *Normalize to Reprolevel* method you have to define a reproduction level (see figure 71) for each spectral range. The spectral distances will be divided by this reproduction level. This is the reason why the spectral distances are indicated as reproduction level units, which means that you can set a threshold for the identity test. For example, if the *Hit Quality* is below 1 in case of a test spectrum, the sample is regarded as being *Identified*. If, however, the spectral distance is above 1, the spectrum cannot be assigned to any reference spectrum.
- 5) Irrespective of the method used, spectral distances can be weighted for each single spectral range (see chapter 7.3), according to the following equation:

$$D = \frac{\sum w_j \cdot D_j}{\sum w_j} \quad (6-6)$$

Spectral distances calculated by the *Normalize to Reprolevel* algorithm may be above 2000, if reproduction levels are other than 1. When using the *Scaling to 1st Range* method, the values have to be between 0 and 2000.

6.3 Data Preprocessing

OPUS provides several data preprocessing methods.

6.3.1 Vector Normalization

The maximum value of the *Hit Quality* has to be defined only if *Vector Normalization* was used to preprocess data. If you use a preprocessing method other than *Vector Normalization*, no upper limit for the *Hit Quality* has to be defined, i.e. you can use any numerical value. The maximum spectral distance is 2 (maximum difference of the spectra) in case of *Vector Normalization*, provided you have selected *Standard* method.

Vector Normalization first calculates the average y value of spectra and only uses data points within the selected spectral ranges. The average value calculated will then be subtracted from the spectrum, which causes the spectrum to be centered at around $y = 0$. This is followed by calculating the sum of squares of all y values, and the respective spectrum is divided by the square root of this sum. The vector norm of the result spectrum is 1:

$$a_m = \frac{\sum_k a(k)}{N} \quad (6-7)$$

$$a'(k) = a(k) - a_m \quad (6-8)$$

$$a''(k) = \frac{a'(k)}{\sqrt{\sum_k (a'(k))^2}} \quad (6-9)$$

$$\sum_k (a''(k))^2 = 1 \quad (6-10)$$

If vector normalized spectra are represented in n dimensional space and n being the number of selected data points, all spectra are on the unit sphere (n dimensional sphere around the coordinate origin with radius 1, see figure 46). The maximum distance between two spectra is the diameter of the unit sphere, i.e. 2.

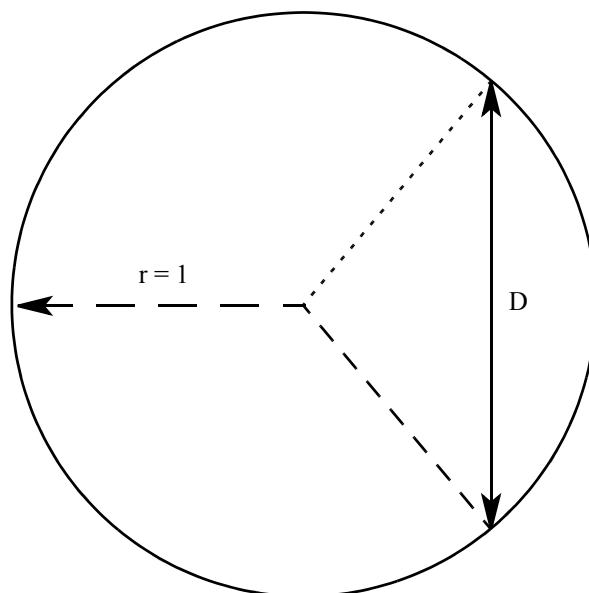


Figure 46: Two vector-normalized spectra on the unit sphere

To explain this in more detail, create a new spectrum. Invert one reference spectrum from the example library, i.e. multiply the spectrum by -1 using the *OPUS Spectrum Calculator*.

Compare the inverted spectrum with the reference spectra. Select an IDENT method that preprocesses data by *Vector Normalization*. Figure 47 shows the original spectrum (top) and the inverted spectrum (down). Figure 48 shows the identity test result. The last hit in the result list is the original reference spectrum with a spectral distance of 2 compared to the test spectrum.

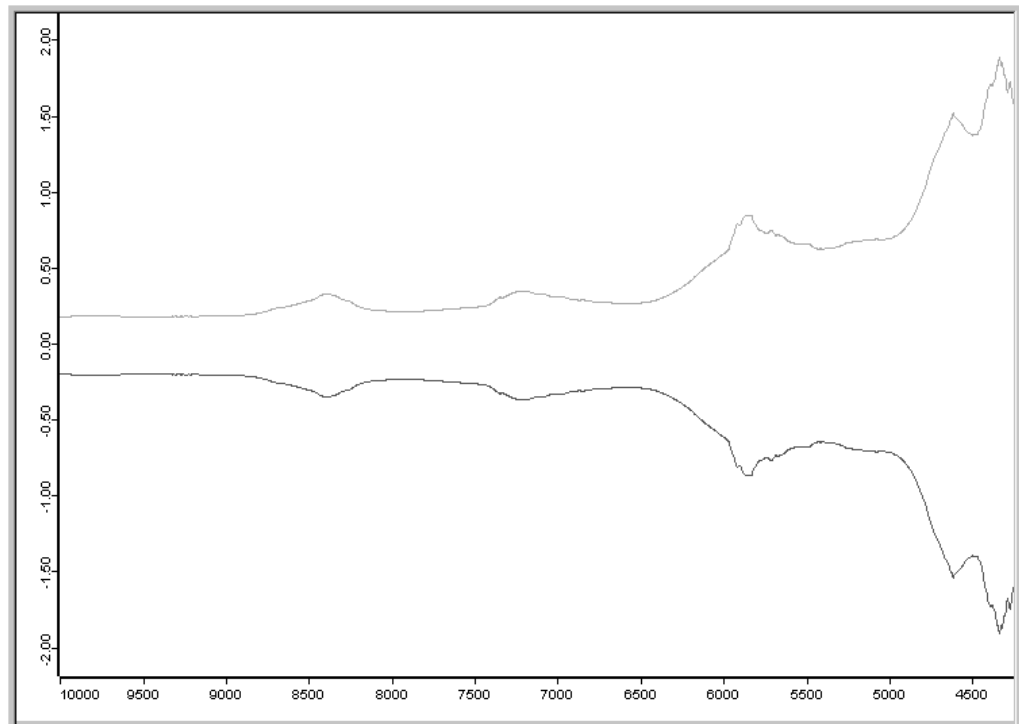


Figure 47: Original and inverted spectrum

One advantage of *Vector Normalization* is that the range of values for *Hit Quality* is known (from 0 to 2). This simplifies the interpretation of the identity test result. Additionally, using *Vector Normalization* as data preprocessing provides an even more important aspect.

Report of Correlation Search		Values
Method file:	\\w5\ident\example2\Standard.FAA	
from (date):	24/05/00	
(time):	15:10:08	
Description:		
IDENTITY NOT CHECKED:	0	
Hit quality with expected reference:	0.000000	
No Threshold avail:	0.000000	
Threshold calculation:	- depends on each ref.-spectrum -	
Algorithm:	Standard	
Vector normalized spectra:	Yes	
Order of Derivative:	0	
Smoothing points:	1	
No. of used factor sp.:	0	
15 hits of 15		
X-Ranges:	1	
Class Name:		
Class Test NOT PERFORMED:	0	
Using residuals:	No	
Order of Internal Derivative:	0	
Smoothing Points for Internal Derivative:	1	
Reduction Factor:	1	

Hit No.	Hit Quality	Sample Name	File Name	Threshold
1	1.812099	000009 Sample Glucose H2O Av. of 11	AN000009.100	0.026233
2	1.849116	000008 Sample Glucose H2Ofrei Av. of 11	AN000008.100	0.090445
3	1.854669	000013 Sample Mannit Av. of 11	AN000013.100	0.017529
4	1.877448	000012 Sample Sorbit Av. of 11	AN000012.100	0.029502
5	1.883937	000010 Sample Fructose Av. of 11	AN000010.100	0.030262
6	1.885355	000011 Sample Xylit Av. of 11	AN000011.100	0.032523
7	1.978973	000007 Sample L-Tryptophan Av. of 11	AN000007.100	0.051395
8	1.982136	000006 Sample DL-Tryptophan Av. of 11	AN000006.100	0.022966
9	1.992922	000014 Sample DL-Methionin Av. of 11	AN000014.100	0.072976
10	1.993975	000005 Sample L-Alanin Av. of 11	AN000005.100	0.073910
11	1.994812	000004 Sample DL-Alanin Av. of 11	AN000004.100	0.011690
12	1.997453	000015 Sample L-Methionin Av. of 11	AN000015.100	0.061757
13	1.997778	000002 Sample DL-Isoleucin Av. of 11	AN000002.100	0.033536
14	1.999440	000003 Sample L-Isoleucin Av. of 11	AN000003.100	0.023866
15	2.000000	000001 Sample L-Leucin Av. of 11	AN000001.100	0.021304

Figure 48: Ident Result of inverted spectrum searching

Vector Normalization also reduces the differences between each single measurement of the same sample. Figure 49 shows 11 spectra acquired from one single sample. As the substance has been powder the single spectra differ substantially from each other. These differences can be considerably reduced by using *Vector Normalization*.

Note the different scaling of the ordinate (y axis) in figure 49. Spectra derived from the same sample have to show only very small differences. Therefore, *Vector Normalization* is highly recommended in these cases. Further data preprocessing methods are *First and 2nd Derivative*.

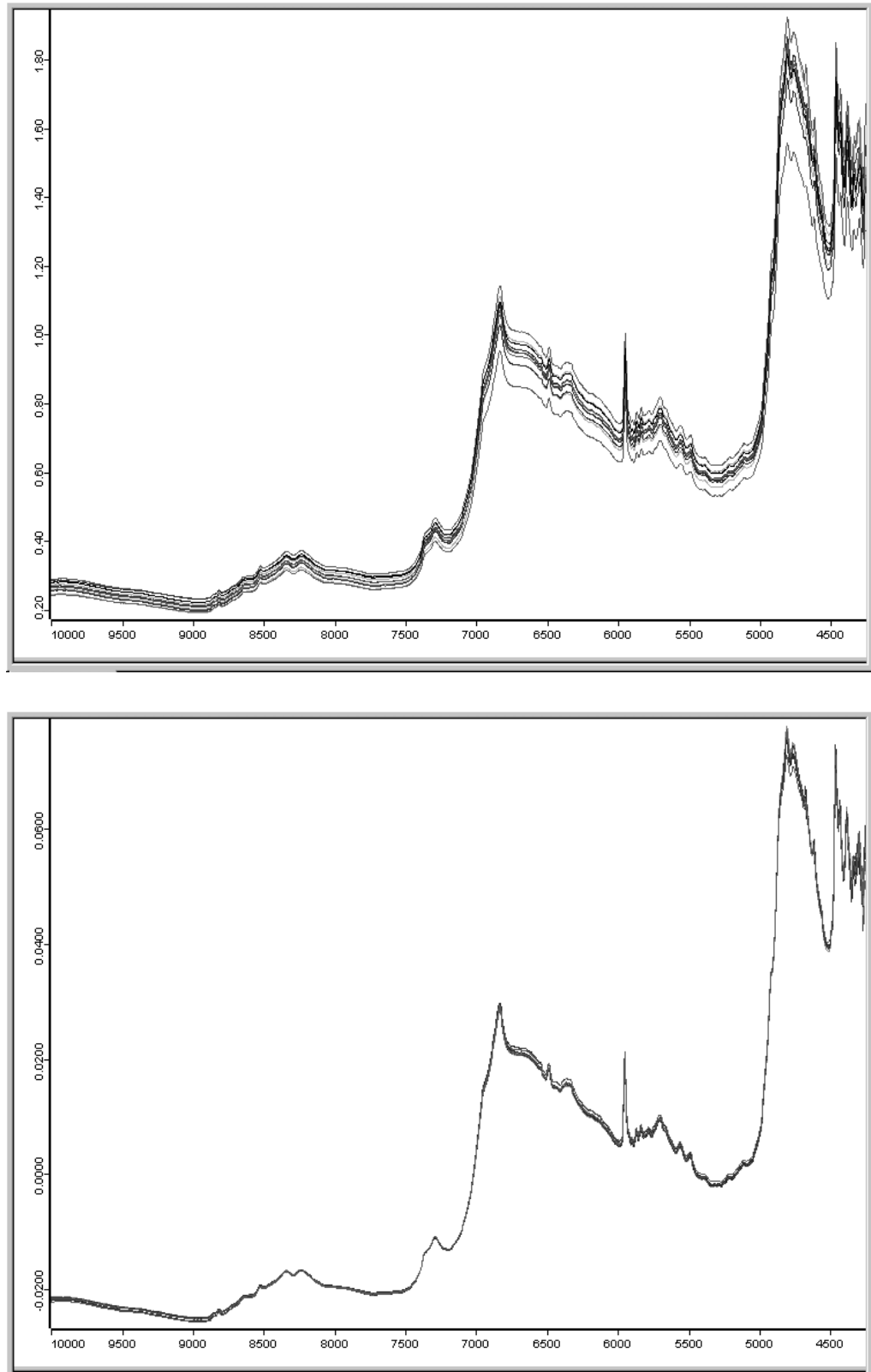


Figure 49: Original and vector-normalized spectra

Repeat the identity test for the first analytical example and select *Vector Normalization* to preprocess data. Figure 50 shows the result of the new analysis. Compare these results with the ones shown in figure 40.

Report of Correlation Search		Values
Method file:	\\WS\Ident\example2\New.FAA	
from (date):	29/05/00	
(time):	12:15:57	
Description:		
IDENTITY NOT CHECKED:	0	
Hit quality with expected reference:	0.000000	
No Threshold avail:	0.000000	
Threshold calculation:	- depends on each ref.-spectrum -	
Algorithm:	Standard	
Vector normalized spectra:	Yes	
Order of Derivative:	0	
Smoothing points:	1	
No. of used factor sp.:	0	
10 hits of 15		
X-Ranges:	1	
Class Name:		
Class Test NOT PERFORMED:	0	
Using residuals:	No	
Order of Internal Derivative:	0	
Smoothing Points for Internal Derivative:	1	
Reduction Factor:	1	

Hit No.	Hit Quality	Sample Name	File Name	Threshold
1	0.009404	000001 Sample L-Leucin Av. of 11	An000001.100	0.694965
2	0.324822	000004 Sample DL-Alanin Av. of 11	An000004.100	0.952755
3	0.338249	000015 Sample L-Methionin Av. of 11	An000015.100	0.883473
4	0.350834	000002 Sample DL-Isoleucin Av. of 11	An000002.100	0.751950
5	0.355342	000007 Sample L-Tryptophan Av. of 11	An000007.100	0.995592
6	0.414823	000014 Sample DL-Methionin Av. of 11	An000014.100	0.834952
7	0.477705	000005 Sample L-Alanin Av. of 11	An000005.100	0.819521
8	0.493052	000003 Sample L-Isoleucin Av. of 11	An000003.100	0.949495
9	0.527658	000006 Sample DL-Tryptophan Av. of 11	An000006.100	0.904797
10	0.608630	000010 Sample Fructose Av. of 11	An000010.100	0.775521

Figure 50: Ident Report – example of vector normalization

The lowest spectral distance is 0.009, the next (0.33) is higher, i.e. by about a factor of 30. This factor is higher than the one obtained without using *Vector Normalization* (factor 6).

The reference spectrum of Hit No. 1 is L-Leucin. This identification is correct as the test spectrum has been measured from the same sample. In general, however, the question is how far the spectral distance (*Hit Quality*) may increase to be still within an acceptable threshold to correctly identify the test spectrum?

To define such a threshold, it is not sufficient to measure only one single spectrum per reference substance. You must measure several spectra and determine this threshold from spectral fluctuations (spectral differences). The

average spectrum calculated from each measurement is then added to the IDENT library as reference spectrum.

6.4 Determining Threshold Value for Identity Test

There are two possibilities to define the limit value for an IDENT group:

1) Fixed Algorithm (*Maximum Hit + x SDev.*)

The threshold is calculated from the worst hit (i.e. the largest *Hit Quality* value in the average report) and the standard deviation S_0 :

$$\text{Threshold } D_T = D_{Max} + S_0 \cdot x \quad (6-11)$$

whereas the default x value is 0.25.

The threshold is selected so that all original spectra used to create the reference spectrum (average spectrum) have a lower distance than this threshold to the reference spectrum.

If the analysis of a sample spectrum produces a spectral distance which is larger than this threshold, the sample spectrum will be defined as not being identical.

2) Confidence Level

Two parameters are derived from the spectral distances (see above) to define the confidence region for the average spectrum:

The mean distance D_M

$$D_M = \sum_i \frac{D(i)}{n} \quad (6-12)$$

The standard deviation S_0

$$S_0 = \sqrt{\frac{\sum_i D(i)^2}{n-1}} \quad (6-13)$$

with n being the number of original spectra. Note that S_0 is the standard deviation from zero and not the standard deviation S_M from the mean value.

The standard deviation from the mean distance S_M can be calculated from D_M and S_0 :

$$S_M = \sqrt{\frac{\sum_i (D(i) - D_M)^2}{n-1}} = \sqrt{S_0^2 - \frac{D_M^2 \cdot n}{n-1}} \quad (6-14)$$

The threshold is calculated by multiplying the standard deviation S_M with a factor f and adding the mean distance D_M :

$$D_T = D_M + f \cdot S_M \quad (6-15)$$

The factor f is calculated from the probability Φ which can be chosen between 95% and 99.9999%. Example: if you choose $\Phi = 97.7\%$, factor f will be 2. The spectral distances are assumed to be distributed according to a normal distribution. Note that the probability value is calculated for a single-sided limit.

If you select $\Phi = 95\%$, then 5% of the original spectra are outside the confidence region (the spectral distance to the average spectrum is larger than the threshold in question). Whether this is actually correct or not can be tested by a validation. For details see chapter 6.7.

3) Abs. Threshold

This option allows you to define the threshold for each reference spectrum. This threshold, however, will only take effect if the group consists of several spectra. If you define the threshold, make sure to consider the spectral differences of the original spectra.

6.5 Identity Test

The identity test routine generates a hit list which is stored in the IDENT report, sorted by ascending spectral distances. Provided the expected reference has been selected automatically from sample name, or has been defined by the user the following three categories of identity test results are possible: *Identical*, *Not Identical* or *Can Be Confused With* (see below).

If a group consists of only one spectrum, the identity test result will be: *Identity Not Checked*. If no expected reference has been defined, the identity test results will be: *Identified As*, *Not Identified*, *No Unique Identification Possible*.

Identical to (in case of expected reference):

The first hit must be the expected reference. The *Hit Quality* of the first substance (i.e. the spectral distance between a test spectrum and the first reference spectrum in the report) has to be smaller than the threshold of the expected reference, and the spectral distances between the query spectrum and all other average spectra are always larger than the corresponding thresholds.

Identified As (in case of no reference defined):

All threshold values are taken into account. In case of *Identified As*, the *Hit Quality* of the first substance (i.e. the spectral distance between a test spectrum and the first reference spectrum in the report) has to be smaller than the corresponding threshold. But the *Hit Quality* of all other substances has to be larger than the corresponding threshold.

Not Identical to (in case of expected reference):

The spectral distance between the test spectrum and expected reference spectrum is larger than the threshold value.

Not Identified (in case of no reference defined):

In this case the *Hit Quality* of *all* reference substances has to be larger than the threshold.

Can be Confused with (in case of expected reference):

This result indicates that the spectral distance of the test spectrum to *at least* one other average spectrum is smaller than the confidence region. If, e.g., the *Hit Quality* of 4 substances is smaller than the corresponding threshold and one of these substances is identified to be the expected reference.

No Unique Identification Possible (in case of no reference defined):

The *Hit Quality* of *more* than one substance is smaller than the corresponding threshold. Code numbers are assigned to the individual test results. Therefore, the results can be easily evaluated in an OPUS macro.

Can be confused with <N> Other Hits (in case of expected reference)

The code number 2 is used if all reference spectra with a *Hit Quality* less than the threshold have the same *Sample Name* (or *Sub Sample Name*) as the query spectrum. The value of 2 may only occur if the *Selected automatically from Sample Name* option button on the *Expected Reference* dialog (figure 87, page 104) has been activated.

Table 2: Code Numbers of Identity Test

Identity Test Result	Code
Identity not Checked (no threshold available)	0
Identical To/Identified As	1
Can be Confused with/No Unique Identification Possible	-1
Not Identical/Not Identified	-2
Can be confused with <N> Other Hits	2

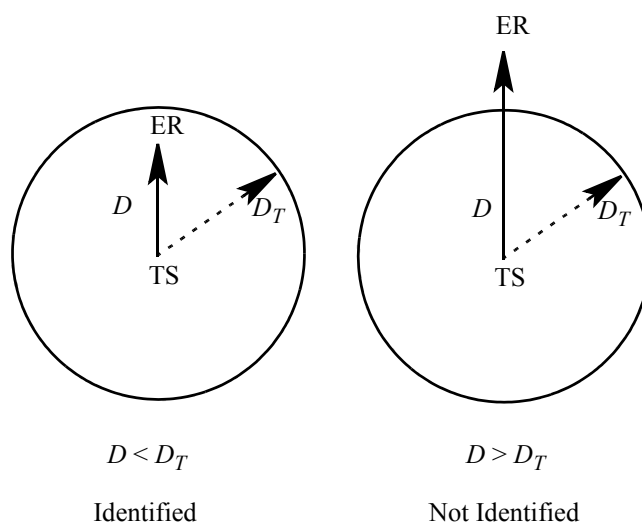


Figure 51: Schematic representation of Identity Test Results - Test Spectrum (TS), Expected Reference (ER)

6.6 Class Test

Ideally, all reference spectra are well-distinguishable from each other and the thresholds are so small that their confidence regions do not overlap. This situation is shown in figure 52. The reference spectra are dots (in a mathematical sense) in the n dimensional space, n being the number of data points selected. The confidence regions can be depicted by *spheres* which centers represent the reference spectra.

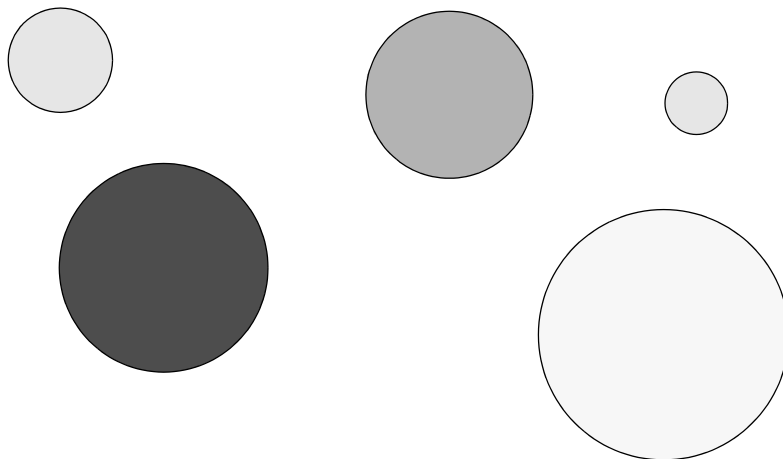


Figure 52: Library with well-distinguishable reference spectra

However, it may occur that the confidence regions of some reference spectra do overlap, see figure 53. As you can see, the confidence regions of the *A*, *B*, and *C* reference spectra clearly overlap.

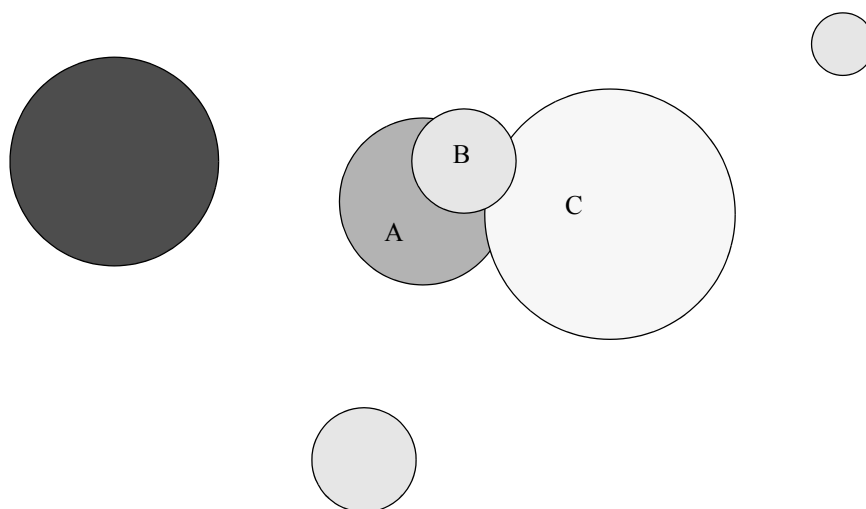


Figure 53: Three reference spectra with overlapping confidence regions

It is possible to define several reference spectra which are members of one class, and to perform a class test during the IDENT analysis. Both in case of an expected reference as well as in case of an analysis with no reference defined, the class test determines whether all reference spectra with their *Hit Quality* below the corresponding threshold are members of the same class. If so, the result will be *Class Test OK*. Otherwise, the result will be *Class Test NOT OK*. If the expected reference spectrum is not part of any class, the IDENT report says *Class Test NOT PERFORMED*.

Make sure to define only one class name for one group in the IDENT method. If you load a previously used method including extraordinary members, only the first class name will be considered.

All class members use their individual thresholds. If you load a previously used method when using the *Setup Identity Test Method* command from the *Evaluate* menu, the exclamation mark (!) next to the old class name is automatically deleted.

Compare Spectra		class.FAA , Signed by: Bruker Optik, 2002/02/28,		
Method file:	D:\test42\validation\Ident\methods\class.FAA			
from (date):	3/03/03			
(time):	16:54:37			
Description:	CAN BE CONFUSED WITH 1 OTHER HITS : -1			
Hit quality with expected reference:	0.002891			
Threshold for expected reference:	0.011451			
Threshold calculation:	- depends on each ref.-spectrum -			
Algorithm:	Standard			
Vector normalized spectra:	Yes			
Order of Derivative:	0			
Smoothing points:	1			
No. of used factor sp.:	0			
15 hits of 15				
X-Ranges:	1			
Class Name:	KLARA			
Class Test OK:	1			
Using residuals:	No			
Order of Internal Derivative:	0			
Smoothing Points for Internal Derivative:	1			
Reduction Factor:	1			

Hit No.	Hit Quality	Sample Name	Group	Threshold
1	0.002891	000004 Sample DL-Alanin Av. of 11	000004	0.011451
2	0.063071	000005 Sample L-Alanin Av. of 11	000005	0.079217
3	0.110633	000002 Sample DL-Isoleucin Av. of 11	000002	0.036291
4	0.132710	000003 Sample L-Isoleucin Av. of 11	000003	0.028581
5	0.135557	000001 Sample L-Leucin Av. of 11	000001	0.025936
6	0.154416	000015 Sample L-Methionin Av. of 11	000015	0.067953
7	0.219237	000014 Sample DL-Methionin Av. of 11	000014	0.081288
8	0.227688	000006 Sample DL-Tryptophan Av. of 11	000006	0.023606
9	0.248521	000007 Sample L-Tryptophan Av. of 11	000007	0.053122
10	0.593082	000010 Sample Fructose Av. of 11	000010	0.029839
11	0.611039	000011 Sample Xylit Av. of 11	000011	0.031277
12	0.613600	000012 Sample Sorbit Av. of 11	000012	0.029364
13	0.679362	000013 Sample Mannit Av. of 11	000013	0.016202
14	0.690536	000008 Sample Glucose H2Ofrei Av. of 11	000008	0.092360
15	0.733616	000009 Sample Glucose H2O Av. of 11	000009	0.026632

Figure 54: IDENT Report with class test performed - Test OK

Compare Spectra		class1.FAA , Signed by: Bruker Optik, 2002/02/28, 1			
Method file:	D:\test42\Validation\Ident\methods\class1.FAA				
from (date):	4/03/03				
(time):	9:29:25				
Description:	CAN BE CONFUSED WITH 11 OTHER HITS : -1				
Hit quality with expected reference:	0.339112				
Threshold for expected reference:	0.694965				
Threshold calculation:	- depends on each ref.-spectrum -				
Algorithm:	Standard				
Vector normalized spectra:	Yes				
Order of Derivative:	0				
Smoothing points:	1				
No. of used factor sp.:	0				
15 hits of 15					
X-Ranges:	1				
Class Name:	HANS				
Class Test NOT OK:	-2				
Using residuals:	No				
Order of Internal Derivative:	0				
Smoothing Points for Internal Derivative:	1				
Reduction Factor:	1				
Hit No.	Hit Quality	Sample Name	Group	Threshold	
1	0.339112	000001 Sample L-Leucin Av. of 11	000001	0.694965	
2	0.339817	000015 Sample L-Methionin Av. of 11	000015	0.883473	
3	0.390038	000004 Sample DL-Alanin Av. of 11	000004	0.793804	
4	0.453472	000007 Sample L-Tryptophan Av. of 11	000007	0.995592	
5	0.456483	000002 Sample DL-Isoleucin Av. of 11	000002	0.991531	
6	0.514776	000014 Sample DL-Methionin Av. of 11	000014	0.885595	
7	0.599937	000005 Sample L-Alanin Av. of 11	000005	0.819521	
8	0.624271	000006 Sample DL-Tryptophan Av. of 11	000006	0.904797	
9	0.679802	000003 Sample L-Isoleucin Av. of 11	000003	0.920193	
10	0.757945	000011 Sample Xylit Av. of 11	000011	0.725310	
11	0.760448	000010 Sample Fructose Av. of 11	000010	0.775521	
12	0.922144	000008 Sample Glucose H2Ofrei Av. of 11	000008	1.091812	
13	0.934589	000012 Sample Sorbit Av. of 11	000012	0.791596	
14	1.023140	000013 Sample Mannit Av. of 11	000013	1.032565	
15	1.061457	000009 Sample Glucose H2O Av. of 11	000009	0.976118	

Figure 55: IDENT Report with class test performed - Test Not OK

Code numbers are assigned to the individual class test results. This causes the results to be easily evaluated in an OPUS macro.

Table 3: Code Numbers of Class Test

Class Test Result	Code
Class Test OK	1
Class Test not Performed	0
Class Test not OK	-2

6.7 Validation

When setting up an IDENT library, you have to check whether the IDENT method parameters are optimized for all reference spectra in one library. This can be done by *Validation*, which compares original spectra with average spectra.

Figure 56 shows a validation report, using the *Standard* method. All original spectra have been compared with the average spectra of the library. The *Abs. Threshold* values are used as confidence region in the validation process. The results are either *Uniquely Identified*, *Not Identified* and *Can Be Confused With*, similar to the *Identity Test* (see chapter 6.5). The number of spectra which belong to the respective class is indicated at the end of the validation report. The total of all spectra has to be equal to the total of the original spectra tested.

In case of overlappings the *Detailed* and *Result* reports provide additional information on which groups should be assigned to a new common sub-library.

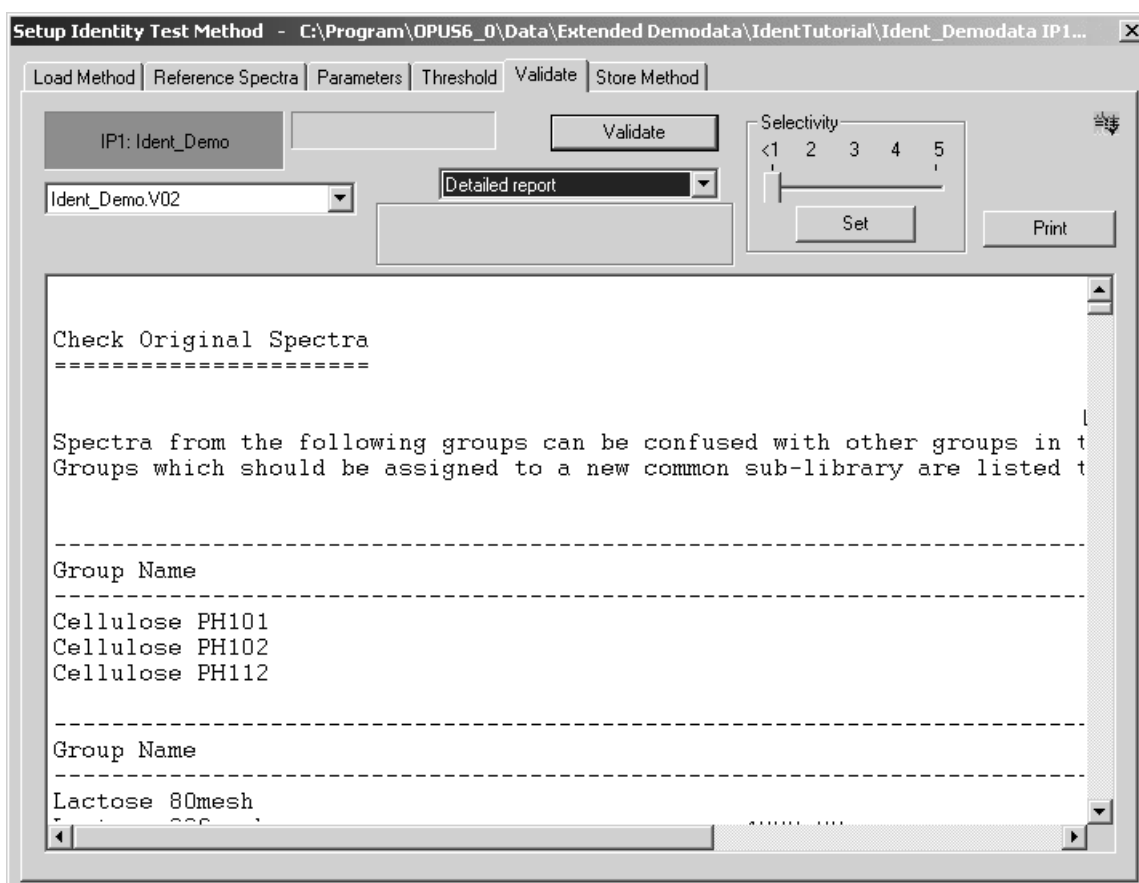


Figure 56: Validation Report showing assignment recommendation

An original spectrum is *Uniquely Identified* if the spectral distances between this spectrum and the corresponding average spectrum is smaller than the threshold, while the spectral distances between the original spectrum and all other average spectra are larger than the corresponding confidence region. The original spectra which are *Uniquely Identified* are not listed in the report.

If a spectrum is *Not Identified*, the spectral distance between the original spectrum and average spectrum is larger than the threshold. In this case the report indicates the corresponding average spectrum, sample name and threshold (confidence region) specified for this substance. The file names of the original spectra which have not been identified, and the spectral distances between these original spectra and the average spectrum (Hit) are listed under the *Original Spectra Outside Confidence Region* definition.

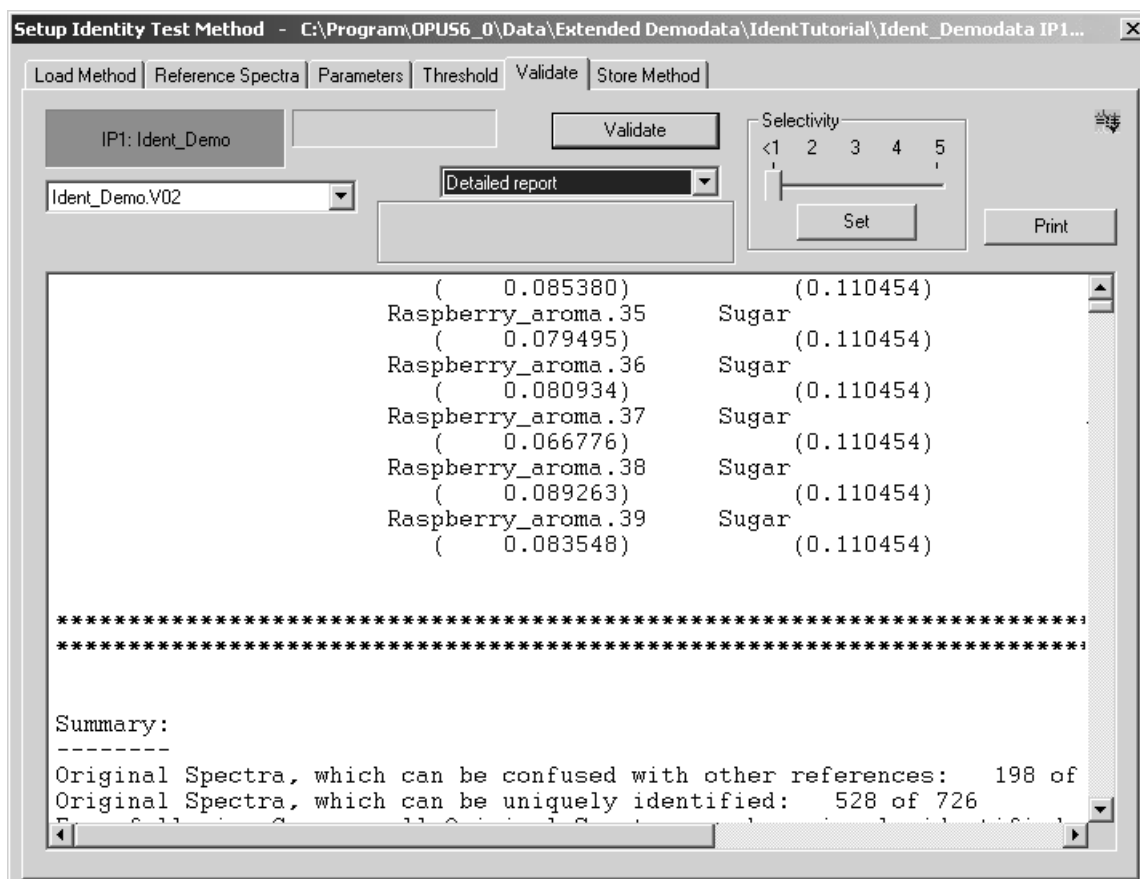


Figure 57: Validation Report – Spectra identified

The *Can Be Confused With* result indicates that the spectral distance of an original spectrum to the corresponding average spectrum is smaller than the confidence region, while one or more spectral distances between the original spectrum and other average spectra are even smaller than the corresponding confidence regions.

If an original spectrum is tested to be *Can Be Confused With* other references, first, its average spectrum, sample name and confidence region are listed in the report, followed by the name of the original spectrum and the threshold. In addition, the name of the average spectrum, the sample name and the spectral distance (Hit) between this average spectrum and the original spectrum are listed under the *Overlapping With* definition. The spectral distance is smaller than the threshold (confidence region). If this original spectrum overlaps several average spectra, they will all be listed in the report.

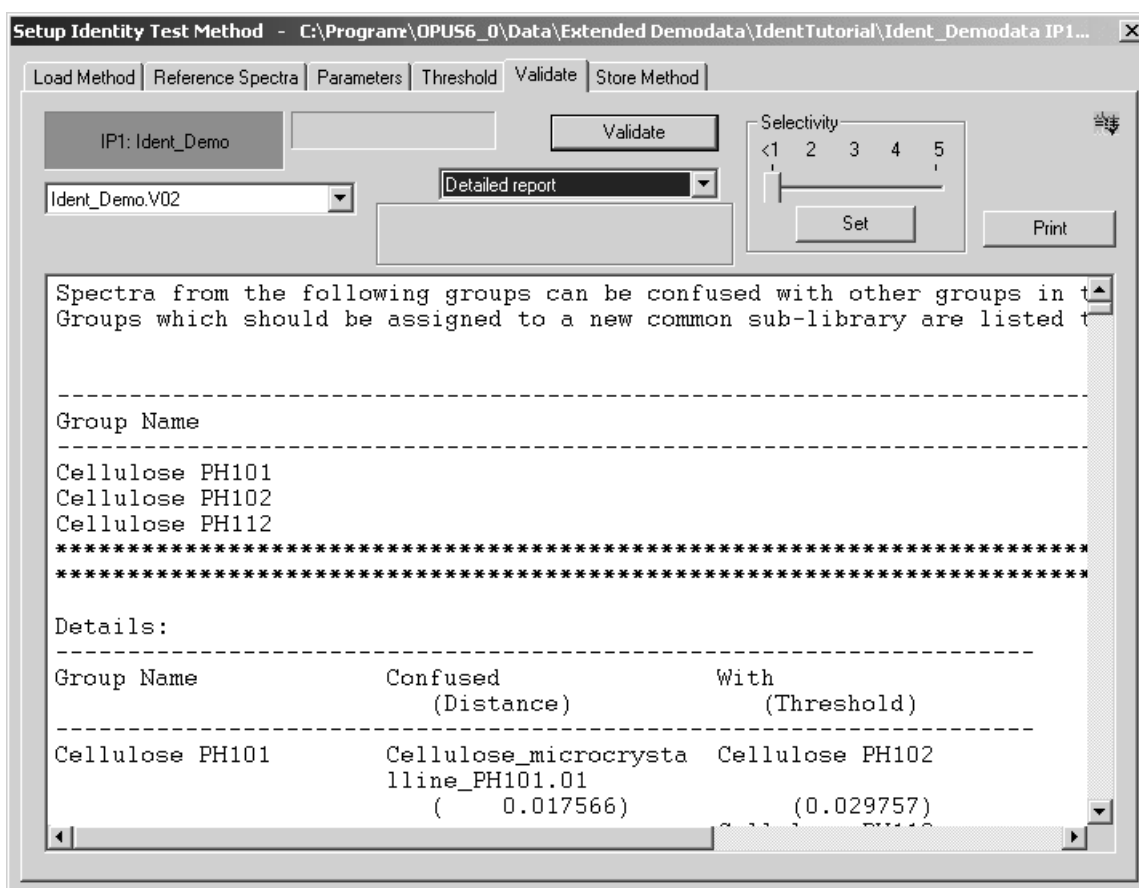


Figure 58: Validation Report – spectra can be confused with

If all reference spectra are selected, each original spectrum will be compared with all average spectra. Even if you only select some of the reference spectra, the original spectra belonging to these reference spectra are tested against **all** average spectra. This option is extremely useful if an existing library is to be extended by new reference spectra and only these new spectra have to be tested.

If you have activated the *Always use lowest IP level* check box on the *Parameters* tab, the detailed report will only list the results of the lowest IP level for all spectra of each group.

7 Reference Section

Before being able to identify spectra by means of OPUS IDENT you have to create an IDENT method first. Select the *Setup Identity Test Method* command from the OPUS *Evaluate* menu.

7.1 Setup Identity Test Method - Load Method

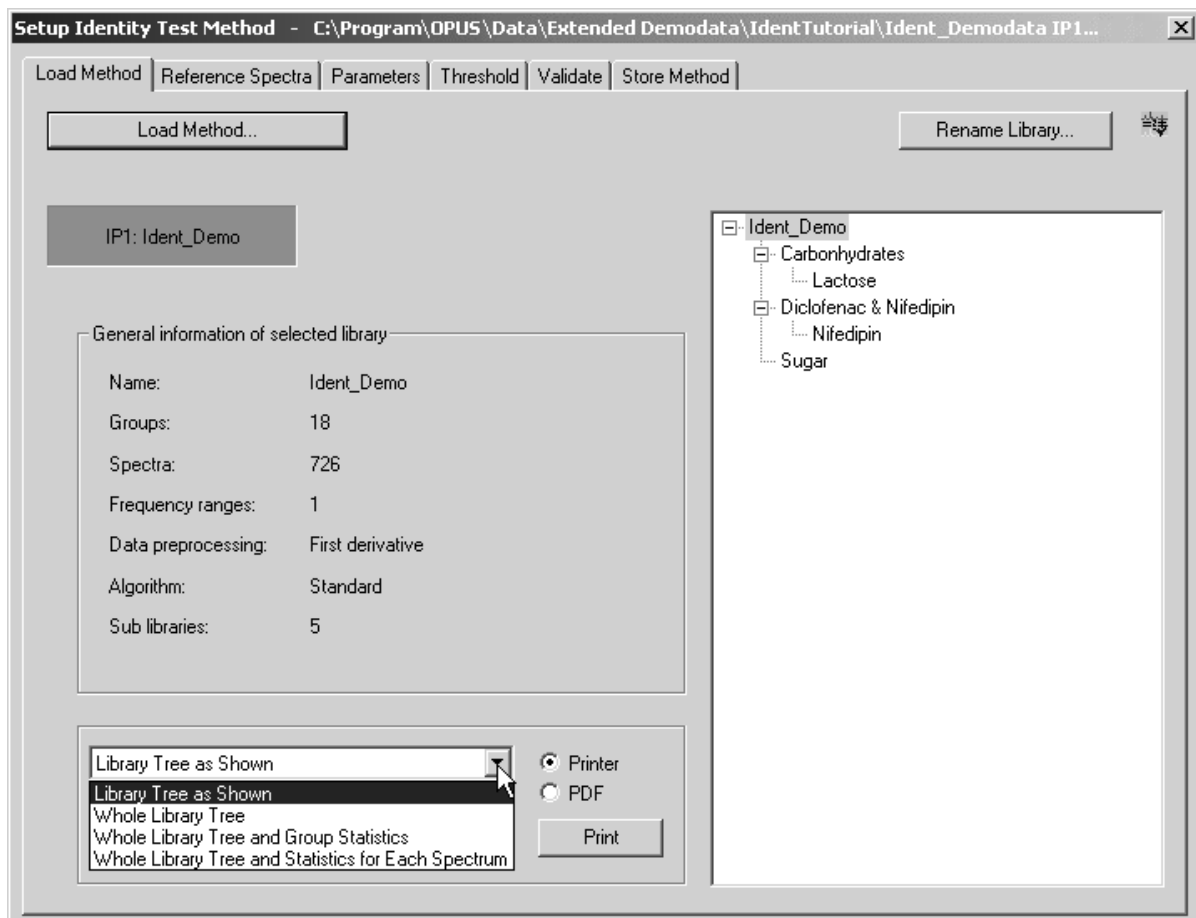


Figure 59: Setup Identity Test Method – Load Method tab

Use the *Load Method* button to load an existing IDENT method. IDENT method files have the extension **.FAA*. It is also possible to load IDENT method files created by OPUS-OS/2. However, if you store such an OS/2 method using OPUS IDENT, you will not be able to load the method by OPUS-OS/2 IDENT again. To solve this problem, store the modified OPUS-OS/2 IDENT file by using a different file name.

The *General information of selected library* group field provides statistical information on the existing method file. The number of spectra used for the method, and the number of frequency ranges included are displayed.

You will get additional information on the data preprocessing method, the algorithm used for the identity test, and how many sub-libraries are part of the reference library.

For further details on how to rename libraries, refer to chapter 1. To print the library structure use the drop-down list to select the respective printing option and click on the *Print* button. For further details see chapter 1.

7.2 Setup Identity Test Method - Reference Spectra

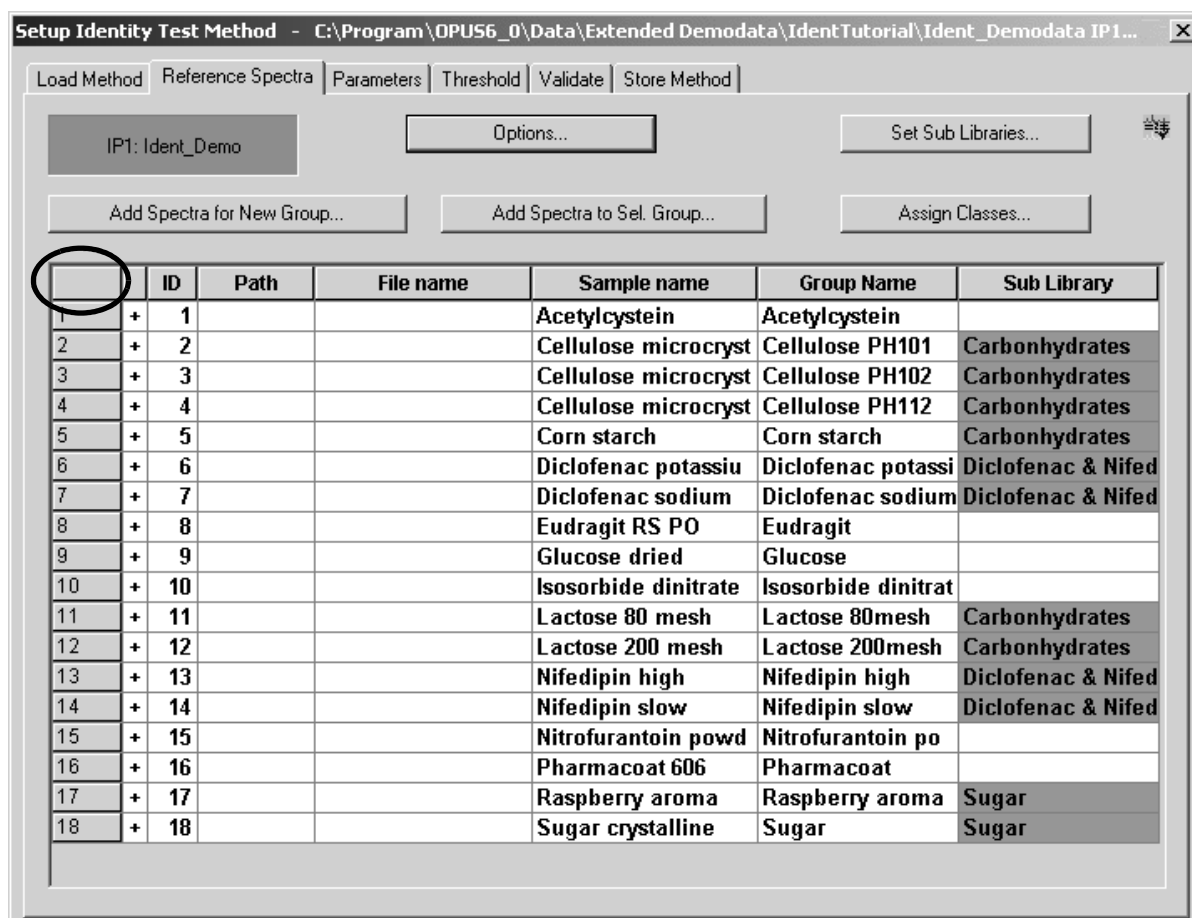


Figure 60: Setup Identity Test Method – Reference Spectra tab

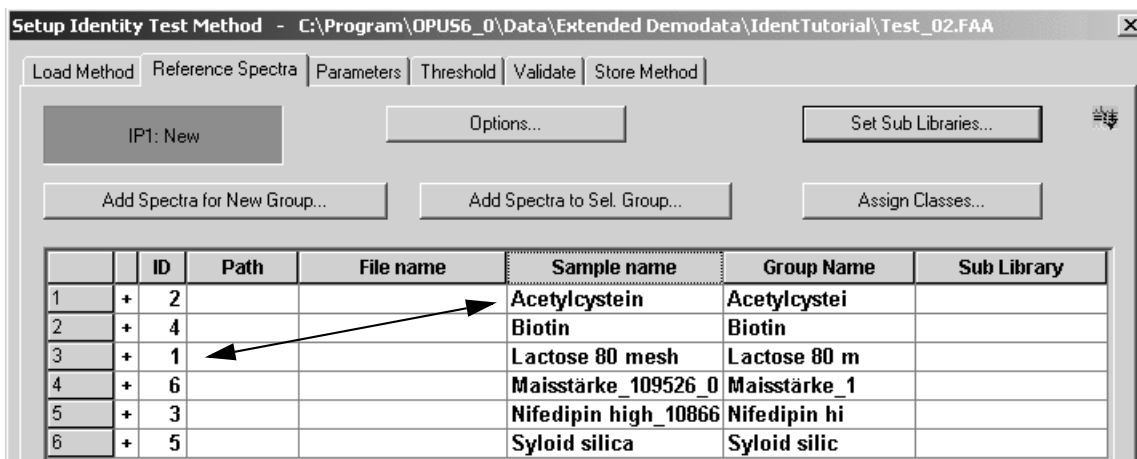
The spectra table lists the spectra groups and each single spectra, including the sample *ID*, *Path*, *File Name*, *Sample Name*, *Group Name* and *Sub Library*. You can have the groups displayed as well as each single spectra of one group. Click on the **+** sign in the first column. The respective line with the group selected opens, and shows the single spectra. To close the list again, click on the **-** sign.

Click on the numbered tiles on the left side of the table to select one spectrum or several spectra. Select the whole table by clicking on the tile on the left side of the table header (see mark in figure 60). Remove spectra from the table by selecting one spectrum or more spectra and pressing the *DEL* key on the PC keyboard.

If you click on the *Add Spec. for New Group* button, a dialog box opens to be used to load one or more files into the spectra list. Spectra loaded simultaneously will be merged into one group. Use the *Add Spec. to Sel. Group* button to add a spectrum to a group selected. This is useful if you want to add a spectrum to a group later.

7.2.1 Sorting Reference Spectra

When working with IDENT methods consisting of a large number of groups it takes quite some time to find a particular group. Therefore, you can sort the reference spectra by ID, sample or group name and sub-library in ascending or descending order. Double click on the respective column. Note that if you sort the spectra, e.g., by sample name, the ID number in the first column, however, will keep the original order (see figure 61).



	ID	Path	File name	Sample name	Group Name	Sub Library
1	+	2		Acetylcystein	Acetylcystei	
2	+	4		Biotin	Biotin	
3	+	1		Lactose 80 mesh	Lactose 80 m	
4	+	6		Maisstärke_109526_0	Maisstärke_1	
5	+	3		Nifedipin high_10866	Nifedipin hi	
6	+	5		Syloid silica	Syloid silic	

Figure 61: Setup Identity Test Method – Reference spectra sorted in descending order

7.2.2 Missing Reference Spectra

If you load an IDENT method, it may occur that certain spectrum files listed in the particular method are missing in the data path. To be able to localize the missing spectra the group name which the missing spectra belong to will be highlighted in red. If you click on the **+** sign of the particular group, the name of the missing spectra will also be highlighted in red.

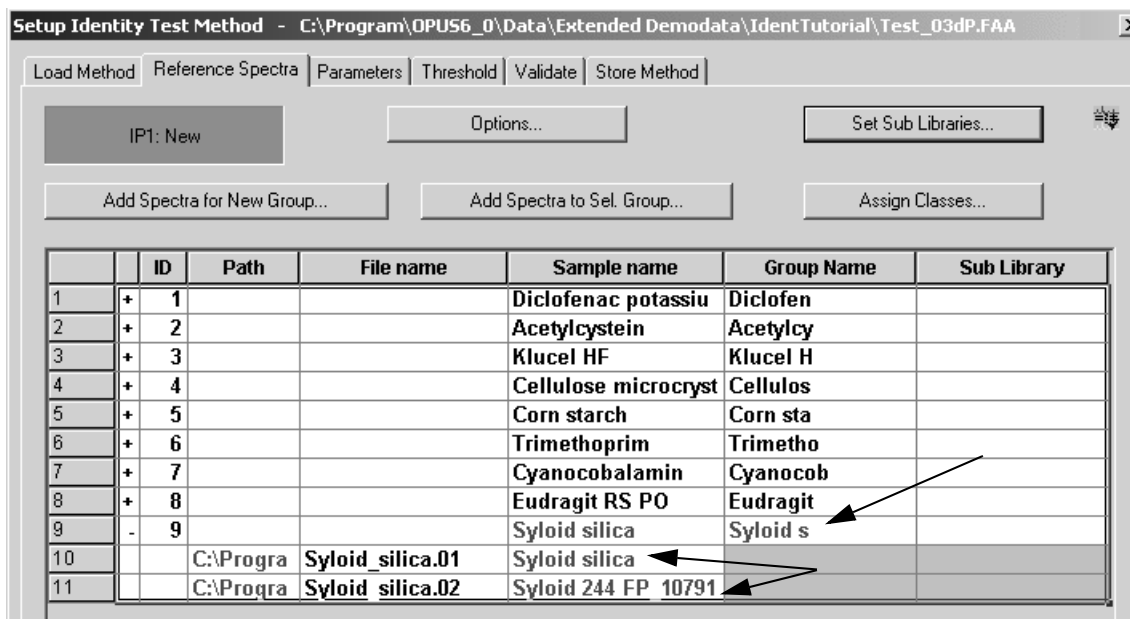


Figure 62: Setup Identity Test Method – Missing reference spectra

Missing reference spectra may be due to file renaming or, as exemplified in figure 62, to a different data path. In this case store the missing reference spectra into the right path and load the IDENT method again. This is important as you cannot perform any calculation using an IDENT method with missing spectra. If you want to know the total number of spectra included in a particular group, click on the *Threshold* tab. The *Spectra* column includes the total number of spectra per group.

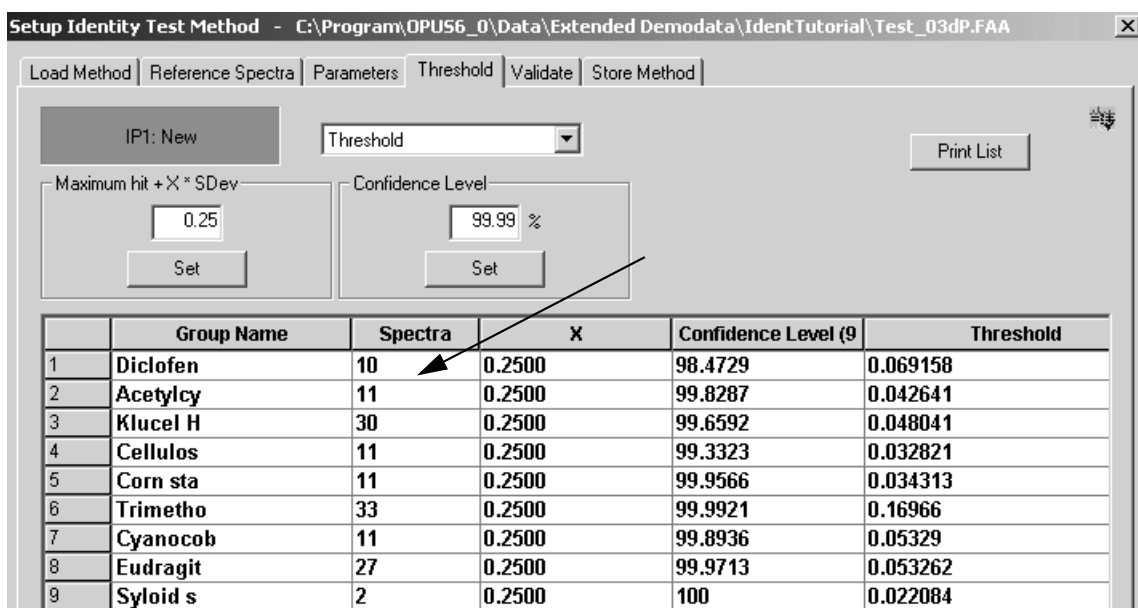


Figure 63: Total number of spectra within one group

7.2.3 Options

It is also possible to add average spectra generated by previous OPUS IDENT versions without requiring to reconstruct the original spectra which the average has been generated from. Each average spectrum loaded represents one group.

To switch between original and average spectra generated by previous OPUS versions, first click on the *Options* button. The following dialog opens:

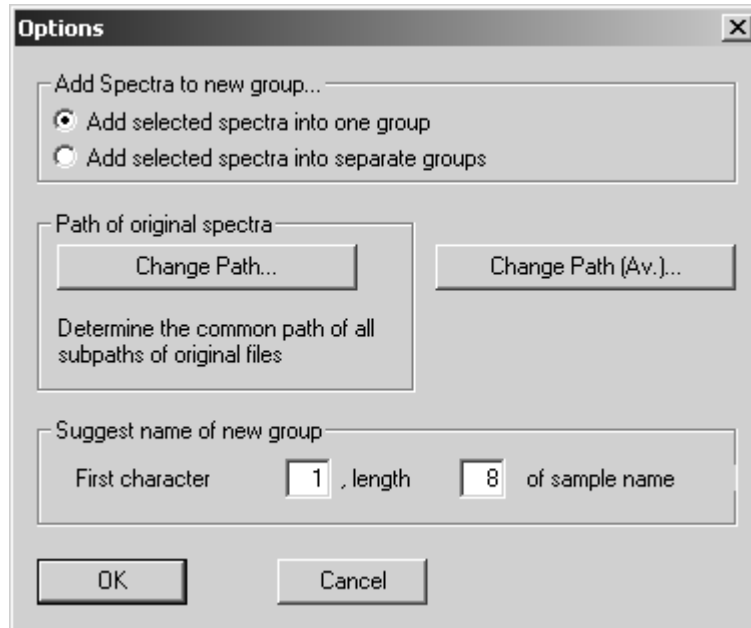


Figure 64: Options

The *Add selected spectra into one group* option button is activated by default. In this case all spectra commonly selected during loading are added to one group.

In the *Options* dialog box you define the paths for the original and average spectra. Additionally, you can define the group name which is derived from the respective sample name. Enter the position of the first character as well as the length of the sample name.

Note: Once defined, the group name should not be changed any more to avoid confusion.

7.2.4 Set Sub Library

The *Set Sub Library* button on the *Reference Spectra* tab enables you to add sub-libraries to the current main method. Select the *New* option from the drop-down list to enter a unique sub-library name for the spectra groups defined. Choose the group(s) which have to be assigned to this new sub-library from the *Select groups for...* selection field. There are only groups available which have not yet been assigned to another sub-library on the same library level. Click on the *Assign* button.

To delete sub-libraries select them first and click on the *Delete Sub Library* button.



Figure 65: Set Sub Library

7.2.5 Assign Classes

To assign classes to an existing method click on the *Assign Classes* button. The following dialog opens:



Figure 66: Assign Classes

Select the *New* option from the drop-down list to enter a unique class name for the spectra groups defined. Choose the group(s) which have to be assigned to this new class from the *Select groups for...* selection field. There are only groups available which have not yet been assigned to another class on the same library level. Click on the *Assign* button.

To delete classes select them first and click on the *Delete Class* button.

7.3 Setup Identity Test Method - Parameters

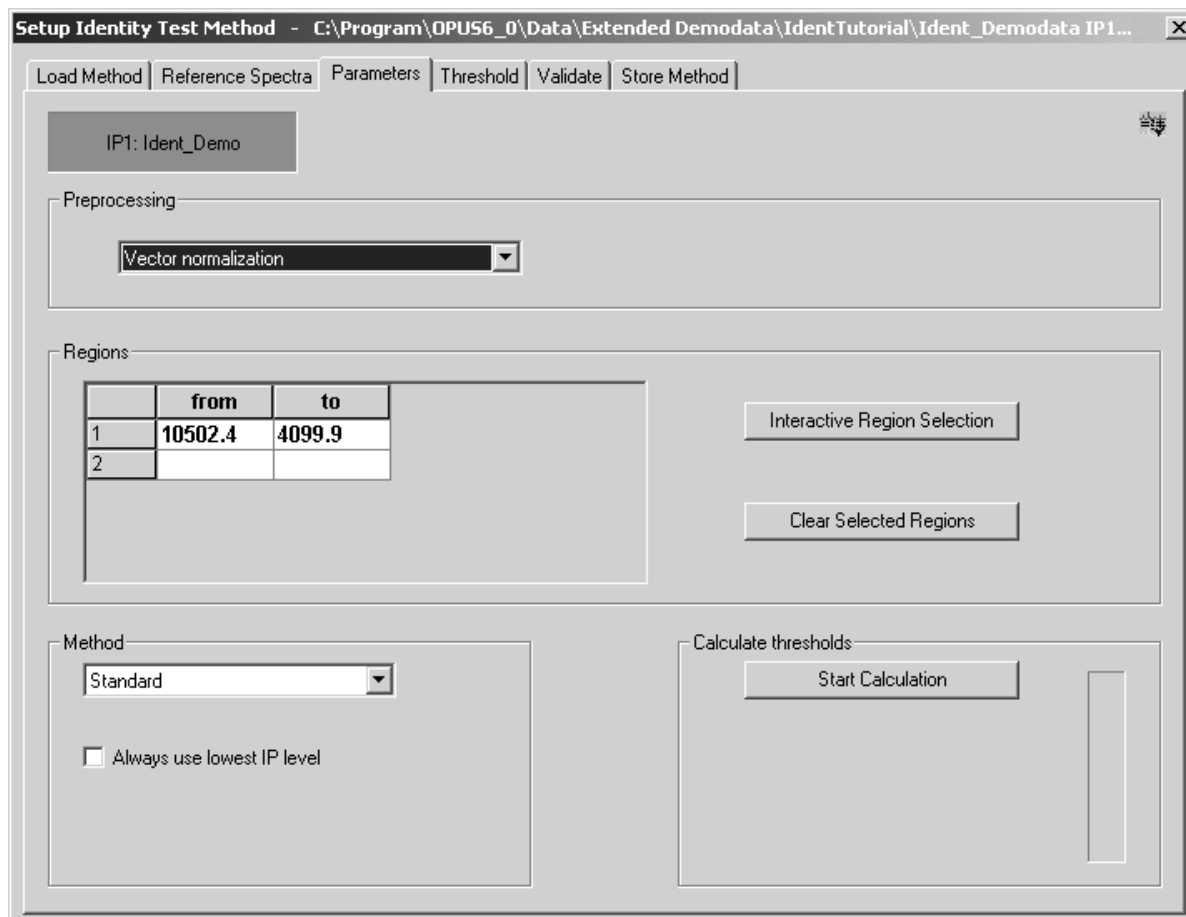


Figure 67: Setup Identity Test Method – Parameters tab

7.3.1 Preprocessing

You can select several data processing methods from the drop-down list: *Vector Normalization*, *First* and *2nd Derivative* as well as combinations of both methods.

- **Vector Normalization**

The *Vector Normalization* data preprocessing normalizes a spectrum, i.e. the average y value is calculated first and subsequently subtracted from the spectrum. Then, the sum of squares of all y values is calculated and the spectrum is divided by the square root of this sum.

This method is used in case of different optical thickness to compare the samples with each other. The form of the different spectra will be preserved, which facilitates the interpreting of spectra. However, the result extremely depends on the spectral region selected, i.e. specific differences of one region are distributed to all data points.

- **First Derivative**

Calculates the first derivative of the spectrum by interpolation. Steep edges of a peak become more important compared to flat structures. This method is mainly used to preprocess pronounced, but small features which are overlaid by a high and broad background.

In case of this method the window size selected is very important. The smaller the window size, the more spectral details are shown, with the spectral-to-noise ratio being apparently higher.

- **2nd Derivative**

This method is similar to *First Derivative*, but it allows to evaluate extremely flat structures.

If you use one of the derivative methods, an additional selection field will be displayed to define the amount of smoothing points. You can select between 5 and 25 points. The optimal number of smoothing points, however, has to be evaluated empirically.

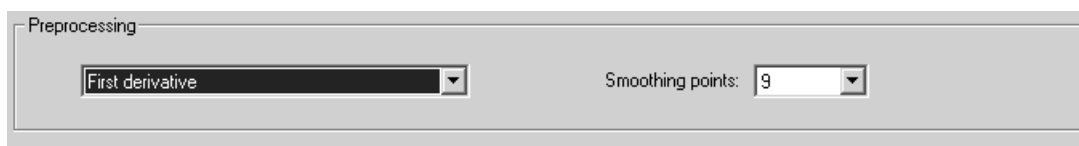




Figure 68: Defining smoothing points

7.3.2 Regions

The *Regions* table allows to limit data to one or several spectral regions to be considered for identification. The frequency limits for the spectral regions can either be entered manually or selected interactively.

7.3.3 Interactive Region Selection

The spectral region shown on the white background will be processed and evaluated. You can also modify the spectral regions displayed. Place the cursor on the boundary between the gray and white area. Press the left mouse button and move the regions. It is also possible to move the entire spectral region. If you position the cursor on the white area, the cursor changes from  into . Press the left mouse button and move the spectral region. To delete a region, right click on the white area and select *Remove* from the pop-up menu.

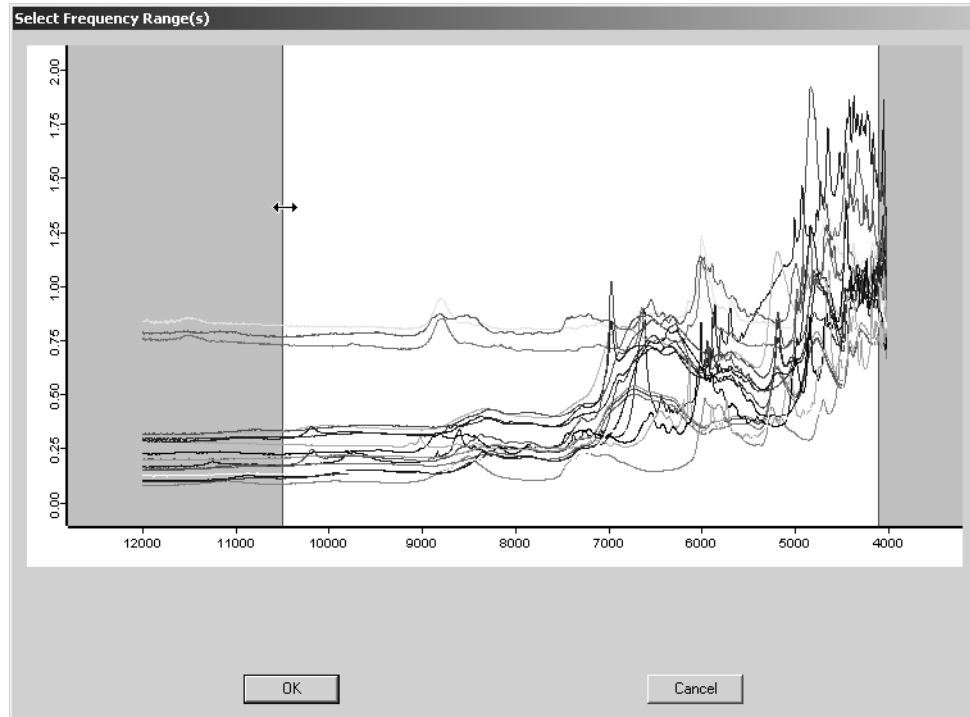


Figure 69: Select Frequency Range(s)

If you click on the *Interactive Region Selection* button, a separate window opens and displays the reference spectra. You can add a new spectral region by right-clicking on the window and selecting the *Add Region* option from the pop-up menu. The pop-up menu also includes the *Zoom* and *Crosshair* option. These options allow to easily define the value of a specific data point.

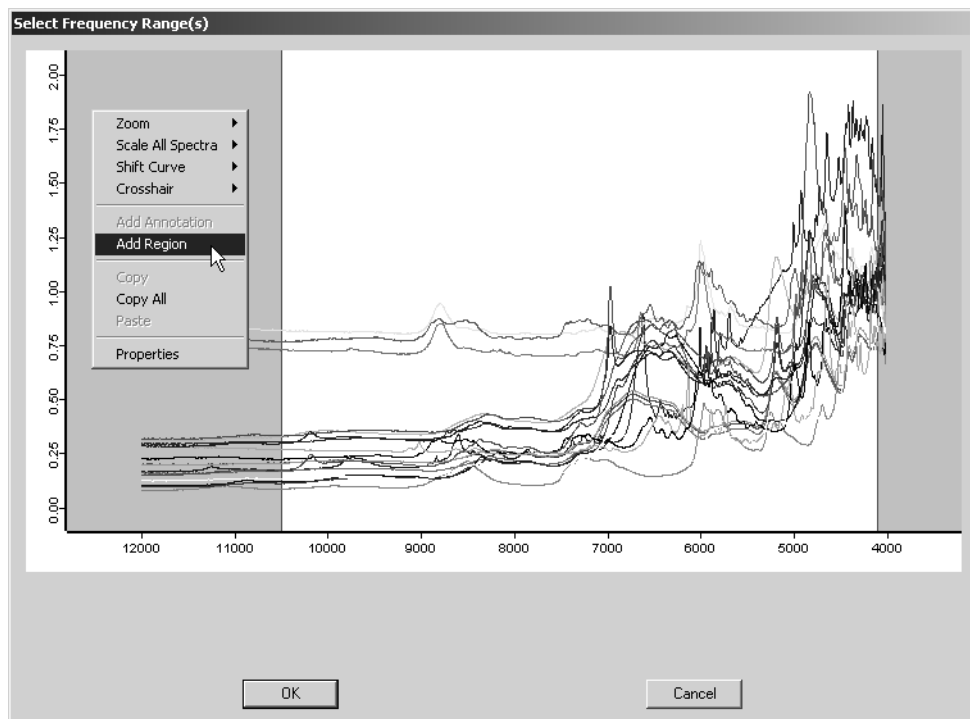


Figure 70: Select Frequency Range(s) with pop-up menu

7.3.4 Clear Selected Regions

You can delete an entry from the *Regions* table by selecting the specific entry and clicking on the *Clear Selected Regions* button or pressing the *DEL* button on your keyboard.

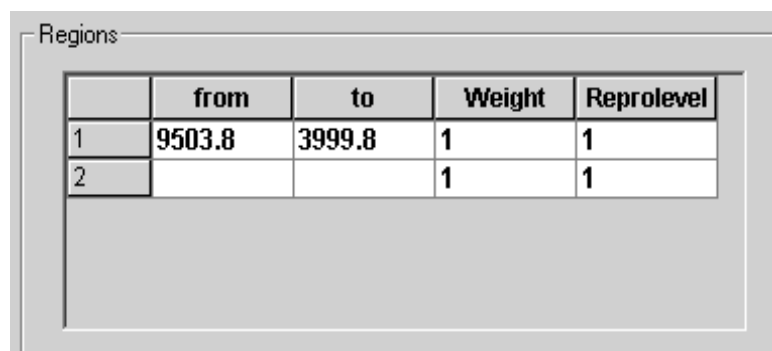
7.3.5 Method

Use this drop-down list to select an identification method. During the identity test the test spectrum is compared with all reference spectra. The result of this comparison is the spectral distance which is also called *Hit Quality*. The more similar two spectra are, the smaller the spectral distance. Four methods are available to calculate the spectral distance. Use the *Method* drop-down list to select one. The basic theory of each method has been described in chapter 6.

To enforce an IDENT analysis on the lowest IP level during the identity test the *Always use lowest IP level* check box has to be activated. This check box is only enabled on the very first library level as this a global setting for the entire library structure, and is deactivated by default. See also chapter 1.

IDENT methods which have been stored created and stored by previous OPUS without this algorithm can be loaded as well to perform this kind of analysis.

Depending on the method selected, additional columns are added to the *Regions* table, e.g. *Weight* and *Reprolevel*.



	from	to	Weight	Reprolevel
1	9503.8	3999.8	1	1
2			1	1

Figure 71: Regions list with weight and reprolevel columns

You can select between the following identification methods:

- **Standard**
The *Standard* method calculates the Euclidean distance between the test and reference spectra.
- **Factorization**
The factorization is performed on average spectra of the respective groups. The spectra are first represented as linear combination of the factor spectra and the resulting coefficients are used to calculate the spectral distance.
- **Factorization (orig. specs)**
The factorization is performed on all original spectra of the respective groups.
- **Scaling to 1st range**
Performs the *Scaling to 1st range* algorithm. For details, see section 6.2.1.
- **Normalize to Reprolevel**
Performs the *Normalize to Reprolevel* algorithm. For details, see section 6.2.1.

7.3.6 Calculate Thresholds

If you click on the *Start Calculation* button, you start the calculation. If you select the *Factorization* or *Factorization (orig. specs.)* algorithm the additional *Set Factors* button will be displayed. Click on the *Set Factors* button to open the following dialog:

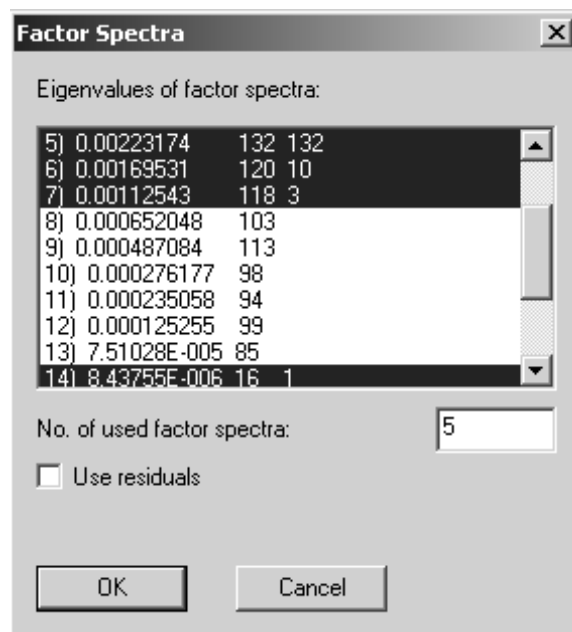
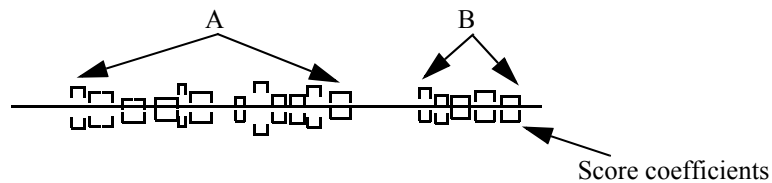


Figure 72: Factor Spectra

Selecting the optimum number of factors is not easy. OPUS facilitates this procedure by highlighting those factors which consists of the most non-overlapping score coefficient clusters, as shown in figure 72. The following graph is a simplified depiction of such non-overlapping score coefficient clusters of two groups A and B:



Example:

The factor selection field in figure 72 has to be interpreted as follows:

	A	B	C
5) 0.00223174	132	132	
6) 0.00169531	120	10	
7) 0.00112543	118	3	
8) 0.000652048	103		
9) 0.000487084	113		
10) 0.000276177	98		
11) 0.000235058	94		
12) 0.000125255	99		
13) 7.51028E-005	85		

Figure 73: Factor spectra highlighted

- A) Eigen values
- B) Number of separated score coefficient clusters (it is always started with the factor which includes the most groups separated)
- C) Number of separated score coefficient clusters additional to the factors selected before. In case of factor 6, e.g., there are 120 clusters separated, with 10 of these 120 clusters not being separated by factor 5. In case of factor 7 there are 118 clusters separated, with 3 of these 118 clusters not being separated by factor 5 nor factor 6.

Select the factor spectra from the list, which are used to calculate the spectral distance. It is not advisable to accept this value without performing a validation first. We recommend to perform several validations of the IDENT library using different numbers of factor spectra. Select the optimum number of factor spectra according to the result obtained.

It is not necessary to use a consecutive sequence of factor spectra. You can select factor 2, 3 and 5. Delete factor 1 if you do not want to get any baseline information.

If you want to calculate spectral distances by using spectral residuals, activate the *Use Residuals* check box.

7.4 Setup Identity Test Method – Threshold

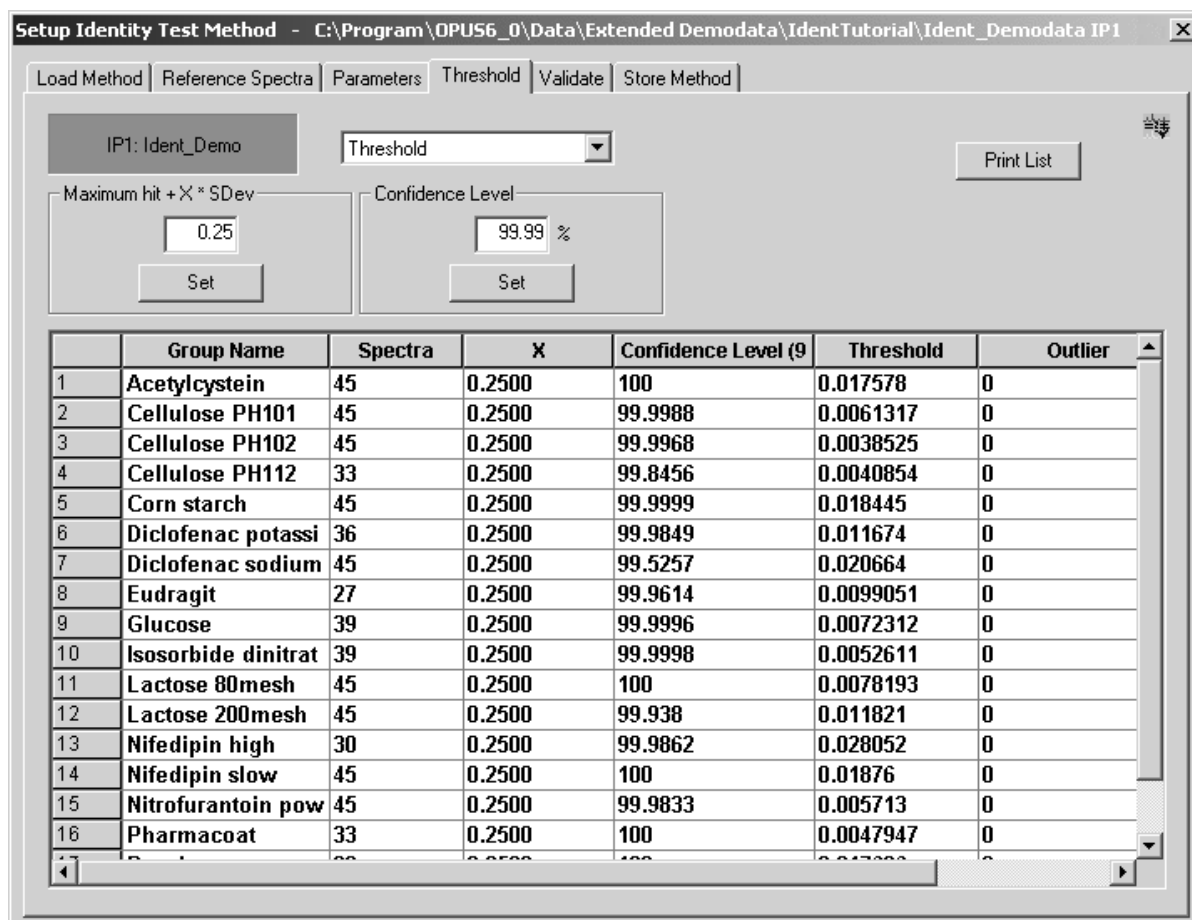


Figure 74: Setup Identity Test Method – Threshold tab

The threshold of a reference spectrum is the sum calculated by the maximum distance (maximum *Hit Quality*) listed in *Group Statistics*, plus the amount resulting from standard deviation (*SDev*) and a user-defined *x* factor. The threshold values, listed in the *Threshold* column are indicated for each reference spectrum.

7.4.1 Maximum Hit + X*SDev

This formula is used to calculate the threshold value. You can enter any value for *X* in the entry field, with 0.25 being set as default. To confirm your entry click on the *Set* button. This causes the new value to be set for all reference spectra, and the *Threshold* values will be updated.

7.4.2 Confidence Level

Two parameters are derived from the spectral distance to define the confidence level. You can enter any factor between 95 and 99.9999% into the entry field, 99.99% is set by default. To confirm your entry click on the *Set* button. See also chapter 1.4.

7.4.3 Set

If you click on the *Set* button, all changes made will be displayed on the table list. If you have selected only part of the column, only the values of the lines selected will be changed.

7.4.4 Group Statistics

The *Group Statistics* parameter includes detailed information on the *Group* and *File Name*, *Hit Quality*, *Standard Deviation* and *Mean Distance*.

Two parameters are derived from the spectral distances between original spectra and the average spectrum to define the confidence region for a group:

D_M mean distance:

$$D_M = \sum_i \frac{D(i)}{n} \quad (7-1)$$

S_0 standard deviation:

$$S_0 = \sqrt{\frac{\sum_i D(i)^2}{n-1}} \quad (7-2)$$

with n being the number of original spectra.

Select the *Group Statistics* option from the selection box. The following dialog opens:

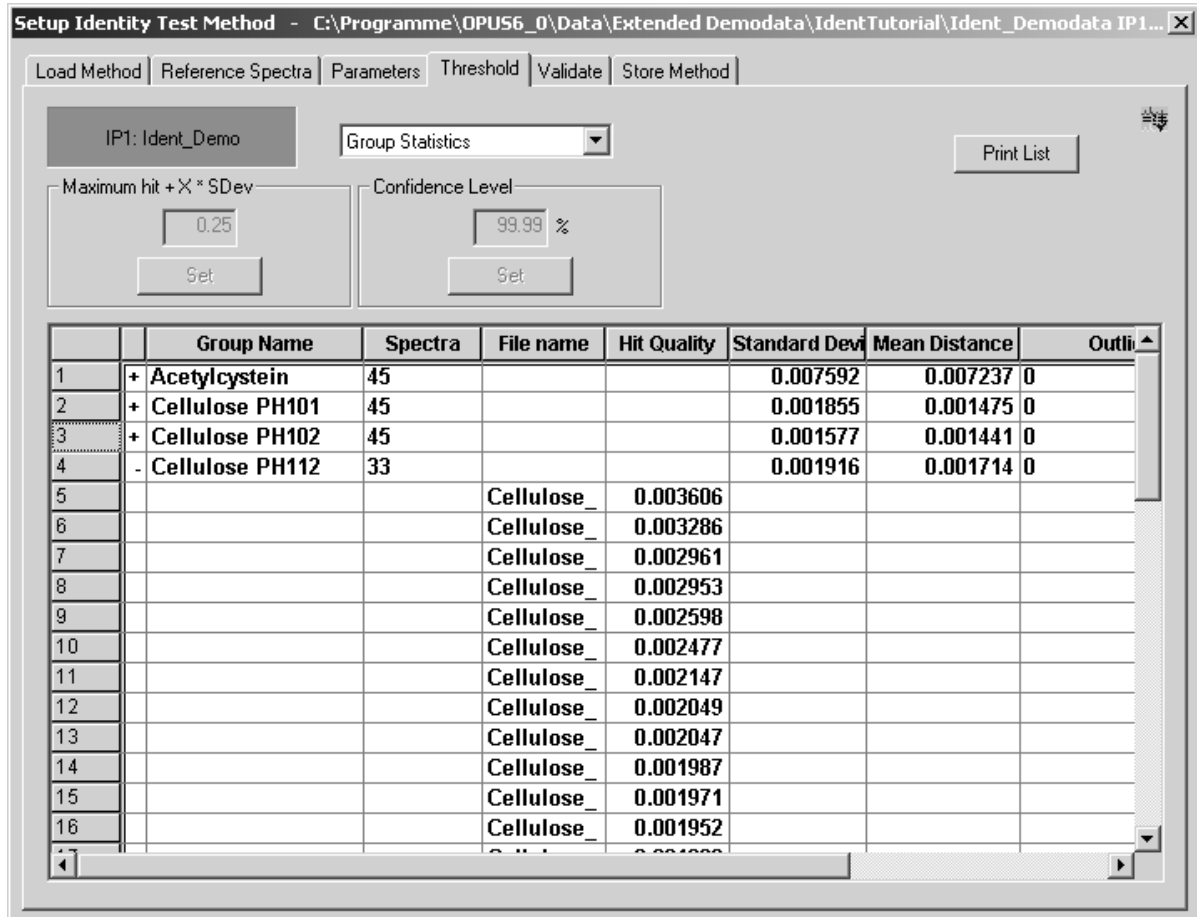


Figure 75: Setup Identity Test Method - Group statistics

If you use the confidence level to calculate the threshold, the *Outlier* column is very helpful to see at once the number of spectra per group, which are outside the threshold. To have the *Hit Quality* and *File Name* displayed click on the **[+]** button of the respective *Group Name* line.

7.5 Setup Identity Test Method – Validate

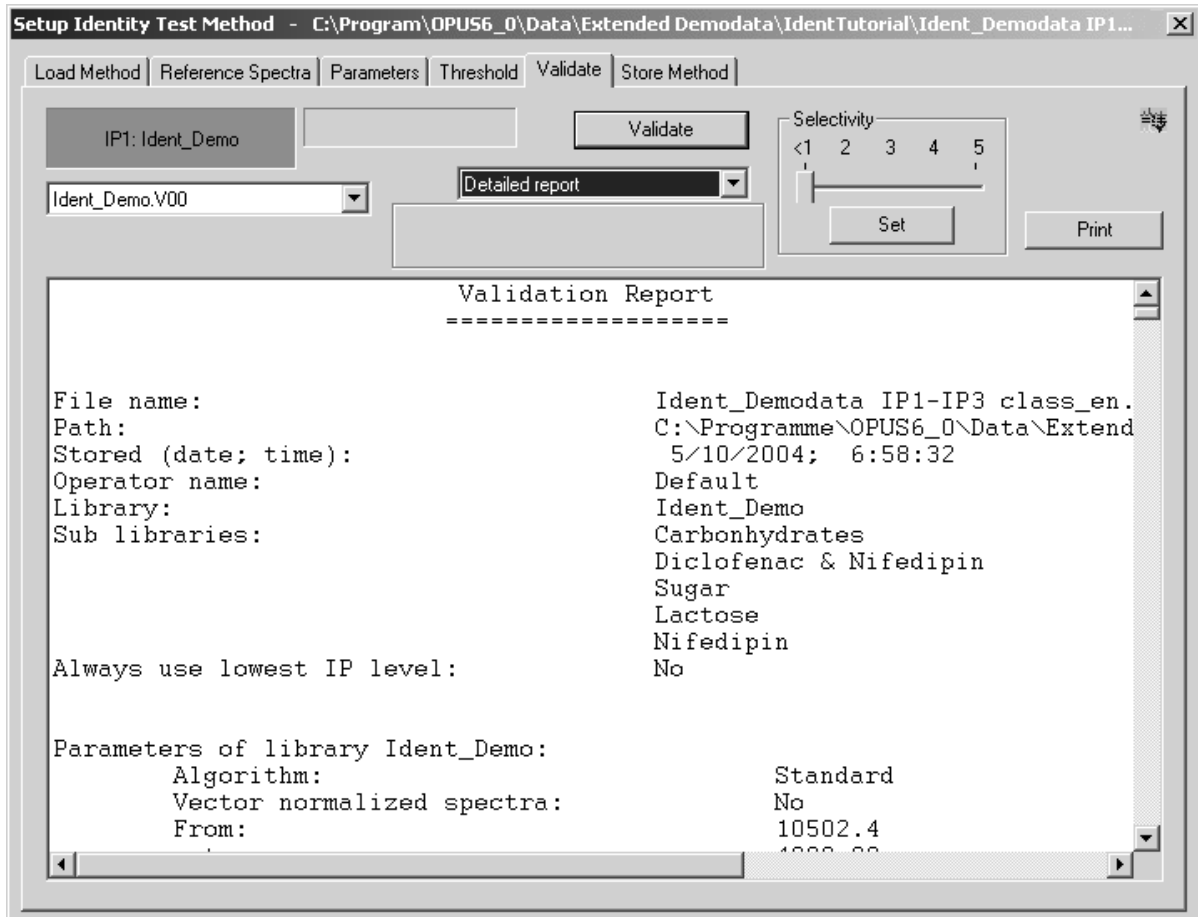


Figure 76: Setup Identity Test Method – Validate tab

Start validation by clicking on the *Validate* button. The following menu pops up:

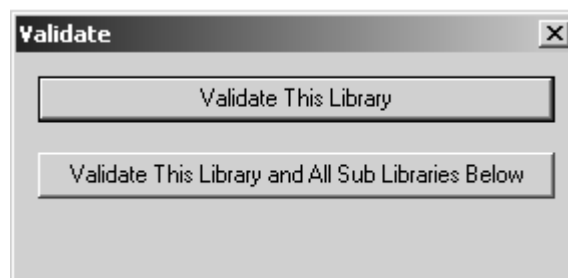


Figure 77: Validate pop-up menu

Two validation options are available:

- **Validate this library**
This option validates the main and sub-library you are currently working with.
- **Validate this library and all sub-libraries below**
This option validates the main and sub-library you are currently working with, and all additional sub-libraries belonging to the current library.

The progress of the validation process is indicated by the status bar.

7.5.1 Validation Report

Validation reports will not be overwritten and the report file name is always the same defined for the main or sub-library. Groups assembled in common classes are not considered to overlap. Nevertheless, the IDENT reports include the groups assigned to classes. You can select between the following reports which are based on single spectra, except for the selectivity report and histogram:

- **Summary Report**
The *Summary Report* outlines all important information on the current IDENT method, e.g. path and file name, sub-library names, date, time, operator name and comments. It includes all the groups which can be confused with other groups.
- **Result Report**
The *Result Report* outlines all important information on the current IDENT method, e.g. path and file name, sub-library names, date, time, operator name and comments. It gives additional information on which groups should be assigned to a new common sub-library.
- **Detailed Report**
The *Detailed Report* includes additional information on the algorithm used and frequency ranges defined of all sub-libraries, the order of internal derivative and smoothing points for internal derivative. Furthermore, this reports specifies all thresholds of the overlapping groups as well as the distances of all single spectra which overlap. The report provides additional information on which groups should be assigned to a new common sub-library.
If you have activated the *Always use lowest IP level* check box on the *Parameters* tab, the *Detailed Report* will include only the results of the lowest IP level for all spectra of each group.
- **Selectivity Report**
This report is based on average spectra. If you use the selectivity slider on top, you can get a more detailed report. For example, if you set the slider to a selectivity of 3, all spectra are shown in the spectral distance between 1 and 3.

- **Selectivity Histogram**

The histogram is a summary of the selectivity report.

In general, validation reports directly compare one spectra group with the adjacent spectrum to see which clusters overlap and where. The selectivity report compares *Material 1* and *Material 2*, as explained in the chart in figure 78.

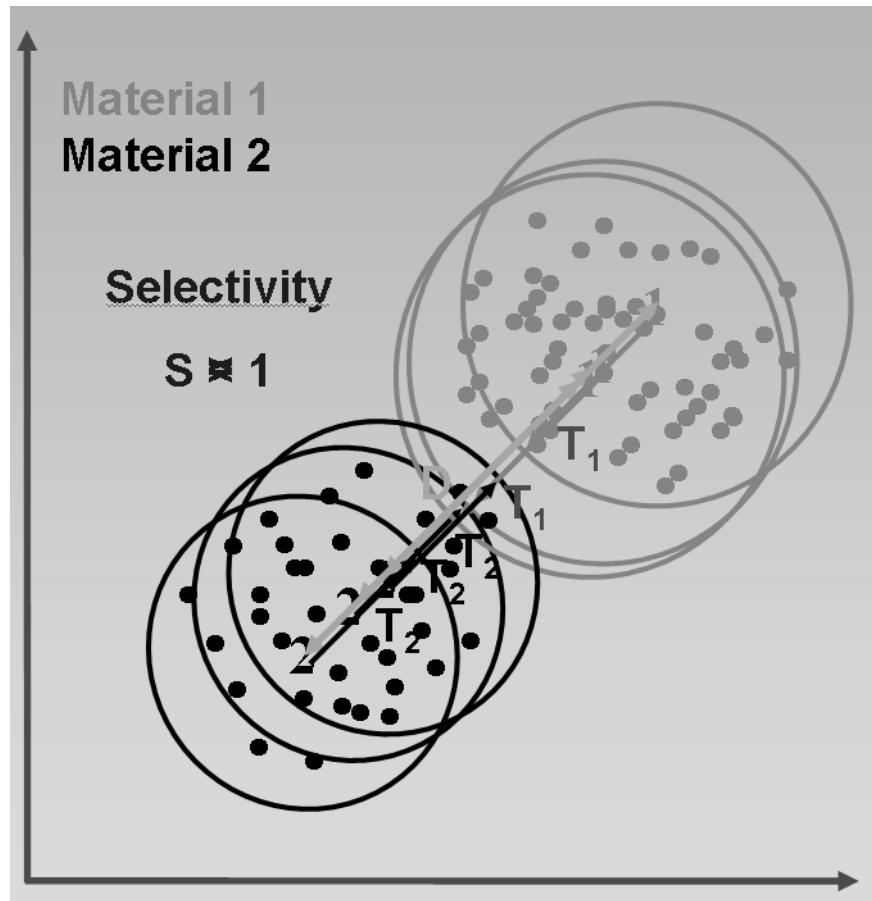


Figure 78: Calculating selectivity

The selectivity will be calculated as follows: $S = \frac{D}{(T_1 + T_2)}$ with S being the ratio of distance D between average spectra and the sum of threshold values T_1 and T_2 (cluster radii). This results in the following:

- $S < 1$: overlapping
- $S = 1$: cluster in contact
- $S > 1$: cluster separated

Figure 79 exemplifies a selectivity report.

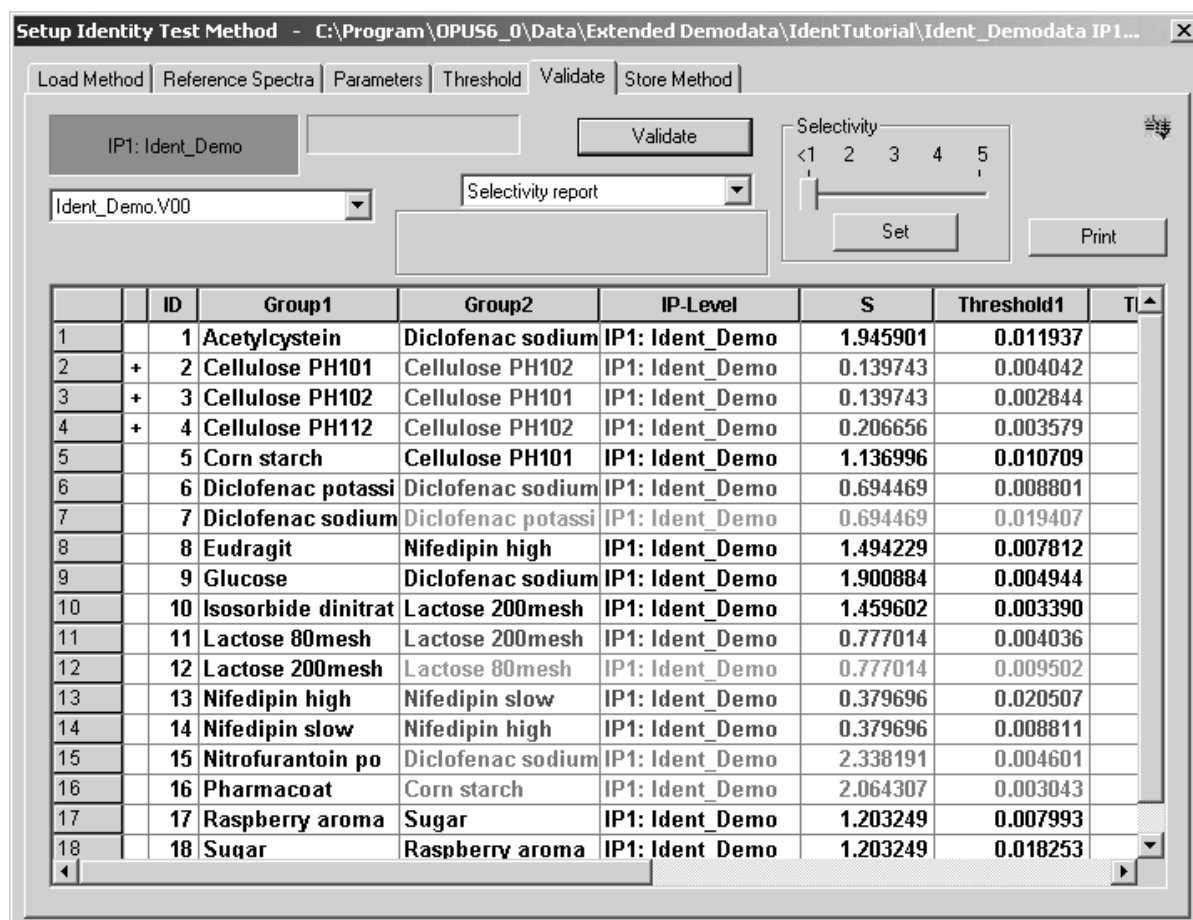


Figure 79: Selectivity Report

The selectivity report displays the result in different colors:

- Red: spectra with a selectivity of <1
- Green: spectra with a selectivity of >2
- Black: spectra with a selectivity between 1 and 2
- Gray: spectra with a selectivity of <1, without single-spectra overlappings in the validation report

Note: Generally, single spectra are not relevant in case of selectivity, however, they are indicated in the validation report. If there are no single-spectra overlappings in this report, the respective group is displayed in gray in the selectivity report.

It may occur that in case of libraries with reference spectra not any of these spectra is within the intersection of two clusters. Therefore, the validation would yield to a non-overlapping result. However, the selectivity does indicate the geometric overlapping, as exemplified in figure 80.

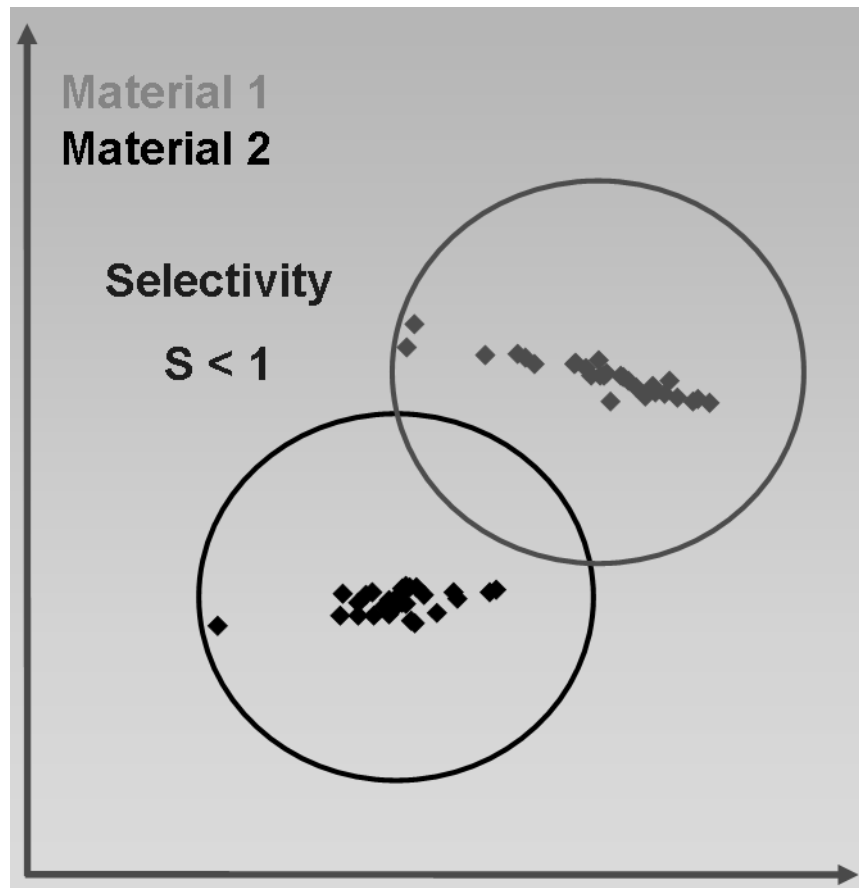


Figure 80: Selectivity - Geometric overlapping

The selectivity report can also be read out as score plots in 3D-format, which is indicated by the Factor View ... tab. This kind of representation is based on the factorization of single reference spectra or average spectra, shows the distribution of spectra and supports the selecting of meaningful factor spectra.

Select *Factorization* from the *Method* drop-down list on the *Parameters* tab and click on the *Start Calculation* button. Define at least 3 factors from the *Factor Spectra* dialog which will serve as a basis for the 3D factor view. Subsequently, validate the library by clicking on the *Validate* button on the *Validate* tab.

To display the cluster of one group select the respective group from the *Selectivity Report* list. The number of neighboring groups can be set by means of the *Selectivity* slider. You can activate or deactivate the *Opaque* check box (see figure 79). In both cases the real threshold of the cluster is shown in all dimensions, i.e. x, y and z axis. If you deactivate the *Opaque* check box, the spectra (A in figure 81) of each cluster can be seen as the clusters will be transparent.

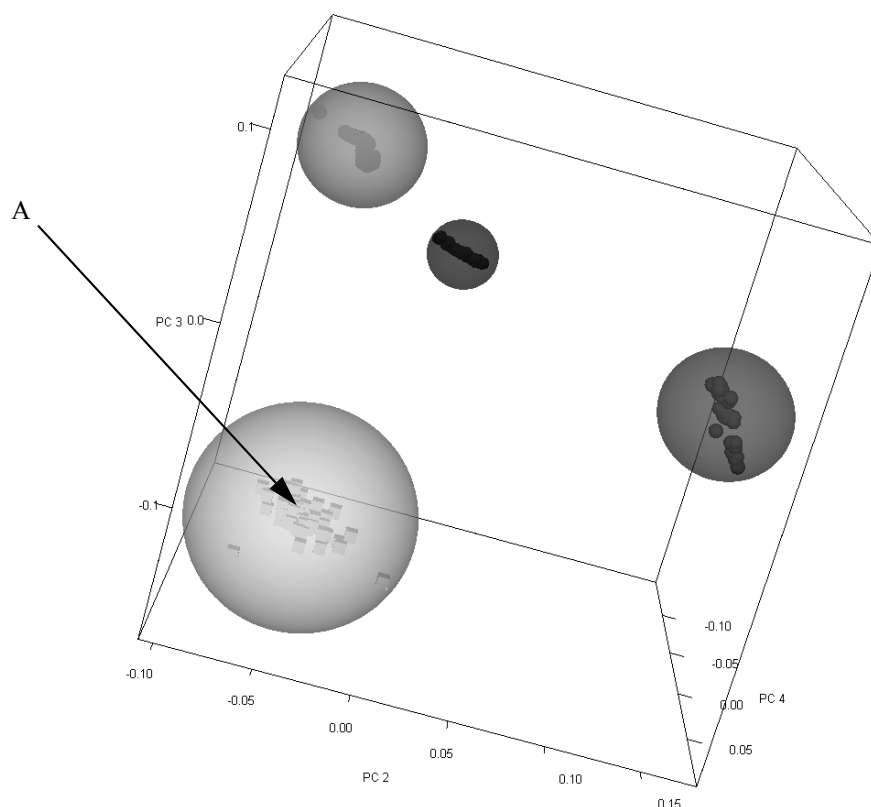



Figure 81: 3D factor view - Transparent clouds and data points

If you position the mouse on one specific spectra, the file name and group name will be displayed. To improve the factor view, you can rotate the box. If you position the mouse on the edge of the box, the cursor changes into . To rotate the box press the left mouse button and move the mouse to the position desired.

Right clicking somewhere on the 3D display pops up the *Properties* button. If you click on this button, the *View properties* dialog is displayed which allows further plot settings. To optimize the 3D factor view you can select additional factors by means of the *Factor* drop-down list. Always the next 3 factors of the original selection can be used for each dimension.

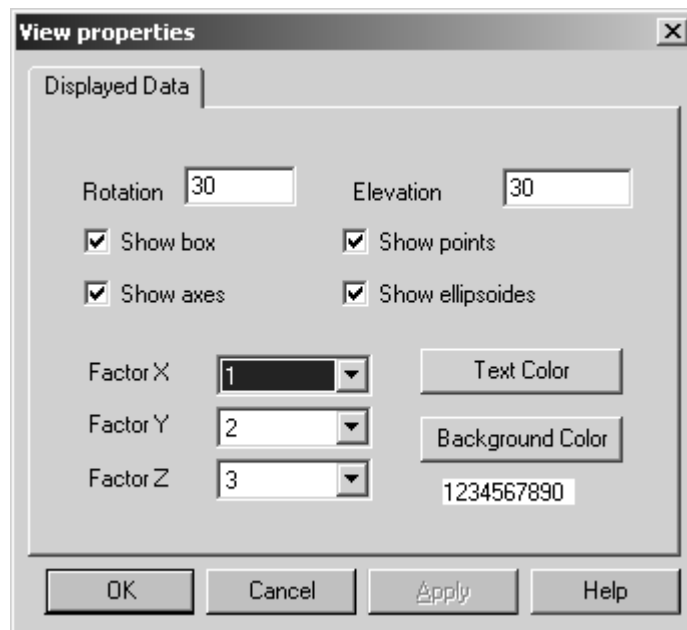



Figure 82: View properties

Sometimes the clusters may be displayed as ellipsoides. This is due to the scaling of the axes, and not a result of the original calculation.

To zoom the box displayed press the left mouse button and draw a frame around the position desired. Double click into the zoomed area to undo the zoom setting. To return to the IDENT setup close the factor view by clicking on the  icon.

7.5.2 Print

To print the report, click on the *Print* button. This starts the *Windows Notepad* program which you can also use to reformat the text, if desired. Use the *Notepad* print function to create a printout.

Note: It is recommended to select a small font in *Windows Notepad* to avoid extraordinary long reports. A proportional font may lead to a confusing display of the results. Therefore, it is advisable to use a monospace font, e.g. *Courier New*, 10.

7.6 Setup Identity Test Method - Store Method

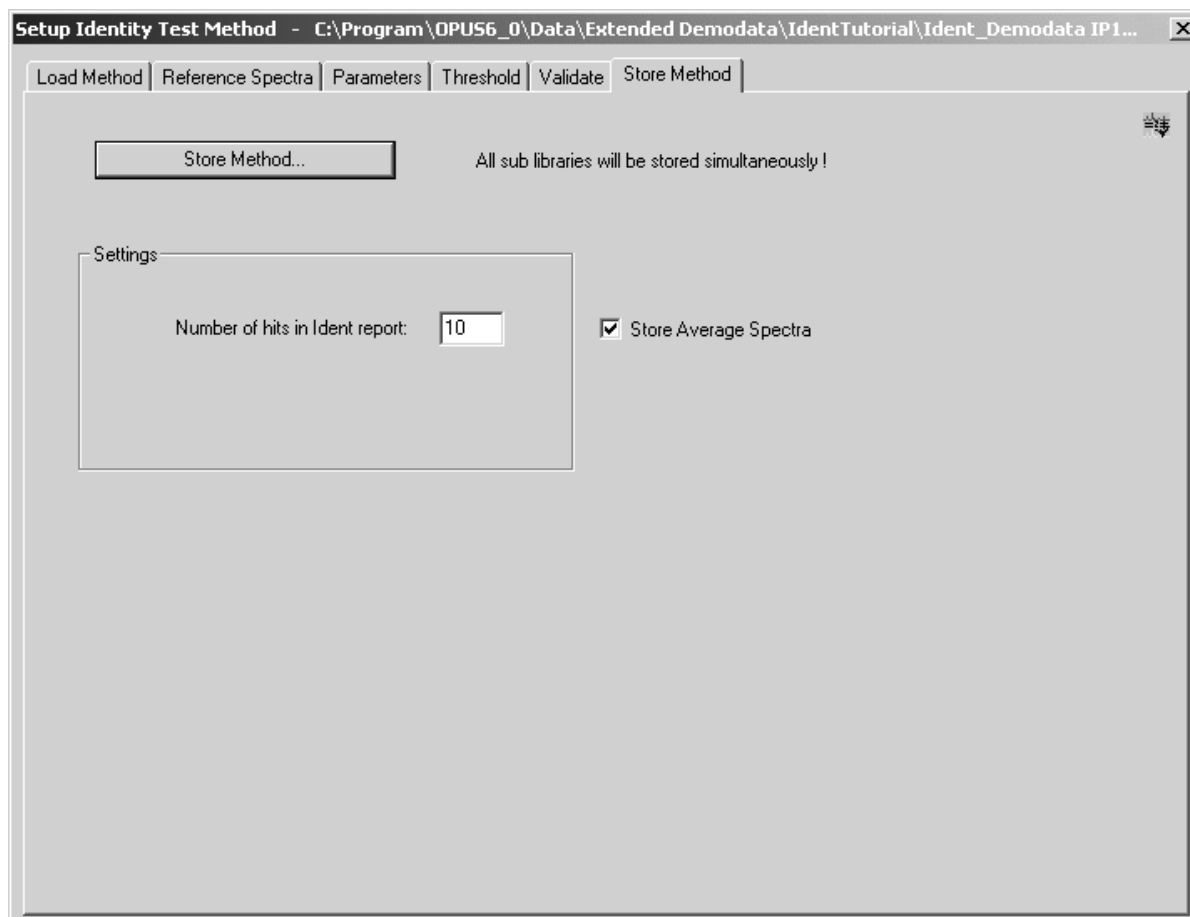


Figure 83: Setup Identity Test Method – Store Method tab

This dialog allows to store a method file created. The parameter you can define is the number of *Hits* to be stored in the IDENT report. By default, the *Store Average Spectra* is activated. Click on the *Store Method* button to open the standard *Save File* dialog box. The method file has the extension **.FAA*. All sub-libraries will be stored simultaneously.

7.7 Identity Test

To start an IDENT analysis select the *Identity Test* command from the *Evaluate* menu. The following dialog box opens:

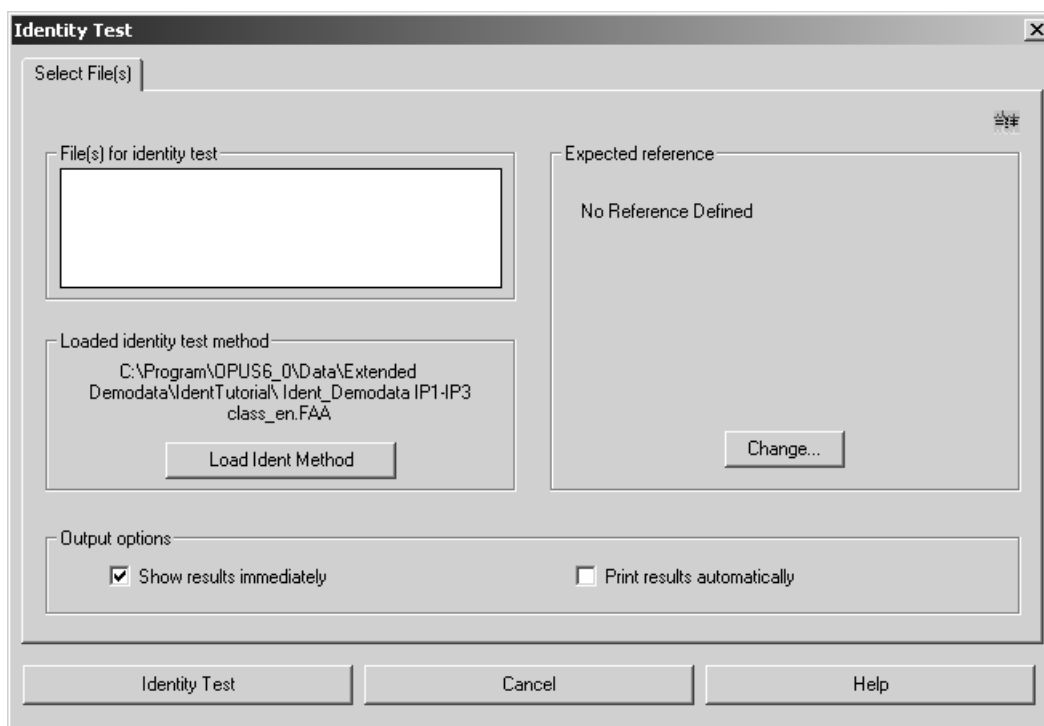


Figure 84: Identity Test - Select File(s) tab

Select a test spectrum and drag and drop the spectrum absorption block from the OPUS browser window into the *File(s) for identity test* selection field.

To load or modify an IDENT method click on the *Load Ident Method* button and select the desired method from the dialog that opens. If an identity test method has already been loaded (e.g. if you have created one prior to starting the analysis) the path and method name will be indicated in the *Loaded Identity Test Method* field.

The *Show results immediately* check box in the *Output options* group field is activated by default and the identity test results will be shown in a special evaluation result display, see figure 85 and 86.

Result of IDENT Evaluation:

Sample: C:\OPUS\Daten_Ident_Kurs\Cellulose microcryst\Cellulose microcryst.2
Method File: C:\OPUS\Daten_Ident_Kurs\DemoMethod.FAA
Date and Time: 2002/08/27 15:29:27 (UTC-1)

Hit No.	Sample Name	Hit Qual.	Threshold	Group
1	Cellulose microcryst.\n	0.00823	0.01046	Cellulos
2	Starch from rice\n	0.14093	0.01013	Starch f01
3	Starch from corn\n	0.17524	0.00905	Starch f
4	Starch soluble\n	0.19326	0.01146	Starch s
5	Starch from wheat\n	0.20849	0.00602	Starch f02
6	Starch from potato\n	0.22265	0.02762	Starch f00
7	Lactose monohydrate\n	0.25061	0.00951	Lactose 01
8	Lactose 80 mesh_110250_01_KFO	0.26170	0.04676	Lactose 00
9	Lactose 200 mesh_110251_02_KFO	0.29578	0.04622	Lactose
10	Glucose monohydrate\n	0.29989	0.00736	Glucose

IDENTIFIED AS Cellulos


 **OK**

Figure 85: IDENT Evaluation Result display - Result OK

In the lower part of the display the *Identity Test* result is indicated. A green check mark and the description *OK* would indicate that the test has passed, i.e. the product has been identified.

A red cross and the description *NOT OK* would indicate that the comparison has failed, i.e. the product has not been identified.

Result of IDENT Evaluation:

Sample: C:\OPUS\Daten_Ident_Kurs\Cellulose microcryst\Cellulose microcryst\Cellulose.5
Method File: C:\OPUS\Daten_Ident_Kurs\DemoMethod.FAA
Date and Time: 20/08/2001 14:43:52

Hit No.	Sample Name	Hit Qual.	Threshold	Group
1	Cellulose microcryst\n	0.05012	0.01046	Cellulos
2	Starch from rice\n	0.13808	0.01013	Starch #01
3	Starch from corn\n	0.16486	0.00905	Starch f
4	Starch soluble\n	0.17751	0.01146	Starch s
5	Starch from wheat\n	0.19599	0.00602	Starch #02
6	Starch from potato\n	0.20626	0.02762	Starch #00
7	Lactose monohydrate\n	0.24137	0.00951	Lactose 01
8	Lactose 80 mesh_110250_01_KFO	0.25294	0.04676	Lactose 00
9	Lactose 200 mesh_110251_02_KFO	0.28417	0.04622	Lactose
10	Glucose monohydrate\n	0.28957	0.00736	Glucose

NOT IDENTIFIED


 **NOT OK**

Figure 86: IDENT Evaluation Result display - Result not OK

To print the evaluation result display of the identity test, activate the *Print Results Automatically* check box in the *Output Options* group field.

7.7.1 No Reference Defined

If there is no expected reference defined, click on the *Change* button. There are three possibilities to define an expected reference:

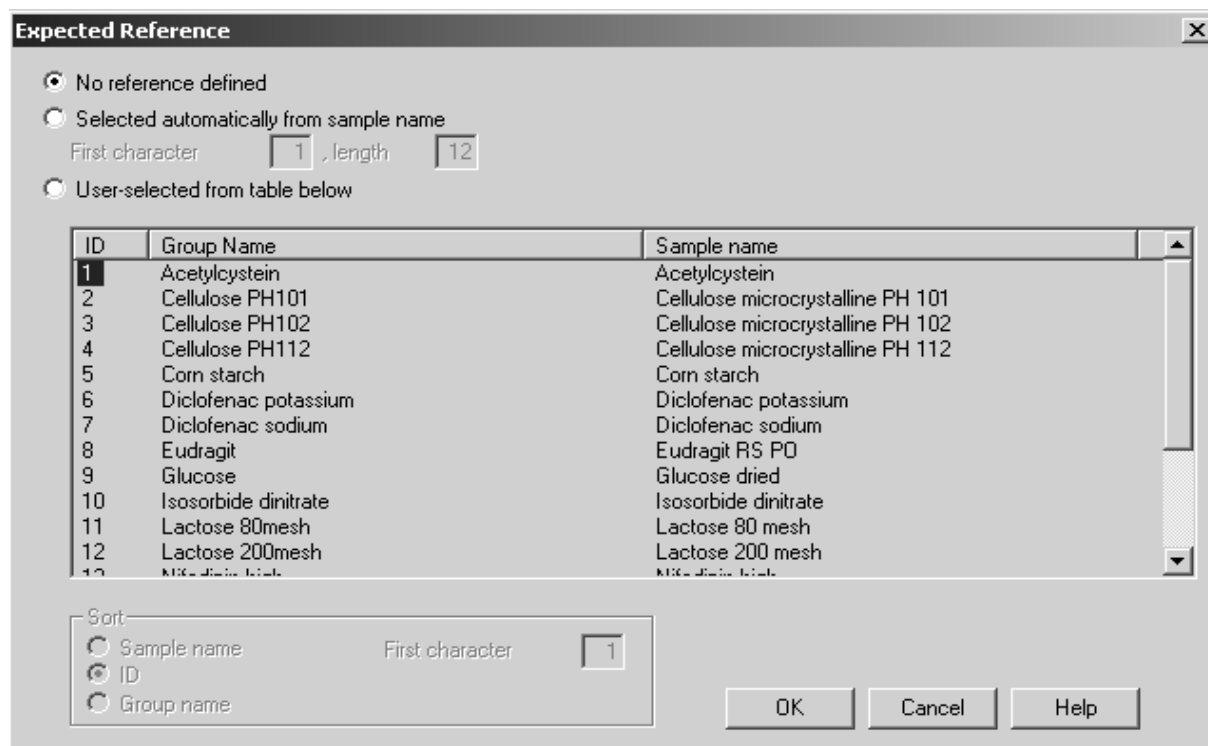


Figure 87: Expected Reference

a) No reference defined

The identity test analysis works without a pre-defined reference spectrum.

b) Selected automatically from sample name

The reference spectrum is determined by comparing the sample name of the test spectrum with the sample names of the library spectra. Usually, the sample name will not be completely used, but partially.

First character:

Indicates at which sample name character the character comparison will start. In this example the comparison starts at character 1.

Length:

The *Length* indicates how many characters will be taken into account during comparison. Example: the sample name of the test spectrum is *000002, Sample DL-Isoleucin*. It is sufficient to use the first six characters for a definite selection: *First = 1* and *Length = 6*.)

c) User-selected from table below

If you activate this option button, all reference spectra will be listed. Each line contains the sample name (e.g. *000002, Sample DL-Isoleucin*), the number of the reference spectrum in the library (e.g. *ID=2*) and the group name. Select the spectrum which you expect to match the test spectrum. The test spectrum can have any sample name.

In the *Sort* group field you define how to sort the list. You either sort according to *Sample name*, *ID* or *Group Name*. If you check *Sample name*, you can additionally define the character number of the sample name, which you start sorting with.

7.8 Cluster Analysis – Load Method

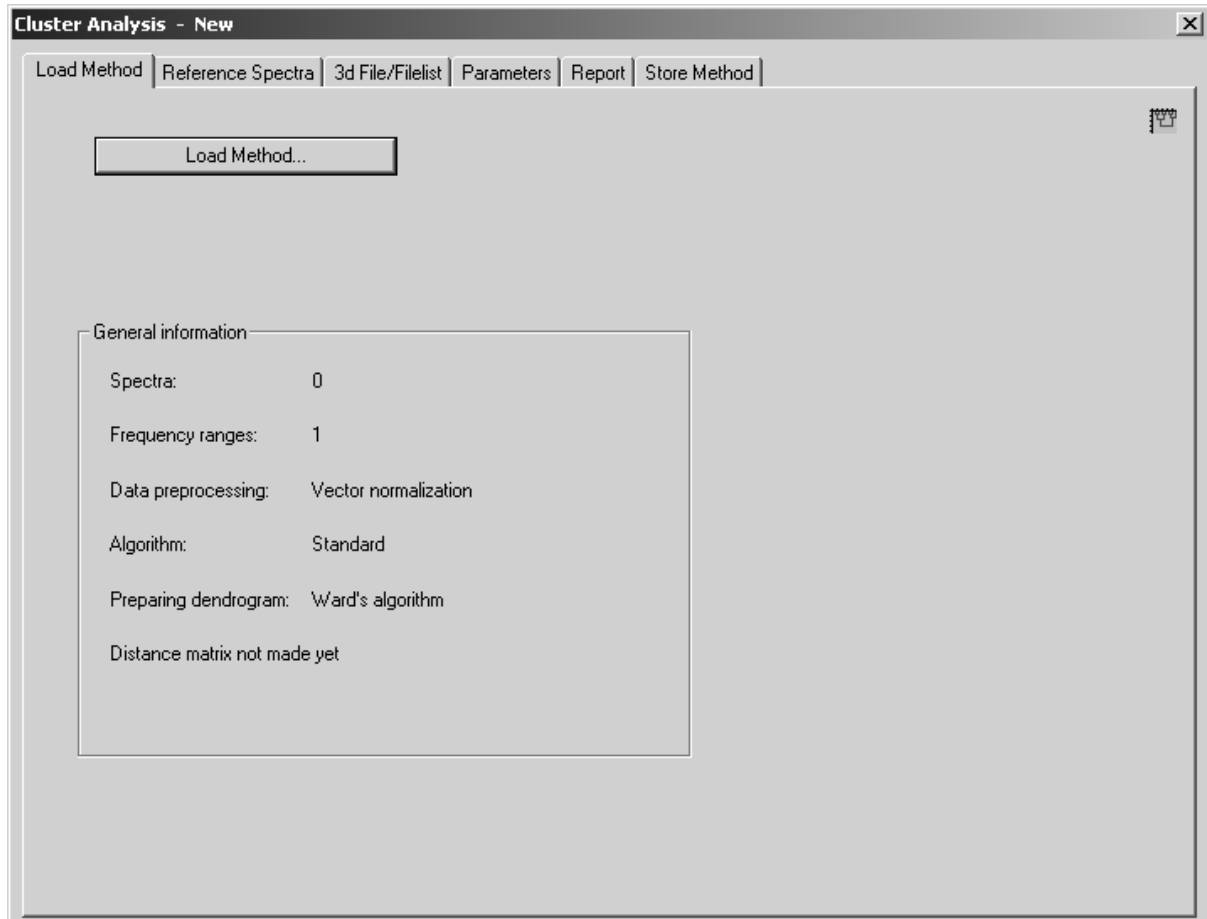


Figure 88: Cluster Analysis – Load Method tab

7.8.1 Load Method

Use the *Load Method* button to load an existing cluster analysis method. Cluster analysis method files have the extension **.CLA*. It is also possible to load cluster analysis method files created by OPUS-OS/2 IDENT. However, if you store such a method using OPUS/IDENT, you will not be able to load the method by OPUS-OS/2 IDENT. To avoid this, store the modified OPUS-OS/2 IDENT file by using a different file name.

7.8.2 General Information

The *General information* group field provides statistical information on the existing method file. The number of spectra used for the method and the number of frequency ranges included are displayed. You will get additional information on the data preprocessing method, the algorithm used for the identity test and dendrogram, and whether a distance matrix has been generated.

7.9 Cluster Analysis – Reference Spectra

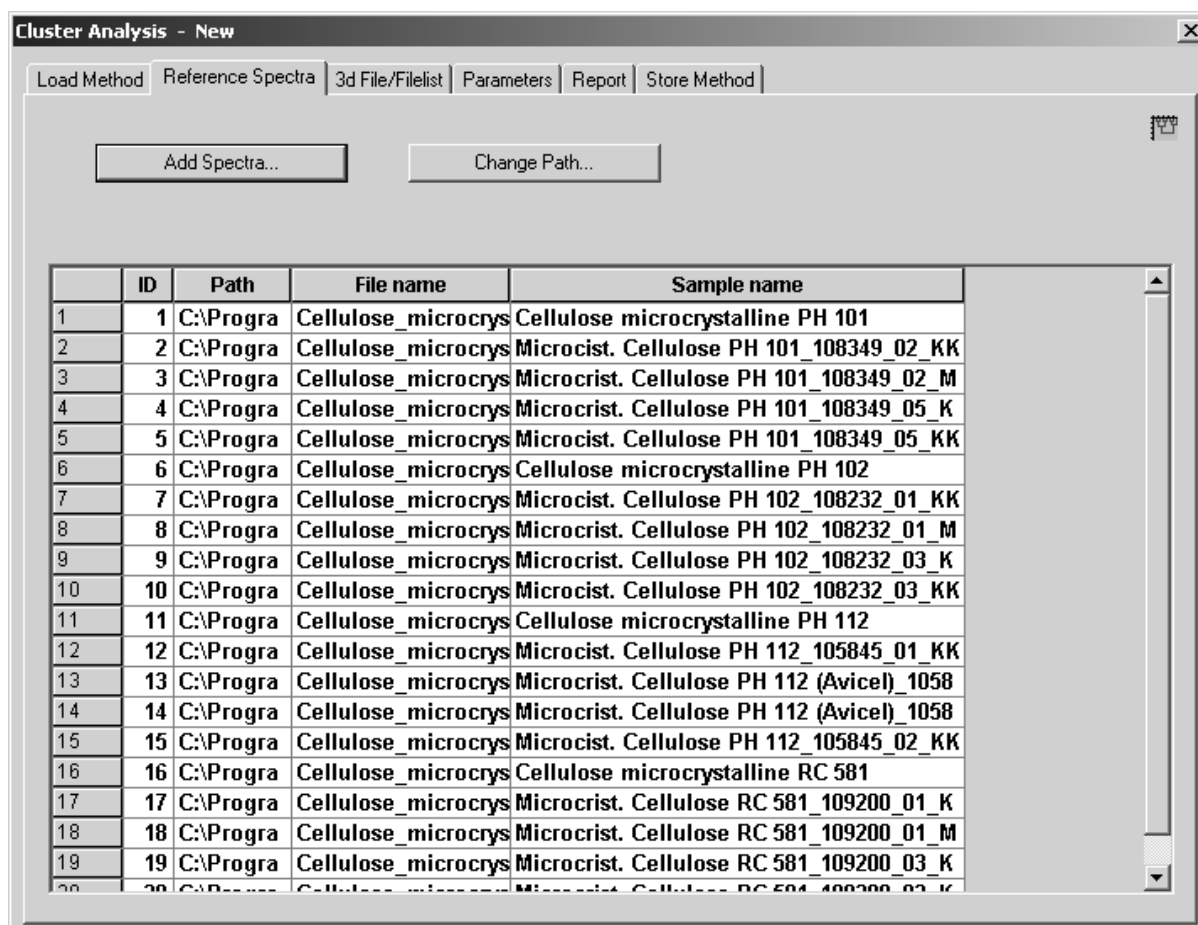


Figure 89: Cluster Analysis – Reference Spectra tab

This dialog box is the same as the *Reference Spectra* dialog box of the *Setup Identity Test Method* command and has been described in section 7.2.

7.10 Cluster Analysis – Parameters

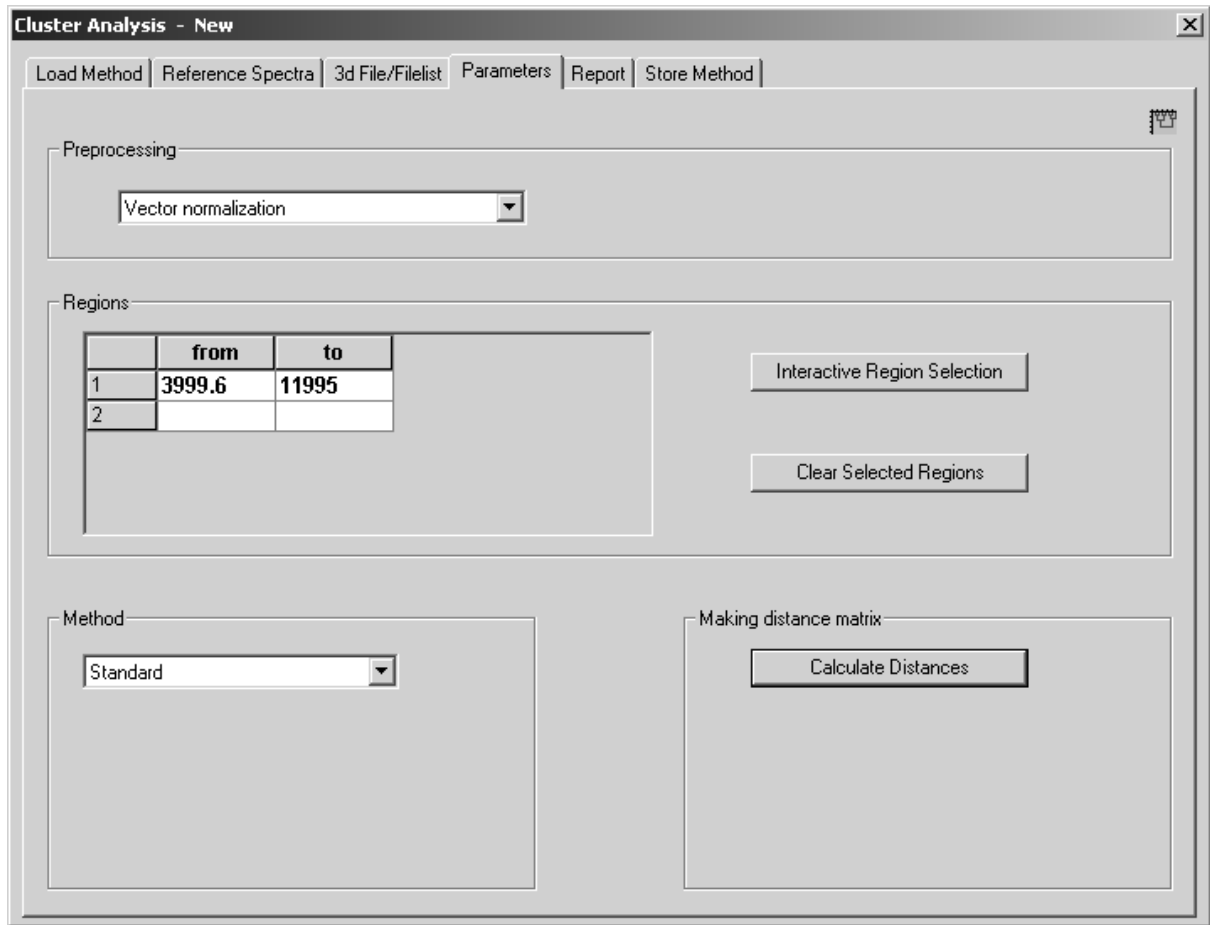


Figure 90: Cluster Analysis – Parameters tab

7.10.1 Preprocessing

The cluster analysis uses the same data preprocessing methods as described in section 7.3.

7.10.2 Regions

The *Regions* table allows to limit the data to one or several spectral regions to be considered for the cluster analysis. The frequency limits for the spectral regions can either be entered manually or selected interactively.

7.10.3 Method

Select an algorithm to identify the spectrum (see section 6.1 and 7.3).

If you select the *Factorization* method, you have to specify the number of factor spectra to be used to calculate the spectral distances, in the *Factor Spectra* dialog box. This dialog box opens automatically after clicking on the *Calculate Distances* button.

In contrast to the identity test, the *Use Residuals* option is not available during cluster analysis. Spectral residuals are not taken into account when calculating spectral distances. The calculation of factor spectra is not necessary during cluster analysis, as test spectra will not be analyzed. To determine the spectrum-to-spectrum distance (see section 6.1.2) only Z (covariance matrix) and L (Eigen vectors) will be calculated. Click on the *OK* button to return to the *Parameters* dialog box.

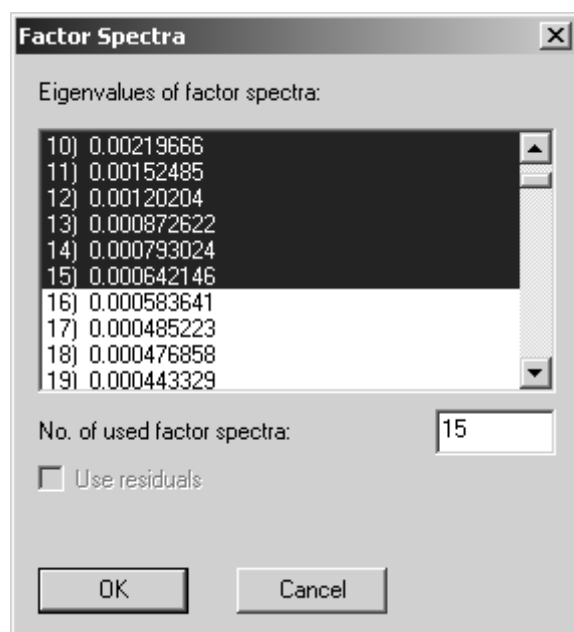


Figure 91: Cluster Analysis – Factor Spectra

7.10.4 Calculate Distances

Click on the *Calculate Distances* button to start the calculation of the spectrum-to-spectrum distance. If the calculation has been finished, it is recommended to store the method before you generate a dendrogram.

7.11 Cluster Analysis – Report

Click on the *Report* tab to have the analysis results displayed.

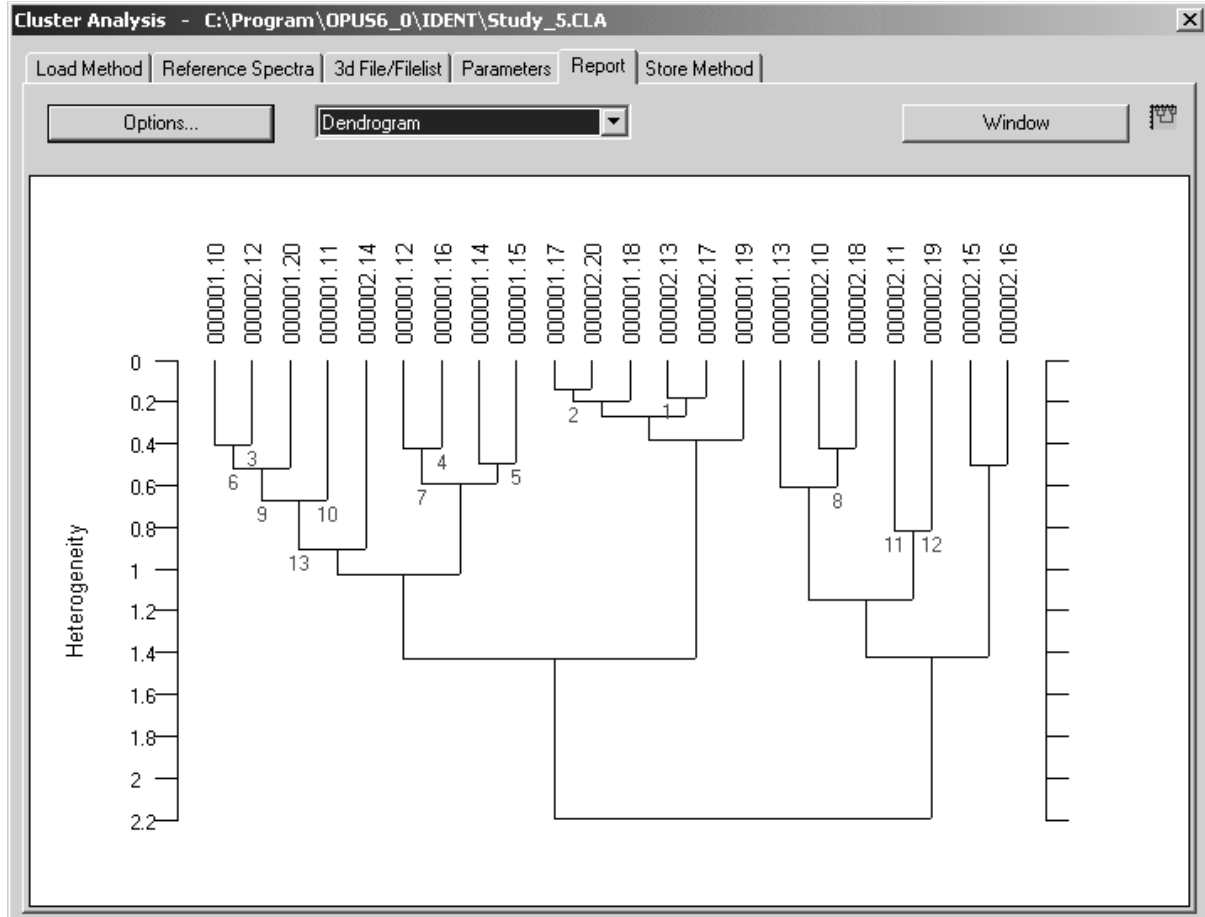


Figure 92: Cluster Analysis – Report tab

Use the drop-down list to define the form of report. You can have the results displayed in the form of a dendrogram, histogram or diagnosis. *Dendrogram* is set by default.

- **Dendrogram**

A dendrogram includes the spectral distances of all reference spectra. Right click on the dendrogram and a menu pops up displaying different options.

- **Histogram**

This kind of report is not intended to be used to analyze clustering. Instead, the spectrum-to-spectrum distances between reference spectra are analyzed. Such distances can be represented in the form of a symmetrical $n \times n$ matrix (n being the number of reference spectra).

The mean value and standard deviation are calculated, and the distance values are displayed in the form of a histogram and divided

into classes. The first class, e.g. ranges from 0 to 1, the second class includes spectral distances from 1 to 2 etc. Each class is represented by a bar in the histogram. This bar indicates the percentage frequency of spectral distances compared to the total number of distances considered.

In this context *Class* means something different than in case of clustering, where cluster can also be referred to as *Class* or *Group*. Besides graphical representation, the *Histogram* includes statistical information.

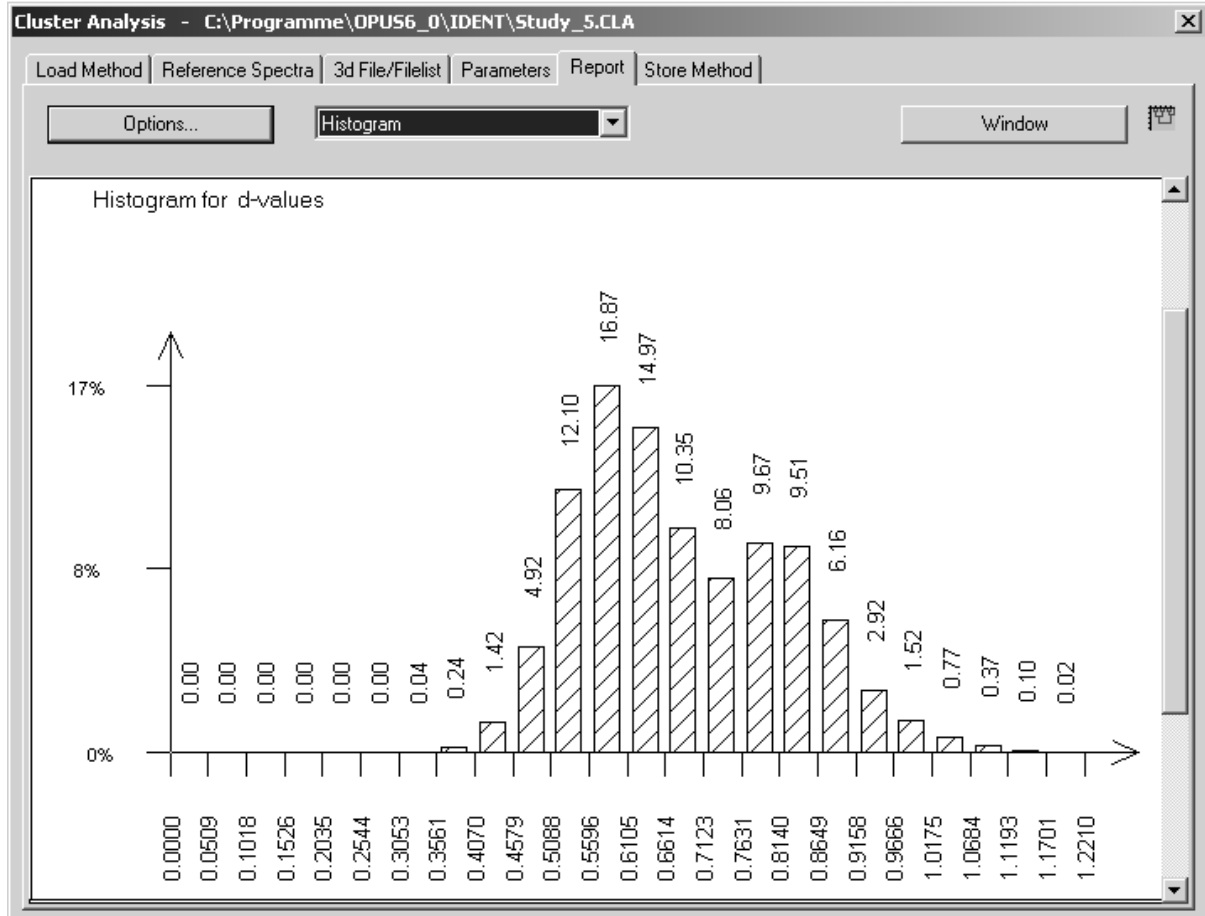


Figure 93: Cluster Analysis – Histogram

- **Diagnosis**

This view produces a horizontal cross section of the dendrogram. Specify the number of classes to create a list which includes the members of each single class. The spectral distance of the last clustering will be displayed for each cluster.

7.11.1 Score Plot

The cluster analysis report can also be read out as score plots in 3D-format which is indicated by the **Factor View** tab. The additional *Score Plot* button will only be displayed if you have selected *Factorization* as analysis method and *Diagnosis* as report before.



Figure 94: Cluster Analysis - *Score Plot* button

Select *Factorization* from the *Method* drop-down list on the *Parameters* tab and click on the *Start Calculation* button. Define at least 3 factors from the *Factor Spectra* dialog which will serve as a basis for the 3D factor view. Subsequently, select *Diagnosis* from the drop-down list and define the number of classes you want to see in the score plot by using the *Options* button. For details on the *Options* dialog see chapter 7.11.2. Click on the *Score Plot* button.

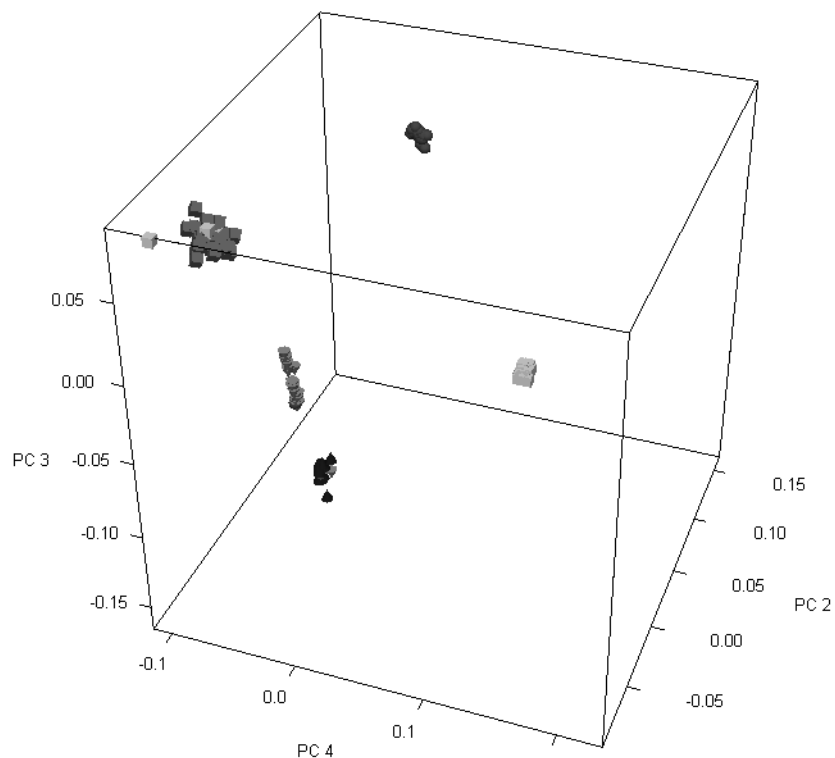



Figure 95: Cluster analysis - 3D factor view

If you position the mouse on one specific spectrum, the file name and group name will be displayed. To improve the factor view, you can rotate the box. If you position the mouse on the edge of the box, the cursor changes into . To rotate the box press the left mouse button and move the mouse to the position desired.

Right clicking somewhere on the 3D view pops up the *Properties* button. If you click on this button, the *View properties* dialog is displayed which allows further plot settings.

7.11.2 Options

Click on the *Options* button to open the *Cluster Analysis - Options* dialog. You can define the algorithm used to calculate spectral distances between different clusters. In addition, you specify the *Number of Classes* used in the diagnosis, and the parameters required for the histogram.

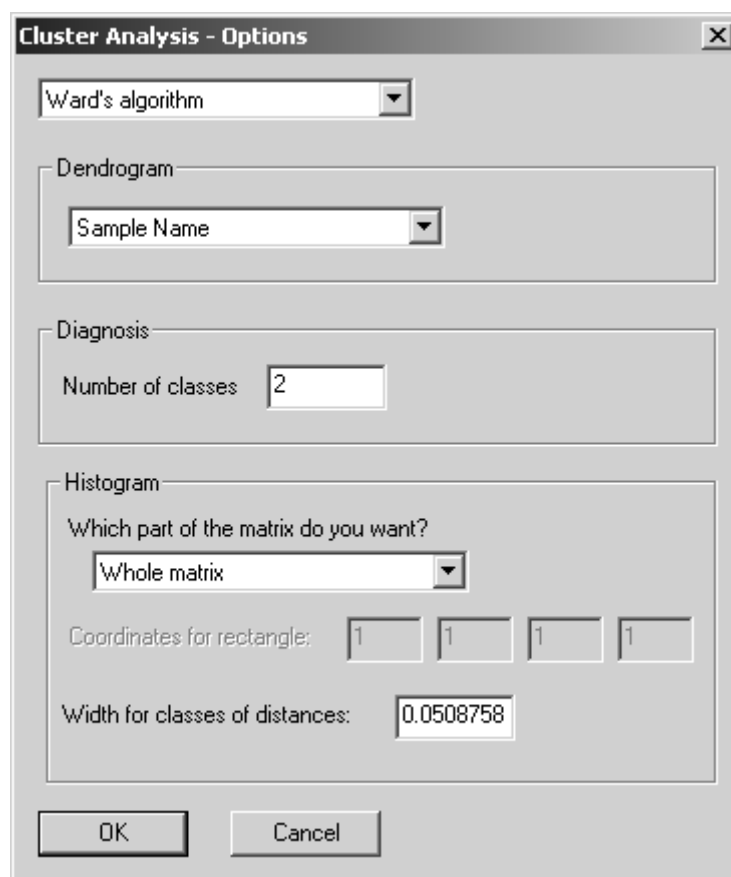


Figure 96: Cluster Analysis – Options

You can select between 7 different algorithms to calculate spectral distances between clusters:

- Single Linkage
- Complete Linkage
- Average Linkage
- Weighted Average Linkage
- Median Algorithm
- Centroid Algorithm
- Ward's Algorithm

For details, see section 4.1.2.

Use the *Dendrogram* drop-down list to define the kind of labeling. Dendrograms are labeled vertically. There are 4 possibilities:

- *File Name* of the reference spectra
- *Sample Name* of the reference spectra
- *File Number* (file sequence in the list of reference spectra)
- *No Name Markers* (no labelling at all)

A text file is automatically created for each dendrogram. This file has the same name as the cluster analysis method and the extension **.DEN*. The file includes the dendrogram and exact clustering levels.

Specify the number of classes you want to test. If you test, e.g., original spectra used to generate average spectra in an identity test, you have to enter the number of average spectra into the *Number of Classes* field.

Define which part of the matrix you want to include into the histogram:


- **Whole Matrix**
In this case all distances will be used. With the matrix being symmetrical and diagonal elements being 0, only a triangular matrix without diagonal elements is used. The matrix size is $(n \cdot (n - 1))/2$.
- **Only Pairs (for repro tests)**
The data record is divided into pairs and the distances between the first and second spectra (first pair), third and fourth spectra (second pair) etc. are calculated. The number of distances being considered is $n/2$. This value can be used to determine the reproduction level of measurements which have been repeated twice.
- **Only Triplets (for repro tests)**
Same as above, but this time the data record is divided in triplets. The number of distances used for statistics is $(n/3) \cdot 3 = 3$.
This value can be used to determine the reproduction level of measurements which have been repeated three times.

- **Only Reference (i.e. the last column)**
This option only considers distances between the last spectrum indicated in the list and all other spectra. The number of distances is $n-1$.
- **A Given Triangle**
The distances are calculated for those spectra between k and l (rows) of the list. This results in a triangle within the matrix. You have to enter the k and l parameters. For $k = 1$ and $l = n$ the result is identical to the *Whole Matrix* option result.
- **A Given Oblong**
The distances are determined for those spectra between the k_1 and l_1 position (rows) and between the k_2 and l_2 position (columns) of the list. This corresponds to a rectangle within the matrix.

The value specified in the *Width for Classes of Distances* field determines the number of classes. The default value corresponds to a division of 20 classes, i.e. the range from 0 up to the maximum distance is divided into 10 equal areas. You can change this value. The maximum number of classes is 20. If you enter an invalid number, the value will automatically be corrected.

The *Window* button opens the dendrogram, diagnosis or histogram within a *Report* window. You can have the report printed out using the print options from the *OPUS Print* menu.

7.12 3D File/Filelist

You can also set up a cluster analysis using either a 3D file or a file list. When working with a 3D file you first have to load such a file. OPUS automatically opens the 3D display indicated by the  tab.

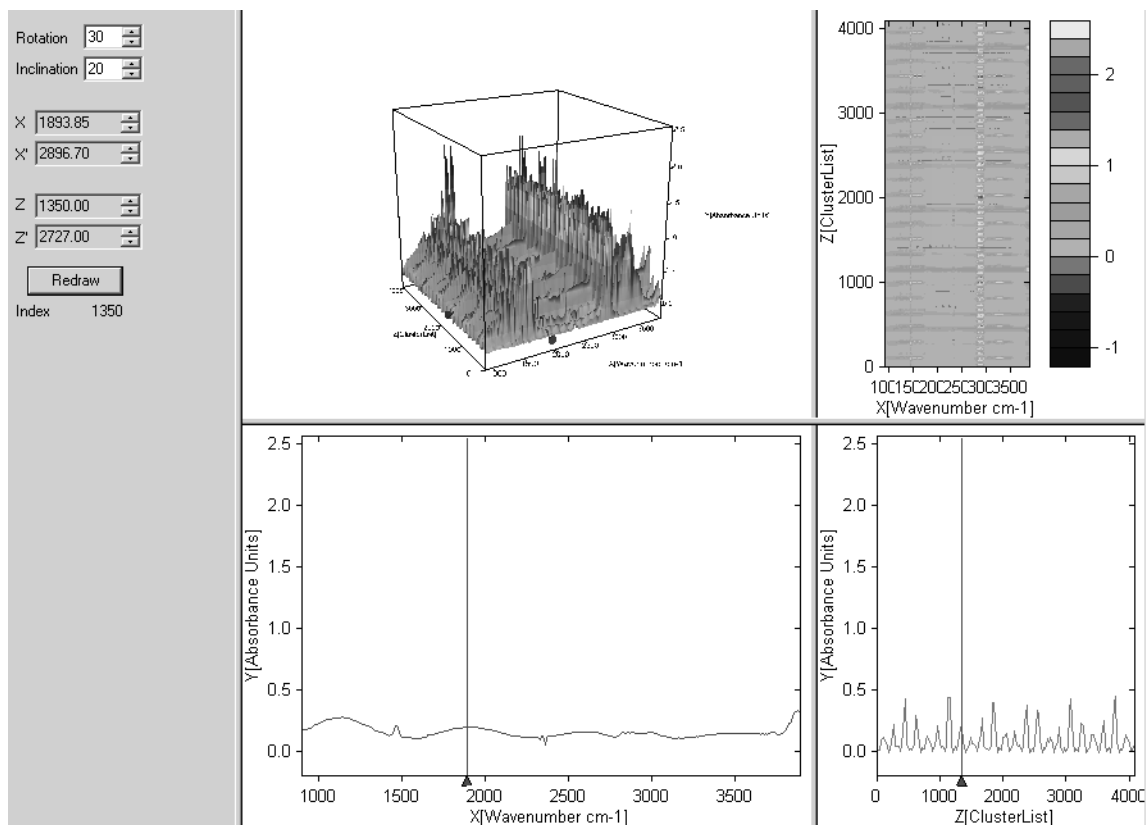


Figure 97: 3D view

For further details on the 3D window settings in OPUS refer to the 3D manual. Now, select the *Cluster Analysis* command from the *Evaluate* menu and click on the *3d File/Filelist* tab. The following dialog opens:

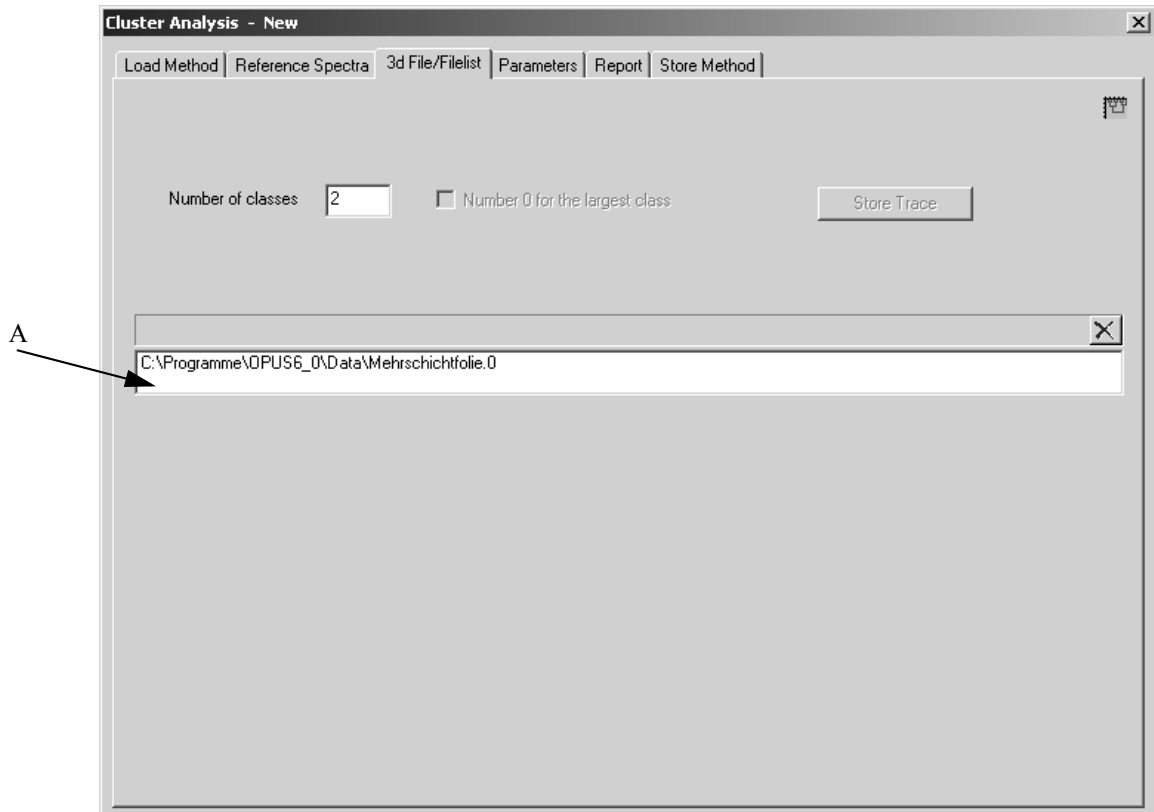



Figure 98: Cluster Analysis - 3d File/Filelist tab

Drag & drop the spectra data block of the 3D file into the entry field (A in figure 98). You cannot load more than one 3D file for a cluster analysis. To remove the spectra data block, select the spectra file in the entry field and click on the  button.

Now, click on the *Parameters* tab, define the frequency regions and select an identification method. Click on the *Start Calculation* button. Depending on the number of spectra the calculation procedure can take quite some time. If the calculation has been finished, click on the *3D File/Filelist* tab again. Define the number of classes and click on the *Store Trace* button which is now enabled.

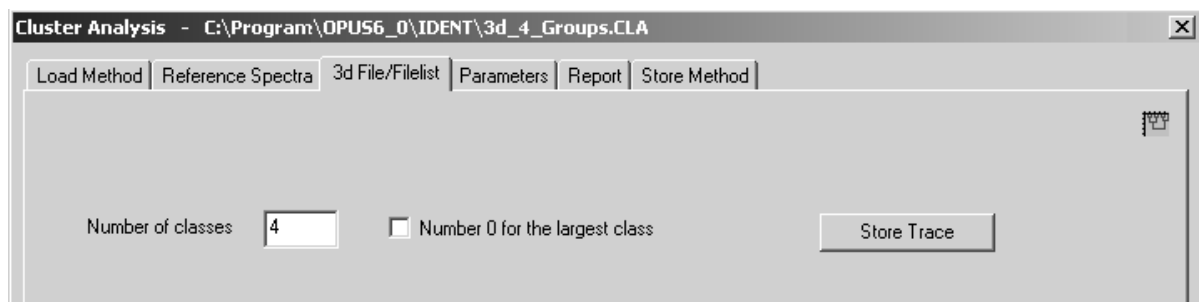



Figure 99: 3d File/Filelist - Defining number of classes

Note: If you activate the *Ignore Rest* check box, the biggest cluster will get the number 0 in the trace report.

The TRC data block () is added to the file displayed in the OPUS browser window. Right click onto this data block to have the corresponding report displayed.

Micron	1490_1420B	1700_1500B	1800-1700B	ClusterList	ClusterList	ClusterList
0.000000	0.418906	-0.120632	-0.173669	1.000000	1.000000	1.000000
1.000000	0.278322	-0.196559	-0.116866	3.000000	3.000000	1.000000
2.000000	0.151217	-1.032597	-0.395969	3.000000	3.000000	1.000000
3.000000	0.101227	-0.717422	-0.241355	3.000000	3.000000	1.000000
4.000000	0.277226	-0.700929	0.020516	3.000000	3.000000	1.000000
5.000000	0.547875	-0.505013	-0.018080	3.000000	3.000000	1.000000
6.000000	0.644494	-0.847821	-0.129880	3.000000	3.000000	1.000000
7.000000	0.671650	-1.464416	-0.143565	3.000000	3.000000	1.000000
8.000000	1.109604	-1.766877	-0.075675	3.000000	3.000000	1.000000
9.000000	1.526494	-2.598635	-0.043315	3.000000	3.000000	1.000000
10.000000	2.988753	-2.730210	-0.076485	3.000000	3.000000	1.000000
11.000000	4.816987	-2.387223	-0.199955	4.000000	0.000000	1.000000
12.000000	8.534338	-2.149328	-1.021528	4.000000	0.000000	2.000000
13.000000	11.658361	-2.408520	-1.096437	4.000000	0.000000	2.000000
14.000000	13.718537	-1.306824	-1.135268	4.000000	0.000000	2.000000
15.000000	14.933686	-0.474837	-1.368319	4.000000	0.000000	2.000000
16.000000	16.100191	0.333270	-2.150514	4.000000	0.000000	2.000000
17.000000	16.300209	0.140034	-2.622877	4.000000	0.000000	2.000000
18.000000	16.168112	0.241585	-2.715865	4.000000	0.000000	2.000000
19.000000	16.493189	-0.654120	-3.486097	4.000000	0.000000	2.000000
20.000000	17.045324	-0.416201	-3.244622	4.000000	0.000000	2.000000

	1490_1420B	1700_1500B	1800-1700B	ClusterList	ClusterList	ClusterList
INSS	0	0	0	0	0	0
INSR	0	0	0	0	0	0
IRun	0	0	0	0	0	0
INPT	4096	4096	4096	4096	4096	4096
INoGoodFW	0	0	0	0	0	0
INoGoodBW	0	0	0	0	0	0
INoBadFW	0	0	0	0	0	0
INoBadBW	0	0	0	0	0	0
dHFL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
dLFL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
dHFFL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
dLFFL	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
IFilterSize	0	0	0	0	0	0
IFilterType	0	0	0	0	0	0
dFFP	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
dFLP	4095.000000	4095.000000	4095.000000	4095.000000	4095.000000	4095.000000
dMin	-1.450400	-8.559641	-9.014781	1.000000	0.000000	1.000000
dMax	26.899309	181.258301	16.887852	4.000000	3.000000	4.000000
dSCF	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
dpka_fw	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
dpka_bw	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 100: Cluster analysis - TRC data report

As the TRC data report in figure 100 exemplifies the *Cluster List* column includes the allocation to the classes defined before. To have the traces scored in a 3D plot open the *Map+Vid+Spec* window by the *New Registered Window* command from the *Window* menu. *Drag & drop* the TRC data block into the first sub-window.

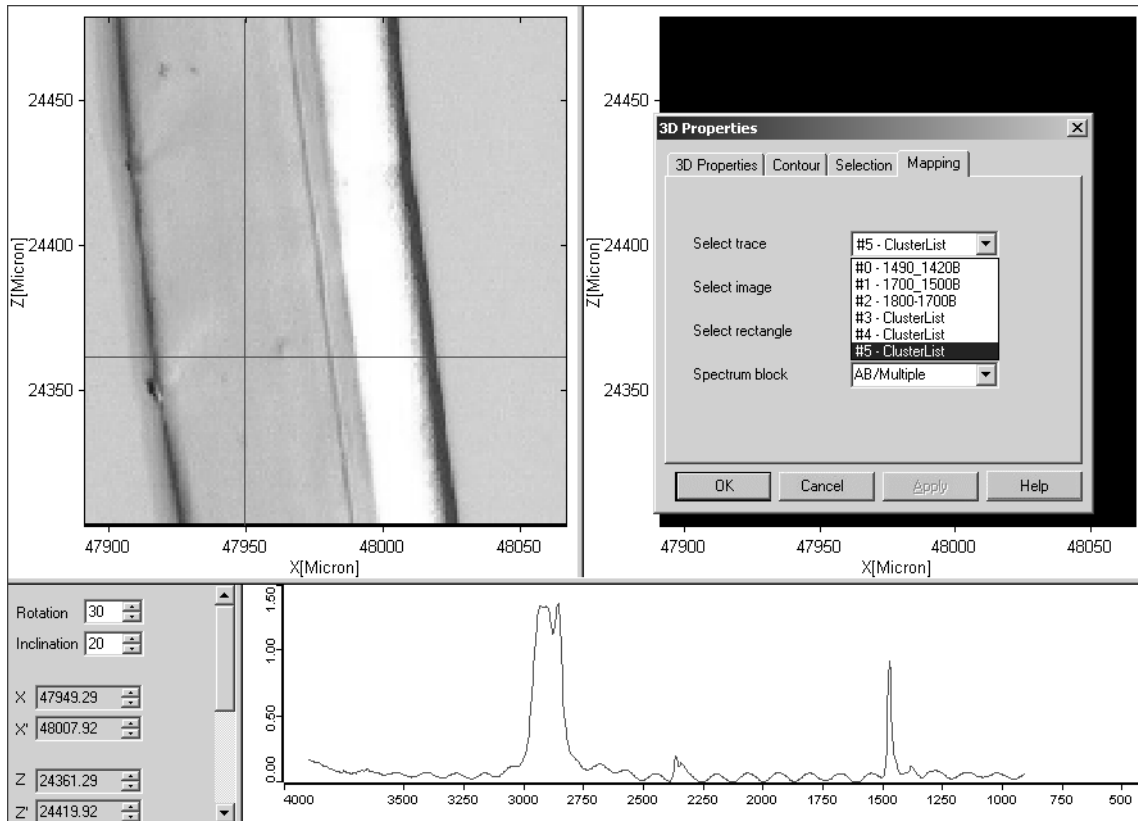


Figure 101: 3D plot of TRC data block

To have the clusters displayed in the second sub-window right click onto the window and select the respective cluster list from the *Select trace* drop-down list on the *Mapping* tab. For all the other plot options refer to the 3D manual.

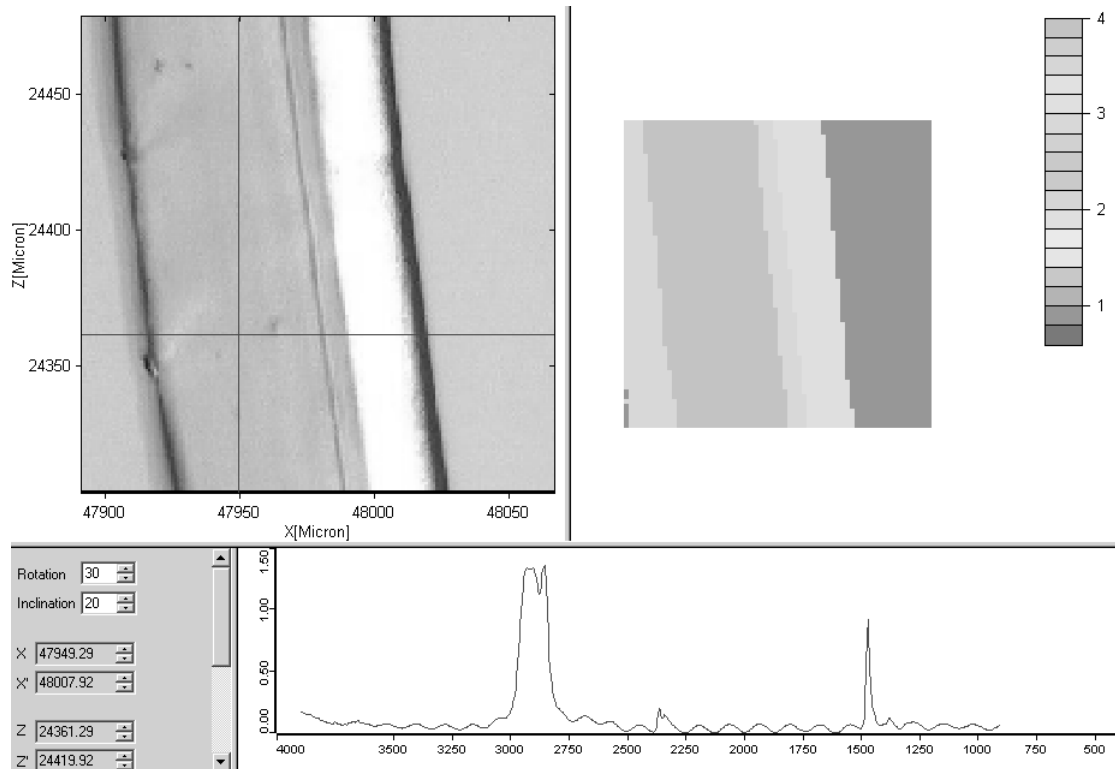


Figure 102: 3D plot of TRC data block with clusters

7.12.1 File List

If you use a file list several spectra of a particular spectrum type are combined into one common file list. First, create a file list by the *Setup File List* command in the *Edit* menu and store it. Drag & drop the LIST data block (LIST) into the entry field (A in figure 98).

Now, click on the *Parameters* tab, define the frequency regions and select an identification method. Click on the *Start Calculation* button. You can also store traces before you have defined the number of classes on the *3d File/Filelist* tab. To open the TRC data block right click on the file list name in the OPUS browser window and select *Show Parameters* from the pop-up menu. The TRC data block is now added to the file list in the browser window. Click onto the data block to be able to see the trace results.

7.13 Cluster Analysis – Store Method

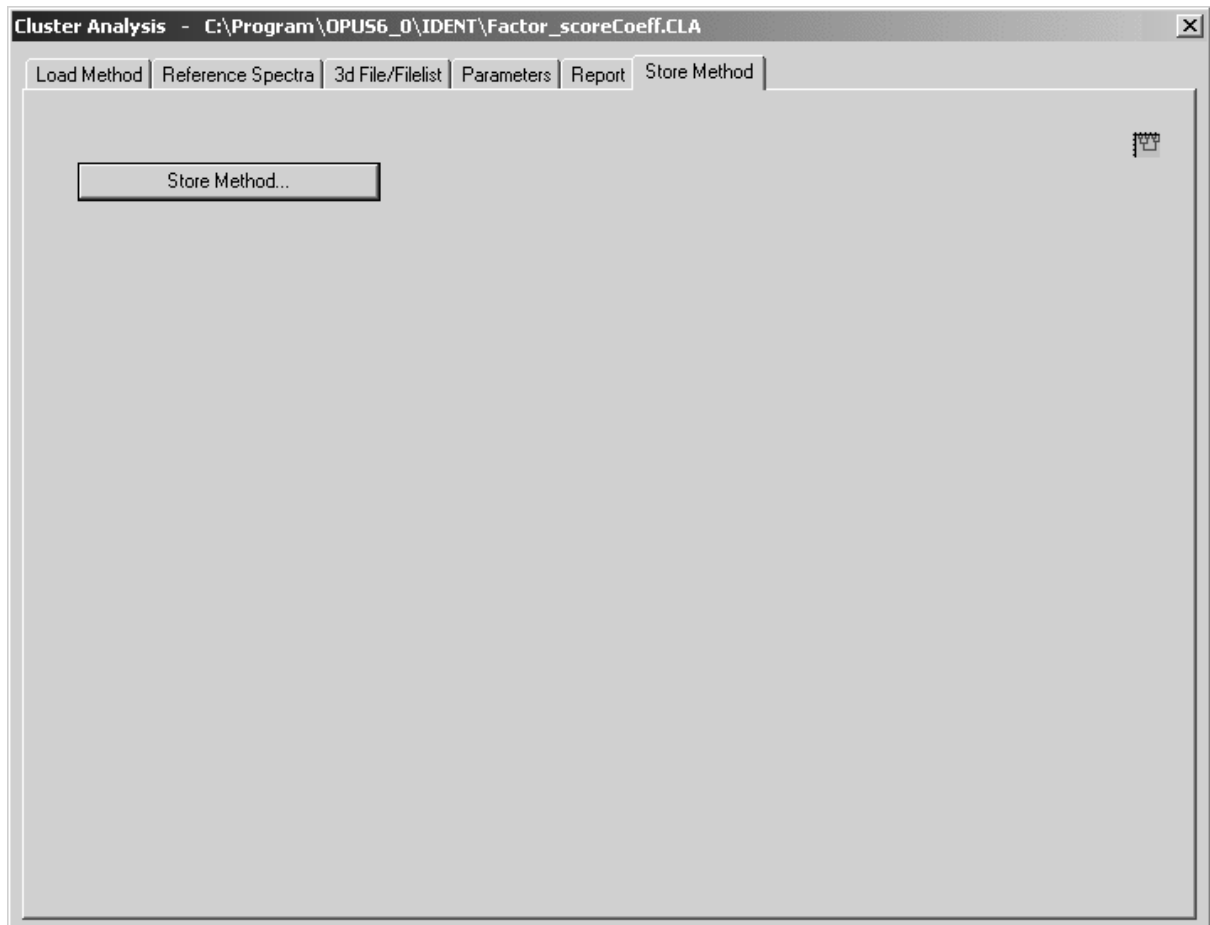


Figure 103: Cluster Analysis – Store Method tab

This dialog box allows to store the cluster analysis method created. Click on the *Store Method* button to open the standard *Save File* dialog box. The method file has the extension **.CLA*.

Index

Numerics

3D Files 114

A

A Given Oblong 114
A Given Triangle 114
Abs. Threshold 68
Add Average Spectra 6
Add Region 12, 86
Add Spectra 6
Artificial Spectrum 29
Average Linkage 27, 29, 30, 113
Average Spectrum 51

C

Calculate Distances 33, 108
Can Be Confused With 69, 74
Centroid Algorithm 113
Centroid Technique 30
Class Test 70
Class Test NOT OK 71
Class Test NOT PERFORMED 71
Class Test OK 71
Classes 25
 Assign 9, 83
Clear Selected Regions 87
Cluster Analysis 25, 105
 Performing 31
Clusters 25
Complete Linkage 30, 113
Composing 88
Confidence Band 44
Confidence Level 67
Conformity Data Block 49
Conformity Index 39, 46
 Maximal 46
Conformity Index Limit 43, 45
Conformity Test 39
 Performing 48
 Report 49
 Setup 39

D

Data Preprocessing 10, 61, 107

Data Processing 84
Dendrogram 25, 36, 109
Derivative 85
Detailed Report 94
Diagnosis 110

E

Eigen Vectors 23
Eigenvalues 23, 54
Euclidian Distances 28
Expected Reference 21

F

Factor Spectrum 57
Factorization 23, 28, 53, 88
 Original Spectra 88
Factorization Method 58
File List 114
Fixed Algorithm 67
Frequency Regions 11, 60, 85, 107

G

Group Statistics 91

H

Histogram 28, 109
Hit Quality 19, 29, 51, 61, 87

I

IDENT Analysis 51
IDENT Report 21
Identified As 68
Identity Test 19
 Result Display 102
Identity Test Limit 13
Identity Test Method 20
Identity Test Report 21
Interactive Region Selection 12, 85

L

Labels 36
Load Method 105

M

Main Library 4
Maximum Distance 90
Maximum Hit 90
Mean Distance 67, 91

Median Algorithm 113
Median Technique 30

N

No Reference Defined 104
Normalization to Reprolevel 28, 59, 88
Not Identified 69, 74

O

Only Pairs 113
Only Reference 114
Only Triplets 113

P

Parameters 10, 32
Pearson's Correlation Coefficient 60

R

Reference Spectra 6
Reference Spectrum 31, 51, 106
Regions 85
Report 34, 109
Reprolevel 87
Result Report 94

S

Scaling to First Range 28, 59, 88
Score Coefficients 57
Score Plot 111
Second Derivative 85
Selectivity 95
Selectivity Histogram 95
Selectivity Report 94
Setup Identity Test Method 31, 77
Single Linkage 29, 113
Spectral Distance 25, 51, 60, 62
Spectral Residuals 54
Standard Deviation 67, 90, 91
Standard Method 21, 28, 51, 54, 88
Store Method 17, 100
Sub-Library 4, 81
 Setting 7
Summary Report 94

T

Threshold 21, 67, 74, 90
TRC Data Block 117

U

Uniquely Identified 68, 74

V

Validation 14, 74, 94
 Report 15
Vector Normalization 61, 84

W

Ward's Technique 29, 30, 113
Weight 59, 87
Weighted Average Linkage 30, 113
Whole Matrix 113