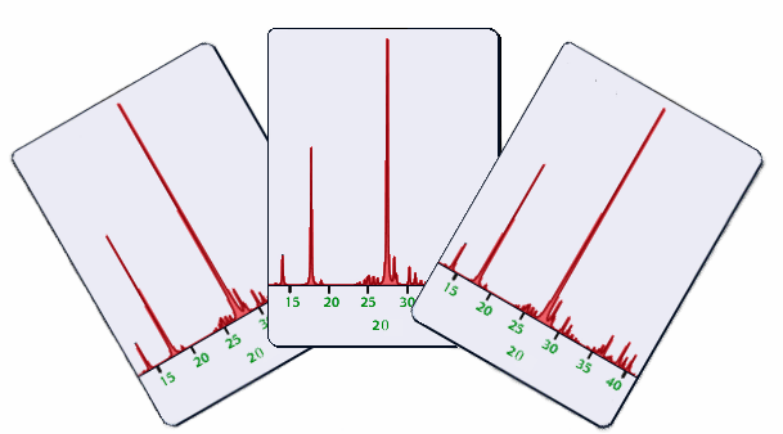
PolySNAP2

High-Throughput Cluster Analysis of Multiple Data Sets



Program Manual and Tutorial

University of Glasgow

Version 2.1
May 2008

Credits

PolySNAP 2:

Systematic Non-parametric Analysis of Patterns

is a product principally of

Gordon Barr, Chris Gilmore, Wei Dong and Jonathan Paisley

of

WestCHEM
The University of Glasgow
Glasgow
Scotland G12 8QQ
United Kingdom

Support

Any problems, comments or questions should be directed to:

Email: snap@chem.gla.ac.uk

Fax: +44 (0)141 330 4419

References

Barr, G., Dong, W. & Gilmore, C.J (2004). *J. Appl. Cryst.* **37**, 658-664.
Gilmore, C.J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 665-668.
Gilmore, C.J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231-242.
Barr, G., Dong, W. & Gilmore, C.J (2004). *J. Appl. Cryst.* **37**, 243-252.
Barr, G., Dong, W. & Gilmore, C.J (2004). *J. Appl. Cryst.* **37**, 635-642.
Barr, G., Dong, W. & Gilmore, C.J (2004). *J. Appl. Cryst.* **37**, 874-882

Documentation Production Notes

Written by G. Barr, with additional material by J. Ferguson, G. Tate and A.Lamarque.
PolySNAP logo by R.Thatcher.
Created using Adobe Framemaker. Set in 12 point Times New Roman. Last modified: 23-5-2008

Disclaimer

Table of Contents

CHAPTER 1

Introduction and Installation

1.1 Program Overview

PolySNAP 2 is a software package designed to match and analyse full profile spectral and other numeric data. The use of the full-profile allows for more flexible and accurate identification of samples, even when data quality is low or preferred orientation effects are significant.

The software provides an easy to use interface to several powerful and novel statistical methods to rank samples in order of their similarity to any other selected sample, allowing unknowns to be quickly identified. In quantitative mode, given a mixture pattern and potential pure phase patterns, it can identify which patterns are in the mixture, and quantify their proportions quickly and easily using a non-Rietveld based approach.

The matching procedure can be automated for computer-controlled high-throughput analysis.An unlimited number of patterns can be pre-screened, and PolySNAP allows for datasets of up to 1,500 patterns with four different datatypes to then be analysed in a single run, and provides highly flexible graphical output to summarise and visualise the results. This highlights any unusual data, and means that time is not wasted looking at the many patterns that behave exactly as expected. It can work with or without the provision of reference patterns, and includes additional features such as an automated report writer and a time/date stamped logfile to assist with audit trail procedures.

1.2 Introduction

PolySNAP can be run either interactively (the default mode when launched by a user) or automatically (the default mode when launched via a command-line).

Four main stages are involved in a standard program run:

1. Import and processing of data files.

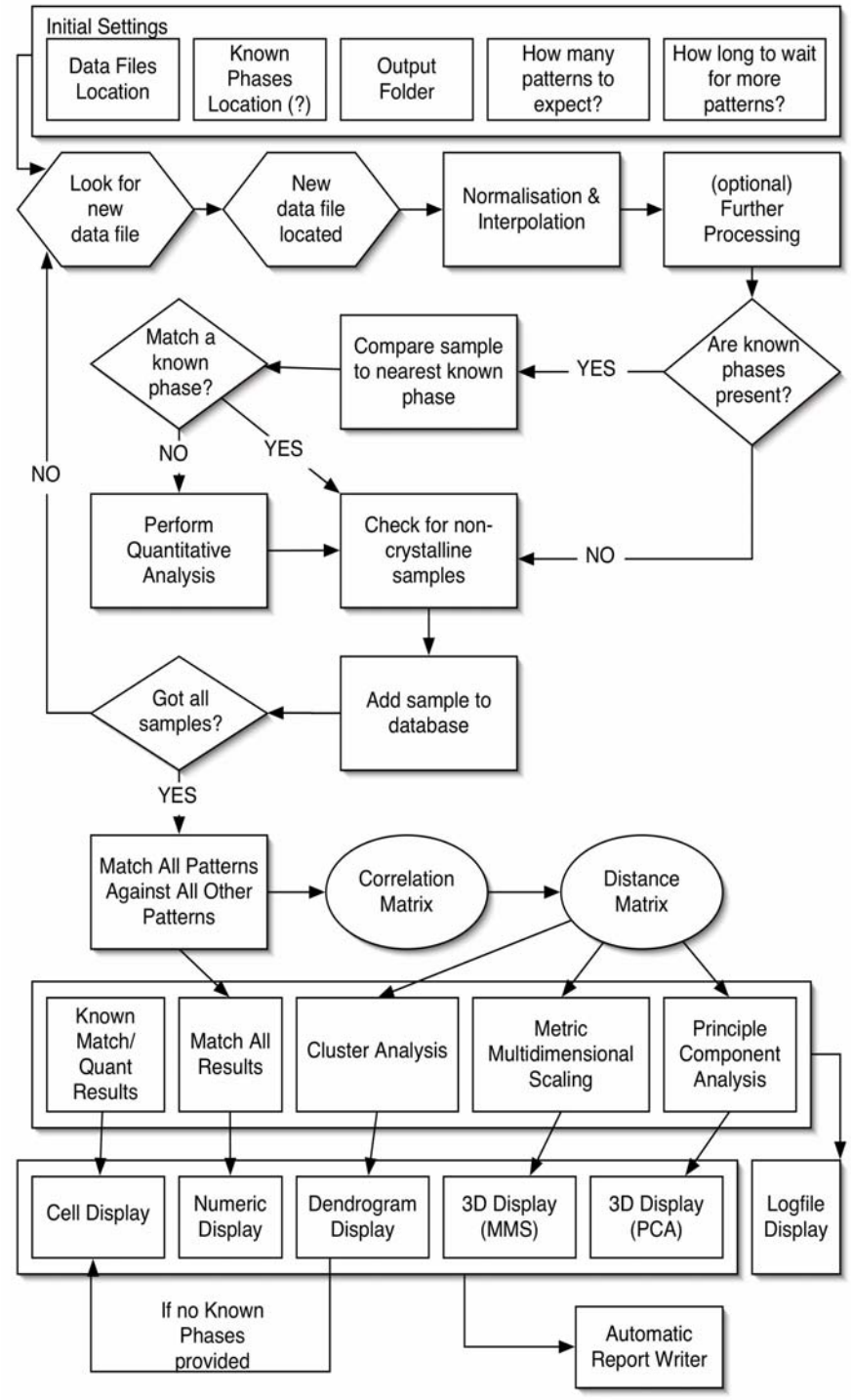
2. Match all data files against all other data files.

3. Perform cluster, quantitative and other analyses.

4. Output results to file and graphically to screen.

A flowchart representation of the main PolySNAP processes is shown overleaf.

1.3 Program Operation Overview



1.4 Program Requirements

PolySNAP requires a modern, high-specification PC running Microsoft Windows XP SP1 or later. Additionally, a monitor with minimum 1024 x 768 resolution at 32 bit colour depth is needed, as is

an active connection to the Internet to validate the software license key.

1.4.1 Note on run-times

On a 2.4Ghz Intel Xeon with 512 Mb RAM, running Windows 2000 Professional SP4, some approximate average program run-times are shown below.

Times are for a single dataset, assuming there are no delays waiting for data, and are measured from launching the program to the results display screen being displayed:

No. of Patterns	Time
50	10 seconds
96	20 seconds
200	68 seconds
400	5 minutes
600	15 minutes
800	30 minutes
1000	1 hour 5 minutes

These times are approximate, and may vary depending on the analysis options selected for a given program run - in particular, allowing for an offset in the matching calculations will greatly add to the run time.

1.5 Installation

If there has been a previous version of the software installed, then it is recommended to install the new version in a different location than the older one; do not install a new version directly over the top.

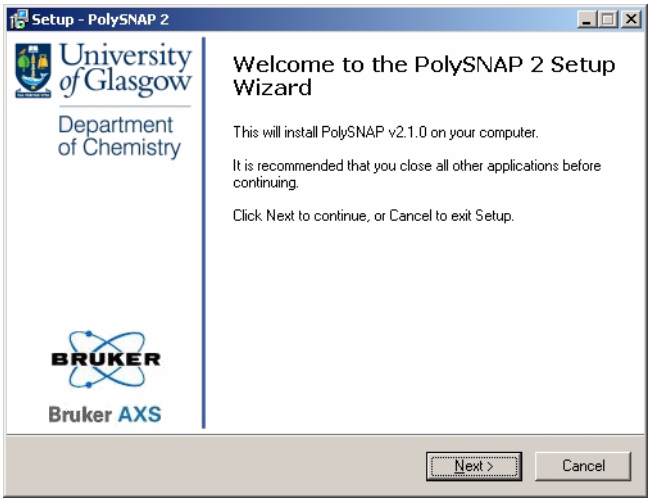
Insert the CD containing the software in the CD-ROM drive of your computer. On the CD (normally drive *D:* on most PCs), open the folder

PolySNAP Install

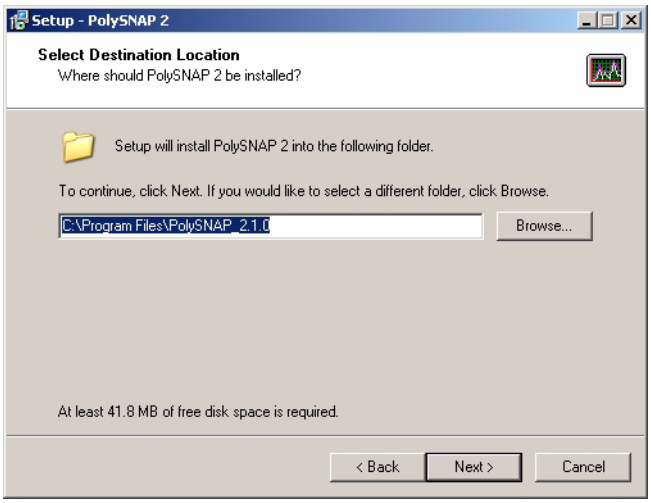
and launch the program *Setup.exe* by double-clicking it.

Note that to install the software a system administrator password will be required.

It will display a welcome window:



Followed by a dialog box allowing the user to control where the program is installed:



The default path,

C:\Program Files\PolySNAP2

should be suitable for most environments, but a different location may be chosen if required by clicking on the *Change Directory* button (please note that running PolySNAP from a remote network drive is not a supported configuration).

The next two screens control if a shortcut to the software is added to the Start Menu and/or the Desktop. The final screen displays a summary of the selected options; click *Install* to start the installation process.

Installation should then proceed automatically, and once completed, the installer will quit. Depending on the version of Windows, a restart may be required at this point; if so the installer will notify that this is the case.

1.6 Launching PolySNAP

Assuming a default installation, the program may be accessed in one of the following two ways:

- Run the shortcut to *PolySNAP2* which the installer will have placed on the desktop by double-clicking its icon.

- From the Start Menu, select the *Programs* sub-menu, followed by the *PolySNAP2 folder*, and then the *PolySNAP2* option. (It is usually installed at the very bottom of the list of programs).

1.7 Registering PolySNAP

The copy of PolySNAP that is now installed needs to be registered before it can be used. A dialog box will appear asking for a Product Key.

If you have purchased a copy of PolySNAP, you will have been provided with a product code of the form:

PSNAPx-xxxx-xxxx-xxxx

Enter this code now (case-sensitive, including dashes), and click OK.

If you do not have such a code, and are wishing to use PolySNAP on a temporary 30-day trial basis, please email *snap@chem.gla.ac.uk* to request a demo license key.

The program will now try to connect to the internet to validate the registration. It will display a message warning you of this, and a progress dialog box will appear letting you know that this is happening.

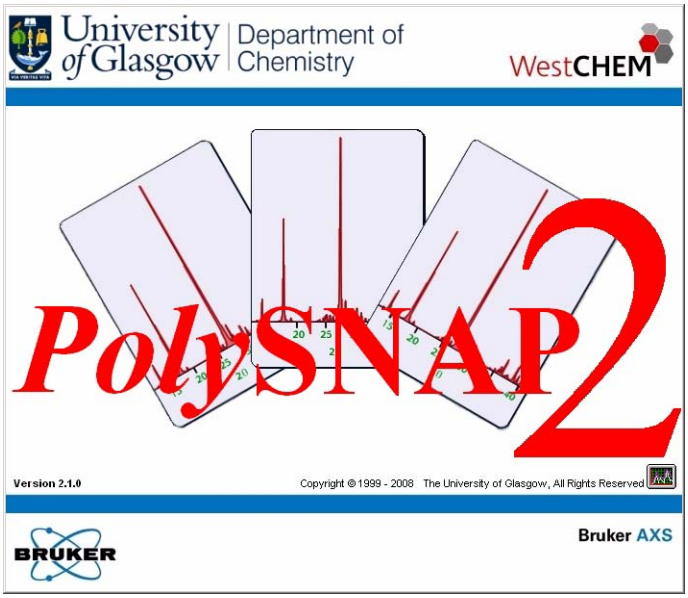
You will then be asked to fill in some registration details. If you have purchased a license, you will only be asked for your name and email address. If you are running in a Trial mode, you will be asked for some more details. Click OK when complete.

NB: your email address and other information will not be passed to any third party and will not be used for any other purpose. Purchased copies, once validated, do not require any further internet access.

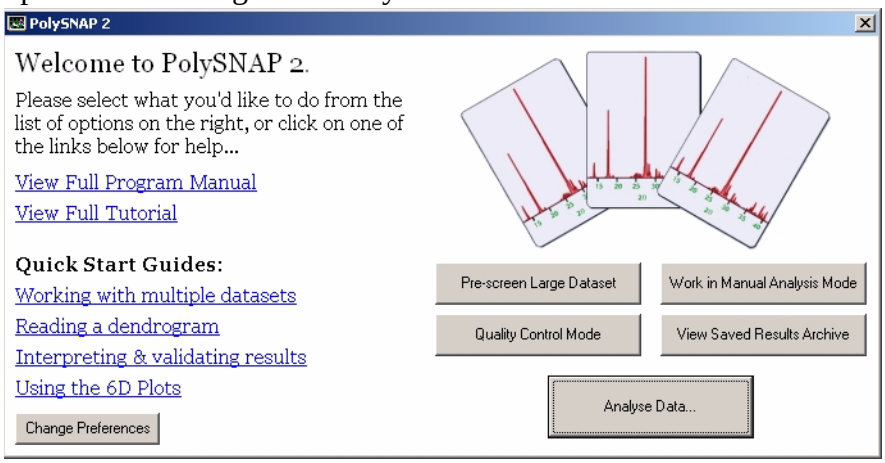
If you have any problems with the process, or are unable to connect to the internet, please contact us directly at

snap@chem.gla.ac.uk

Once this registration process is complete, the program logo screen will appear for a few seconds, before being automatically dismissed:



The main PolySNAP window will then appear, and will by default fill the entire screen. A welcome window will also open providing options for starting to use PolySNAP:



You are now ready to begin working with the software.

1.8 Obtaining Help and Information

1.8.1 View Manual

At any point in the program, select *View Manual* from the *Help* menu. The PDF reader application will be launched, and an on-line version of this manual displayed. It is possible to search the full text for particular keywords of interest.

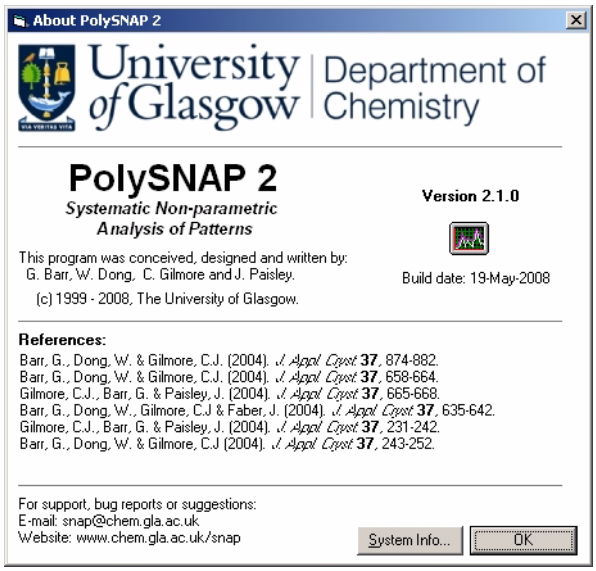
Chapter and sub-headings are displayed on the left as bookmarks. Clicking on one brings up that particular selection.

If a 'File Not Found' or other error message is displayed, ensure that Adobe Acrobat Reader or another PDF viewer is installed on the system.

1.8.2 About PolySNAP

Select *About PolySNAP* from the *Help* menu.

A credits box is displayed:; right click on the References to copy them to the clipboard.



Clicking *System Info* brings up a window giving information on the machine the program is installed on.

1.9 Installation Troubleshooting

Problem

Program registration was not completed, because PolySNAP was

unable to connect to the internet.

Solution
If the registration process could not connect to the internet to validate the install, then it will have displayed an error and a corresponding unique Installation Code for your copy. Please either email this number, along with your Product Code, to **snap@chem.gla.ac.uk** or using another computer visit

www.copyminder.com/activate.php

to obtain an activation code manually.

Problem
A message box displaying an error '51' is displayed during program installation. The program will not launch correctly and may display further errors after this occurs.

Solution
Search the Windows System Folder and its subfolders (usually *C:\WINNT* for Windows 2000, or *C:\Windows* for Windows XP), for copies of the file *CCMOVE32.DLL*. Move this file to a different location (*e.g. C:\temp*), then re-run the program installer.

Problem
When PolySNAP is launched, a Microsoft Office 2000 Installer window keeps repeatedly appearing and cannot easily be dismissed.

Solution
This problem occurs on a system where the currently logged in user has never previously run any of the installed Microsoft Office applications on this machine. Keep clicking *Cancel* until the window finally goes away, and then quit PolySNAP. Launch any installed Office application, *e.g.* Word, and the same installer window will probably appear. This time allow it to run to completion, and click *OK* when requested. Once Word has fully loaded as normal, exit the program and then run PolySNAP again. It should now run normally without any further interruptions from Office.

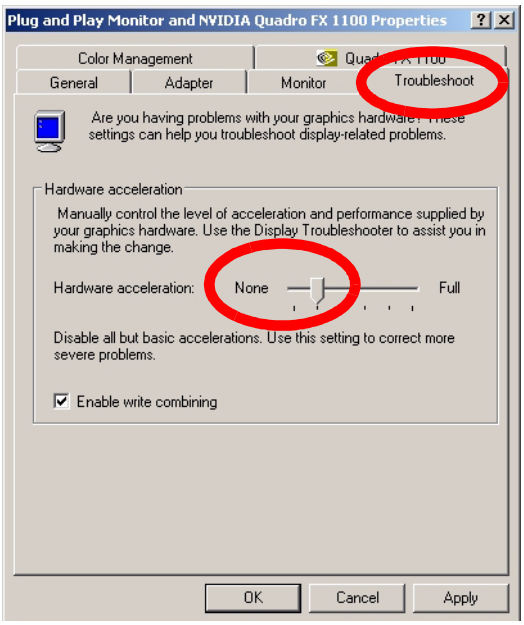
Problem
Problems may occur with the graphics panes – for example, strange artefacts may appear on the results display screen (such as some of the lines in the dendrogram not being visible), the program may freeze when a graphics pane is being interacted with (*e.g.* rotated, zoomed *etc*.) or the program may hang when attempting to first dis-

play the results screen at the end of a program run.

Solution

This problem can occur on systems that have graphics cards that are not 100% compatible with the standard OpenGL libraries. As a workaround, first quit PolySNAP, then go to

Start Menu -> Settings -> Control Panel -> Display..



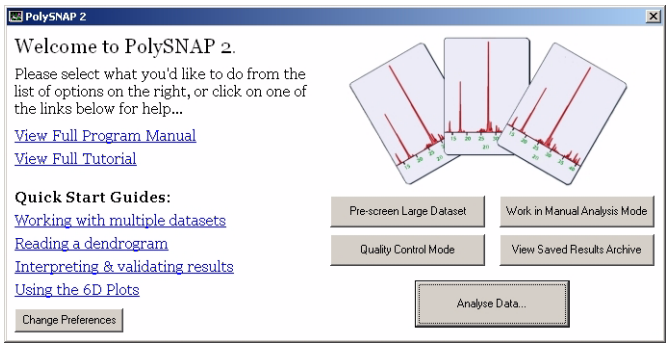
Under the *Settings* tab, click the Advanced button on the bottom right of the pane. In the window that appears, select the *Troubleshooting* tab, and move the Hardware Acceleration slider down to one notch above the far left hand side (“None”). Click OK, then OK again. Restart PolySNAP and see if the problem has been resolved

CHAPTER 2

Using PolySNAP: Automatic Mode

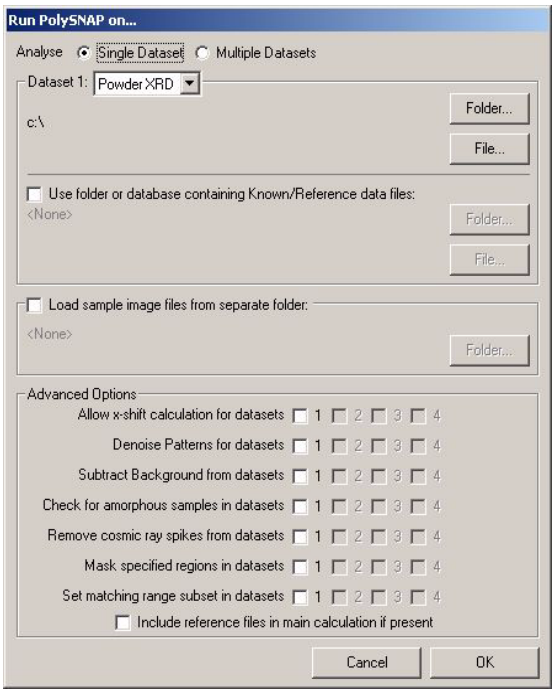
2.1 Starting Automatic Mode

To begin working in automatic mode select *Analyse Data* from the welcome window, or simply pressing *Enter* on the keyboard while the welcome window is open.



Alternatively automatic mode can also be started by selecting *Automatic mode -> Analyse Data* from the *File* menu. This is only available when no other databases or match windows are open.

An input window will open allowing the files containing the sample data to be loaded into the program.

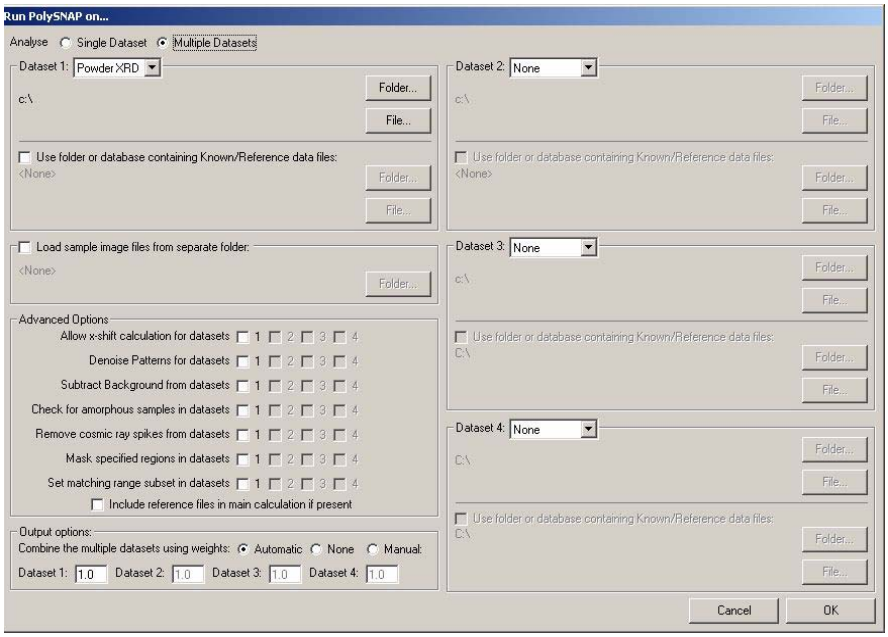


There is a section for inputting the location of the dataset and any known reference files. There is also a separate section where images of the sample can be included. Finally there is a section providing *Advanced Options* to control how the data is processed. All of these features are explained in *Section 2.3.*

2.2 Using Multiple Datasets

By default the input window will appear with space to enter a single dataset. To enter up to four different datasets and run them for analysis together select *Multiple Datasets* from the top of the input window,

This input window will expand and there will be space provided to enter a further three datasets.



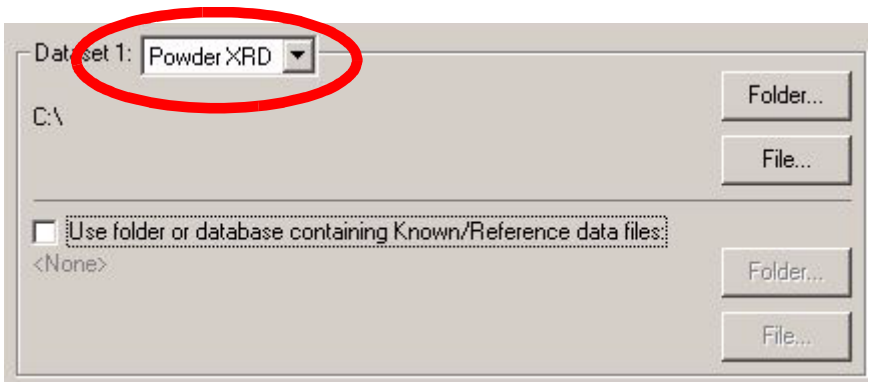
There are now four identical sections allowing input files to be entered. There is also another options section in the bottom left of the display, providing *Output Options*. These are all detailed in Section 2.3.

2.3 Inputting Dataset Locations

The input window for running automatic mode has several different sections which are described below:

2.3.1 Type of dataset

The input files are loaded through the data input sections. These are the same for all datasets.



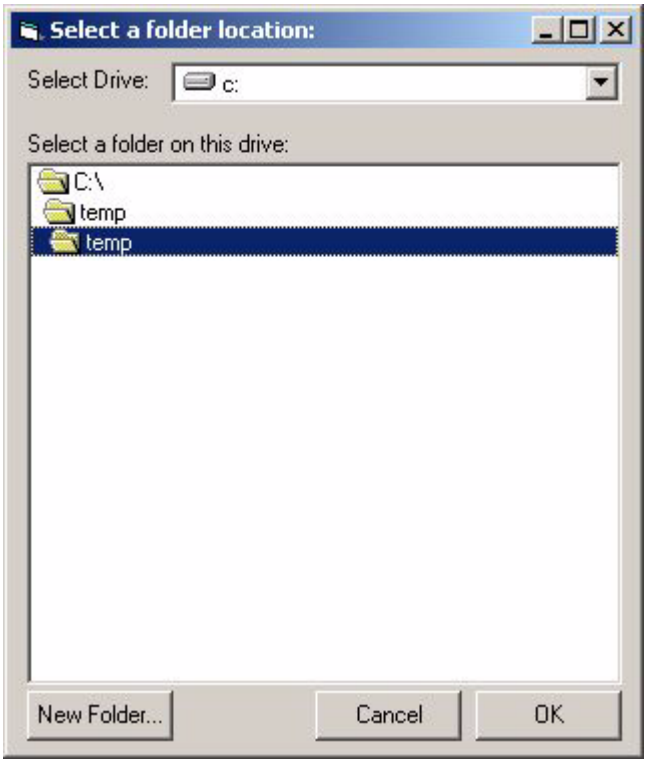
There is a drop down menu which allows the user to select which kind of dataset is going to be input.The options are *Powder XRD* (powder X-ray data), *Raman*, *DSC* (Differential Scanning Calorimetry), *IR* (infra-red spectroscopy), *Numeric* or *Other*. The *Other* option allows the user to include a dataset of a type not specified above.

The same type of dataset can be entered in more than one field. While it is possible to compare different types of data on the same sample, it is also possible to compare different sets of the same type of data on the sample. For example, a series of Raman patterns taken at different times, or a series of IR data taken at different temperatures could be compared.

When entering multiple datasets, the second, third and fourth datasets can be specified as *None* which leaves that section empty. This means the dataset is turned off and will not be used in the analysis. It is not necessary for all sections to include a dataset and any number between one and four datasets can be run.

2.3.2 Location of dataset

Once the type of data has been set the location of the dataset can be input in two ways. If the dataset consists of separate files in a folder then clicking on the *Folder...* button will the following window:



This window allows selection of an entire folder, rather than a single file within the folder.

From the upper selection area any available local or network drives can be selected, and doing so updates the lower region to display all folders contained in that drive. By double-clicking on a folder name, it may be opened, and any folders contained within that folder are then shown.

To select a folder to use as the main folder, double-click on it to open it. Clicking *OK* results in the selected folder path being displayed in that section of the main window.

If the dataset consists of a single file then this can be accessed by clicking on the *File...* button. This opens a standard *Open* window. Navigate to the location of the file and select it by double-clicking on it or by clicking on it once and clicking on *Open*. The file being input can either be a pattern database file (*.par*) or a numeric data file (*.txt*, *.dat*).

The second section contains a check-box to state where reference or known data files are to be included if they exist for the data. To use this option, click on the check-box, which then allows either a folder or file location to be specified as before.

The same process is then repeated for the other datasets that have to be analysed. Four is the maximum number of datasets that can chosen for any one run.

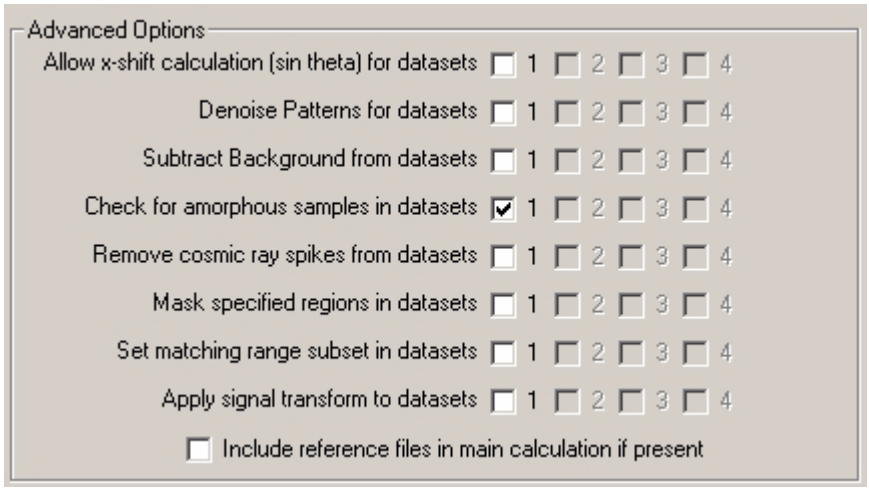
2.3.3 Location of image files

This section allows the user to specify a folder containing image files of the sample.

To select this option click on the checkbox then click on the *Folder...* button and follow the same procedure as detailed before to select the appropriate folder. If image files are provided then these are shown in the results display once analysis is complete. For proper display, they must have the same filenames as the corresponding datafiles, but with a *.jpg* extension.

2.3.4 Advanced Options

At the bottom right of the input window there is a section providing advanced options for the initial analysis.



Some of the check-boxes are initially grey and unselectable to begin with, but become white and clickable once that dataset has been turned on. As an example, if the location of Raman data was provided in the second dataset section, all of the grey checkboxes next to the number 2 that are currently inactive would become activated, like those for section 1.

The options provided are as follows. (More details on how they work are available in Chapter 4). Each of these will only apply to the datasets that they have been selected for:

Allow x-shift calculation for datasets:

This option controls how patterns will be shifted in an attempt to maximise the correlations between them. This can be done in a number of different ways depending on the reasons for the shift. The default value is a sin theta shift; others can be selected in the *Options -> Matching* screen (see Chapter 4).

Denoise patterns for datasets:

Removes noise from a signal using wavelet smoothing.

Subtract Background from datasets:

Removes background signal where present. It is important that the correct data type is specified before using this option, as the type of background subtraction applied is different for Raman datasets, compared to those used for X-ray datasets.

Check for amorphous samples in datasets:

Applicable mainly to powder X-ray diffraction datasets. When selected this option monitors the patterns for any samples that appear to amorphous or non-crystalline. These samples are then labelled as such in the cell display, and can optionally be discarded.

Remove cosmic ray spikes from datasets:

Applicable mainly to Raman datasets. When selected this option monitors the Raman pattern for cosmic ray spikes (characterised by an extreme intensity peak with a very narrow range) and removes them from the pattern.

Mask specific regions in datasets:

This option allows the user to set a sub-region of the pattern to a value of zero. This tool is useful when the pattern contains a spurious peak or unwanted standard, or if for some reason the user wishes to only compare a certain region of the pattern. Up to three separate regions of a pattern can be masked out.

Set matching range subset in datasets:

This option allows the user to select the sub-range within the pattern that will be used for the matching process, disregarding the other sections of the pattern. This is used when there is only a specific region of the pattern that is of interest.

Apply signal transform to datasets:

Selecting this option brings up a dialog box allowing the selection of either a fourier transform, or first or second order derivative, to be applied to the selected spectrum.

The final option, *Include reference files in the calculation if present*, will apply to all datasets where reference files have been provided. This allows the reference files to be plotted in all the graphical displays along with the samples, for a visual comparison of how close the references files are to the sample files. When this option is deselected the reference files are still used to analyse the samples, however are not included in any of the displays.

2.3.5 Output Options

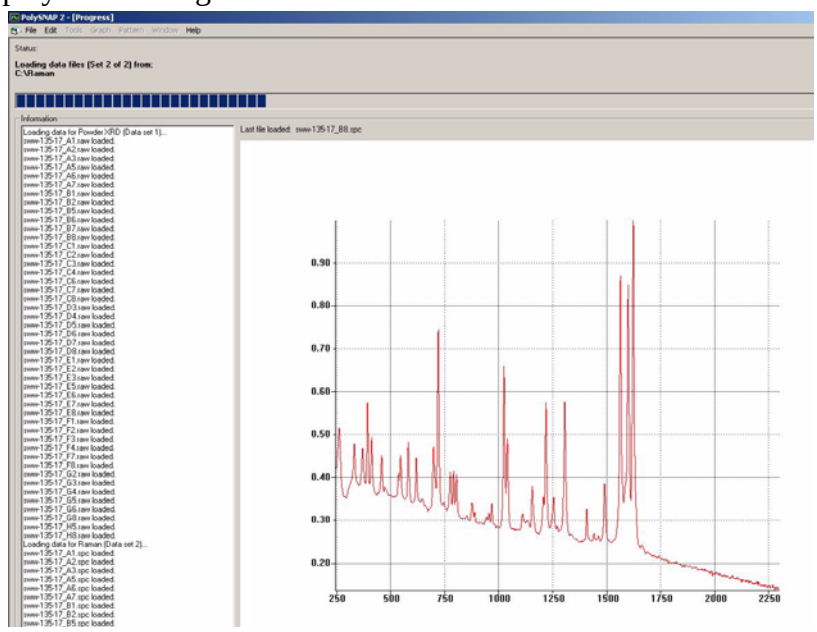
In Multiple Dataset mode, the final section allows weights to be assigned when combining the datasets.

Using the default *Automatic* option allows *PolySNAP* to combine the results in the manner it calculates as being most appropriate. This can be overridden by the user by selecting the *Manual* option and entering numerical values for the weighs manually. Note that weights can only be entered for datasets that have been turned on. Otherwise the field for entering weights in unselectable (as with *Dataset 3* and *Dataset 4* in the image above).

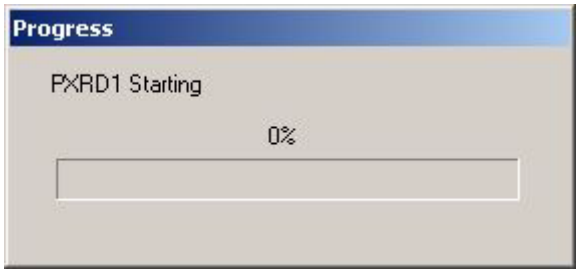
Selecting *None* in the *Output Options* means that the datasets will not be combined and the results will only be presented as individual cases.

2.3.6 Data input

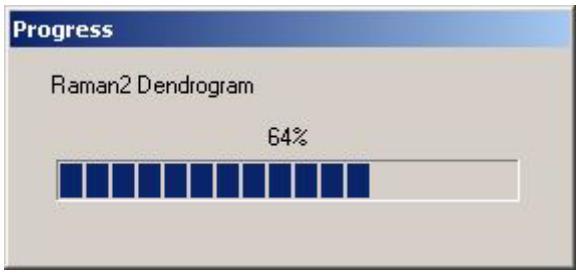
Once all of the appropriate settings have been selected in the input window click *OK* to start the analysis. A progress window will open, with a progress bar running along the top, a list of imported files in the pane on the left-hand side and a large graph of the current file displayed to the right.



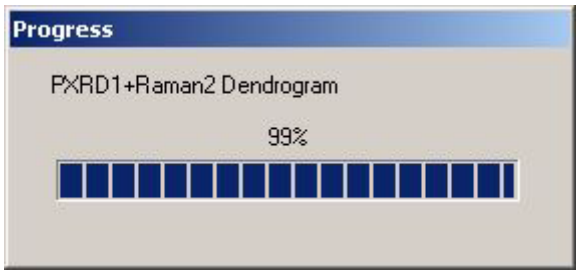
Once all of the files have been imported and processed, a small dialog box will open, displaying the progress of the analysis. To begin the first dataset is analysed on its own.



This will close when the individual analysis of the first dataset is complete and a new window will open for the analysis of the second set.



Once all of the provided datasets have been analysed in this way a new dialog box will open as analysis is carried out by combining the datasets:



This will be performed for every combination of datasets in turn. When analysing more than one dataset simultaneously *PolySNAP*

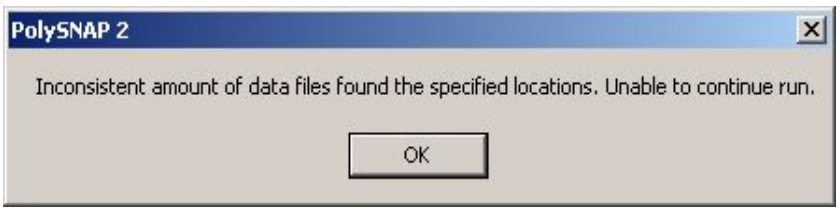
Number of datasets used	Number of individual sets of results
1	1
2	3
3	7
4	15

Once this is complete all progress windows are closed and replaced the results display.

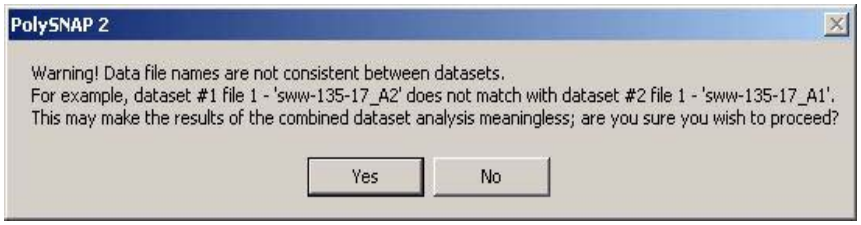
2.3.7 Potential problems

Running automatic analysis on multiple datasets at the same time requires coordination of all the individual datasets in order for the results to make sense. For this reason there are problems that can occur when there are inconsistencies between the datasets.

Each dataset should have the same number of files, which should all have the same filenames. Should there be an inconsistent number of files between datasets then an error message will be displayed and *PolySNAP* will be unable to continue the analysis:



Should there be an equal number of files, but they have been inconsistently named, then a warning message will be displayed:



The user can choose to continue with the analysis, or to abort. If analysis is carried out then the user is warned to treat the combined results with caution as the program can no longer be certain if the correct files have been matched together when creating the combined results. The individual results should be unaffected.

Finally if there are inconsistencies in the image files provided then no warning message will be displayed, as this is not considered a critical error. However as a result image files will only be displayed for the samples that were found in the expected form.

2.4 File formats

PolySNAP 2 can import data files in a variety of formats. These are detailed below. Most types of data that will be analysed will be 2D x-y spectral data, but other types of raw numeric data can also be imported.

2.4.1 Spectral Data (e.g. Powder XRD, Raman, etc.)

2.4.1.1 Text-based File Formats

Tab-delimited Text files (.txt or *.prn)*

These should be standard ASCII text files. The angle should be the x data, and the corresponding intensities the y data. The x-y data must be in the format:

x1	y1
x2	y2
.	.
.	.

etc.

The fields must be tab-delimited.

Note that the data-step size (x2 - x1) *must* ***be constant.***

If the first line in the file begins with either a '#' or the string 'ID: ', the rest of that line is used as the 'Pattern Name' for the pattern.

Any subsequent lines beginning with the '#' character are ignored.

Comma-separated value files (.csv)*

These should be standard ASCII CSV files, containing the x-y data in the format:

x1,y1
x2,y2
.,.
.,.

etc.

Note that the data-step size (x2 - x1) *must* ***be constant.***

If the first line in the file begins with either a '#' or the string 'ID: ', the rest of that line is used as the 'Chemical Name' for the pattern.

Any subsequent lines beginning with the '#' character are ignored.

Powder CIF (Crystallographic Information Format) files (.cif)*

The program contains a CIF-format translator that reads standard Powder CIF files. Although the only data necessary for SNAP is the

x-y intensity data, the rest of the CIF information - such as chemical names and formulae, author names and addresses *etc*. are also retained in the database for reference purposes, and may be viewed from the Pattern Editor window. Additionally, unit cell dimensions and contents are read in if present for use in Quantitative Analysis mode.

For more information on CIF format files, see the IUCr website at:

*http://www.iucr.org/iucr-top/cif/pd/index.htm*l

MDI ASCII format files can also be read in (*.*mdi*).

2.4.1.2 Binary File Formats

Bruker RAW Format (PXRD) (.raw)*

The program can import data from Bruker RAW format. There are several different types of RAW format; PolySNAP can import all versions up to Version 4. Multi-range files can be used, but only a single range is imported from them.

Although the only data necessary for PolySNAP is the x-y intensity data, much of the rest of the information stored in the file - such as chemical names and formulae, author names and addresses etc. are also retained in the database for reference purposes, and may be viewed from the Pattern Editor window.

PolySNAP Pattern files (.pat)*

Patterns in PolySNAP databases can be individually exported in the program's own *.pat* format. This format contains the original x-y raw data, any processed data - *e.g.* the profile after noise removal, marked peak positions and any other data fields present. Once saved as separate files, they can then be re-imported to the same or a different database at a later date.

Bruker OPUS Raman (*.1, *,2) and *Thermo Scientific SPC* (*.spc) files can also be imported.

2.4.2 Raw Numeric Data Files

2.4.2.1 Numeric Data

Raw numeric data should be in a format where there each line corresponds to one sample. There can be multiple entries on each line

for different measurements or variables corresponding to that sample. Different entries should be tab-delimited. Optionally, a header line with variable label names, and an initial column consisting of sample labels can be used.

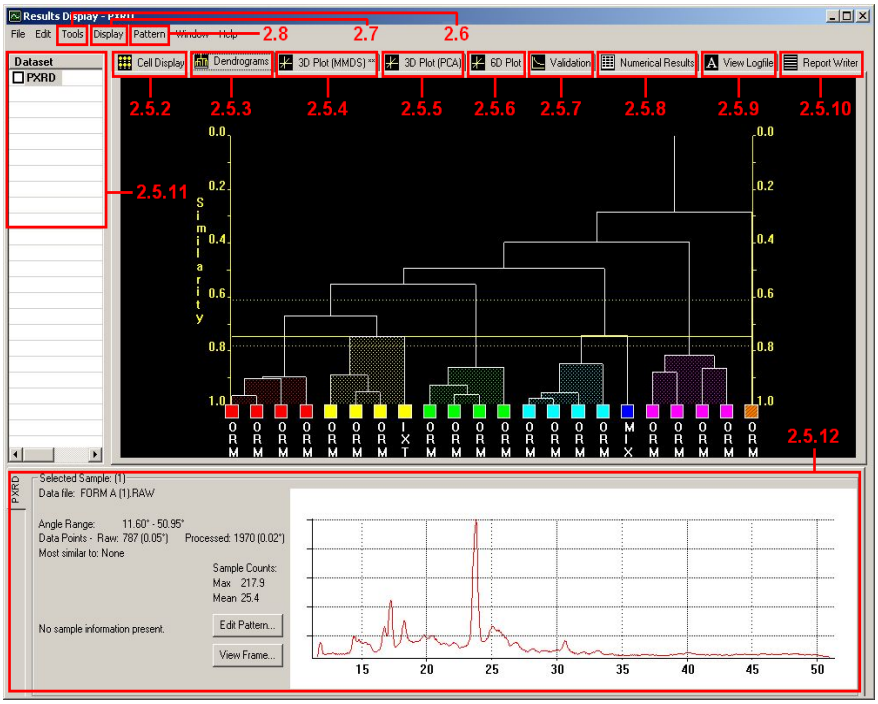
2.4.2.2 Correlation Matrices

A square, n x n correlation matrix, in tab-delimited ASCII format, with a diagonal of 1.0, can be used as input. Optionally, a separate file containing labels can also be imported along with it to make identification of entries easier. This should contain one label entry per line, with n lines for an n x n matrix. Spaces are not permitted in label names.

2.5 Results display

A list of available results screens is shown on the left hand side (Dataset1, Dataset 2, ... etc.) The main part of the results screen shows the results for the currently selected dataset.

When first opening the results display the user will be met with many different features. A quick guide to where to find the relevant section in this manual for some of the main features is given below:

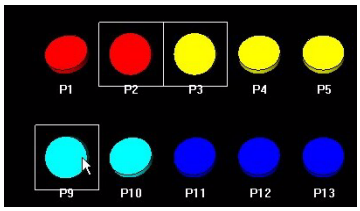


2.5.1 General features

The various graphical display panes - the Cell Display, Dendrogram, Screen and two 3D plots - all share many similarities in their options and controls, and are hence initially described together here.

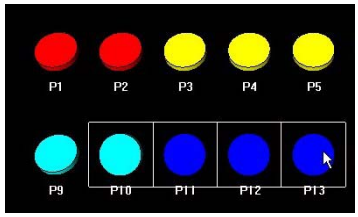
2.5.1.1 Multiple Selection

Multiple non-contiguous selection is achieved by clicking on multiple patterns with the *Control* key held down on the keyboard, for example:



Individual patterns can be de-selected in a similar manner, and their profiles will be removed from the graph.

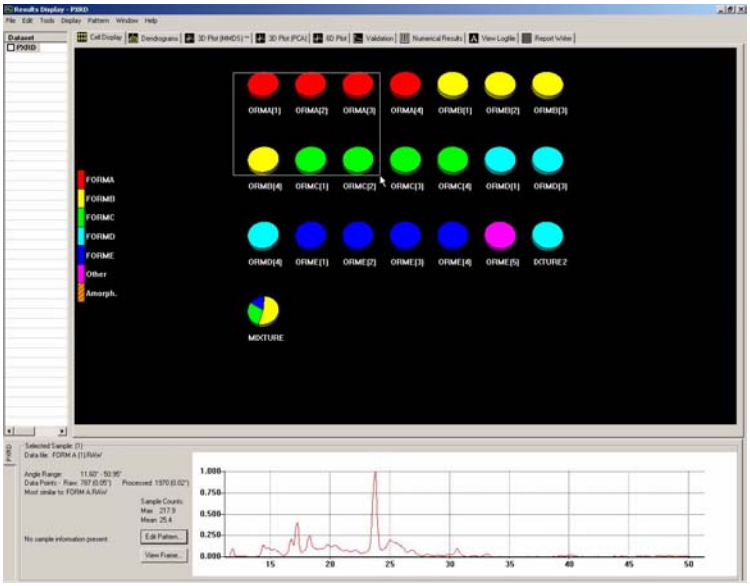
Alternatively, a continuous number of consecutively displayed patterns may be selected by holding the *Shift* key down and clicking on the first and then last pattern in the desired range, for example:



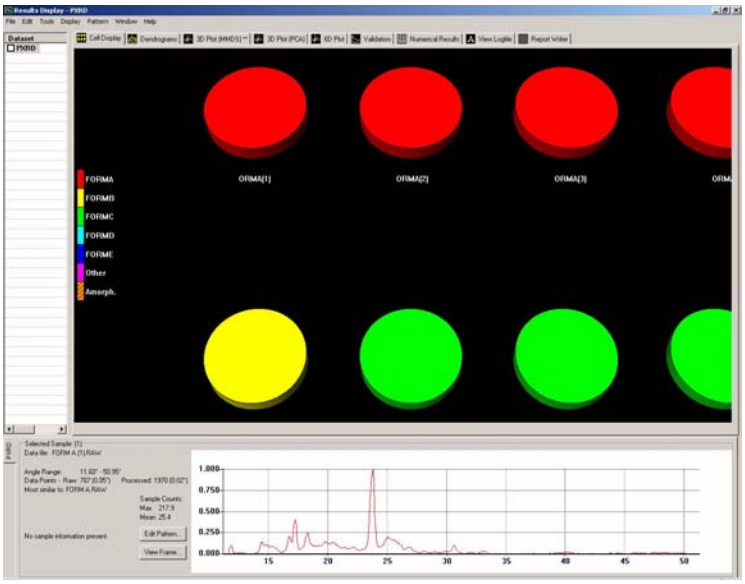
When multiple patterns are plotted, selecting *Toggle Mode* from the right-click menu for the graph pane displays all but the most recently selected pattern in the same colour, to allow easier comparison between one particular pattern and several others.

2.5.1.2 Display Controls Common to All Modes

To zoom in to a region of a graphical display, with the left mouse button held down, drag a rectangle over the region you wish to zoom:



The screen will then redraw with the contents of the rectangle filling the display area:



To move the contents of the display window, for example to move the contents up to see more results than will fit in the window by default, hold down the *Alt* key on the keyboard, and drag the mouse in the desired direction of movement.

2.5.1.3 Additional General Options

Right clicking on the graphics pane causes the following menu to appear:

Reset View - This feature will return to the original view of the display if it has been moved or zoom has been activated.

Zoom In - Will zoom in on the centre area of the current display.

Zoom Out - Will zoom out from the centre of the current display.

Toggle Mode - This feature switches between viewing the patterns as Pie Charts or Stacks when in Cell Display mode, or between the full and simplified views of the Dendrogram in that mode.

Centre Selection - The currently selected item will be centred in the display.

Deselect all - Any patterns that have been highlighted in the display will be de-selected with this option.

Show Toolbar - The optional toolbar at the top of the graphics pane can be hidden or shown with this feature. The toolbar provides access to most of the options available through the right-click menu; for a full description see Section 2.5.1.4.

Print - The standard Windows print dialog box will appear, allowing the current graphics display region to be sent to a printer.

Copy - The whole of the current graphics display will be copied to the clipboard, and then can be pasted into either the Report Writer pane of PolySNAP, or any other standard Windows program - for example, *Microsoft Word.*

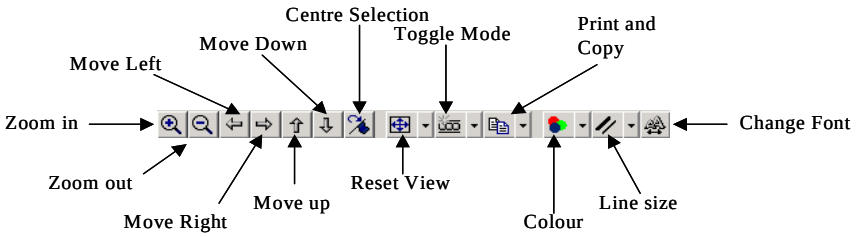
Copy selection - This option allows smaller specific regions of the graphic display to be copied. Click on *Copy Selection,* and then drag

a rectangle over the area to be copied: This area only will now be copied to the clipboard, and can then be pasted in elsewhere as required.

Other mode-specific display options (e.g. *Show Grid/Show Labels/ Show MRP Marks/Find Item/Objects Colour* are discussed under the relevant display section).

2.5.1.4 The Toolbar

The basic functions of the toolbar are illustrated below:



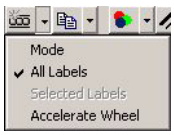
Reset View - This feature will return to the original view settings of the display.

Clicking on the small arrow next to the *Reset View* button allows access to the options for both *Reset View* and *Deselect All.*

Zoom In - Will zoom in on the centre area of the current display.

Zoom Out - Will zoom out from the centre of the current display.

Toggle Mode - This feature switches between viewing the results as Pie Charts or Stacks for Cell Display Mode.



Clicking on the small arrow next to the *Toggle Mode* button allows access to the options for use on graphics panes where there is the option to show all labels or, in cases where a series of objects have been highlighted, only those which are selected. The *Accelerate Wheel* option can also be accessed from this menu.

Centre Selection - The currently selected item will be centred in the display.

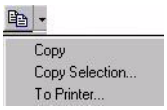
Move Left - The current display will move one unit to the left.

Move Right - The current display will move one unit to the right.

Move Up - The current display will move up one unit.

Move Down - the current display will move down one unit.

Print - The standard Windows print dialog box will appear, allowing the current graphics display region to be sent to a printer.



Clicking on the small arrow next to the *Print* button can be further used to access the *Copy* and *Copy Selection* options described above.

Line size - Clicking on the arrow next to the *Line Size* button opens a pull-down menu with a series of options for line size numbered 1 to 5, with 1 being the thinnest and 5 being the thickest. A dot appears next to the thickness currently selected. The line size options changes the thickness of lines on the graphical displays which may be useful when preparing images for publication.

Colour - This option allows the colours of both the foreground and background to be altered to suit the user's requirements. Note that these will be reset on the next launch of *PolySNAP*:

> *Background*: The default colour is black, simply click on the down arrow next to the colour box and select *Background*:

This will now display the colour palette from which a colour can be chosen.

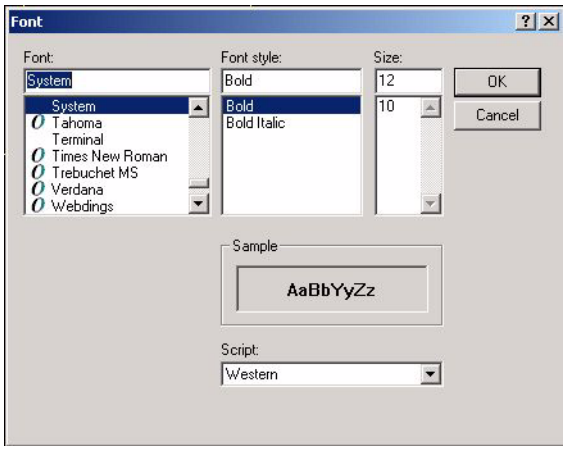
The colour selected here will then be applied to the background of the current view. Note that it may be useful to change the colour of the background to white before printing.

Foreground: The foreground selection is performed in the exact same way as the background, only this time selecting *Foreground* from the options presented:

The foreground option will alter the colour of the text or axes used in the current display.

The *Axes* option allows the user to change the colour of the axes on displays such as the Dendrogram and 3D plots where the *Objects* option opens a new window, which is described in the 3D plot section where it most features.

Font - Finally clicking on the *Font* button will open a standard font options dialog box:



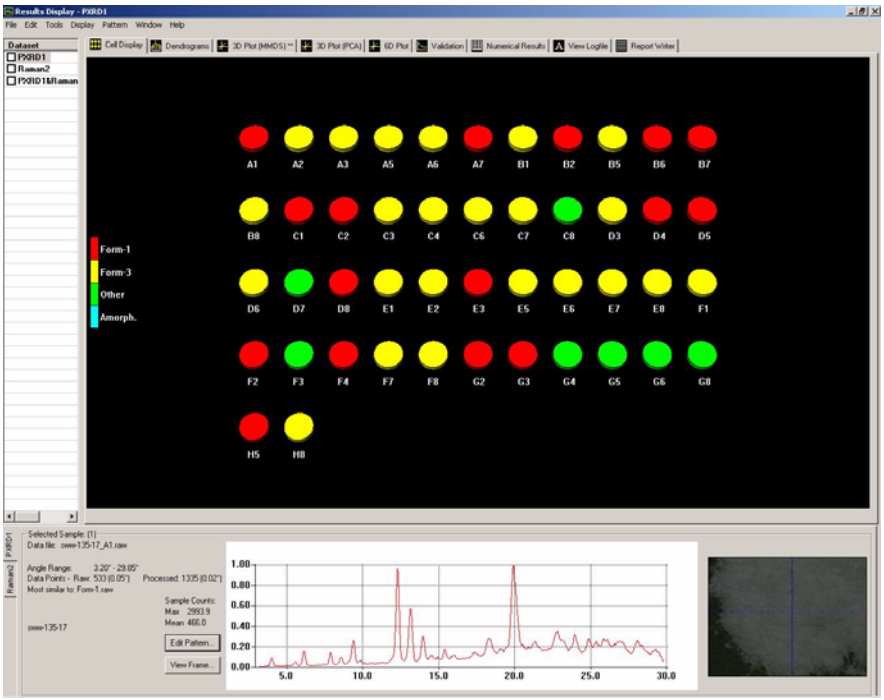
This can be used to select the format of the text that will appear on axes labels, plot headers and pattern labels.

2.5.1.5 The Display Menu

The display menu contains options that control how the results are displayed in all display modes. These options are described in Section 2.7.

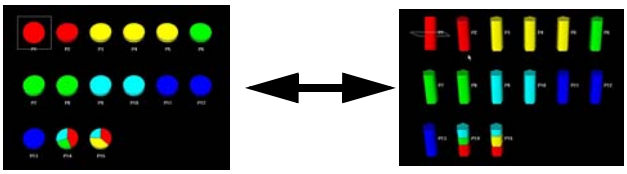
2.5.2 Cell display

The results displays default to the cell display.



This comprises of a pane in which the patterns loaded are each represented by an individual pie chart. Each pie chart is colour coordinated to group together similar patterns.

The cell display can be shown in either of two modes - as standard pie charts (the default), or as 'stacks':

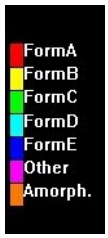


To switch between the two view modes, right-click on the display, and select *Toggle Mode* from the resulting menu.

Depending on the presence or absence of known phases for a given run, the colour-coding is obtained from two different sources.

2.5.2.1 If known phases are available...

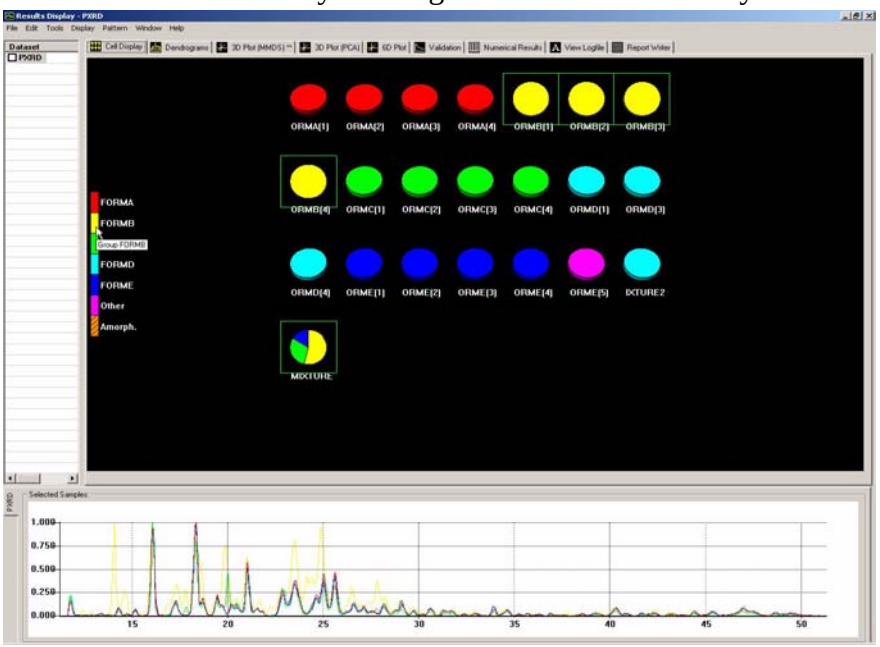
When a database of known phases were provided to compare the new samples to, the 'key' on the left hand side of the display shows a list of those known phases, with the labels shown generated from the relevant pattern filenames:



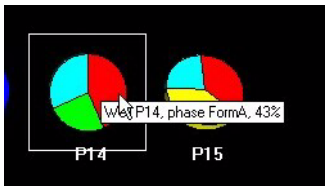
Each known phase has a unique colour assignment; and in addition colours are shown for patterns considered to be either non-crystalline/amorphous (*Amorph.*) or unlike any other known pattern provided (*Other*).

Each individual pattern that corresponds to a known phase (*i.e.* one that gave good matching statistics when being compared to a known phase) is given the same colour as that known phase - for example, in the screenshot above, the patterns that matched well to Form B are all

the same colour of yellow. All of the patterns that match to Form B can be selected at once by clicking on the colour in the key:

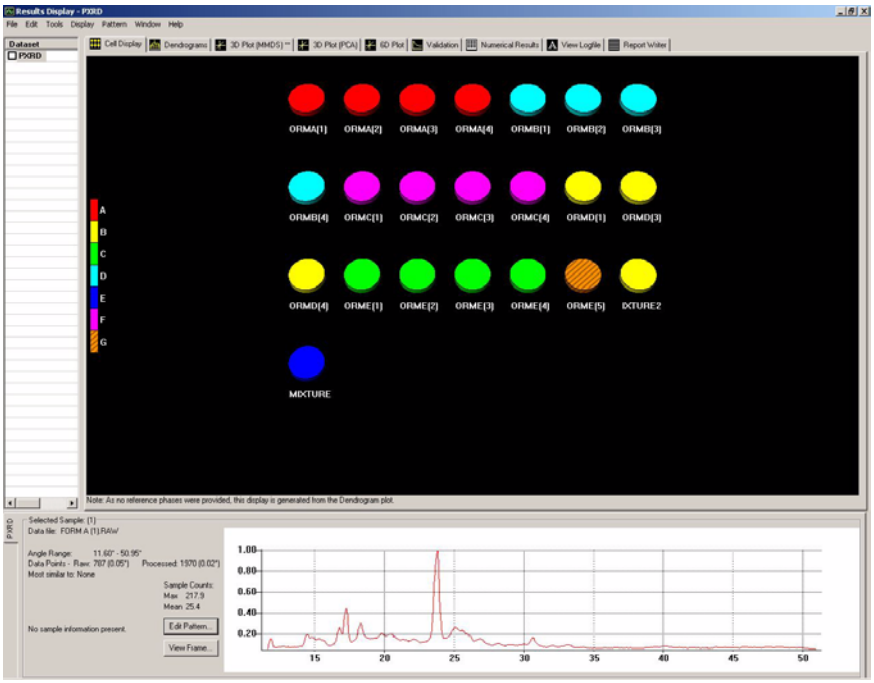


A pie that contains multiple colours (*e.g.* sample MIXTURE above) represents a pattern which is thought to be a mixture of two or more of the known patterns, and the colours within the pie-chart again correspond to the phases thought to comprise it. Allowing the cursor to hover over a particular component of a mixture brings up a tooltip describing which phase and in what amount



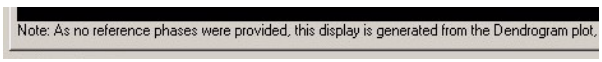
2.5.2.2 If no known phases are available...

On the other hand, if no known phases are available what *appear* to be similar results are displayed:



However, there are subtle and very important differences. The colour of the pies obviously no longer correspond to those of known phases, the colours are now merely representative of patterns which are similar to each other. In the example above Pies 1, 2, 3 and 4 are similar to each other as are 5, 6, 7 and 8. Pattern 21 is dissimilar to all the other samples, as it is a colour not shared by any other samples.

This information is generated from the cluster analysis results as presented in the dendrogram view (see the next section). As a result of this, a warning message is displayed that the results shown are no longer determined separately, but are merely clustering results presented differently:



Because clustering results are being used, the groups shown are entirely dependant on the cut-level used on the dendrogram display. The cut-level is discussed in more detail in Section , but it is important to note that when no known phases are present, altering the cut-level or otherwise editing the dendrogram display will cause the cell display colours to be altered and updated accordingly.

This display mode can also be accessed when known phases are present, by means of the *Show Pseudo Cell Display* option in the *Display* menu.

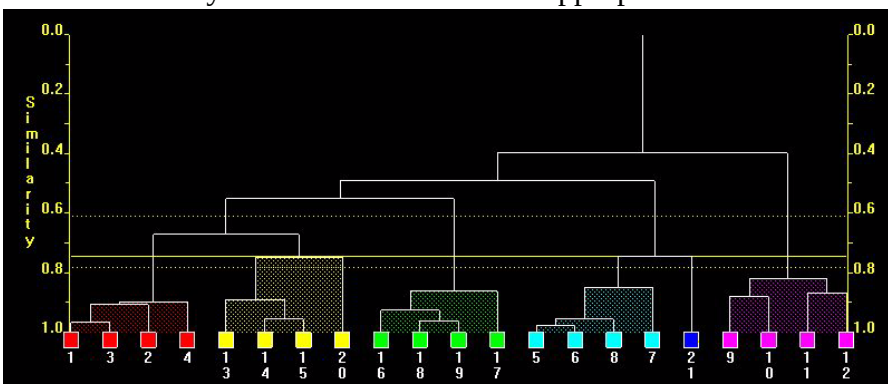
Note that once a particular well is selected, the wells on either side can be selected in turn by means of the left and right arrow keys, thus allowing for quickly scanning through multiple patterns.

2.5.3 Dendrogram

The dendrogram display can be accessed by clicking on the Dendrogram tab along the top of the display window.

The dendrogram provides a visual means to display the results of the hierarchical method of data classification using cluster analysis. The dendrogram itself takes the form of a tree-diagram in which each single terminal branch is representative of a single object (in this case an individual pattern from the data input).

The initial cut point is set by the program, and is shown by the yellow horizontal line. Upper and lower confidence limits on this cut-level are shown with yellow dotted lines where appropriate.



Each pattern is numbered along the bottom axis of the dendrogram. Each number is the same as in the other displays - for example, number 1 on the dendrogram is the same pattern as number 1 in the cell display.

Each pattern can be selected by clicking on the box above its number. When a pattern is selected, the sample information along with its pattern profile is displayed in the bottom half of the display window. Multiple patterns can be selected by holding down the *Control* key and clicking on different patterns. A series of consecutive patterns can be quickly selecting by clicking on the first pattern then holding the *Shift* key and clicking on the last pattern. This selects all the patterns in the range in one step.

Note that once a particular pattern is selected, the patterns on either side can be selected in turn by means of the left and right arrow keys, thus allowing for quickly scanning through multiple patterns.

The view of the dendrogram can be zoomed in on by dragging a rectangle over the relevant area with the left-hand mouse button down as with the graph and cell displays.

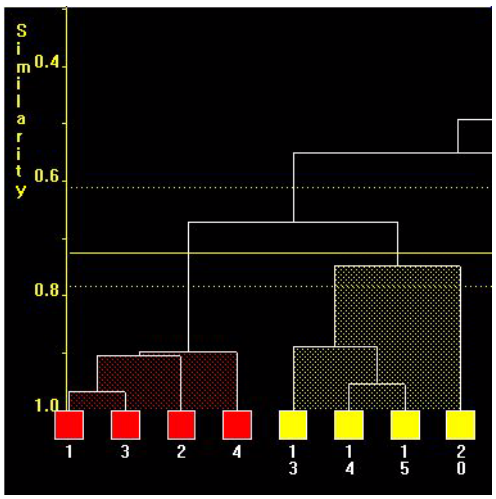
The position of the dendrogram position on the horizontal axis can be altered by holding down the *Alt* key and moving the mouse left or right. This can be useful if the zoom has been used and the whole tree no longer fits on the screen at one time.

Patterns are joined together by a series of lines. The further up the similarity axis (y-axis) the patterns are joined, the less similar they are. Therefore, in the screenshot above, patterns 1 and 3 are joined at a high level of similarity (nearly 1.0), and are therefore very similar, whereas patterns 1 and 12 are not joined until a similarity of less than 0.4, indicating a large difference between them.

Given the calculated similarity between patterns, it is then possible to categorise similar patterns as belonging to the same cluster. This is done by drawing a horizontal line across the display at a given similarity level - this is called the cut-level.

The optimum cut-level is determined by PolySNAP using a combination of several different techniques in order to determine the number of clusters that statistically best represents the data given. These techniques include principle component analysis, metric multidimensional scaling, the C-H test, gamma statistics, *etc*.

The cut-level is then drawn on the dendrogram, and different patterns which are grouped together below this line are considered to be similar enough to be thought of as being in the same cluster:



In the screenshot above, the horizontal cut-point is set at around 0.72, and is therefore considering patterns 1, 3, 2 and 4 to be in one cluster, whereas 13,14, 15 and 20 are in a separate and distinct one.

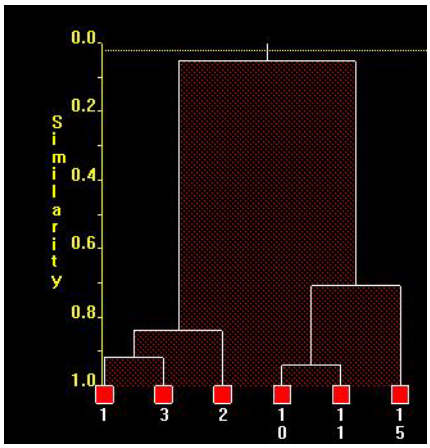
The different clusters are colour coded, and if no known phases are present, these dendrogram results are used to generate a pseudo-cell display (see Section 2.5.2). In this case the colours used here will correspond to the colours in that display. Optionally, the same colour-coding can be used to help interpret the results in the various 3D plots.

The *Toggle Mode* option, accessed through the right-click menu, redraws the dendrogram with only the first, last and middle patterns of each cluster shown. This allows easier interpretation of a crowded display when many patterns are being analysed. Note that while showing the simplified dendrogram, the cut-level and other modifications cannot be made. A label is displayed on the top-left of the display to indicate that the display is in simplified mode.

It is possible to toggle the display of the axes on and off using the *Show Axes* option, accessed through the right-click menu.

2.5.3.1 Changing the Cut-level

If the program-calculated cut-level is not considered to be correct, it can be overridden by the user. This is done by holding down the *Control* key and left mouse button, while dragging the mouse up or down. The cut-line on the dendrogram display will move, and the cluster colouring will update in real time. For example, moving the cut-level up to around 0.1 results in the following:



Note that the colour assignations have updated accordingly. If the user chooses to retain this change when closing the dendrogram display or switching to another display pane, then the pseudo-cell display will be updated accordingly, and the modification noted in

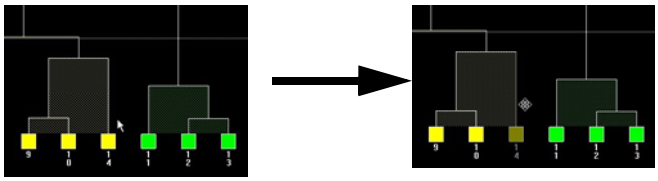
the program logfile. If the change is not retained, the previous value is kept.

[In addition to this method, if a mouse with a scroll wheel is used, it is possible to just click once in the display area with the left mouse button, and then move the mouse wheel up or down to move the cut-level indicator.]

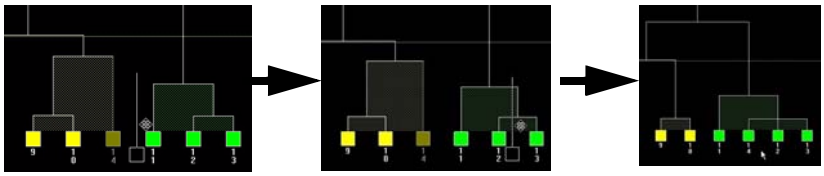
2.5.3.2 Manually changing the contents of clusters

PolySNAP makes its assignation of the contents of clusters using a combination of powerful statistical techniques. It should not normally be necessary to override its results, but for the occasions when this is necessary, it is possible to reassign either a single pattern or group of patterns manually.

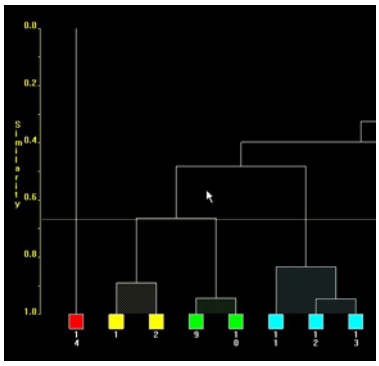
To do this, hold down the *Shift* key and click and hold down on the vertical line attached to the pattern or sub-cluster required to be moved:



Continuing to hold both the *Shift* key and mouse button, drag the unit to the desired location. To add it to an existing cluster, release the mouse when the cursor is over the area the sample is to be added to:



To create an entirely new, separate cluster, drag it to an empty space between existing clusters:



To cancel a drag operation part-way through, it is only necessary to release the *Shift* key.

A one-step *Undo* is available for this method of altering the dendrogram, if for example a user changes their mind, or a cluster is incorrectly joined. To do this, right click on the dendrogram, and select *Undo* from the pop-up menu. Note that only the most recent operation can be undone.

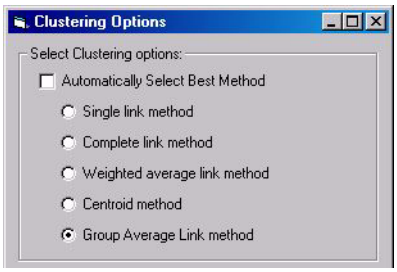
The only way to be able to undo multiple operations is to choose not to save the changes to the dendrogram when closing or switching the window; that way the original dendrogram will be retained and shown the next time the results are examined.

When changes to the dendrogram are saved, the changes and the resulting new clusters and corresponding component patterns are listed in the program logfile.

2.5.3.3 Changing the Clustering Method

PolySNAP can be set to select what it calculates to be the most appropriate clustering method for a given problem. The user can also choose to re-run just the clustering part of the analysis to experiment with how different individual clustering methods would affect the results.

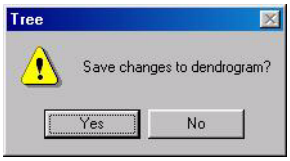
Select *Change Cluster Method* from the *Tools* menu. An options dialog box will appear:



To select an individual cluster method, deselect the *Select Cluster Method Automatically* checkbox, choose from the seven options then presented, and click *OK*. The recalculation process may take several minutes on larger data sets.

2.5.3.4 Saving and Reverting Changes

When another tab is clicked on after the dendrogram has been altered in one of the ways described above, the following message is displayed:



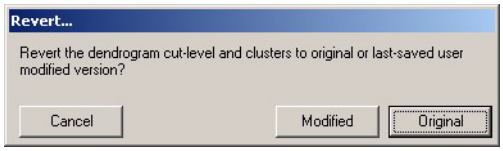
Selecting *No* will cause any changes to be discarded, and the previous version of the dendrogram, prior to any manual changes, will be retained.

Selecting *Yes* causes the changes made to the dendrogram to be kept, and also be recorded in the program logfile. Changes to the dendrogram can also be saved at any point using the *Save Modified Tree* option in the right-click menu.

2.5.3.5 Reverting to a previous dendrogram

Because the program retains earlier saved versions of the dendrogram, it is possible to revert to them if required at a later stage. This function is accessed through the *Tools* menu option *Undo dendrogram modifications...* Selecting this option brings up the following dialog box.

If more than one set of changes to the dendrogram have been saved, the user is offered the choice between the original, program generated tree, and the most-recently saved previously modified version of the dendrogram:



Selecting *Cancel* retains the current version with no changes. Selecting *Original* reverts to the original version. Selecting *Modified* causes the current, modified dendrogram to be replaced with the most recently modified saved changes.

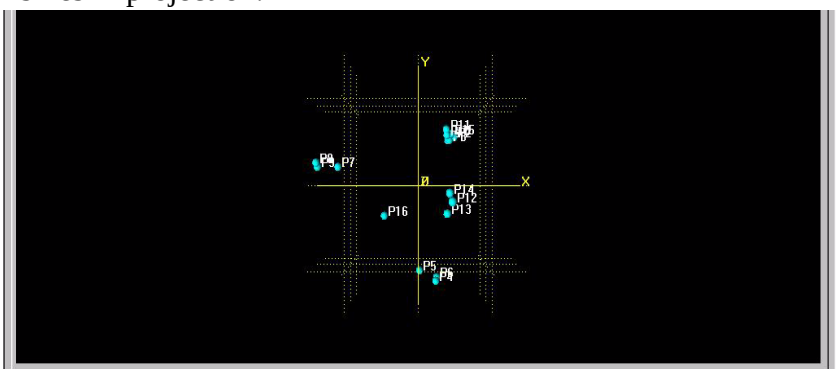
2.5.4 3D Plot (MMDS)

This option displays the results of metric multidimensional scaling which make use of distances between objects calculated from the correlation matrix generated by the matching process to produce a

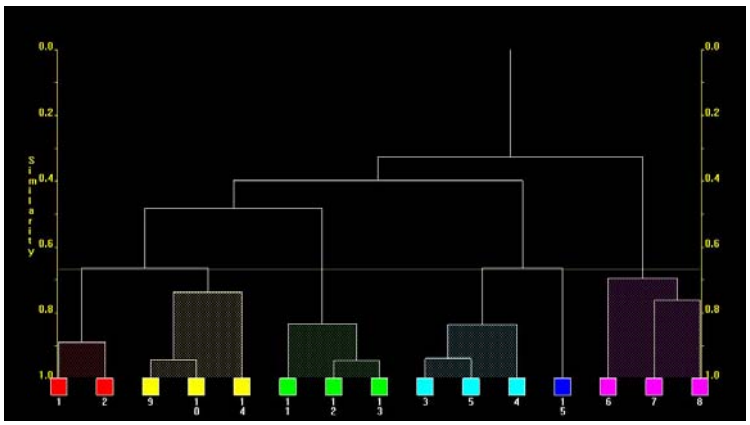
three-dimensional spatial representation of the samples. Each point which appears on this spatial representation corresponds to one of the patterns. The closer two points appear on the plot, the more similar the patterns are, and the more different an object is to another the further apart they will appear. Therefore groups of similar patterns appear to cluster together.

The multidimensional scaling performed is based on calculated proximities rather than observations. First for *n* patterns, we generate an *(n x n)* distance matrix *D* based on dissimilarities, δ_{rs} *= 1,2,....,n*, computed from the correlation matrix. Each object is compared against itself and every other object. The result of an object being paired against itself gives a dissimilarity of zero, which corresponds to the diagonal of the matrix. The goal of this method is to derive a set of underlying dimensions, with co-ordinates that should create a Euclidean distance matrix, which in turn should be the same or very close to the δ_{rs} of the original *D.*

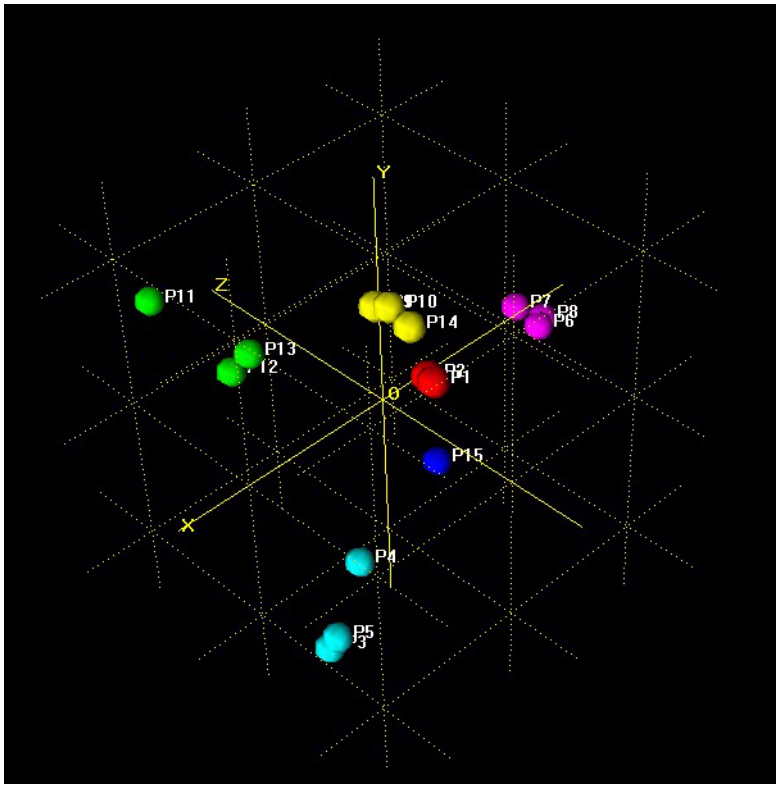
The initial view of the 3D plot shows only the X and Y axis - the Z axis lies in projection.



In the *Display* menu, the option *Use Dendrogram Colours on 3D Plots* is available. When this item is selected, a check mark appears next to it in the menu, and the individual points plotted on either of the 3D plots are coloured to correspond to the colour groupings shown in the Dendrogram plot - *e.g.* if samples 11, 12 and 13 are all in the same cluster according to the dendrogram, they will all be the same colour - green:



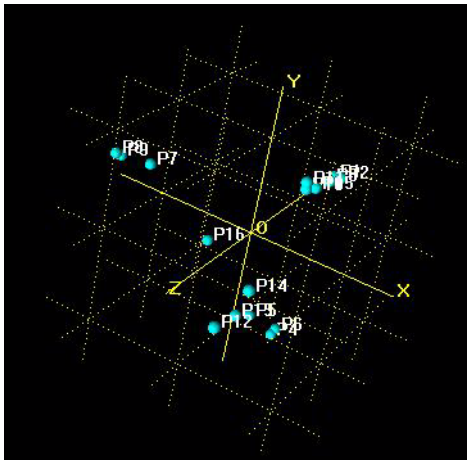
In the 3D plot, they will all still be coloured similarly:

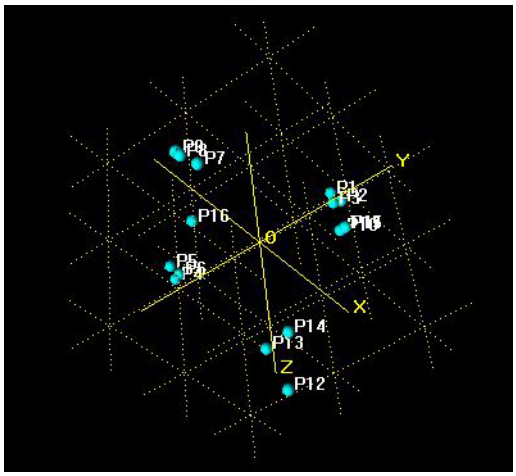


Changing the dendrogram cut-level causes the colours in the 3D plots to be updated. A small numerical label in the top-left corner of the display gives an indication as to the goodness of fit of these results. Numbers close to 1.0 suggest that it is a good fit, and low numbers suggest that caution may be required, or that the program had trouble adequately partitioning the data.

Note that as the number of samples increases, the average GOF score will decrease.

The orientation of the 3D plot can be altered by holding down the *Shift* key and dragging the mouse in any direction as desired; the plot rotates as shown below:

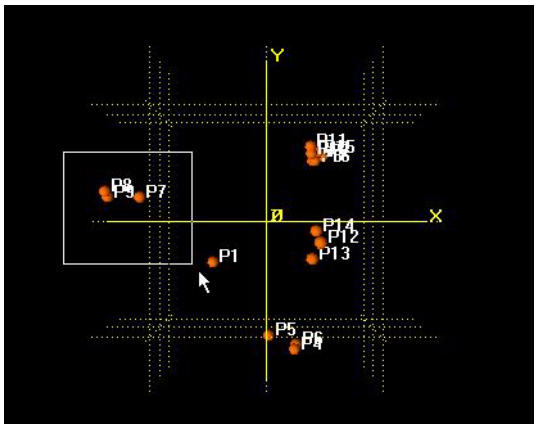




In the above plots it can be seen that there are 4 or 5 clusters depending on P16 and whether it can be considered to belong to an adjacent cluster or not.

A variety of views can be achieved to gain a better understanding of the distribution of the pattern data points in the three-dimensional space.

As in the other graphics screens, any particular area of the 3D view can be zoomed in by dragging a rectangle over the relevant region:

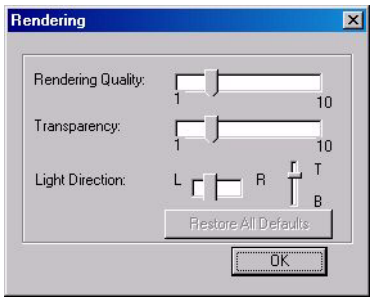


The pattern corresponding to each display point can be selected by clicking on it, this in turn updates the pattern information display in the lower portion of the window.

The points representing the patterns can be enlarged or shrunk to suit any zoom level by holding down the *Ctrl* key and moving the mouse either up or down. An upward movement will reduce the size of the spheres, a downward movement will increase the size.

The 3D plot position itself can be translated by holding down the *Alt* key and then moving the mouse in any direction as required.

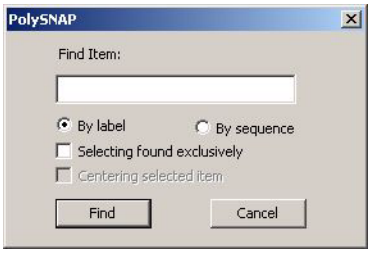
The drawing quality of the spheres can be altered if needed - with many points plotted, working with the display can be much faster if the rendering quality is reduced (this is especially the case with lower-powered graphics cards). To do this, click on the 3D display and press the *F12* button on the keyboard:



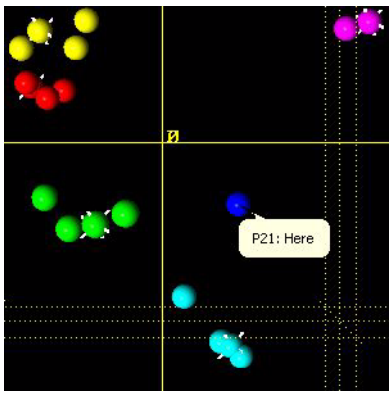
Moving the slider causes the drawing of the spheres to update in real time, so the effects of a particular setting can be easily seen. The chosen quality setting is saved and used from that point on. The higher quality settings can be useful when preparing screenshots for use in reports or for publication.

Further options can be accessed by right clicking on the display area to show the standard pop-up menu. Options relevant to the 3D plot include:

Find Item...



This brings up a dialog box allowing a particular pattern of interest to be located on the display by means of its index number, or label as appropriate.

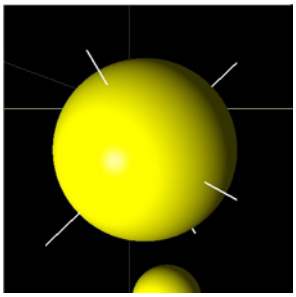


Additional options allow the pattern to become automatically selected when found - thus updating the pattern information display, and optionally centred in the display. This is useful when dealing with large numbers of patterns, as locating an individual pattern of interest on a crowded plot can be difficult.

Show Grid - the grid which appears in the 3D plot can be hidden or displayed.

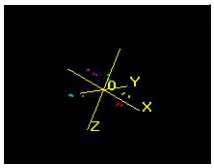
Show Labels - the labels which appear next to the plotted points can be turned on or off with this option. This may aid in seeing an overall clustering pattern when the display is crowded with many points.

Show MRP Marks - the Most Representative Pattern in each cluster can be highlighted if required with this option. it appears on the display as a normal pattern with several 'spikes' coming out of it:

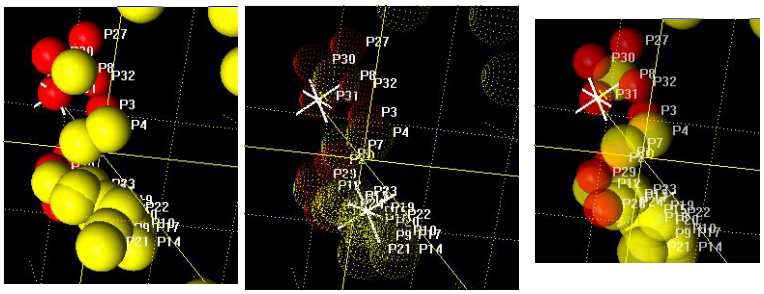


These spikes can be hidden or shown by means of this menu option. Clicking on an MRP sphere brings up a dialog box containing information about the mean pattern-pattern distance for that particular cluster. The smaller the distance, the tighter the cluster.

Show Top View - this option brings up a small simplified overview of the plot in the lower right hand corner. It can be useful for orientating yourself when zoomed into the display:

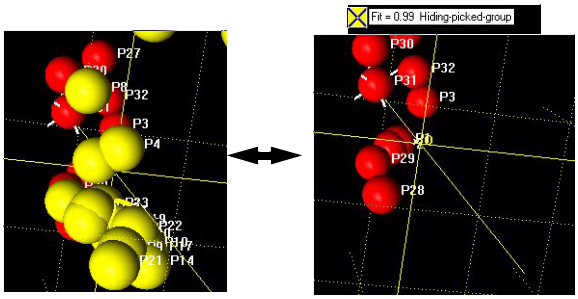


Render as Dots and *Transparent* alter the way the spheres are plotted as shown in the diagrams below. This can be useful to identify if for example, a single pattern of one colour is hidden within a group of another patterns:



Mask Group

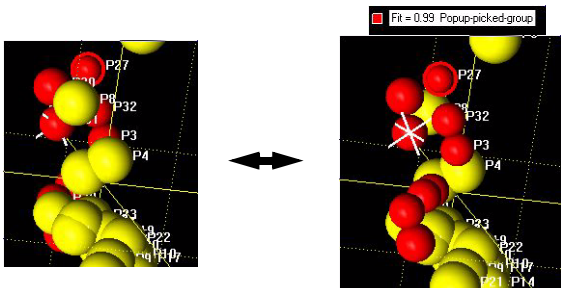
This option is useful in a crowded display with lots of different clusters overlapping each other - select the *Mask Group* option, and the next cluster you click on is temporarily removed from the display. A key in the top left corner shows the colour of the hidden cluster. To restore the cluster, deselect the option from the menu. Alternatively, click on a different colour, and that will be hidden instead. Only one cluster at a time can be hidden.



With the *Mask Group* option selected, pressing the spacebar toggles between hiding the selected group and showing everything else, and hiding everything else and showing just the selected group. The left and right arrow keys can also be used at this point to cycle through and have each group selected in turn.

Popup Group

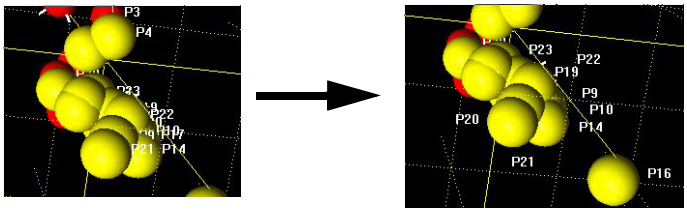
This option can be used to bring the whole of a particular coloured cluster to the front of the display, so it can all be seen at one time. A key in the upper left shows the colour of the selected group. Deselecting this option returns the display to normal.



Drag Labels

With this option toggled on, the display is fixed and cannot be rotated or zoomed. Therefore, it is necessary to use the normal zooming and rotation controls to select a suitable view angle before selecting this option. Once selected, all of the pattern labels on the display can be moved individually by clicking on them and dragging to the required

new location. This allows for neater diagrams to be created when required for a report.



2.5.5 3D Plot (PCA)

The plot drawn here is based on the results from principle component analysis of the modified correlation matrix. The use and interaction options available for this plot are identical to those for the 3D Plot (MMDS) described in Section 2.5.4.

2.5.6 6D Plot

PolySNAP provides the facility to allow either of the two standard three-dimensional plots discussed earlier to be augmented by up to three additional user-specified dimensions, as described here.

These additional dimensions are used to represent information recorded about each sample regarding its method of preparation. Available information that can be plotted is by default (assuming it is available in the pattern data files):

Mass
Total Volume
Counterion
Stirrer Rate
Sample Presentation
Solvent
Antisolvent
Initial Temperature
Isolation Temperature
Cooling Rate
Heating Rate
Reaction Time
Antisolvent Volume

The 3 additional plotting dimensions that are available to represent these are:

Point Size
Point Shape

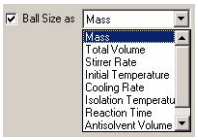
By using this approach, it can be possible to discover if there is a connection between a particular combination of sample preparation conditions and the resulting clusters (note however that there are some restrictions as to which fields may be plotted as which dimension).

To take advantage of this feature, select the *6D Plot* tab. The following options toolbar will appear:

The first option presented is to select which of the two different 3D plots available is to be used as the basis for this user-modified plot. The selection is between the MMDS and PCA plots, each of which were discussed in more detail earlier in Section 2.5.4 and Section 2.5.5 respectively.

Once this has been done, select which information field is to be plotted as which dimension. Up to three dimensions may be plotted at any one time, as chosen by selecting or deselecting the checkboxes as appropriate.

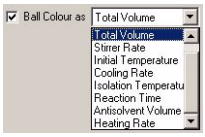
2.5.6.1 Ball Size Dimension:



The default possible information types available to be plotted as the size dimension are: Mass, Total Volume, Stirrer Rate, Initial Temp, Cooling Rate, Isolation Temp, Reaction Time, Antisolvent Volume and Heating Rate.

When each pattern was imported, any sample information fields were parsed, and the maximum and minimum values of each field stored. This data is then used to scale each datapoint to an appropriate size.

2.5.6.2 Colour Dimension:



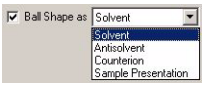
The information types available to be plotted as the colour dimension are: Mass, Total Volume, Stirrer Rate, Initial Temp, Cooling Rate,

Isolation Temp, Reaction Time, Antisolvent Volume and Heating Rate. Dendrogram Colours are also available.

When each pattern was imported, its sample information fields were parsed, and the maximum and minimum values of each field stored. This is then used to plot a range of colours based on the data values.

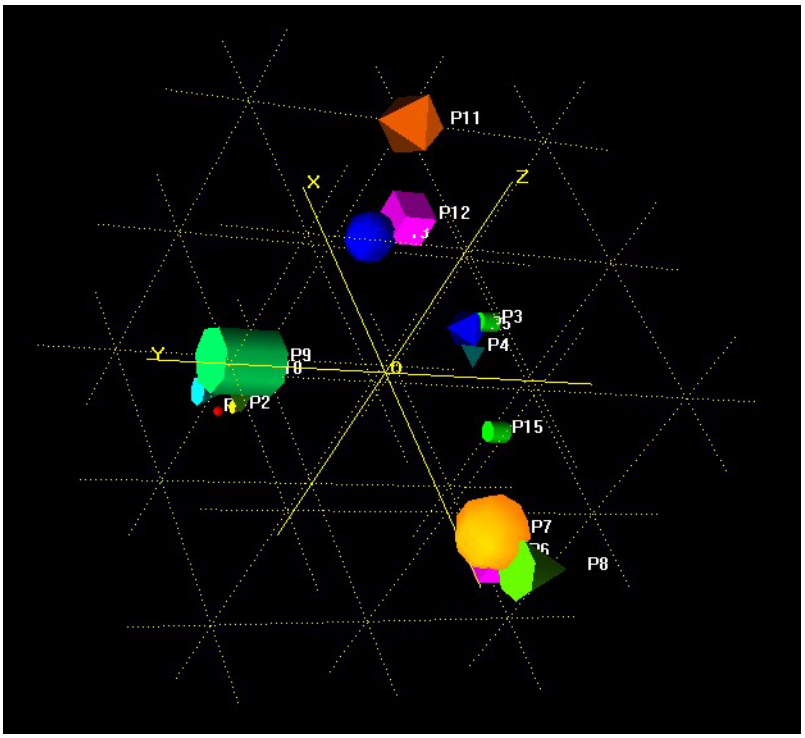
Additionally, the final option *Dendrogram Colours* allows the colour scheme of the dendrogram at the currently saved cut-level to be applied to the plot.

2.5.6.3 Shape Dimension:



This is provided as a means of plotting the more textually orientated information fields, such as solvent and counterion. Up to six different shapes are available (sphere, cone, cylinder, tetrahedron, octahedron and cube), allowing for up to six different solvents, for example, to be represented on the plot.

Once the settings are complete, click on the *Apply* button. The generated plot, is then shown below the toolbar:



The standard graphics controls for moving, zooming and rotating the display *etc.* all apply as normal, as do the controls for selecting

individual patterns of interest, and displaying their details in the pattern information pane.

The settings used to produce the plot are saved to the logfile for future reference.

Note that in 6D plot results need to be manually copied and added to the report pane at present.

2.5.6.4 Editing Sample Preparation information details and location

There is a file located by default in the program folder (usually *C:/ Program Files/PolySNAP/*) called *sample_info_format.txt*. (The location of this file can be altered by changing the option in the program *Options* window). Editing this file enables the user to change the field names, types and order expected for info for the 6D plot. The user can also set program to look for info in a single file (1 line per pattern) rather than in the individual raw data headers.

If stored in the RAW headers, the program looks in the fields for Sample Name and Comment for the information (n that order), which should be semi-colon delimited. If a given field is unused for a particular pattern, it should be left blank, but the semi-colon delimiter should still be there. Fields must be in the correct order as described in the setup file.

For example:

2.2E-5;2.0;ethanol;hexane;;100.0;50.0;2.0;0.0;180.0;residue;

If in a single file, there should be one line of numbers as shown above per pattern. By default, PolySNAP looks for this as 'sample.dat' in the data input directory.

These should be in the same order as the patterns are loaded into the program; line 1 should correspond to Sample 1, and so on. The program loads files in system alphabetical filename order.

```
# File format info
# Where is this info stored - either just HEADER
for in data files header space or
# FILENAME full path\to\file.txt

HEADER
#FILENAME C:/test.dat

# Fields - maximum of 20 fields, 1 per line, in
expected order
# Specify Type; units; format (format is either
textual TEXT or numeric NUMBER)
# e.g. Mass;mg;NUMBER
# if no units just put ;;
```

```
# put END as last line

Mass;mg;NUMBER
Reactor Identity;;TEXT
Total Volume;ml;NUMBER
Solvent;;TEXT
Solvent Ratio;;TEXT
Antisolvent;;TEXT
Antisolvent Volume;ml;NUMBER
Counterion;;TEXT
Stirrer rate;rpm;NUMBER
Initial temp;deg C;NUMBER
Heating Rate;deg C/min;NUMBER
Cooling Rate;deg C/min;NUMBER
Isolation Temp;deg C;NUMBER
Reaction Time;min;NUMBER
Sample Presentation;;TEXT
END
```

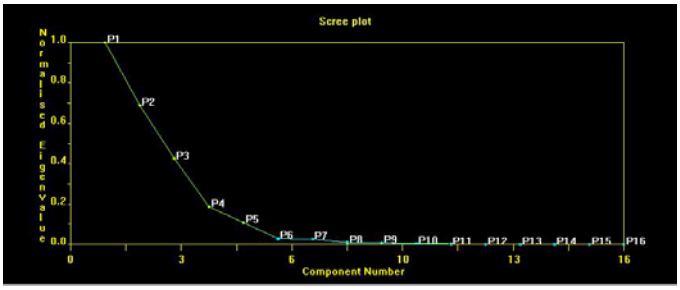
2.5.7 Validation

The validation screen contains six different displays, each with a different method of validating the data. Initially the *Scree Plot* is displayed but the other validation screens can be accessed using the pull-down menu at the top of the display:

These different displays are described below.

2.5.7.1 Scree Plot

The Scree plot is a 2 dimensional graph. Along the x-axis is the *Eigenvalue Number* and the y-axis is made up from the *eigenvalue* itself.



Eigenvalues are derived from the modified correlation matrix, which is first normalised. The eigenvalues are sorted in descending order.

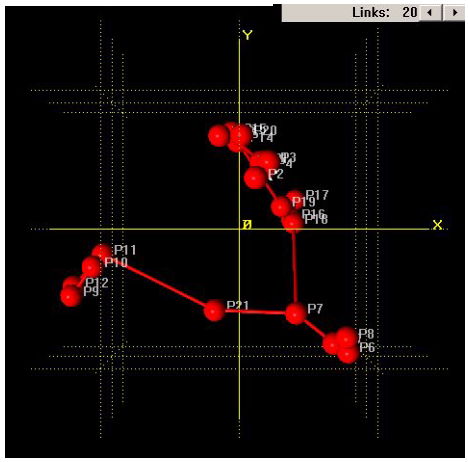
What this represents is the minimum number of clusters that can be used to describe the entire data set being examined. The point where the plotted line changes colour - *e.g.* between 5 and 6 in the example

above, suggests that just 5 clusters are needed to explain over 95% of the variation in the data. A well behaved scree plot should have a reasonably steep initial descent. A gradual, sloping descent indicates difficulty in establishing the number of clusters required, so the program-generated dendrogram cut-level should be examined especially closely.

By holding down the *Alt* key and the mouse button, the graph can be moved about the pane to any desired location. The plot can be zoomed in by dragging a rectangle over the relevant area, and the usual further options can be accessed by right clicking to reveal the pop-up menu.

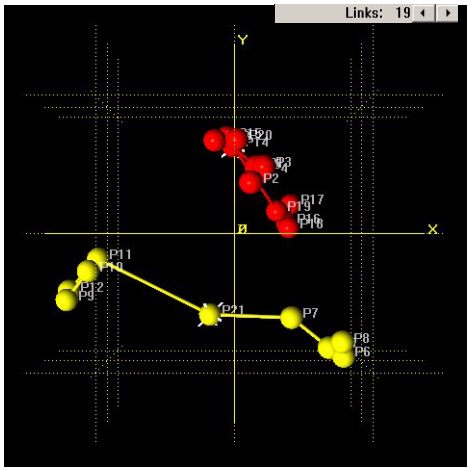
2.5.7.2 Minimum Spanning Trees

The MST display presents a different way of partitioning the patterns into different clusters. All of the patterns are initially joined together by a single line, in the order of increasing distance between them. The initial view shows all of the samples connected in this manner:

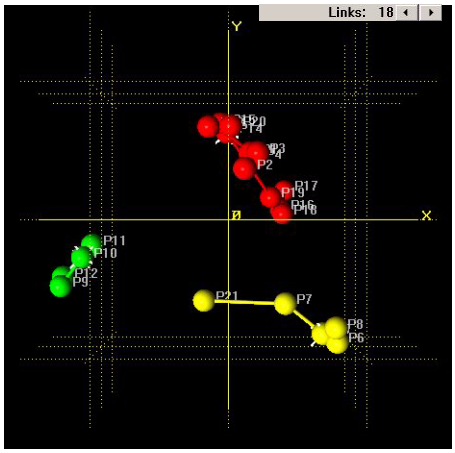


The program then cuts links between the patterns in order of decreasing maximum distance until the estimated number of clusters is reached. The user can then manually adjust this by clicking on the left-facing arrow in the top-right corner, to reduce the number of

links by one, or by clicking on the right-facing arrow in the top-right corner, to increase the number of links by one:



This process can be repeated, cutting the next smallest distance, and creating an additional cluster each time.



The display can be zoomed, rotated and otherwise manipulated in a similar manner to the other 3D plots. Note that each time the number of clusters changes the most representative patterns in each are recalculated. Note that the colours here do **not** correspond to those on the dendrogram.

2.5.7.3 Silhouettes

For each of the current clusters as defined by the dendrogram cut-level, this display shows a histogram. Silhouettes (Rousseeuw, P.J. (1987). *J. Computation & Appl. Math.*, **20**, 53-65.) provide an alternative formalism for assessing the compactness and isolation of clusters, and also for identifying those members of a given cluster which are well established members of the core cluster or outlying,

and thus potentially problematic. If the *i-th* pattern belongs to cluster C_r then we define the silhouette, *s(i)* as follows:

$$a(i) = \frac{\sum_{j \in C_r} d_{ij}}{n_r - 1}$$

$$b(i) = \min_{s \neq r} \left(\frac{\sum_{j \in C_s} d_{ij}}{n_s} \right)$$

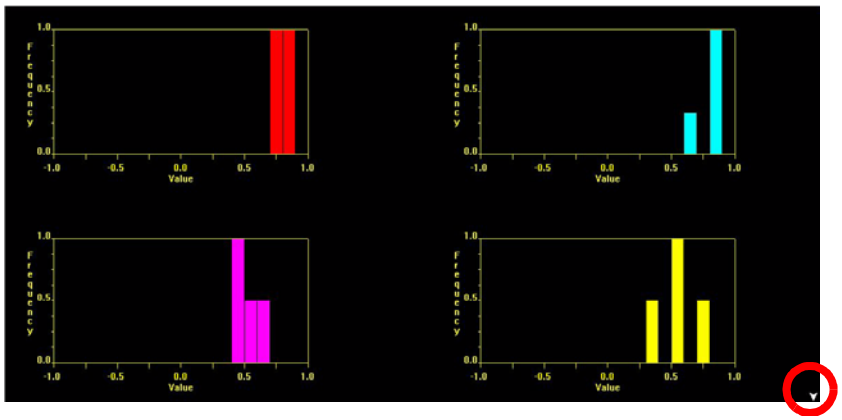
$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

where there are n_r patterns in cluster *r*, n_s patterns in cluster *s*, and d_{ij} is the distance between patterns *i* and *j*. The values of s(i) lie between –1 and +1. The lower the value of the silhouette, the more likely it is that the pattern either:

1.Belongs to a different cluster and you should try altering the dendrogram cut level, or:

2. The pattern is a mixture. The MMDS and PCA plots will be useful here to identify the closest neighbours of pattern *i*.

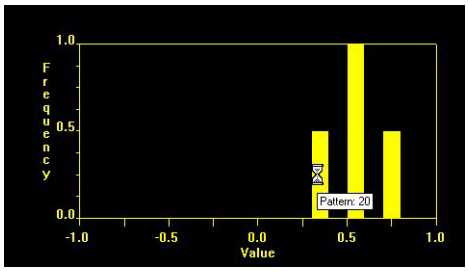
One histogram is shown for each cluster. If there are more histograms than will fit on the screen at one time, a small down-pointing arrow appears in the lower right corner.

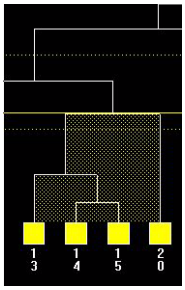


To scroll down to see the next row of histograms, click on the arrow, or use the scroll wheel on the mouse. Similarly, an upwards pointing arrow will appear if there are histograms above the top of the screen.

Allowing the mouse to hover over a particular column shows which pattern numbers correspond to that score. For example, the yellow cluster appears to have a member that has a reasonably low membership score - mousing over it shows that this is pattern 20,

which makes sense when looking at the dendrogram, where it is the least tightly linked member:

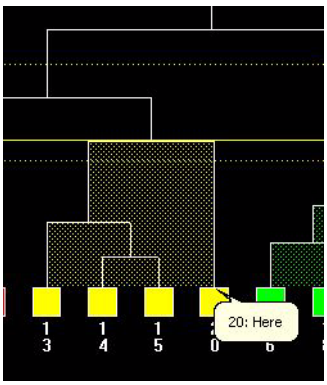




Note that if the dendrogram cut-level has been manually adjusted since the previous time this display was accessed, there may be a short delay as new silhouettes are calculated for the revised cluster membership list.

Clicking on a column selects the patterns presented in that column, allowing them to later be located on the 3D or dendrogram displays. For example, clicking on the column pictured above brings up a dialog:

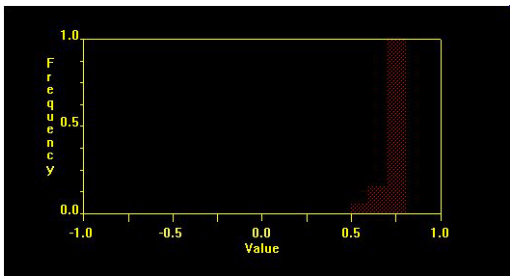
When next viewing the 3D plot or Dendrogram, press F4, and pattern 20 is then automatically highlighted on the display. This allows cross checking between patterns that the silhouettes display may highlight as unusual, and their corresponding position in the other results output.



2.5.7.4 Fuzzy Clustering

This display is superficially similar to the silhouettes, and is navigated in the same manner. With fuzzy clustering, and unlike the standard clustering methods used elsewhere in the program, a single

pattern can be assigned to more than one cluster. Using the concept of membership, a value can be calculated for how well a given pattern fits in a given cluster. A low membership score may suggest that that pattern either does not belong in that cluster, or that it is a mixture. A single pattern having reasonably high scores in more than one cluster may indicate a mixture. Only patterns which fall outside a defined threshold range, and therefore are samples that may need to be examined manually in more detail, are shown in the fuzzy clustering output. If this is not the case for any of the patterns, then no output is shown - this is the ideal case, as all patterns clearly belong to only one cluster.



Checking the output from this calculation should therefore help to highlight borderline or unusual cases that do not fit neatly into one cluster or another. Such examples may possibly be mixtures, or evidence of some other problem such as a different background or a 2θ-shift.

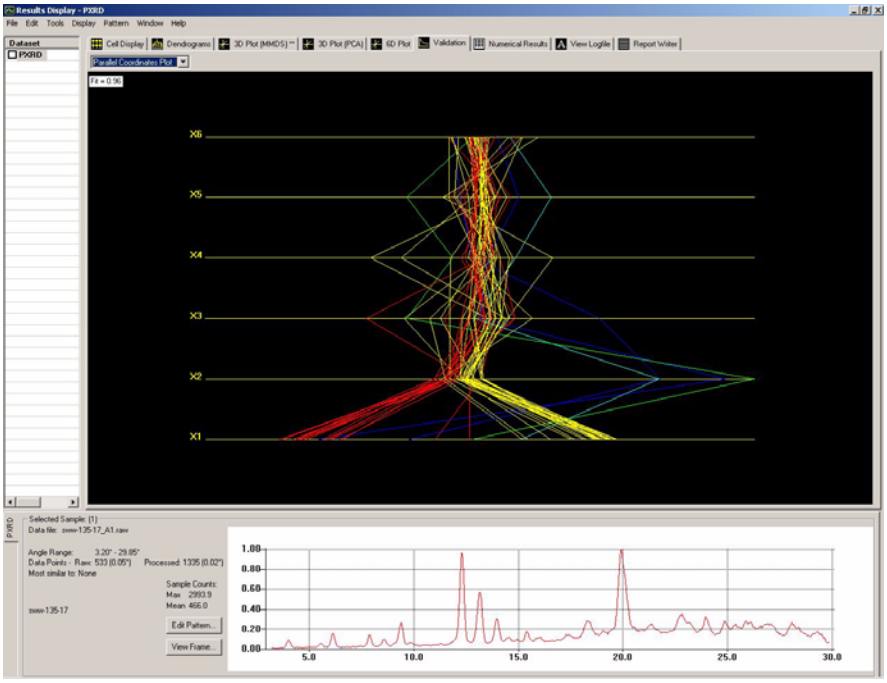
Detailed numeric output from this method is available in the logfile pane.

Pattern selection and automatic find features as described above in the Silhouettes are also available in this display.

2.5.7.5 Parallel Coordinates Plot

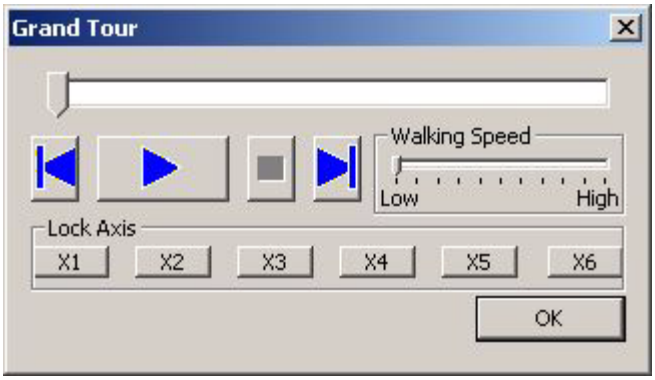
Rather than plotting our patterns in a 3D space, here we plot the first 6 dimensions of the clusters in a linear fashion, allowing the user to see if the cluster separations evident in the first 3 dimensions still hold together in higher dimensional space. Where as the first three dimensions are plotted as x, y and z axes on a 3D plot here they are

plotted as the first, second and third vertical axes on the plot, with the fourth, fifth and sixth dimensions plotted above.



While the three axes on the 3D plot are arranged orthogonal to each other, they are arranged horizontally parallel to each other in this plot. As in the 3D plot each object (fragment or variable) is given a value (between one and zero) for each dimension and this is plotted for each axis. While in 3D space this becomes a data point represented by a sphere, in the parallel coordinates plot this becomes a line joining up the different values for an object as the value varies from dimension to dimension.

The colours are taken from the data space Dendrogram so it is easy to identify the separate clusters. By right-clicking and selecting *Grand Tour* from the pop-up menu, the display can be animated by rotating round the different axes simultaneously. This can also be accessed by using the *Ctrl-G* shortcut.



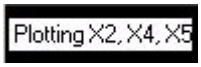
The animation can be played, paused, fast-forwarded to the end, or rewound to the start using the blue play controls. The progress bar at the top indicates how far into the animation the grand tour is. This can also be used to skip directly to different parts of the sequence. The *Walking Speed* option controls how fast the animation proceeds at. Finally the *Lock Axis* allows the user to freeze each axis individually in the state being currently displayed while the rest of the display continues to be animated.

2.5.7.6 Space Explorer

This display resembles the standard 3D plot, with the exception of an additional option on the toolbar.

Like the parallel coordinates plot it allows analysis of higher dimensions of data however here they can be visualised on orthogonal axes. For this reason only three dimensions can be displayed at the one time.The display opens with the dimensions 1, 2 and 3 plotted as in a normal 3D plot. The extra toolbar button allows the user to change the axes being plotted. Clicking the button once changes to plot to display the first, second and fourth dimensions.

By iterating through the various combinations of the first six dimensions by means of the *Next* and *Previous* options from the toolbar dropdown menu all of the other combinations can be accessed. The legend in the upper left corner updates to show which dimensions are currently being plotted:



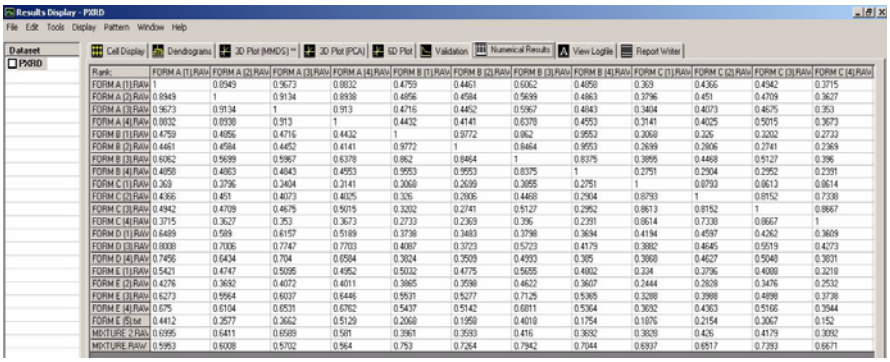
This allows the user to see if the cluster separations evident in the first three dimensions still hold together in higher dimensional space.

By right-clicking and selecting *Grand Tour* from the pop-up menu, or using the *Ctrl-G* shortcut, the display can be animated for easier interpretation. The Grand Tour controls work in the same manner as in the Parallel Coordinates Plot.

2.5.8 Numerical Results

This pane is used to display the pattern correlation matrix. If there are a large number of patterns then the entire matrix may not fit within

the display, and scroll bars will be displayed along the side and bottom of the table.



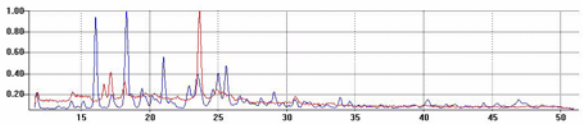
If the program was run with the *Allow Offsets* option turned on, then the results from this can be seen by selecting a particular cell, and then allowing the mouse to hover over that cell for a second, until a 'tool-tip' appears. This contains two numbers in the form (a_0,a_1), where a_0 is the amount of linear offset applied, and a_1 the amount of non-linear offset applied.

A diagonal line of 1 should be present to show the result of each pattern matched against itself. Clicking on the *1* for each pattern will produce the profile and information for that pattern in the relevant area below.

For comparison purposes two patterns can be overlaid on each other by clicking on a number above or below the line of *1*.

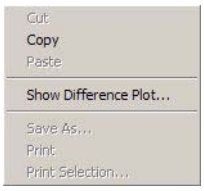
Rank:	FORM A (2)	FORM A (3)	FORM A (4)	FORM B (2)	FORM B (3)
FORM A (2)	1	0.8592	0.9038	0.3387	0.4366
FORM A (3)	0.8592	1	0.8885	0.3184	0.4531
FORM A (4)	0.9038	0.8885	1	0.2914	0.4855
FORM B (2)	0.3387	0.3184	0.2914	1	0.8482
FORM B (3)	0.4366	0.4531	0.4855	0.8482	1

For example if 0.2914 is clicked on as highlighted above, the graph of FORM A (4) and FORM B (2) will be displayed, one overlaid on the other in the graph pane:



This allows a visual comparison to help decide if the matching results displayed are sensible or not.

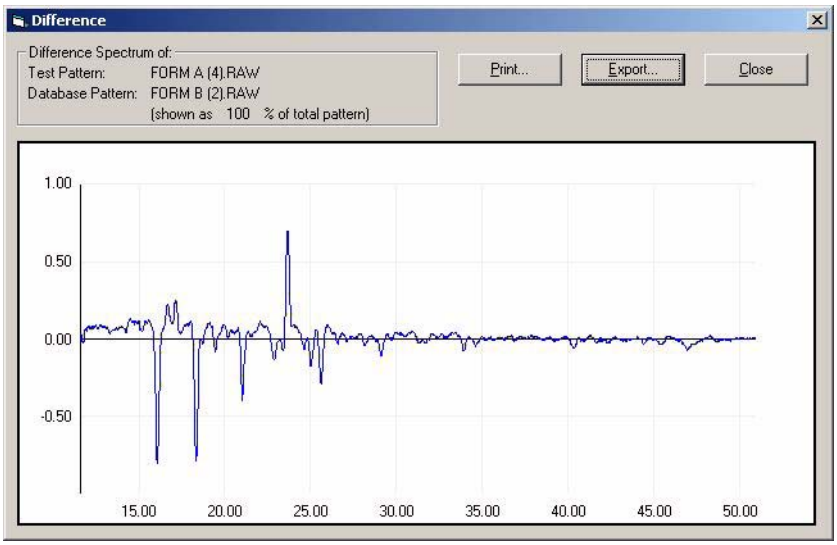
Right clicking on the matrix of results will produce the following options:



Clicking on *Copy* (or alternatively selecting *Copy* from the *Edit* menu) will copy the contents of any selected cells to the clipboard. To copy only part or all of the results, drag a rectangle over the desired numbers, a blue highlight region will appear indicating the numbers selected. Now right click to produce the options and click on *Copy.*

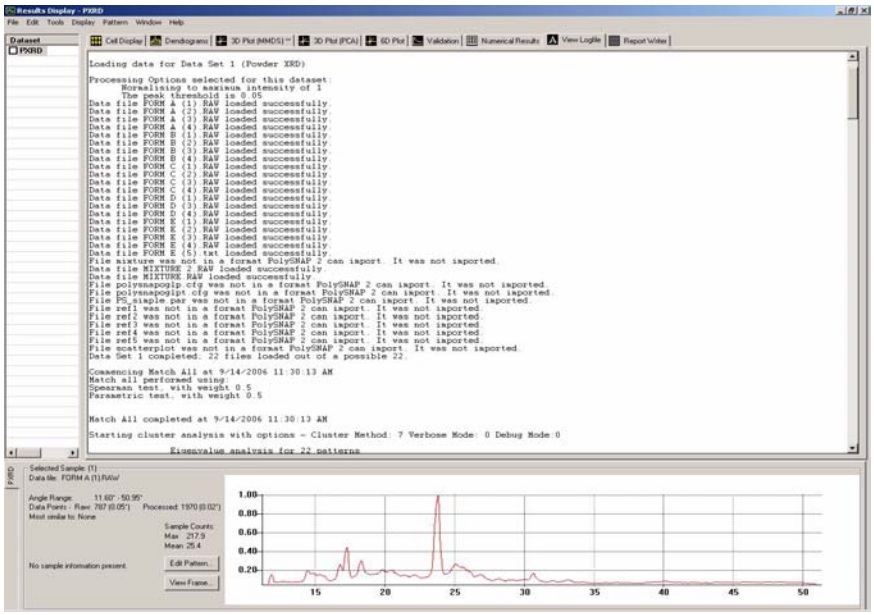
Note that the numbers in this window are not editable in any way.

The *Show Difference Plot...* option brings up a new window with a plot showing the difference trace between the two currently selected patterns, for example:



2.5.9 View Logfile

The results of the file import, processing, pattern matching, clustering *etc.*, and any subsequent changes to the results are displayed here in the View Logfile Pane:

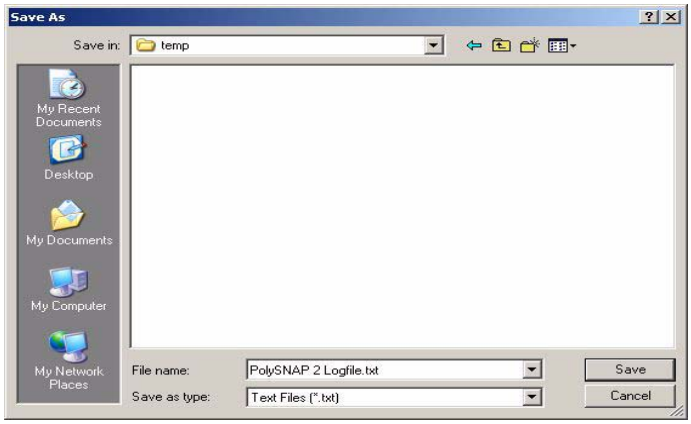


A scroll bar appears on the right hand side of the pane to allow the user to view all of the text, as the output is normally quite long.

The text can be copied to the clipboard and pasted to other applications by selecting the relevant text to highlight it and then right-clicking in the text pane:

Click on *Copy* to copy the text (alternatively choose the *Edit* menu and click on *Copy*). Note that the *Cut* and *Paste* options are unavailable, as the logfile cannot be edited manually.

The *Save As...* option causes a standard file saving dialog box to appear



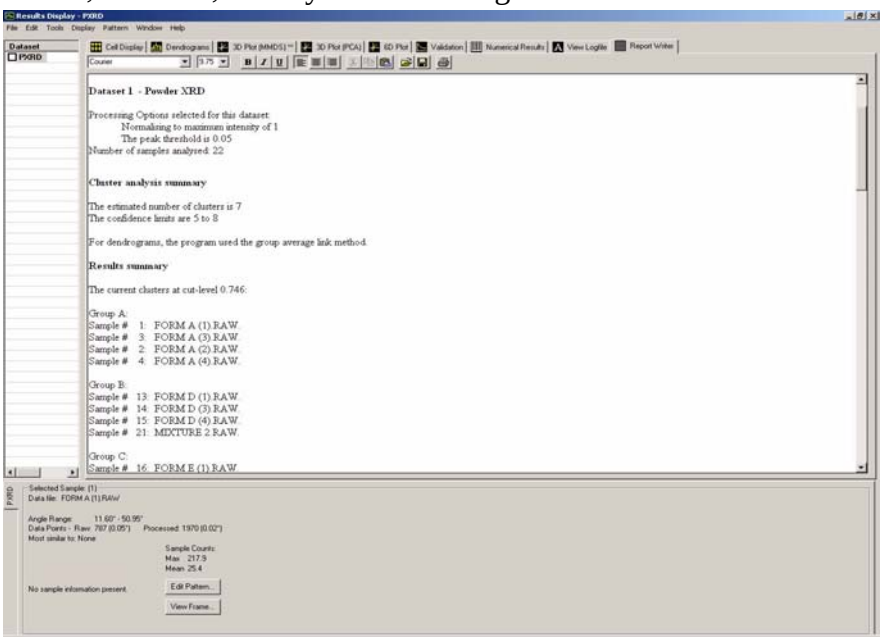
This allows a copy of the current logfile to be saved to a new file. Select the desired location for the file, edit the filename if required. The format for saving is an ASCII text file (*.txt)

The *Print...* option brings up the standard Windows print dialog box, allowing the current selection of the logfile output to be printed.

2.5.10 Report Writer

The report writer is an area that functions as a basic word-processor, in which the user can produce reports containing information from any of the output displays, or any other text or bitmapped graphics that can be pasted from the clipboard as desired.

The pane itself initially consists of a white screen with standard text-editing toolbar is provided at the top of the window, allowing choice of font, font size, text style and text alignment in the usual manner:



Right clicking on the report pane will produce the following options:

Cut - Highlight an area by holding the left mouse button down and dragging over a relevant area. Selecting *Cut* causes the selection to be removed from the screen and copied into the clipboard.

Copy - A selected region of the report pane is copied in to the clipboard, and can then be pasted into another application.

Paste - An object or text which has previously been copied into the clipboard is pasted at the current insertion point.

Save As - Clicking on *Save as...*opens a standard Save As dialog box. The report can be saved in one of three file formats:

RTF Files (*.*rtf*) - Rich Text Format

HTML Files (*.*html*)

Text Files (*.*txt*) - ASCII Text Format.

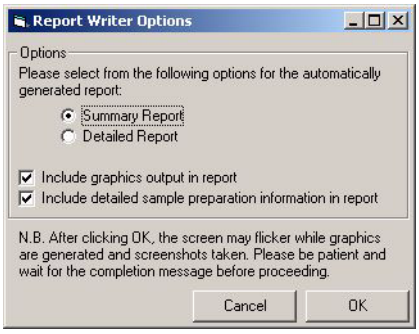
RTF format is recommended for most uses, as any images added to the report are saved together within the RTF file, whereas HMTL export produces a HTML file and separate files for each embedded graphic.

N.B. Saving as ASCII text retains only the raw text; any embedded images are not saved.

The *Print...* option brings up the standard Windows print dialog box, allowing the current selection of the logfile output to be printed.

2.5.10.1 Automatic generation of reports.

Selecting this option brings up a dialog box with several options to control the automatically generated report:



The difference between the Summary and Detailed report is as follows:

The summary report includes information as to the run start and end time, lists the processing and matching options selected, states the calculated number of clusters, lists the components of each cluster, and the results of any mixture analysis.

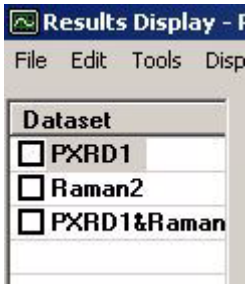
In addition to all of the above, the detailed report also includes lists of files imported, more detailed output on which cluster methods were selected, and a summary of any sample preparation information provided.

With the *Include graphics output in report* option selected, for Summary mode this adds the Cell Display and Dendrogram output to the report, and for Detailed mode it adds copies of the output from all of the main graphics modes. The *Include detailed sample preparation information* option adds a semi-colon delimited table of the sample information to the report. This can later be converted into a proper table in an external word processor such as Word.

Clicking *OK* starts the process, during which the screen will flicker as various displays are generated and information retrieved as required. A message is displayed once generation is complete, after which the report can be viewed, edited and saved in the *Report* pane of the main window.

2.5.11 Datasets pane

In cases where more than one dataset has been run the datasets pane is the interface to accessing the different datasets.

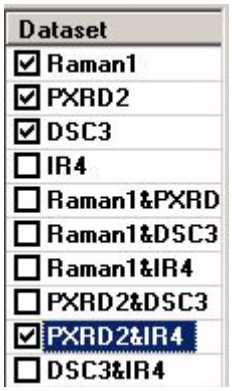


If only a single dataset has been run then the datasets pane will still be displayed, but will only contain one dataset. For multiple datasets there will be a list of all the available datasets.

By clicking on the different dataset names in the list (not on the checkboxes next to them) the results from each of these can be displayed in turn, including the results from combining the datasets.

2.5.11.1 Comparing Results display

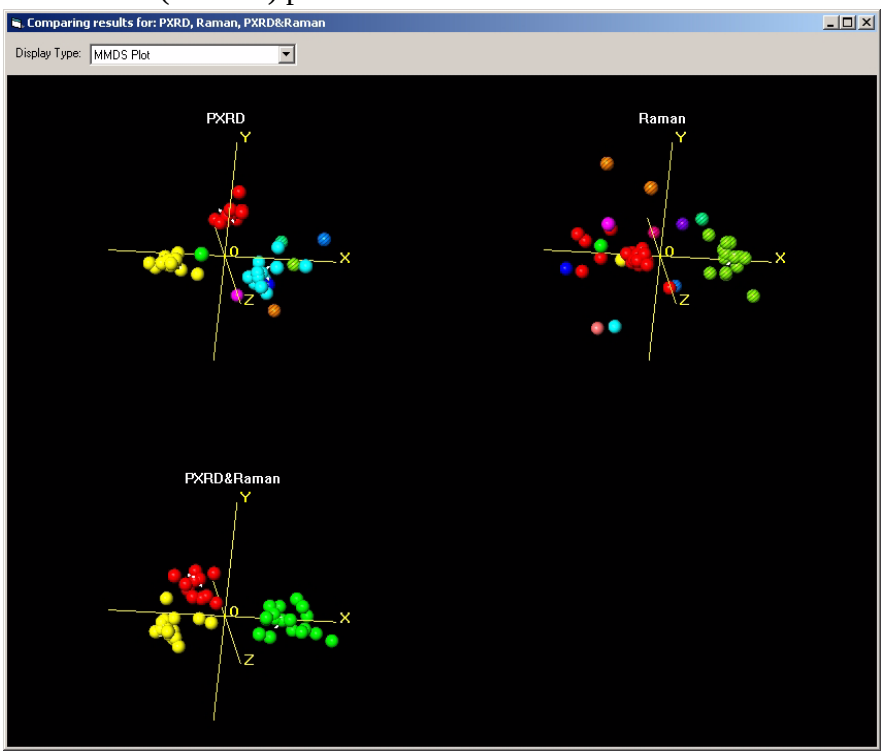
There is a special display window for comparing the results from different datasets. This option is only available when multiple datasets have been analysed together. To use this display the user must select the datasets they wish to compare. This is done by clicking on the checkboxes next to the dataset labels in the left-hand tab bar.



Up to a maximum of four datasets can be chosen and compared against each other. This is not restricted to the four original datasets that were input into *PolySNAP*. Combined results can be selected and compared using this option, like in the case above where combined X-ray and IR data are entered in to be compared.

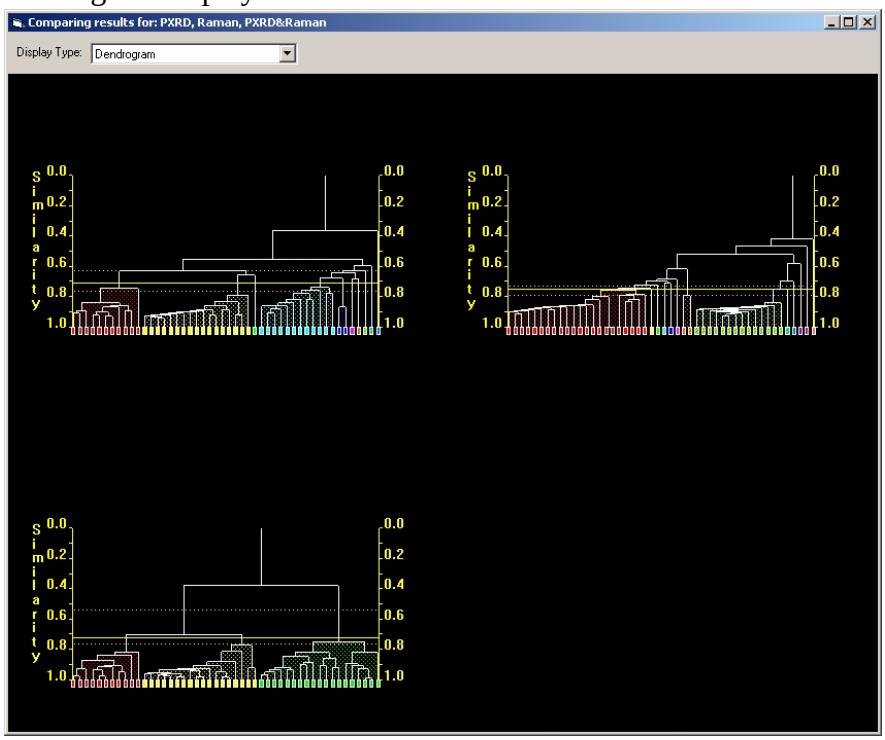
Once all of the relevant datasets have been highlighted select *Compare Results...* from the *Tools* menu. If more than the maximum number of four datasets have been selected then the comparison will not run and a message will display showing that this is the cause.

This opens a new window where the visual results from the three selected datasets are displayed next to each other. By default it opens with the 3D (MMDS) plot.



In this way the features of the different 3D plots can be directly compared, as the display can be rotated and zoomed into like a normal 3D plot. Here, we can easily see that the clusters for the combined PXRD&Raman dataset seem to be better than for either of the individual original datasets.

There is a pull-down menu in the top-left of the window that allows the 3D plot (MMDS) to be changed either to the 3D plot (PCA) or the dendrogram display.



2.5.12 Sample Information pane

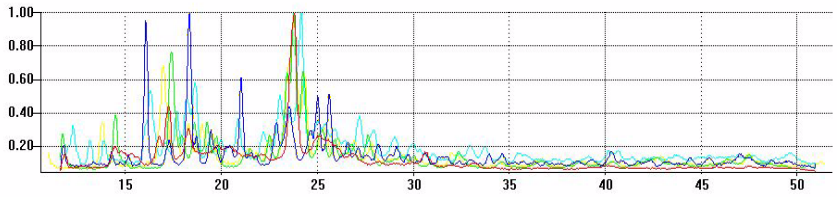
For a selected sample, the sample information pane displays relevant textual information on the left, and a graphical representation of the sample profile on the right. Although this information pane is always available by default, it is possible to hide it if more screen space is required to examine results in the graphical display areas of the window.

Selecting *Hide/Show Pattern Information* from the *Display* menu allows the information region to be hidden or shown as required.

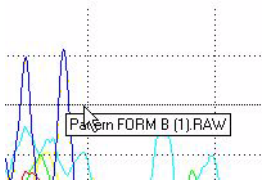
The right-most region shows the current pattern profile; it can be zoomed in by using the mouse to drag a rectangle over the relevant area. The graph pane is part of a larger area that shows information regarding the selected sample.

To the left of the graph appears an area with text information associated with that sample. Typical information displayed includes the data file location, the angle range, the data points for the original raw data and once any processing has been performed, the known phase most similar to this pattern (if present), and finally other relevant information regarding the sample preparation, if saved in the data file headers.

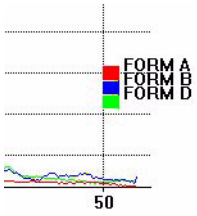
If more than one sample pattern has been selected in a given display screen, the pattern information pane displays a list of the patterns selected, and shows their profiles plotted on top of each other for easy comparison:



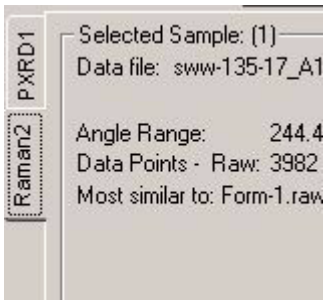
To identify which profile corresponds to which sample, hover the mouse over part of the line, until a tooltip appears containing the sample filename:



Alternatively, selecting the *Show Pattern Key on Graph* item in the *Display* menu brings up a visual colour-coded key:



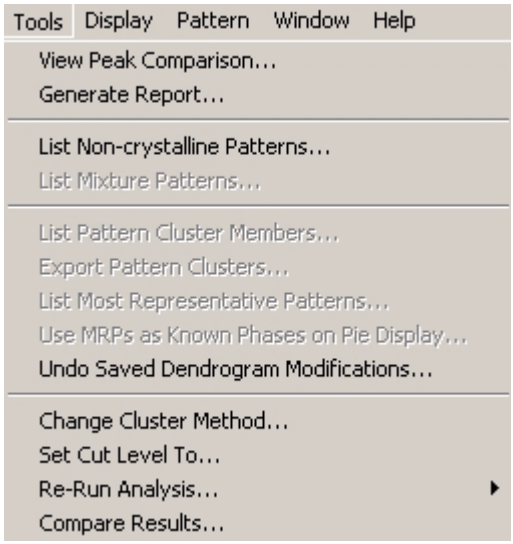
There is also a tab running vertically along the left hand side of the sample information pane.



This changes the dataset from which the sample information is being displayed. Changing the dataset being displayed in the sample information pane has no effect on the main display.

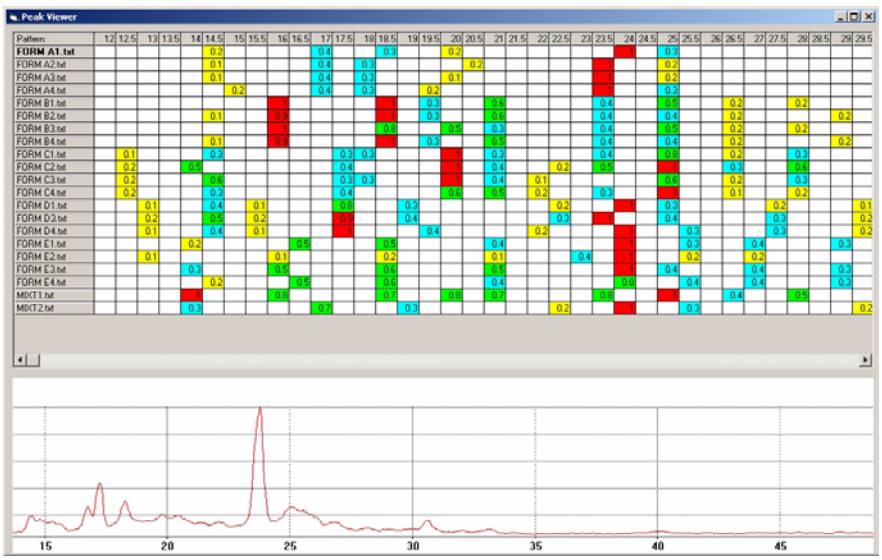
2.6 Tools menu

The *Tools* menu contains a series of further options:



2.6.1 View Peak Comparison

This option brings up a window in which the currently selected patterns are shown on a grid. Each pattern corresponds to one row of the grid, with the marked peak positions of those patterns highlighted according to their intensities. The x-axis is split up into bins of (by default 0.5 degrees), to allow easy visual comparison of peak positions in different patterns.



Clicking on the name of a pattern in the list on the left shows it on the display; clicking again on the name removes it from the display. The first pattern in the list is shown automatically. The colours can be

turned off, and the default angle-range bin size changed in the Display pane of the program *Options* window.

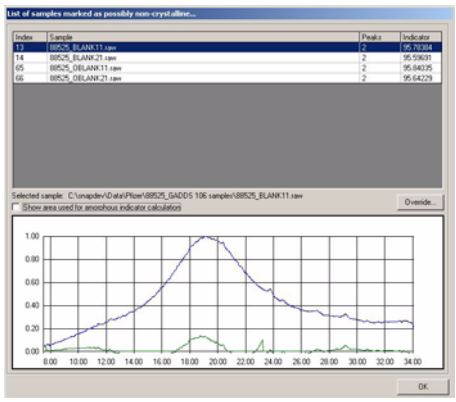
2.6.2 Generate Report...

This feature automatically generates a report in the *Report Writer* pane and is described in Section 2.5.10.1.

2.6.3 List Non-crystalline Patterns...

This option, accessed through the *Tools* menu, allows the user to examine a list of any patterns that the program has marked as being possibly non-crystalline when it has examined them.

Selecting it brings up a display window:



The upper region is a grid showing the pattern names, number of peaks, and approximate percentage amorphous content. The larger this value, the more likely the pattern is to have an amorphous content.

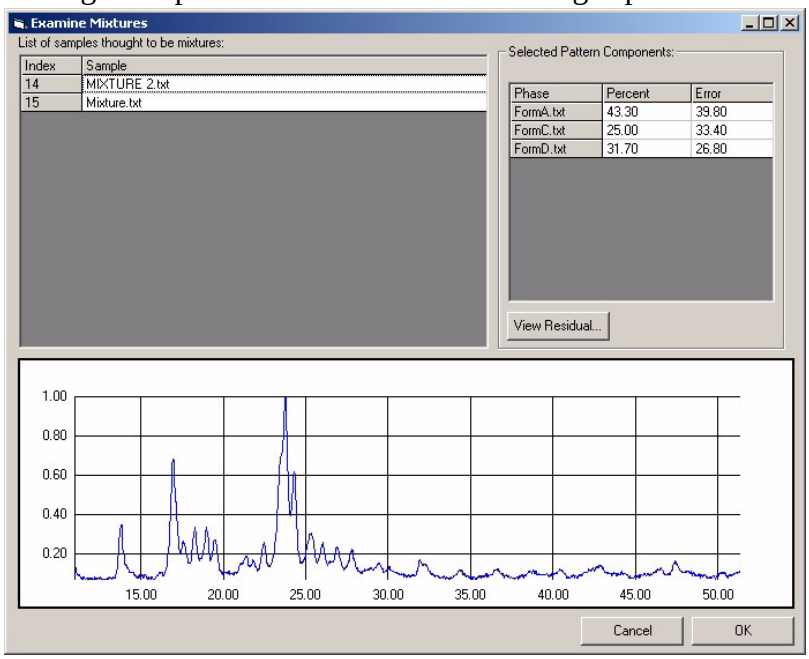
Selecting a pattern in the list (by clicking on it once) updates the profile display in the lower region of the window to show what the pattern profile actually looks like. If the *Show area used for amorphous indicator calculation* checkbox is selected, then a second, green trace shows the rough profile that would be left if the program subtracted out everything it thought was amorphous from the pattern.

If the user considers that the program has made an error in marking a particular pattern as non-crystalline, they may override this by clicking the *Override* button on the lower left. This action is recorded in the program logfile.

It is also possible to do the opposite, and mark a pattern as amorphous; for information on how to do this, see Section 2.5.3.2.

2.6.4 List Mixture Patterns...

Selecting this option from the *Tools* menu brings up a new window:



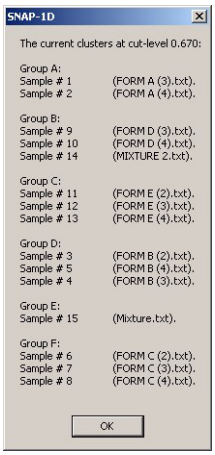
This consists of an upper region listing all of the patterns marked as mixtures, and a lower graph-pane region. Selecting a particular pattern from the list causes it to be displayed in the graph pane, and a list of which known phases are thought to make it up are listed in the display area. Clicking the *Residual* button opens a PolySNAP window to allow comparison of a simulated pattern created from the suggested phase components, and the original sample.

N.B. this option is only available when powder patterns for known phases were provided.

2.6.5 List Pattern Cluster Members...

This option is available only when the dendrogram is being displayed, and generates a list of the current members of each cluster suggested by the cluster analysis, according to the current saved dendrogram cut-level.

If the list of cluster members is short enough, it is displayed as a message box, which can then be dismissed:



If the list of members is too long for a standard message box, it is automatically added to the end of the program logfile, from where it can be copied into a report or other program as required.

2.6.6 Export Pattern Clusters...

This option brings up a dialog box allowing the list of pattern clusters to be saved to a data (*.dat*) file for use elsewhere.

2.6.7 List Most Representative Patterns...

This option is only available when either the MMDS or PCA plots are being displayed. Note that the calculation of MRPs is specific to the type of plot displayed, so different results may be obtained from each of the two methods. Different results will also be obtained if the dendrogram cut-level is adjusted manually.

A list of the pattern indexes and filenames for the most representative patterns in each cluster to have 3 or more members is shown in a dialog box on the screen. A copy of the list is also automatically added to the logfile when this option is selected.

2.6.8 Rerun Analysis...

Note that selecting these options causes the current PolySNAP session to finish, and the program is restarted. It is therefore important to ensure that any changes to the report, dendrogram or other displays are saved prior to selecting it.

2.6.8.1 - Using Only Currently Selected Patterns

This option allows the user to select a subset of the current data set to re-analyse. It closes the current PolySNAP results window, and re-runs the initial analysis of the patterns using the currently selected patterns on the display.

Output from the original run output is not overwritten and can therefore still be examined later using the *View Results* option.

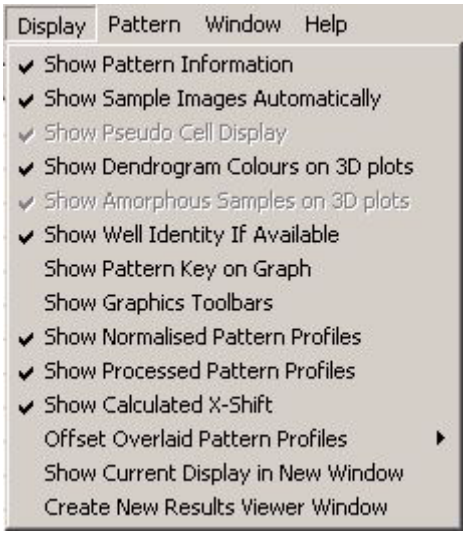
At least 4 samples must be selected on the display to be able to use this option.

2.6.8.2 - Ignoring All Non-crystalline Patterns

This option allows the user to re-analysis the current dataset minus any patterns flagged as being amorphous.

2.7 Display Menu

The *Display* menu contains a series of options that relate to how the results are presented on the display screen.



2.7.1 Show Pattern Information

This option allows the user to display or hide the sample information pane (described in Section 2.5.12) at the bottom of the screen. By default this option is on.

2.7.2 Show Sample Images Automatically

This option is only available when image files have been provided and controls whether the images are automatically displayed in the sample information pane or not. By default this option is on.

2.7.3 Show Pseudo Cell Display

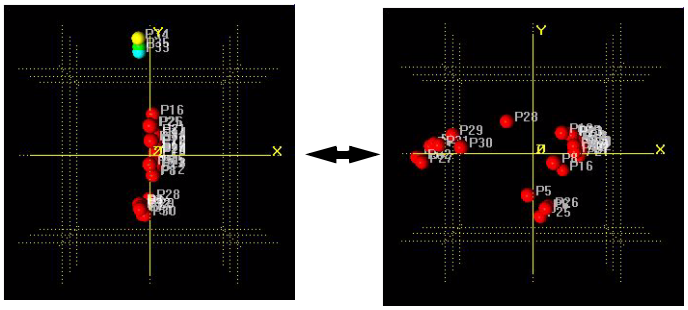
This option is available when known reference files have been provided and allows the user to have the cell display coloured according the dendrogram as if there were there where no known phases.

2.7.4 Show Dendrogram Colours on 3D Plots

This options controls whether the 3D plots are coloured according to the dendrogram. If this option is turned off then all spheres on the 3D plots will have the same colour. By default this option is on.

2.7.5 Show Amorphous Samples on 3D Plots

This toggle determines if samples that are flagged by the program as being amorphous are shown or not on the standard 3D plots. It is only enabled when more than one pattern is labelled as amorphous. In some cases, hiding such samples can make the plots appear quite different - this is often especially the case in the PCA plot.

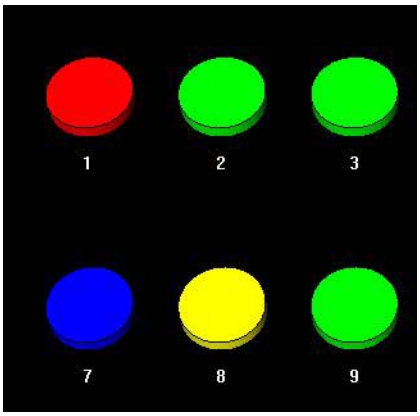


In the examples shown here the majority of the patterns are artificially 'flattened out' by the extreme distance between them and the amorphous; removing the amorphous allows for a better representation of the distances between the remaining patterns.

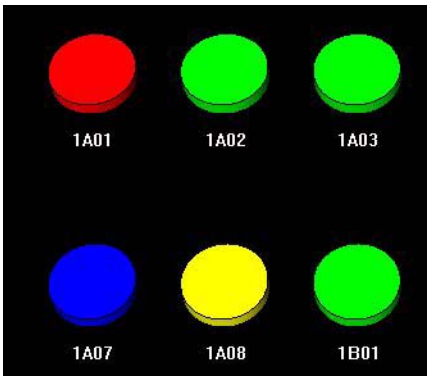
2.7.6 Show Well Identity If Available

If the sample filenames are of the form where some form of sample ID is contained within the filename, then that information can be extracted and displayed to help identify the samples.

For example, with this option turned off, a typical display might resemble:



Whereas with the option turned on, the display would look like this:



The default for this toggle is on.

Note that all the files in a given dataset must have the same consistent format for Well IDs to be displayed. See the *Options and Defaults* section for details.

2.7.7 Show Pattern Key on Graph

This option displays the name or label of the currently selected pattern profiles in a key to the right-hand side of the graph pane display.

2.7.8 Show Graphics Toolbars

This toggles the toolbar displays which are available on some of the graphics panes. The setting is persistent for a given run, but may be overridden temporarily by means of the *Show Toolbar* option available in the right-click pop-up menu of most of the different displays. The features of the toolbar are detailed in Section 2.5.1.4.

2.7.9 Show Normalised Pattern Profiles

This option toggles the graph pane between displaying all profiles on a common intensity scale, where the maximum value is always 1.0, to displaying the profiles with their maximum intensities as recorded in the original data files. The default for this toggle is off.

2.7.10 Show Processed Pattern Profiles

This option toggles the graph pane between displaying all profiles with any processing - such as background subtraction or smoothing - or in their unprocessed original form. The default for this toggle is on.

2.7.11 Show Calculated X-Shift

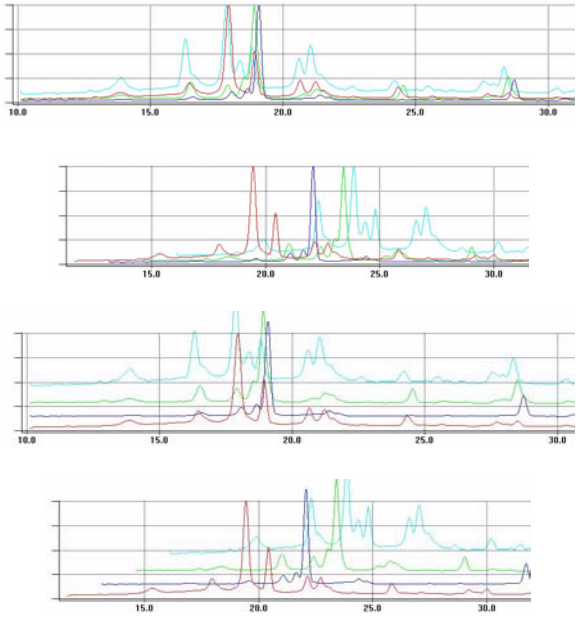
If an optimal x-shift has been calculated for a dataset, with this option on selecting a pair of patterns in the Numerical Results grid displays their profiles with the calculated amount of offset.

2.7.12 Offset Overlaid Pattern Profiles -> *on x-axis, on y-axis*

The two options in this submenu may be selected independently of one another. They control how patterns are displayed when more than one profile is shown on the graph pane at a time.

By default, overlaid profile traces are plotted directly on top of each other. By using these display toggles however, they may be plotted with a slight offset in either the x- or y-axes, or both. This may enable differences or similarities between the patterns to be observed more clearly.

For example, the screenshots below show first, several directly overlaid patterns, then the same shifted in the x, and then in the y, and finally shifted with both options turned on:



Note: it is important to remember to check that the x-shift option is turned off when trying to see if the peak positions of two patterns coincide or not!

2.7.13 Show Current Display in New Window

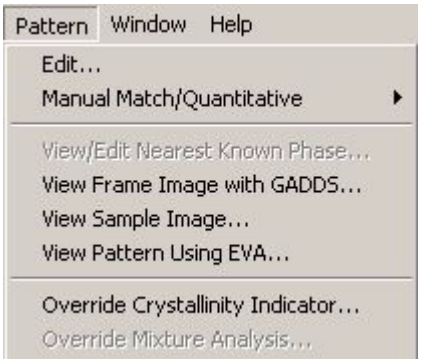
Selecting *Show current display in new window* from the *Display* menu will open a new window displaying only the current contents of the main display area (e.g. the cell display or the Dendrogram). It can be visually manipulated and moved around like a normal display, however will lack the sample information pane, or any of the tab menus to switch between display screens or datasets. This is especially useful when using a computer with more than one monitor.

2.7.14 Create New Results Viewer Window

Selecting *Create new results viewer window* from the *Display* menu will open a complete duplicate window, with all of the features and menu options. Note that this is simply a second display screen and not a new branch of the current run. Any change made to one display screen will take effect as soon as the others are updated. This is especially useful when using a computer with more than one monitor and lets the user compare different datasets of results simultaneously.

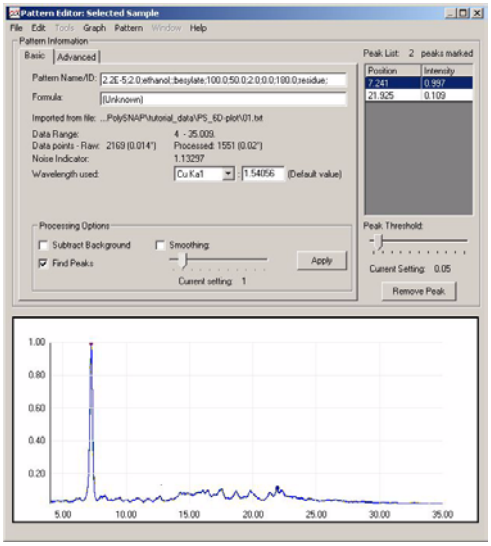
2.8 Pattern Menu

The *Pattern* menu contains a series of options relating to the sample patterns.



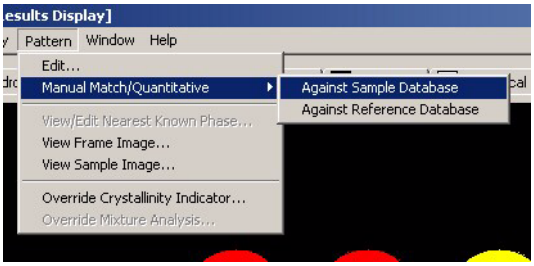
2.8.1 Edit...

Selecting this option, or clicking the *Edit Pattern* button causes the standard PolySNAP pattern editor to appear with the current selected pattern loaded into it:



Refer section 3.3.1 on page 95 for full *Pattern Editor* description and functions.

2.8.2 Manual Match/Quantitative.

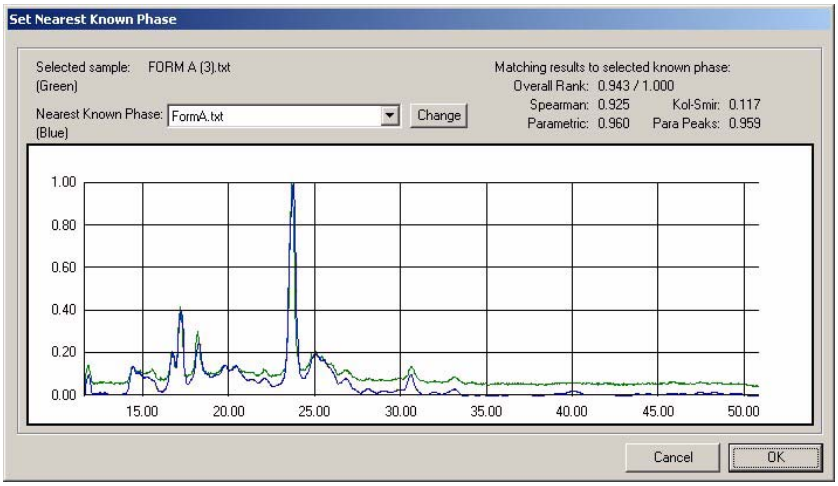


This submenu has the options for manually matching the selected sample pattern to either the rest of the sample database, or the database of known phases (references), if present.

This opens a standard PolySNAP Match window, allowing a manual comparison or analysis to be performed on the selected pattern. See section 3.4.1 on page 111 for more information on the Match window.

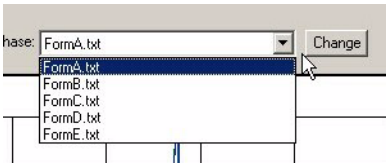
2.8.3 View/Edit Nearest Known Phase...

This option is only available if a comparison with a known-phase database was done during the analysis. If so, selecting a particular pattern and selecting this option brings up the following window:



The selected sample is plotted in green, and the known phase the program considers to be the best match to it is shown overlaid in blue. The actual matching statistics are displayed on the top right.

To see how well the sample compares to any of the other known phases provided, select the name of the phase from the pop-up list, and click *Change*:



The pattern overlay, and matching statistics, will update to represent the newly selected pattern.

To retain a changed assignation of a nearest known phase, click *OK*, and answer in the affirmative to the presented dialog box.

To keep the program-calculated nearest phase, click *Cancel.*

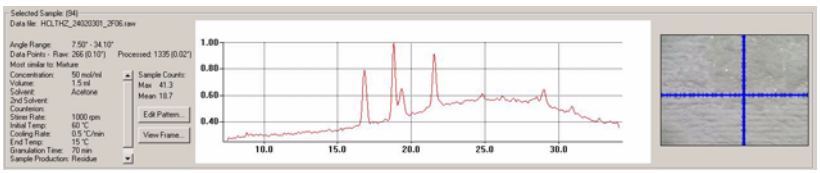
2.8.4 View Frame Image...

If the original GADDS frame file corresponding to the current pattern is available, and was saved in the same input data directory, clicking the *View Frame* button or selecting this menu item will open the Bruker GADDS software to display the frame image.

Note that this option requires both the GADDS software to be installed on the local computer, as well as a script file, *DisplayFrame.slm*, to be placed in the GADDS scripts directory. This should have been done by use of the option *Set GADDS...* in the program options *Display & Advanced* tab, but if this menu item fails to work, check that this file is correctly placed.

2.8.5 View Sample Image...

If the original sample image file corresponding to the current pattern is available, selecting this menu item will launch an external image viewer program. If the program is set to display sample images inline where possible, then clicking on the image preview has the same effect:



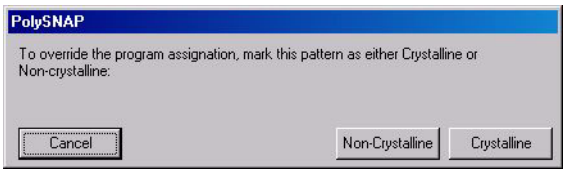
2.8.6 View Pattern Using EVA

If the current pattern was imported from a RAW format file, this option allows it to be opened in the Bruker EVA software package for

examination. Note that EVA must be set up as the default editing software for RAW files, and be installed on the local computer.

2.8.7 Override Crystallinity Indicator...

This brings up a dialog box allowing the user to manually override the program-calculated crystallinity marker setting:



The user can then choose to mark the currently selected pattern as either *Crystalline*, or *Non-crystallin*e. (The criteria for what the program considers to constitute a non-crystalline phase are defined in section 4.6 on page 150).

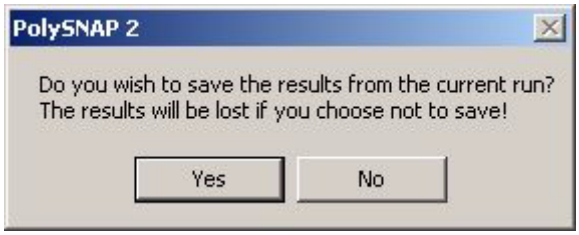
It may be useful to view a list of all samples from a current run that are marked as non-crystalline; this functionality is available from the *Tools* menu option *List Non-crystalline Patterns...* described in Section 2.6.3.

2.9 Other Menu Options

Close Window in the *File* menu closes the automatic analysis results window. The user will be asked if they wish to retain any unsaved changes to either the dendrogram, or the report.

There is also a *Save Results to Archive...* option, which allows the current results and analysis to be saved through a standard save window.

The results can then be recalled and accessed at a later date with no need to repeat all the analysis. The results are saved as a single *PolySNAP* archive file (*.psnaparchive*). *PolySNAP* will also offer the chance to save the results before *PolySNAP* is closed through a dialog box.



Note that if you do not explicitly save the results, they will be lost (although by default an archive of each run is retained in the Program Files / PolySNAP 2 / Archive folder - see Chapter 4).

The *Edit* menu, in addition to giving access to the standard text editing controls (*Cut, Copy, Paste, etc.*) also allows access to the program defaults and preferences through the *Options* item.

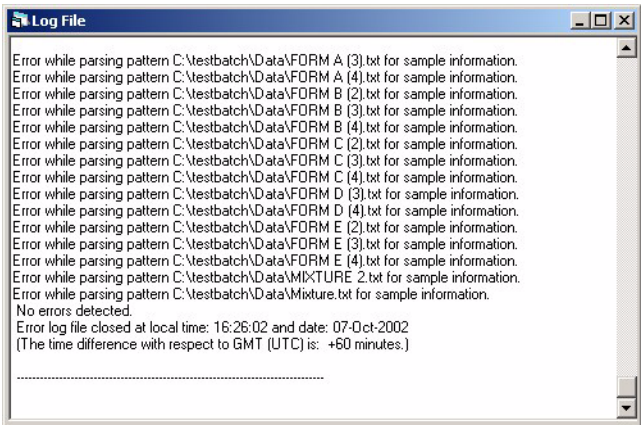
The *Window* menu contains the following items:

Window List (submenu)

This submenu lists all of the currently open windows within the PolySNAP program. Selecting a window from the list brings it to the front.

View Error Log

This option opens a new window containing an edited version of the PolySNAP logfile, which contains only lists of errors or problems that have occurred during a run of the program:



It may be a useful aid in the determination of where any problems are occurring.

CHAPTER 3

Using PolySNAP: Manual Analysis for a Single Dataset

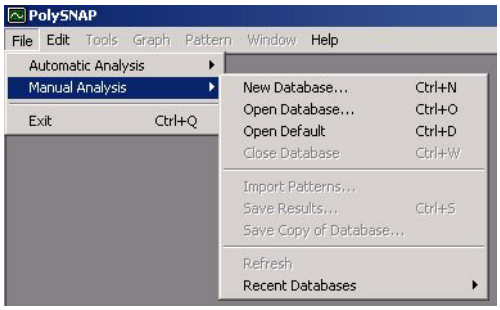
3.1 Database and File Handling

In order to use *PolySNAP 2* in manual analysis mode, the user must first either create a new, empty database and load some patterns into it, or open an existing database. This chapter describes both these processes.

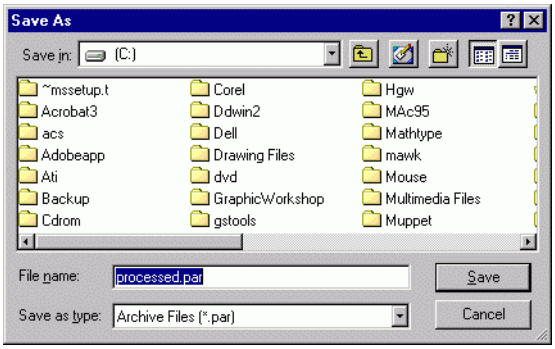
The saving procedure is different in *Manual analysis* mode than that found in *Automatic Analysis*. Instead of saving to a *PolySNAP* archive file a database (.par) file is created to work from, and remains saved after the current running of the program is closed. Once a database has been created to work from there is no need to save results to a separate location.

3.1.1 Database creation

Having launched the program, select the *Work in Manual Analysis Mode* from the *PolySNAP* welcome window. Alternatively this mode can be starting by selecting *New Database...* from the *Manual Analysis* section of the *File* menu.



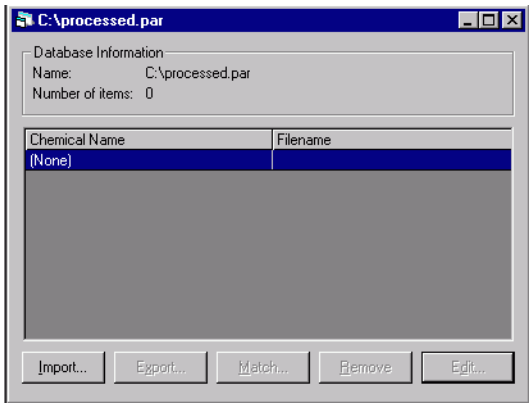
A standard Windows file dialog will appear:



This will initially suggest the filename *processed.par* and a location at the root directory of the C:\ drive. The filename and location can be changed in the normal Windows manner.

Note that the *.par* suffix indicates the database being is saved in the standard PolySNAP database format.

Once created, the database will be opened in a new window:



The top portion of the window provides some basic information on the database, including its name and location, and the number of patterns stored in it. Initially, of course, it is empty.

The central portion of the window will list the patterns included in it. It will be described in more detail on page 88.

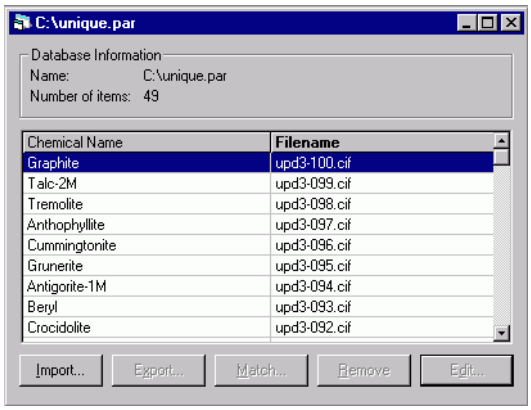
The lower region consists of 5 buttons, used to access the main functions of the program. The buttons all have corresponding entries in the program menus.

3.1.2 Opening an Existing Database

To open a pre-existing database, select *Open...* from the *File* menu. A standard Windows file browser will appear. Locate the database, and

either double-click on it, or click once to select it, and then click the *Open* button.

A database window will open, and the patterns loaded. Depending on the number of patterns in the database, this may take some time. Progress is indicated *via* a progress bar at the top of the window:

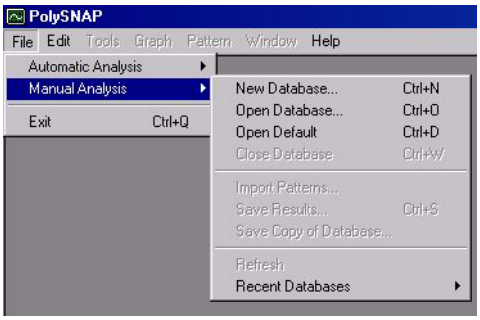


3.1.3 Default Databases

If you commonly open the same database each time you launch the program, this process can be speeded up by defining the path of your database as the program's default. This can then be opened directly by use of the *Open Default* option in the *File* menu, or by the keyboard shortcut *Ctrl-D*.

The default default database is *C:/processed.par.* This can be changed in the Program Options dialog, accessed through the *Edit* menu's *Options...* entry.

3.1.4 Recent Databases



The four most recently opened databases are stored in the *Recent Databases* submenu for easy access. Selecting one of the entries automatically causes that particular database to be re-opened.

Note that if you have moved the file since it was last opened you will have to open it manually.

3.1.5 Importing Files into a New Database

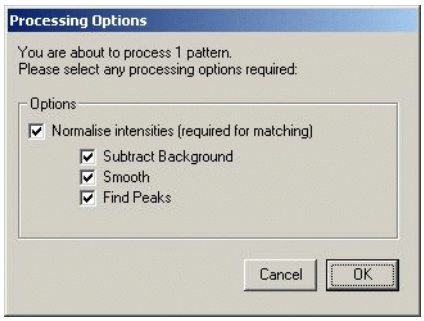
Once a new, empty database has been created and opened, the next stage is to import some patterns into it. This is achieved by clicking on the *Import...* button at the bottom of the database window (or using the *Alt-I* short-cut, or choosing *Import Patterns...* from the *File* menu).

A standard Windows file selection dialog will appear, and either single or multiple pattern files can be selected for import. Multiple files of different types can be easily selected using the shift or option keys, or by dragging the mouse across the names to select them.

Click *Open* to begin the data importation process.

3.1.6 Processing

Once files have been selected for import, a dialog box will then appear, informing you how many patterns will be imported, and offering several data processing options:



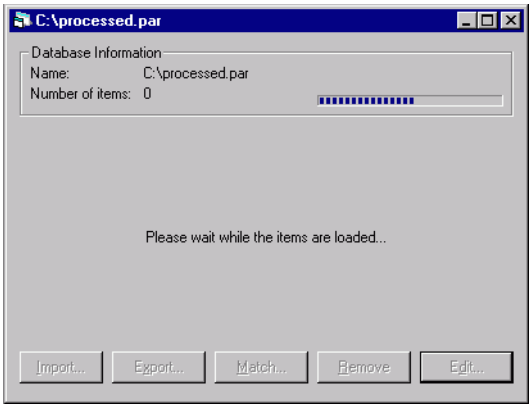
Normally, the default settings at this stage are to perform all three options:

- Background subtraction
- Smoothing and noise removal
- Peak finding.

All of the processing choices are optional, although in order to perform basic matching operations, at least the main *'Normalise intensities (required for matching)'* checkbox should be selected.

Selecting *OK* will begin the import process.

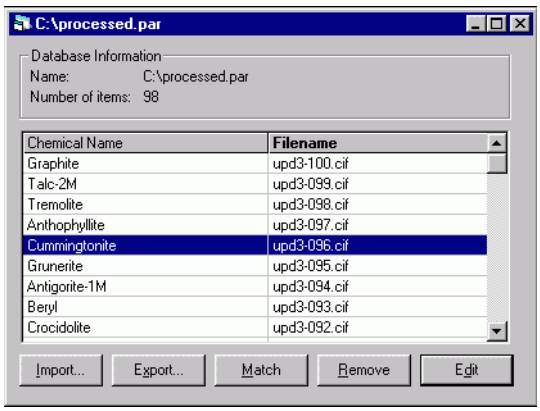
A progress bar indicates progress through the processing:



If any selected patterns cannot be loaded - for example if they are not in the correct format, or a file is corrupted, a warning will be shown:

In such a case, the rest of the selected patterns should still load correctly.

Once the import procedure is completed, the list of patterns contained in the database will appear in the centre of the database window:



Each entry in this list corresponds to one imported pattern. The left-most column contains the pattern name or ID associated with that pattern, if one is known. (This is either read from the CIF data field if present, or can be entered manually in the pattern editor - see Section 3.3.3) If no name is known, this column shows the filename.

The rightmost column lists the filename from which each pattern was loaded.

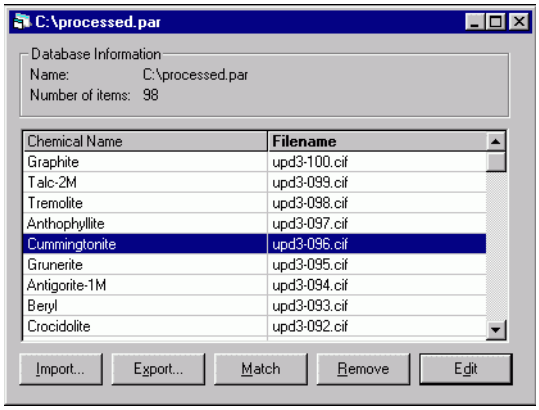
3.1.7 Sorting Database Entries

Initially, this list of patterns is shown in the order that patterns were added to the database. It is often more useful to see the list of patterns sorted in order of either chemical name/ID, or individual filenames. This is done by clicking once on the relevant column header. Clicking again on the same header will reverse the order of the sorting, *e.g.* from ascending to descending, or *vice versa*.

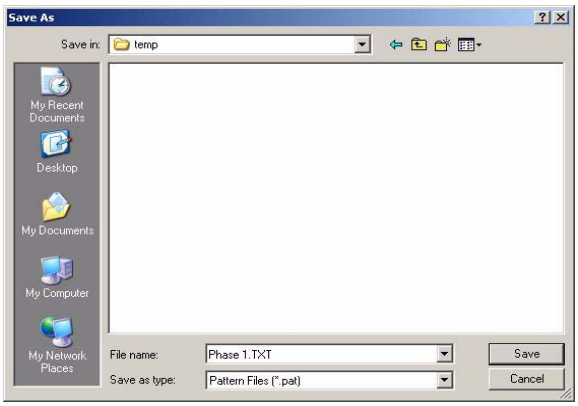
3.1.8 Exporting patterns

Individual patterns contained within a PolySNAP database can be exported to either text or pattern format files, for easy viewing or editing in other programs, or for transferring to other PolySNAP databases. This means PolySNAP can be used as a general purpose pre-processing tool for patterns that are to be used in other programs.

Click once on the pattern you wish to export. It should become highlighted in the list:



Then click once on the *Export...* button (or use the *Export...* option under the *Pattern* menu). A standard file saving dialog box appears:



Select the desired location for the file, and edit the filename if required. The default format for export is a *.pat* pattern file, but a *.txt* text file format option can also be selected from the *Save as type* drop-down menu at the bottom of the dialog box.

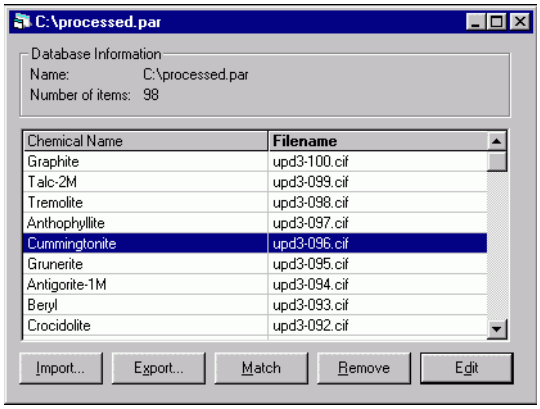
Click *Save*, and the file will be saved to the chosen location.

Note: This process saves a copy of the selected pattern to a separate file, it does not remove the original pattern entry in the database.

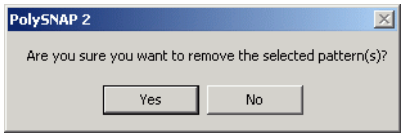
This option saves the *processed* version of the pattern (*i.e*. the version with smoothing, background subtraction *etc*. applied) if one is present, so if the raw pattern is required, remove any processing using the *Edit* window before exporting.

3.1.9 Removing Patterns

Select the pattern or patterns you wish to delete from a database. They should become selected in the list:



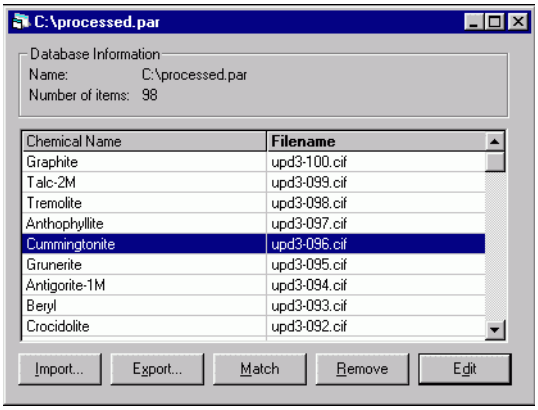
Then click once on the *Remove* button (or select *Remove* from the *Pattern* menu, or press the *Delete* key on the keyboard). A warning appears:



If you change your mind, select *No*. Selecting *Yes* removes the selected pattern(s) from the database permanently. To retrieve them, you would then have to re-import the pattern from the original data file. (The original data files are not affected in any way by the *Remove* command.)

3.1.10 Viewing and Editing Patterns

Click once on the pattern you wish to examine. It should become selected in the list:



Then click once on the *Edit* button (or select *Edit Pattern* from the *Pattern* menu).

The Pattern Editor window should appear. This is described in detail in section 3.3.1 on page 95.

3.2 Matching and Analysing Patterns

If you have a new, unknown pattern you wish to analyse or compare to entries in an existing database, the first step is to import the new pattern to the existing database.

Having done this, select it in the list of patterns, and click *Match* (or select *Open in Match Window* from the *Pattern* menu).

The matching window is described in detail in Section 3.4, Pattern Matching and Analysis.

3.2.1 Refresh Database

This feature is accessed through the menu item *Refresh* found in the *Manual Analysis* section of the *File* menu. This option updates the local copy of the database in memory by reloading all pattern information from the copy on disc.

3.2.2 Save Copy of Database

This feature, accessed through the *Save Copy of Database...* Menu item located in the *Manual Analysis* section of the *File* menu, brings up a standard Windows *Save As* file dialog. The process asks for a new database name, and copies the pattern contents of the existing

database into the new one. Note that you cannot use the existing database name as that file remains open during the process.

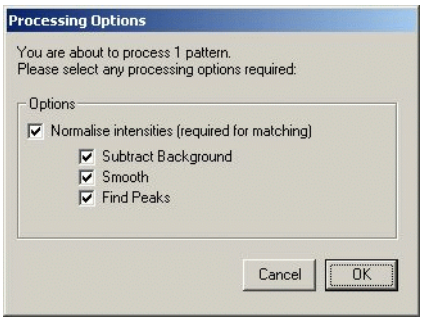
Once completed, the resulting database will be saved with the name selected and can then be opened by PolySNAP in the usual manner.

3.2.3 Apply Processing to Database

This feature is accessed through the *Tools* menu item *Apply Processing...*

If you wish to change the way in which all of the patterns are processed at once, it is possible to do so without having to change them individually, or having to re-import them from the original files.

Select *Process*, then select the new required processing options as required from the resulting dialog:

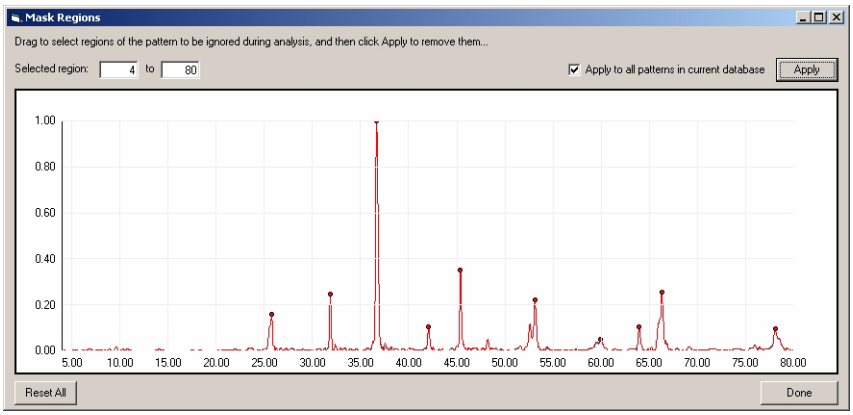


The selected processing will then be applied, which for a large database may take some time.

3.2.4 Mask Regions

This option, accessed under the *Tools* menu, is used to mask a particular region of the diffraction pattern to zero for every pattern in a database. (For more information on masking regions, section 3.3.10 on page 105)

Selecting this option brings up a window which displays the currently selected pattern:



To mask a region, either enter the start and end x-values of the area to be ignored in the text boxed on the upper left, or drag the mouse across an area of the pattern with the left mouse button held down. This will enter the start and end angles for the selected ranges in the text boxes automatically.

Click Apply. By default, the region selected will be set to zero for *all* of the patterns in the current database. Depending on the size of the database, this may take some time. To apply the masking just to the current pattern, deselect the *Apply to all patterns* checkbox.

The image of the current pattern will be updated to show the area now masked.

Multiple regions can be masked by repeating this process; click *Done* when complete.

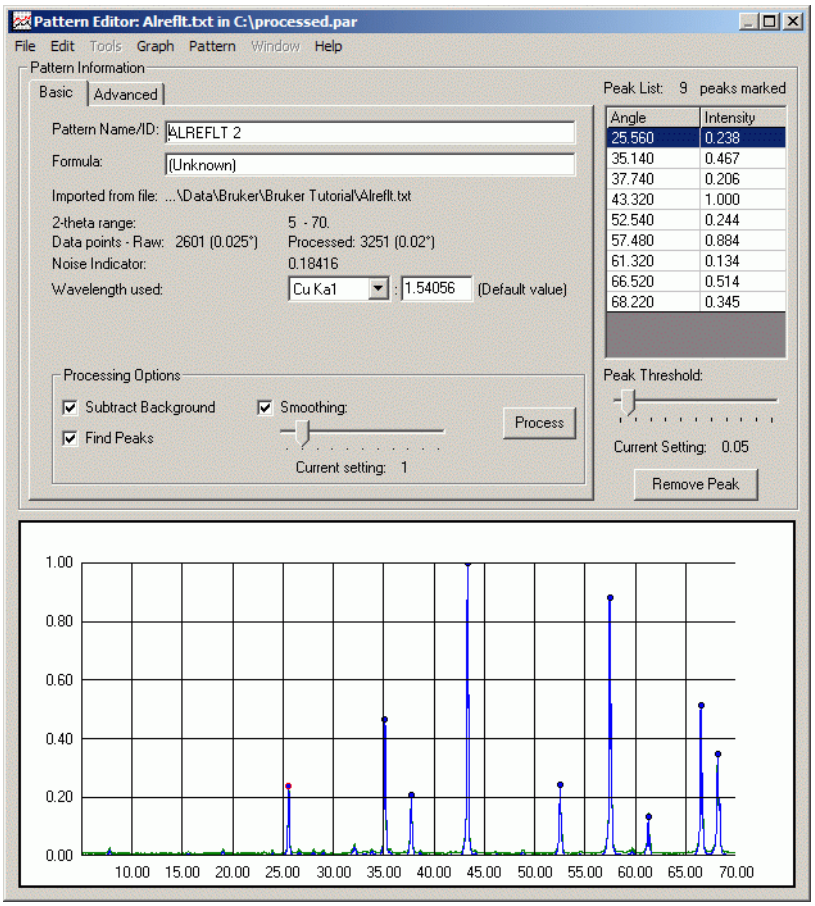
Note that applying any additional processing to a pattern after this point will remove any masked regions. Masking can also be re-applied or changed for an individual pattern in the Pattern Editor (see Chapter 3).

To remove all masking from all patterns for a particular database, select the *Mask Regions* menu option, and click the *Reset All* button in the lower left corner. Again, this may take some time.

3.3 Pattern Editing and Processing

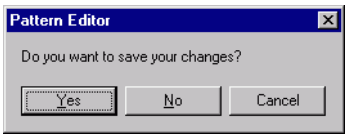
3.3.1 The Pattern Editor

After selecting a pattern in the database, and selecting the *Edit* option in the Database window, the Pattern Editor window will appear:



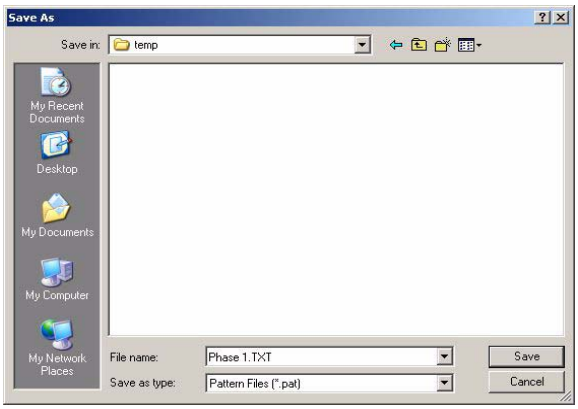
Note that while the editor window is open, no other changes can be made to the open database.

The *File* menu option, *Close Window*, closes the editor and returns to the main database window. (The same result is achieved by clicking the close box on the titlebar of the window.) If any changes to the pattern have been made, a dialog box will appear asking if those changes should be retained or not:



Selecting *Yes* saves the changes and returns to the database listing window, *No* discards any changes made and *Cancel* returns to the Pattern Editor.

One of the other *File* menu option available, *Export...*, performs the same function for the pattern currently being edited as the *Export...* button in the main database window. Selecting it causes a standard file saving dialog box to appear:



Select the desired location for the file, and edit the filename if required. The default format for export is a *.pat* pattern file, but a *.txt* text file format option can also be selected from the pop-up menu at the bottom of the dialog box.

Click *Save*, and the file will be saved to the chosen location.

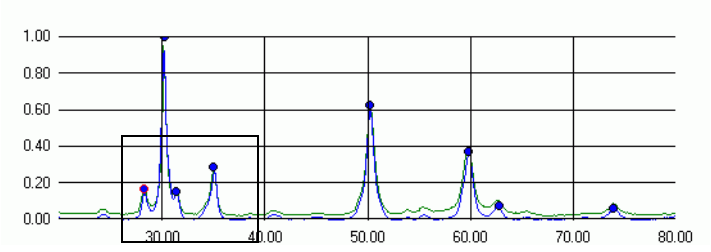
Note: This process saves a copy of the selected pattern to a separate file, it does not remove the original pattern entry in the database.

This option saves the *processed* version of the pattern (*i.e.* the version with smoothing, background subtraction *etc.* applied) if one is present, so if the raw pattern is required, remove any processing by turning off all processing options and clicking *Process* before exporting.

The *Export Peak List* option brings up a save dialog box, allowing the peak list for the current pattern to be saved to an ASCII text file for use in other programs.

3.3.2 Graph Pane

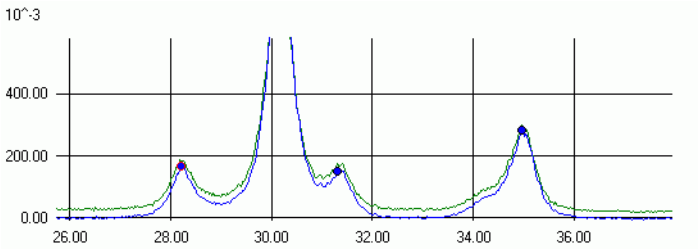
The lower part of the window, where the pattern is displayed, is called the graph pane:



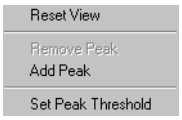
The x-axis is normally displayed as a 2θ range and the y-axis an intensity. For positive data, all values are normally scaled to between 0.0 and 1.0 for use in PolySNAP, unless no processing has been performed at all upon the data.

Individual peaks in the graph pane are marked by peak markers, which are seen as a single large dot at the top of the peak. (For more information on peak markers, See “Find Peaks” on page 102.)

Clicking and dragging to draw a rectangular box on the graph pane using the left mouse button causes the graph to zoom in into the selected region, for example:



It is possible to zoom in repeatedly using this method. Clicking on the graph with the right mouse button causes a pop-up menu to appear:



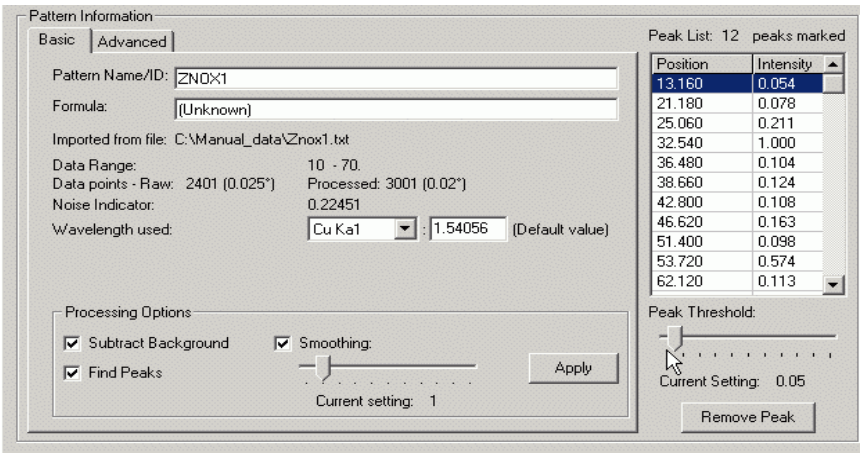
Options in this menu are:

- Reset View - selecting this after the graph scale has been changed returns to the default zoom level.
- Remove Peak - if the click was in the vicinity of a marked peak, the marker will be removed.

- Add Peak - the nearest maxima to the click location is marked as a peak on the graph.
- Set Peak Threshold - sets the y-value of the current mouse position to the minimum allowed peak height.

The same options are also available from either the *Graph* or *Pattern* menus respectively. The central region of the Pattern Editor contains the pattern information controls in two tabs: *Basic* and *Advanced.*

3.3.3 Basic Pattern Information



At the top left corner is the pattern name/ID text box. If no name has been read in for the pattern from the data file, this will normally read "*(Unknown)*". This name can be edited by clicking in the text box, and deleting or adding text as required. It is recommended that pattern names do not exceed 256 characters. Standard text-editing commands, such as *Cut*, *Copy* and *Paste* are all available from the *Edit* menu. This field is used by default to identify different patterns in the database and match windows.

Below this is the *Formula* field, which is also an editable text field. If formula information is present in the datafile, it will be read in and displayed automatically. Otherwise, it would need to be entered manually. This field is not required for operation of the program, but may be useful for reference.

Below that is a text field displaying the path and filename of the original data file the pattern was obtained from. If this path is longer than 100 characters, it will be truncated, and only the right-most 100 characters will be displayed. To see the entire path, leave the mouse cursor hovering over the text for a few seconds to see a 'tool tip' containing the full path.

Next is the range the current pattern spans on the x-axis (normally in degrees.)

Following this is the number of raw data points in the pattern: this corresponds to the number of data points read from the original data file.

The number in brackets after this is the resolution of the raw data calculated using:

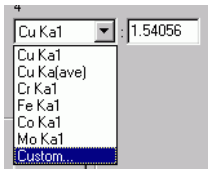
$$\frac{\text{(End angle - Start angle)}}{\text{No. of raw data points}}$$

This corresponds to the magnitude of the x-difference of the difference between two adjacent data points.

Next to this is a similar indicator for the processed data. The processing referred to is the basic processing undergone by every pattern imported into a PolySNAP database. This involves scaling the intensity data between 0.0 and 1.0, and interpolating the data (using 5th order polynomials) so that all of the patterns in the database are the same standard data resolution. This is normally 0.02 degrees per datapoint, or 50 data points per degree.

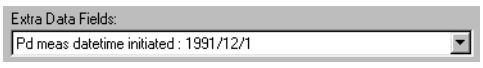
Below this is *noise indicator* for the pattern calculated in the wavelet domain, using the MAD estimator (see section 3.3.8, References). A larger number corresponds to a more "noisy" pattern.

The *wavelength* field reads in a value for the wavelength at which the data was recorded, either from a CIF or RAW data file. If another import format was used, or the information was not present in the file, the default value is used. In this situation, the legend 'Default Value' appears next to the wavelength value to warn the user that it may require to be changed manually.

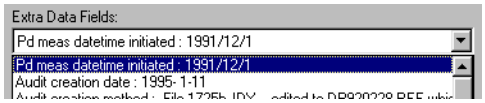


This can be done by either selecting a standard value from the scrolling list, or by choosing the '*Custom*' option at the bottom of the list of choices, and entering a numerical value in the adjacent text box. The pattern display is updated for a new standard wavelength immediately, and for a custom one after the changes to the pattern have been saved.

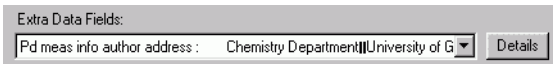
The final pattern information display only appears when the pattern is being edited was originally imported from a CIF or RAW data file.



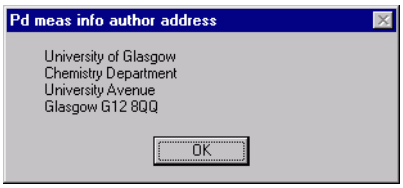
Any other data fields that were imported with the data are able to be examined *via* the list obtained by clicking on the small disclosure triangle:



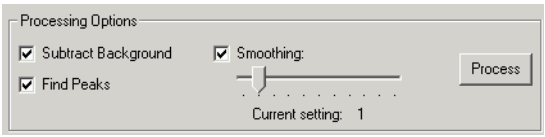
Only the first line of each data field is displayed in the list; if viewing of a multiple line field is required (for example, an author's address), select it from the list. A button will be enabled beside the pop-up region, labelled *Details*:



Clicking this produces a dialog box containing the full text of the field:

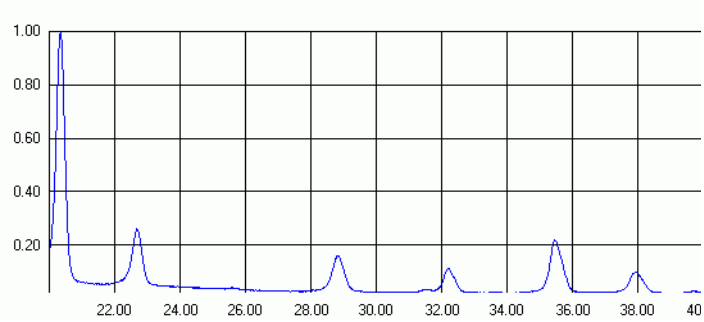


The final region of the of this pane contains the pattern processing options:



Although patterns will normally have been processed automatically upon import, it is easy to override those general processing options for an individual pattern using these controls.

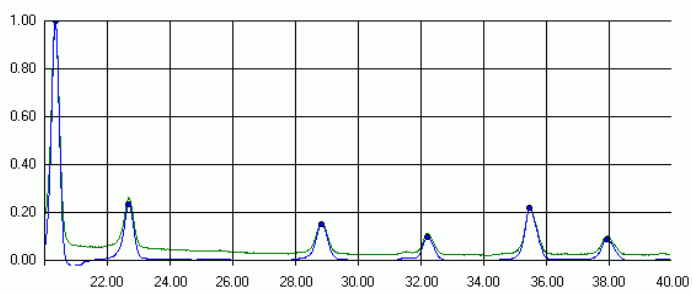
The operation of these controls maybe most easily illustrated upon a pattern that had no processing performed upon import. Such data would look something like the diagram below:



A raw data plot is generally shown in green on the graph pane when processing has been applied. Blue represents the processed profile. It is this processed profile that is used throughout the rest of the program for matching and analysis.

3.3.4 Background Subtraction

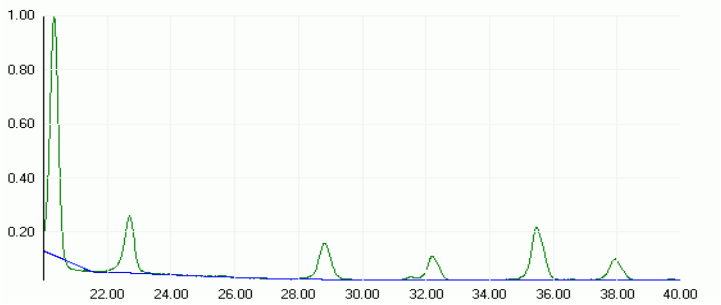
Checking the *Subtract Background* option under Processing Options, followed by clicking the *Process* button, causes a second pattern to appear on the graph pane, generally in blue:



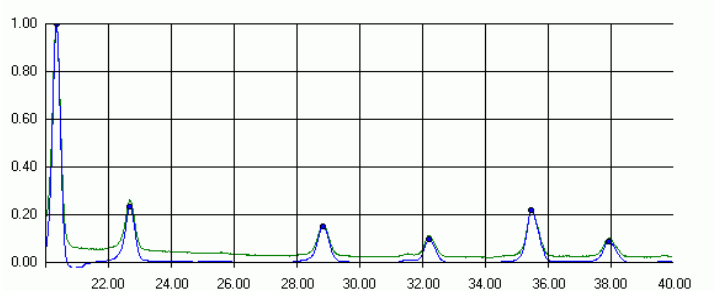
This is the same pattern as the raw data one, but with any unwanted background removed. This process is achieved by best-fitting several linked polynomials to the base of the raw data pattern, and then subtracting them from it.

If results from this are unsatisfactory, this processing option can be switched off by simply unchecking the box, and clicking the *Process* button again.

If the actual curve that has been subtracted is required to be examined, select the *Show Background Curve* option in the *Pattern* menu. A new window opens, and the calculated background region is then outlined in blue above the green raw pattern data.



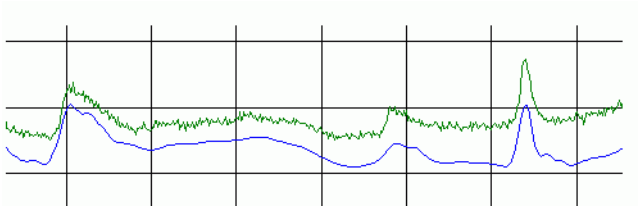
The same pattern with background subtracted:



3.3.5 Smoothing

Selecting the *Smooth* option, and clicking *Process* causes the pattern to be smoothed using a wavelet-based SURE thresholding procedure (see section 3.3.8, References). This allows obvious noise to be removed without the loss of small or fine peaks.

For example, a detail is shown below of the same region of the same pattern, before and after smoothing has taken place.



The upper green line is the original unsmoothed profile, and the lower blue is the same after smoothing has been performed.

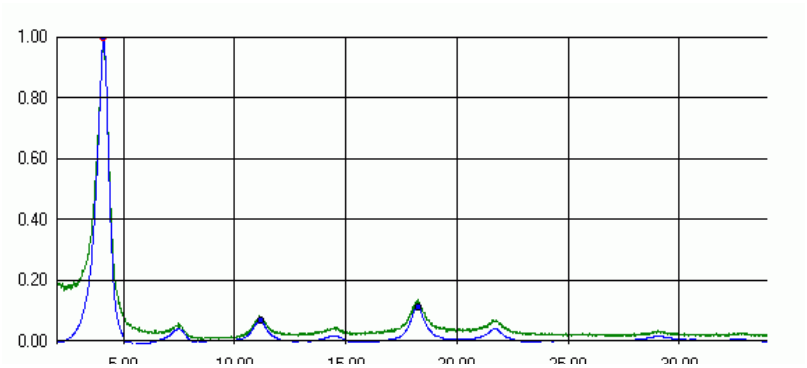
To alter the amount of smoothing applied to the individual pattern being examined, move the slider to the right to increase, or to the left to decrease it. The display is updated automatically.

The amount of smoothing performed by default can be controlled by changing the 'Smoothing Factor' value in the program *Options* dialog.

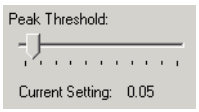
3.3.6 Find Peaks

Finally, selecting the *Find Peaks* option, followed by clicking the *Process* button tells the program to use a simple first-derivative peak search algorithm to locate the main peak maxima in the pattern. These are then marked with coloured circles on the graph pane.

For example:

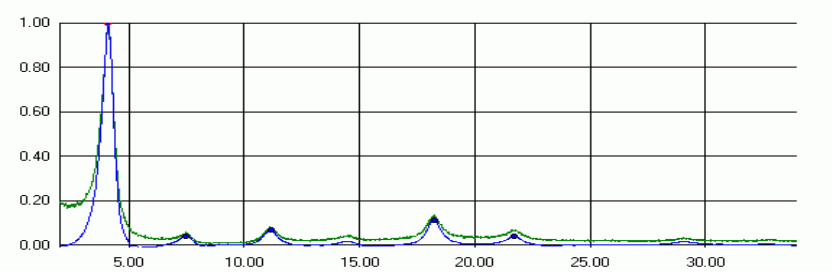


In order to stop small, unwanted peaks being marked, a peak threshold level can be set, *via* the slider on the right hand side of the window:



The default value is 0.05, meaning that no peaks will be marked below 5% of the maximum peak height. A global default value for the peak threshold can be set in the *Options* dialog (see Chapter 4.6, Matching Options). This value is used when processing multiple patterns automatically on import. Changes made to the threshold level in the pattern editor override the global value for the individual pattern in question.

For example, note the differences in the number of marked peaks for the same pattern as above, but with the peak threshold level set to zero:

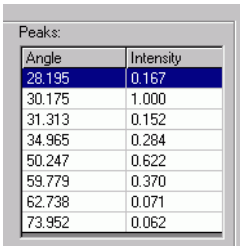


Individual peak markers can be removed by clicking on them once to select them (a differently coloured border will appear round the spot on the graph pane), followed by either clicking the *Remove Peak* button, or by selecting the *Remove Peak* option from the *Pattern* menu. After a peak is removed, the peak immediately to the right of it becomes automatically selected.

It is also possible to remove a peak by clicking on its entry in the Peak List box to select it, and then clicking the *Remove Peak* button.

The number of peaks field reports how many points in the pattern have been marked as peaks. These are shown by round dots on the profile displayed on the graph pane.

A scrolling list of all the marked peaks showing their x-position and corresponding y-value is displayed on the right-hand side of the pattern editor window:



Clicking on a particular peak in this list causes it to become selected, and the corresponding peak marker to be highlighted on the pattern itself to aid in identification.

3.3.7 Notes on processing

Although the processing options (such as subtracting the background) can be made on a pattern by pattern basis, best results are obtained in the matching and analysis sections of the program with databases whose constituent patterns are all processed in a consistent manner. For example, two otherwise identical patterns may appear very different if one has been smoothed and background-subtracted, and the other has not. Consistency in how patterns are processed generally leads to the best matching and analysis results.

3.3.8 References

Wavelet SURE thresholding is discussed in:

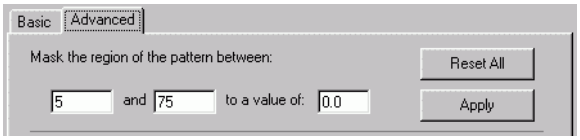
Donoho, D.L.; Johnstone, I.M.; (1994). Adapting to Unknown Smoothness via Wavelet Shrinkage. *Technical Report, Department of Statistics*, Stanford University.

3.3.9 Advanced Pattern Information

The advanced tab has two main regions: one allowing the ability to mask sub-regions of the pattern, and the other to enter additional information and thus calculate various constants for the pattern.

3.3.10 Masking

The upper section in the advanced pane contains the controls to mask selected regions of a pattern.

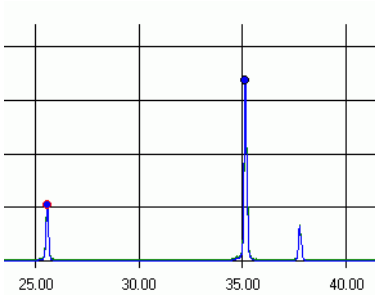


If a pattern contains a spurious peak or unwanted standard, or for some reason the user wishes to only compare a certain region from the pattern, this option allows a chosen sub-range to be set to a particular value. (This is generally zero, although non-zero values may be useful when a high background signal is present.)

To use this option, first decide on the angle range to be masked. Enter the start and end angles in the relevant boxes. The value in the magnitude box sets the signal to that value over that range.

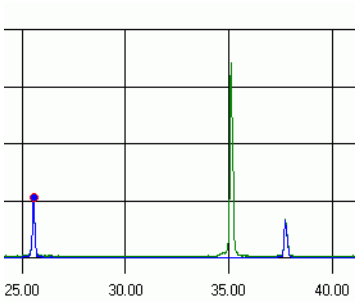
Masked regions can also be applied in one step to every pattern in a database using the *Mask Regions...* option. This can be found under the *Tools* menu, when the database window is frontmost. See Section 3.2.4, Mask Regions.

In the example below, the user wishes to mask the central of the three main peaks:



As it is at approximately 35° 2θ, the start angle could be 34°, and the end angle 36°. There is no noticeable background, so the mask value can remain as zero.

Clicking the *Apply* button results in the following:

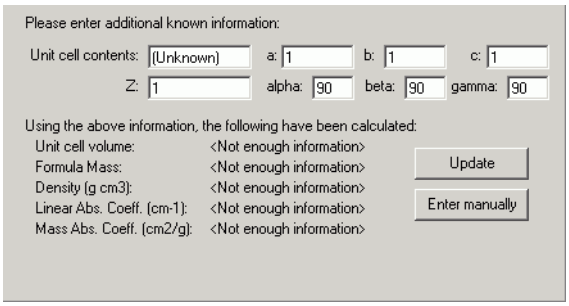


Notice that the peak is still visible in green. This shows that the original data - the unprocessed pattern - has not been modified. However, there is no peak visible in blue - it has been removed from the processed pattern. It will not appear in the Match window or other regions of the program which only make use of the processed pattern profile.

Multiple regions in the same pattern can be masked by repeated application of the same technique.

To reset all masking, and make the processed profile available again, click the *Reset All* button. The same effect is obtained by re-processing the pattern using any of the standard controls in the Basic tab.

3.3.11 Additional Pattern Information



This section of the pattern editor contains fields for entering other information that is known about the sample.

This information is used to calculate various useful constants and numbers about the pattern. The information calculated here is required if you wish quantitative analysis to return results as a weight fraction.

The first time the Advanced tab is used for a pattern, the *Unit Cell Contents* field is taken to be the same as the *Formula* field from the Basic tab. However, they are saved separately in the pattern, and changes to one do not affect the other. The *Formula* field is

completely free format, whereas the *Unit Cell Contents* field has a fairly strict format, as described below.

The unit cell contents should be entered in the form:

<Atom Symbol><No. of that atom><Space or comma><Next atom symbol ... etc.

For example, the following entries would be valid:

C3 H8 02
C3,H8,O2
Al O3 H
c3 H8 o2

Non-integer values are permitted.

The following entries would not be valid, for the reasons described:

C 3 H 8 O 2

-has spaces between the element types and the number.

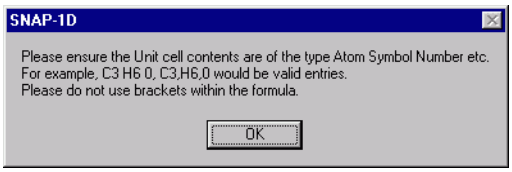
C3H8O2

-has no spaces at all between each separate entry!

Al(OH)3

-Contains brackets.

Any invalid entries should result in an error:



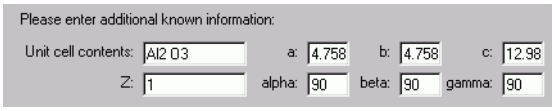
Note however that it is possible to enter a string that does not generate an error message, but that does not produce correct results. *For this reason, entries in this field should be checked carefully.* Also please note that entries are **not** checked for chemical sense during calculation.

The field *Z* represents the number of formula units per unit cell. It should be a numerical value greater than zero. It can be non-integer.

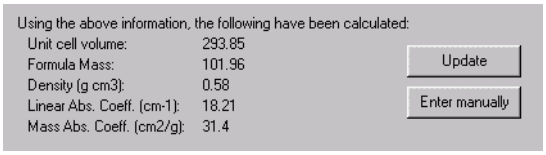
The fields for unit cell dimensions (a, b, c, α, β, γ) all expect numerical non-zero entries. The angles should be in degrees, and the lengths in angstroms (Å).

Once all known information is entered, click the *Update* button. Depending on what information was provided, some or all of the text fields at the bottom of the pane will update with new values.

For example, with the following information entered:



Clicking *Update* results in the following results:



As in this case all of the fields have been successfully calculated, this pattern could potentially be used as part of a quantitative analysis resulting in a weight fraction being calculated.

3.3.12 Additional Calculation Information

For reference, the items above are calculated as follows:

- Unit Cell Volume:
  Given unit cell lengths a, b and c in Å, the volume *V* is then:

  $$V = abc[1-\cos^2\alpha - \cos^2\beta - \cos^2\gamma + 2\cos\alpha\cos\beta\cos\gamma]^{1/2}$$

- Density:
  Given the molecular weight *MW* in atomic mass units, the number of formula units per unit cell *Z*, and cell volume *V* in cubic angstroms, the density ρ may be calculated by:

  $$\rho = \frac{MW \times Z \times 1.66}{V}$$

- Linear Absorption Coefficient:
  Given the cell volume V_c, and the total photon interaction cross-section σ_i in barns/atom for each atom in the unit cell, the linear absorption coefficient μ is calculated by:

  $$\mu = \frac{1}{Vc}\sum_{i=1}^{n}\sigma_i$$

- Mass Absorption Coefficient:
  This is calculated by dividing the linear absorption coefficient μ by the density, ρ:

$$\mu^* = \mu / \rho$$

- The database of atomic information is contained in the file *SNAPdb.dat*, which should be located inside the folder containing the *SNAP.exe* program itself. The format is discussed below:

3.3.13 Atom Info Database Format

There are 11 lines per entry:

Line 1: Format(a3,i4,2x,a12)
Element symbol, charge or oxidation state, full element name

Line 2,3: Format(6f12.6/6f12.6)
X-ray atomic scattering factors in parameterised form:

a1	a2	a3	a4	a5	c
b1	b2	b3	b4	b5	0

Sources:
- Fit parameters of all atoms/ions (with the exception of O1-) from publication "New Analytical Scattering Factor Functions for Free Atoms and Ions", D. Waasmaier & A. Kirfel, *Acta Cryst*. 1995, A51, 416-431)

- Fit for O1- based on the tabulated values of Table 2 (D.Rez, P. Rez & I. Grant, *Acta Cryst*. (1994), A50, 481-497).

- Fits for all other atoms/ions based on the tabulated values of Table 6.1.1.1 (atoms) and Table 6.1.1.3 (ions) (*International Tables for Crystallography, Vol. C*, 1992)]
f(atom)=a(1)exp[b(1)]+a(2)exp[b(2)]+a(3)exp[b(3)]+a(4)exp[b(4)]+a(5)exp[b(5)]+c]

Line 4: Format(f12.6)
Neutron scattering cross section.

Line 5: Format(i6)
Flag for the source of the electron scattering factor.
[In some cases, values for atoms have been 'filled in' for ions. In this case the key is the same except the flag is negative (e.g. flag is -1 if have used scattering factors for a neutral atom from Smith & Burge for an ion).]

Line 6,7: Format(5f12.6/4f12.6)
Electron scattering factors in parameterised form:

a1 b1 a2 b2 a3 b3 a4 b4 c

f(atom)=a(1)exp[b(1)]+a(2)exp[b(2)]+a(3)exp[b(3)]+a(4)exp[b(4)]+c

Line 8: Format(7e12.4)
μ (for the following radiations:
Ti,Cr,Fe,Co,Cu,Mo,Ag
All ions have the same μ value as the neutral element. μ values are taken from *International Tables Volume C* 193-199

Line 9: Format(7f12.6)
df' for the following radiations:
Ti,Cr,Fe,Co,Cu,Mo,Ag

Line 10: Format(7f12.6)
df" for the following radiations:
Ti,Cr,Fe,Co,Cu,Mo,Ag
All ions have the same df' and df" as the neutral element. df' and df" are taken from *International Tables Volume C* 219-221.

Line 11: Format(2f12.6)
Covalent radius, van der Waals radius.

Note: Many Covalent radii and most van der Walls radii are approximate. The ions are given the same van der Walls radii as the neutral atoms. For heavy atoms there is very little knowledge of these parameters

In addition to the items specified above, information is taken from the following sources:
1. “New Analytical Scattering Factor Functions for Free Atoms and Ions”, D. Waasmaier & A. Kirfel, (1995) *Acta Cryst.* A51, 416-431.

2. Rez, D, Rez, P. & Grant, I, (1994) *Acta Cryst.* A50, 481-497.

3. Smith, G.H. & Burge, R.E. (1962) *Acta Cryst.* 15, 182-186.

4. Doyle, P.A. & Turner, P.S. (1968) *Acta Cryst.* A24, 390-397.

5. Jiang, J.S. & Fang-Hua, L. (1984) *Acta Physica Sinica* 33, 845-849

6. A Bondi, (1963) *J. Phys. Chem.*, 68, 441-51.

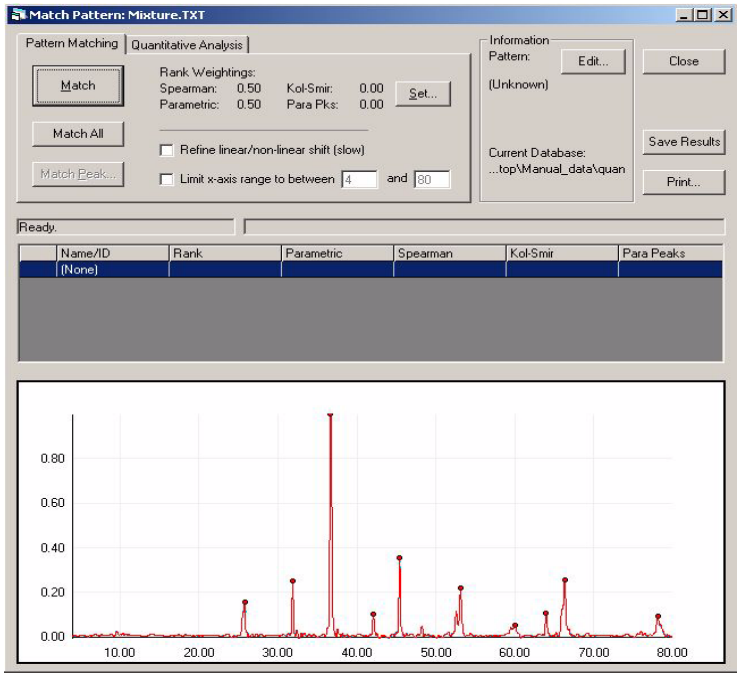
Note: The 'Format' described above refers to Fortran-style fixed format numbers. For example if the format is (3I4,2F8.2), the program expects to find 3 integers, each of 4 characters, followed

by 2 floating point numbers, each of which total width 8 characters (including the decimal point, with 2 places after it).

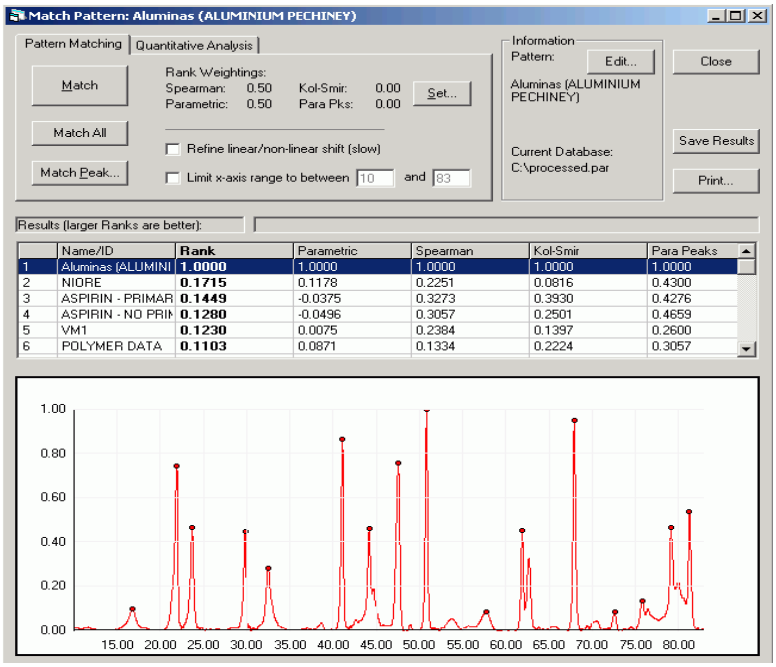
3.4 Pattern Matching and Analysis

3.4.1 The Match Window

After selecting a pattern, and clicking the *Match* option, the Pattern Matching window appears. At first the display is mostly empty like so:



But once the matching analysis has taken place (this can be done in different ways, described later) the window changes to display the match results.:

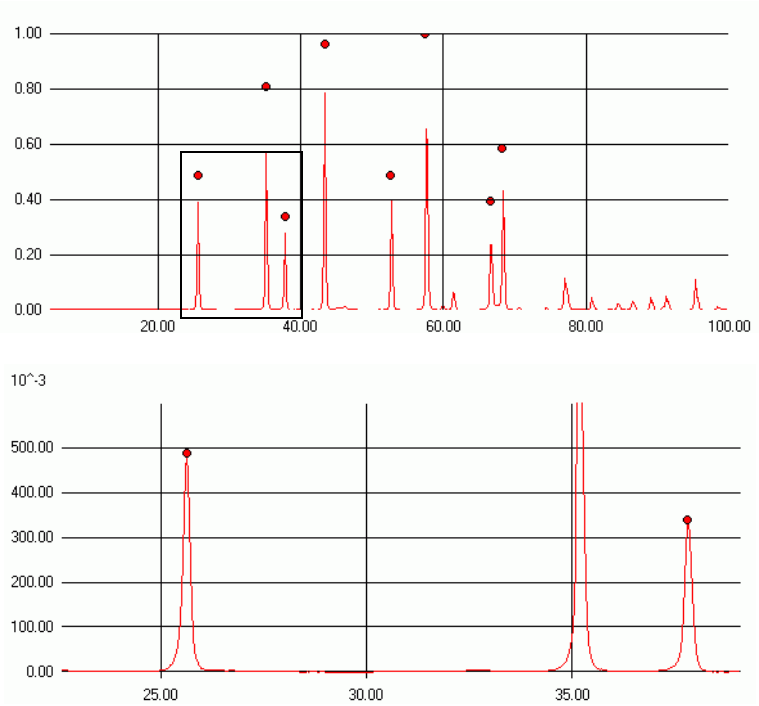


The two main active modes of the program are accessed from this central window.

3.4.2 Graph Pane

The pattern being matched appears in red the graph pane in the lower part of the window. It can be zoomed in and out in the same manner

as in the Pattern Editor, by using the mouse to drag a rectangle over the relevant area:



3.4.3 Results Pane

The central part of the window is a scrolling list where any matching results are displayed. This results pane contains, from left to right, the following fields:

	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	ASPIRIN - NO PRII	**2.0000**	1.0000	1.0000	1.0000	1.0000
2	ASPIRIN - PRIMAF	**1.8248**	0.9217	0.9030	0.3012	0.9258
3	SAMPLE H	**0.3602**	0.0822	0.2780	0.2615	0.2506
4	ALREFHT	**0.3449**	0.2667	0.0783	0.0060	0.4468
5	Aluminas (ALUMINI	**0.2890**	-0.0267	0.3157	0.1729	0.3824
6	SAMPLE EB	**0.2068**	-0.0137	0.2205	0.0733	0.4255

The first field contains the number of the pattern, where No. 1 is at the top of the list. Clicking once in this leftmost column for a particular pattern entry results in a coloured dot appearing, and the corresponding pattern being shown in the graph pane, superimposed upon the matched pattern. The spot is the colour the pattern is displayed on the graph pane.

Clicking the same column again removes the pattern from the graph pane.

The next field is the *Name/ID* field. This normally shows the chemical name, but if this is "(Unknown)", it shows the short-filename of the pattern instead.

Next is the *Rank* value calculated for each pattern. This is obtained by summing the results from the four different matching tests, in a user-defined manner. By default the two are combined in a one to one ratio - see Section 3.4.8.2, Rank Weightings.

The *Parametric* field displays the result of the parametric correlation coefficient.

The *Spearman* field displays the result of the Spearman non-parametric correlation coefficient.

The *Kol-Smir* field displays the result of the Kolmogorov-Smirnov non-parametric statistics test.

The *Para Peaks* field displays the result of the parametric correlation coefficient, applied to marked peak regions only.

These individual tests are described in detail later.

The first two tests are applied to the entire full profile (*i.e*. every single datapoint) of each entry in the database. Consequently, they do not require any peak markers to have been located in the individual patterns. The last two tests are performed only upon individual marked peaks that coincide between the sample and database patterns.

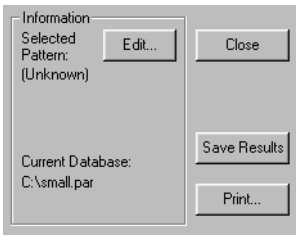
3.4.4 Status Bar

Above the results field is the Status Bar display:

The left side of this displays any status messages or errors. The right hand side consists of a progress bar. This indicates how far through the matching process the program is at any one time.

3.4.5 Mode-independant Controls

The upper-right portion of the window contains the matching controls. These are common to all of the operational modes:



Information display*:*

The top section contains the chemical name for the selected pattern (if known). The *Edit* button allows direct access to the Pattern Editor for the current pattern. This means that, for example, peaks could be added or removed at this stage.

The bottom section shows the name and location of the database the user is matching the unknown pattern against.

If either of these are too long to be displayed in the space, then the details are truncated, but the full information can be accessed by resting the mouse cursor over the appropriate region for a few seconds to reveal a 'tool tip' containing the full details.

On the far right hand side are the window controls:

Close - closes the window and returns to the database display. If any matching has been done, and results from it have not been saved, then the program will check if the user wishes to save them to a file.

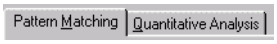
Save Results - once a matching calculation has taken place, this option allows the tabled results to be exported to a either an ASCII text or Microsoft Excel compatible file. Enter the filename and select the filetype from the list in the standard File Save dialog box. This option is also available from the *File* menu. (*N.B.* only the results from the most recent matching run will be saved by this method, hence separate saves would be required to keep results from multiple runs.)

Print - brings up a standard Windows Print dialog, and causes the current contents of the graph pane to be sent to your selected printer. This option is also available under the *Graph* menu.

Note that printing to a postscript printer may be very slow as every single datapoint (usually several thousand) is sent to the printer. Much faster results can be obtained with a non-postscript printer, for example an inkjet.

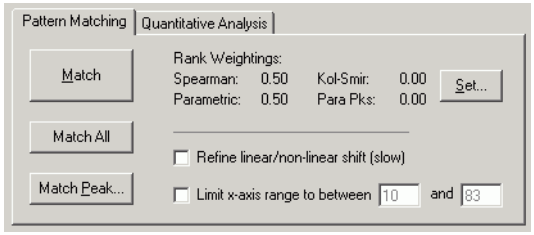
The remaining controls in the window are the mode-dependant controls.

The user can switch between the two matching modes by means of the tabs in the top left of the window:



3.4.6 Pattern Matching

The first tab represents the Pattern matching program mode:



This process involves the repeated comparison of pairs of datasets, the *sample* and the *database* patterns.

The sample pattern is the pattern that is being compared to the rest of the patterns in the database. For each pattern, a comparison is made as follows:

1. The full profiles of the patterns are compared on a point by point basis using the non-parametric Spearman rank-order coefficient test. The result from this appears in the results column headed '*Spearman*'. A score of 1.0 represents a perfect match, 0.0 the worst possible match, and -1.0 a negative correlation (see Section 3.6, References).

2. A parametric Pearson equivalent of the Spearman test is then applied, both to the full intersecting profile (results shown in column *Parametric*), and to individual corresponding peaks (results shown in column *Para Peaks*).

3. If any peaks have been marked in the sample and database patterns, the probability of correlation between the two is also calculated using the Kolmogorov-Smirnov test (see Section 3.6, References).

4. The range of each peak to be tested is taken to be the intersection of the two peaks ranges, calculated by tracing their shoulders until either the intensity falls below a set threshold, or the intensity starts to increase again.

5. The pattern with the greater number of peaks is taken as a reference, (its number of peaks = *m*).

The Kolmogorov-Smirnov test is then performed on each of these peaks, and a probability value *p* returned for each one. This is 1.0 where peaks are identical and zero when a peak is matched against no peak.

Then K-S value for the overall pattern is:

$$(p_1+p_2+p_3+\ldots+p_m)/m$$

which returns a value between 0 and 1.0. Note that the test results for each individual peak can also be examined in the Single Peak Viewer - see Section 3.4.8.6, Match Peak.

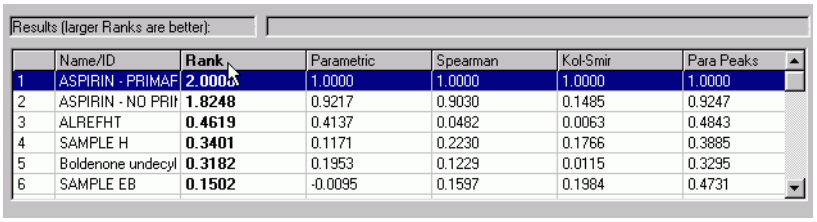
6. Results are reported in the *Kol-Smir* column, where a result of 1.0 corresponds to a perfect match. This is also true of the *Para Peaks* result, which is calculated in a similar manner.

7. Finally, a *Rank* value is calculated for each database sample after comparison, comprised of a weighted mean of each of the available statistics. These weights are user-definable - see Section 3.4.8.2, Rank Weightings.

3.4.7 Non-quantitative Peak Matching:

Click the *Match* button (or press F1). Matching will commence. Progress is indicated by the progress bar and status bar:

Once matching has completed, the results will be displayed in the results pane:



	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	ASPIRIN - PRIMAF	**2.0000**	1.0000	1.0000	1.0000	1.0000
2	ASPIRIN - NO PRIM	**1.8248**	0.9217	0.9030	0.1485	0.9247
3	ALREFHT	**0.4619**	0.4137	0.0482	0.0063	0.4843
4	SAMPLE H	**0.3401**	0.1171	0.2230	0.1766	0.3885
5	Boldenone undecyl	**0.3182**	0.1953	0.1229	0.0115	0.3295
6	SAMPLE EB	**0.1502**	-0.0095	0.1597	0.1984	0.4731

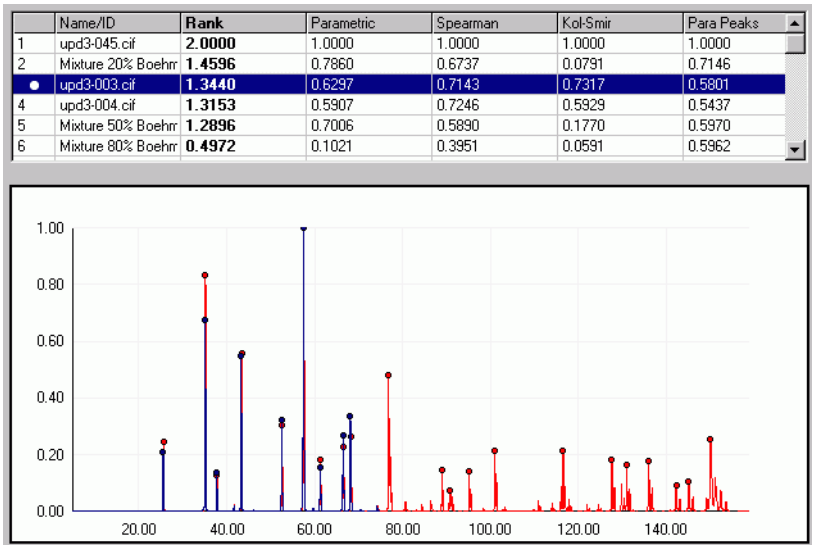
The best overall match is shown at the top of the list, which is sorted by default in descending order of Rank. The best match is always the pattern you are matching against. This provides a useful check that everything is operating correctly, since the unknown pattern matched against itself should give a perfect match in all four tests.

i.e. it should have a

Spearman of 1.0
Parametric of 1.0
Kol-Smir of 1.0
Para Peaks of 1.0

and a rank (calculated from 0.5×Spearman + 0.5×Parametric) of 1.0 (assuming default weightings).

To view a pattern overlayed on the matched pattern, click once in the leftmost column of that pattern (the third row in the example below):



	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	upd3-045.cif	**2.0000**	1.0000	1.0000	1.0000	1.0000
2	Mixture 20% Boehm	**1.4596**	0.7860	0.6737	0.0791	0.7146
●	upd3-003.cif	**1.3440**	0.6297	0.7143	0.7317	0.5801
4	upd3-004.cif	**1.3153**	0.5907	0.7246	0.5929	0.5437
5	Mixture 50% Boehm	**1.2896**	0.7006	0.5890	0.1770	0.5970
6	Mixture 80% Boehm	**0.4972**	0.1021	0.3951	0.0591	0.5962

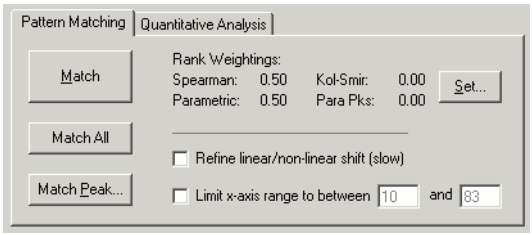
Clicking once in this column for a particular pattern results in a coloured dot appearing, and the corresponding pattern being shown in the graph pane, superimposed upon the matched pattern using the colour of the spot.

Clicking the same column again removes the pattern from the graph pane. Up to seven different patterns can be displayed at once in this manner. To remove multiple displayed graphs in one step, a *Remove Overlaid Graphs* option is available from the *Graph* menu, or the pop-up menu obtained by right-clicking on the graph region.

The sort-order in which the results are displayed can be changed by clicking the header of the column on which you wish the results sorted. Clicking the same header again causes the results to be sorted in that same column, but in the opposite order *i.e.* repeatedly clicking on one header causes the sort order to switch between ascending and descending.

If the top few patterns appear to be a really bad match, the chances are that the list has been sorted in descending order. Click again to re-sort.

3.4.8 Advanced Options



3.4.8.1 *Refine linear/non-linear offset*

Selecting the *Refine linear/non-linear offset* checkbox causes an 2θ-axis offset to be computed when comparing patterns.

For example, if two patterns are very similar, except that one is offset by +0.3 degrees along the 2θ-axis from the other, selecting this option should reveal this.

The maximum amount by which patterns are allowed to be offset can be changed from the default of 0.4 degrees by changing the values in the *Advanced* pane of the program *Options* dialog box. The program attempts to maximise the correlation between patterns by varying the values a_0 and a_1 in the equation $\Delta 2\theta = a_0 + a_1 \sin \theta$, although $\cos \theta$ or $\sin 2\theta$ may optionally be used instead.

After turning the option on, click *Match* to begin the matching. This process will be much slower than the default matching, as many more calculations are being performed for each pattern.

Once complete, the results are displayed as before, with one difference. There is now an additional column on the far right - *Offset used*:

	Name	Rank	Parametric	Spearman	Kol-Smir	Para Peaks	Offset used
1	Aluminas (ALUM	**1.0000**	1.0000	1.0000	1.0000	1.0000	0, 0
2	ASPIRIN - PRIM	**0.2191**	0.0149	0.4233	0.3930	0.4276	-0.4, 0.38
3	ASPIRIN - NO P	**0.2109**	0.0101	0.4117	0.2501	0.4659	-0.4, 0.23
4	NIORE	**0.1867**	0.1792	0.1942	0.0816	0.4300	-0.25, 0
5	FRANK	**0.1399**	0.1709	0.1089	0.1455	0.2872	-0.4, 0.4
6	VM1	**0.1383**	0.0193	0.2574	0.1397	0.2600	-0.4, 0.38

This corresponds to the amount of offset (a_0, a_1) required to maximise the correlation coefficients for each pattern.

As a shift of up to about one degree can be possible when dealing with patterns of the same substance recorded with different experimental conditions, and this option may be useful to reveal matches between patterns that at first appear to be different.

Note that at present the shift only applies to the profile based tests and that this option can slow down the operation of the program substantially.

3.4.8.2 Rank Weightings

For patterns of a particular type, it may be the case that one particular correlation coefficient proves to be more useful than others.

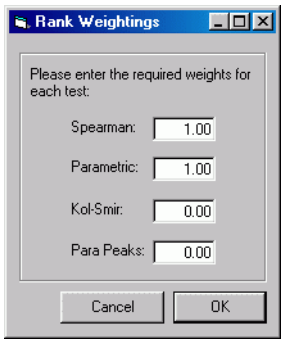
In order to relate this fact to the overall ranking of pattern matches, it is possible to change the way the overall rank is calculated.

Normally, this is of the form:

$$Rank = w_1 \times (Parametric\ score) + w_2 \times (Spearman\ score)$$

$$+ w_3 \times (Kol\text{-}Smir\ score) + w_4 \times (Para\ Pks\ score)$$

where w_1 and w_2 are the rank weightings, normally both 1.0. By varying these weightings.



It is possible to alter the overall rank order of the matched patterns, for example:

Once the weighting have been altered, the *Apply* button is activated. Clicking it applies the new rank weightings to the current list of matched results:

	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	Corundum.txt	**3.0000**	1.0000	1.0000	1.0000	1.0000
2	upd3-003.cif	**2.3966**	0.7482	0.8242	0.1534	0.9124
3	upd3-004.cif	**2.3810**	0.7426	0.8192	0.1700	0.9047
4	upd3-045.cif	**2.2111**	0.5541	0.8285	0.0758	0.7421

If rank weightings are changed from the default values before a match has taken place, the new values are used automatically when the results are displayed.

3.4.8.3 Limit x-axis range...

Turning on the *Limit x-axis range* checkbox has the effect of limiting the analysis calculation to a subset of the entire x-axis range of the unknown pattern. This can be useful if a particular feature of the pattern is causing problems for the calculation.

The default values in the textboxes are the full range of the unknown pattern, although these can be changed by either:

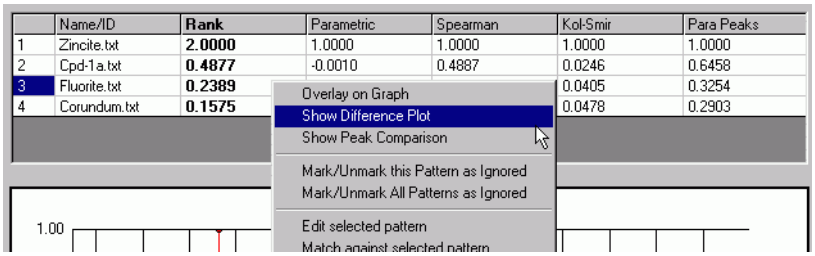
1. Typing in the desired start and end angles of the required range into the textboxes, or:

2. Zooming in to the desired range by clicking and dragging a zoom-rectangle on the graph pane. The start and end ranges of the 2θ-range of the zoomed region of the pattern will be automatically placed in the textboxes.

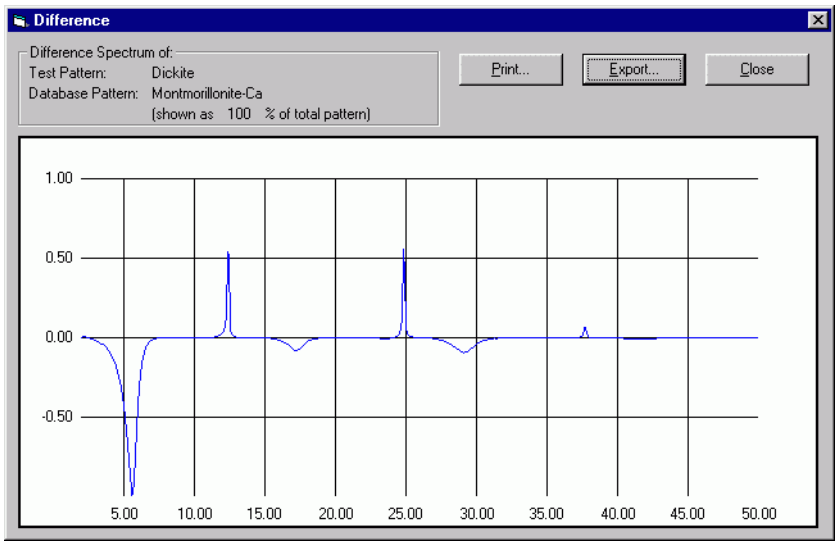
To use this feature, turn the checkbox on, enter the desired start- and end- details, and click the *Calculate* button.

3.4.8.4 Difference Plots

When comparing two very similar patterns it can be useful to get a different kind of visual feedback as to their similarity. One such feedback method is the difference plot, which subtracts the selected database pattern from the unknown pattern and displays the remainder. This feature is accessed by first clicking once to select a database pattern from the list of match results, and then either selecting *Show Difference Plot* from the *Pattern* menu, or clicking once with the right-hand mouse button to access the results pane pop-up menu:



Selecting the *Show Difference Plot* option brings up the following window:



The *Print* button outputs the graph pane to the local Windows printer.

The *Close* button returns to the main Match window.

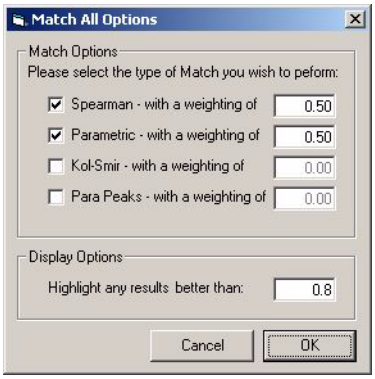
The *Export* button saves the difference profile to an ASCII text file.

3.4.8.5 Match All

This option is only available for a maximum of 1000 patterns in a database. It matches every pattern in the database against every other pattern in one step, and displays a matrix of results.

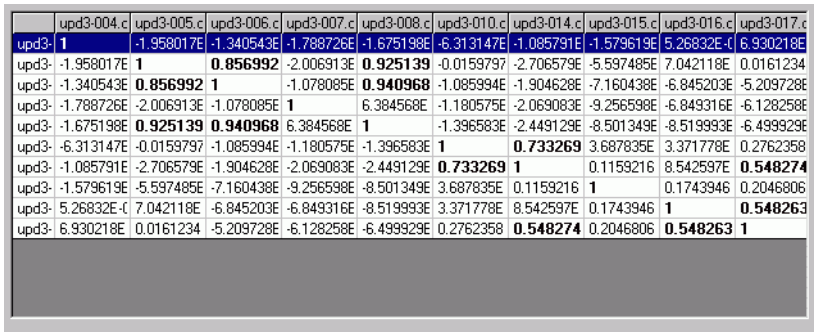
Please note that with this large number of patterns the matching process will take some time to complete, and this will be exacerbated if the allow x-offset option is enabled.

Clicking the *Match All* button produces a dialog in which to select the individual tests to be used:



Select the tests of choice, their required weightings and enter the highlight cut-off value (any results greater than this value will be highlighted in the results table).

Clicking *OK* initiates the matching process. This may take some time, as $n^2/2$-n matching tests are being performed (where n is the number of patterns in the database). When over 50 patterns are being matched, a window with a progress bar is displayed. When complete, the results are displayed as follows:



	upd3-004.c	upd3-005.c	upd3-006.c	upd3-007.c	upd3-008.c	upd3-010.c	upd3-014.c	upd3-015.c	upd3-016.c	upd3-017.c
upd3-	**1**	-1.958017E	-1.340543E	-1.788726E	-1.675198E	-6.313147E	-1.085791E	-1.579619E	5.26832E-(	6.930218E
upd3-	-1.958017E	**1**	**0.856992**	-2.006913E	**0.925139**	-0.0159797	-2.706579E	-5.597485E	7.042118E	0.0161234
upd3-	-1.340543E	**0.856992**	**1**	-1.078085E	**0.940968**	-1.085994E	-1.904628E	-7.160438E	-6.845203E	-5.209728E
upd3-	-1.788726E	-2.006913E	-1.078085E	**1**	6.384568E	-1.180575E	-2.069083E	-9.256598E	-6.849316E	-6.128258E
upd3-	-1.675198E	**0.925139**	**0.940968**	6.384568E	**1**	-1.396583E	-2.449129E	-8.501349E	-8.519993E	-6.499929E
upd3-	-6.313147E	-0.0159797	-1.085994E	-1.180575E	-1.396583E	**1**	**0.733269**	3.687835E	3.371778E	0.2762358
upd3-	-1.085791E	-2.706579E	-1.904628E	-2.069083E	-2.449129E	**0.733269**	**1**	0.1159216	8.542597E	**0.548274**
upd3-	-1.579619E	-5.597485E	-7.160438E	-9.256598E	-8.501349E	3.687835E	0.1159216	**1**	0.1743946	0.2046806
upd3-	5.26832E-(	7.042118E	-6.845203E	-6.849316E	-8.519993E	3.371778E	8.542597E	0.1743946	**1**	**0.548263**
upd3-	6.930218E	0.0161234	-5.209728E	-6.128258E	-6.499929E	0.2762358	**0.548274**	0.2046806	**0.548263**	**1**

The results pane is automatically vertically enlarged to show more of the results. To toggle between this enlarged size, and the standard size, click on the grey region between the results table and the graph pane.

Any test scores that are above the cut-off level selected in the preceding dialog box are highlighted in bold. This procedure may reveal hitherto unnoticed correlations between different patterns in a database.

All the matching results obtained against *upd3-004* are shown in the first column, *upd3-005* in the second, and so on. By looking at the intersections between patterns, an idea of the similarities between any two database patterns may be obtained.

To view a particular correlation of interest graphically, click once on the grid square containing the result. For example, the patterns *upd3-005* and *upd3-008* have a match score of 0.9251 in the above result. Clicking on the square containing the 0.9251 results in the two patterns of interest being superimposed on the graph pane below to allow visual comparison.

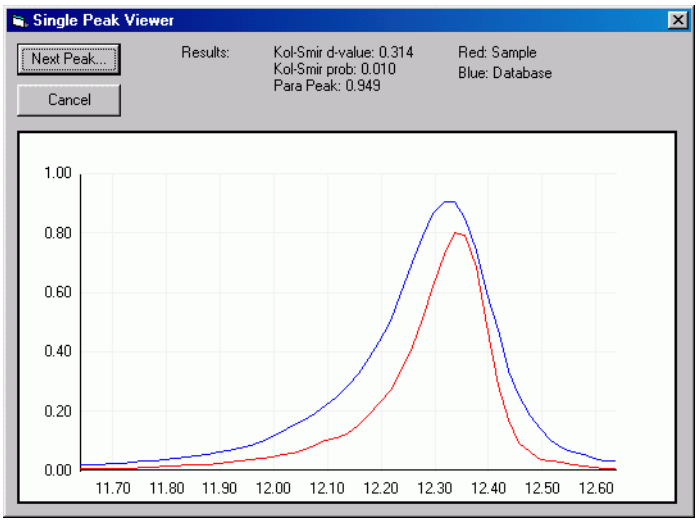
If the option to refine pattern shifts was selected, after clicking on a given correlation, holding the mouse steady for a second brings up a tooltip containing the calculated best offset for the two patterns, in the form (a_0, a_1) as discussed above.

In cases where a large volume of correlations are being managed the display may appear overcrowded. This can be overcome in two ways. One is to enlarge the display by clicking the enlarge button on the top right of the main match window. The other is to save the results as a text file to be viewed outside PolySNAP, which can be done by selecting *Save Results...* from the *File* menu, using the shortcut *Ctrl-S* or by clicking the *Save Results* button in the upper portion of the match window. A window will then open asking for a location and filename for saving the results.

3.4.8.6 Match Peak

Individual corresponding peaks can be easily compared in this mode. This can be particularly useful to spot very small differences between otherwise almost identical patterns: click once in the results list to select the pattern you wish to compare the unknown pattern with.

Then click the *Match Peak...* button (or select *Show Peak Comparison* from the *Pattern* menu). The window will appear for the first of the matched peaks:

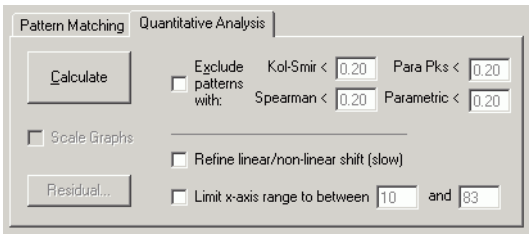


Clicking *Next Peak* closes this window and opens another for the next peak common to both patterns. *Cancel* returns to the match window.

Detailed results of the Kol-Smir and Parametric Peaks test results are shown on an individual peak basis in this window (these individual values are combined to give the result for the whole pattern displayed in the main match window results pane).

3.5 Quantitative Analysis

The second tab represents the second of the main program modes: Quantitative Analysis:



This is used with patterns of mixtures to estimate their percentage composition from pure component patterns contained in the matching database. **Note that for the method to work successfully, patterns corresponding to all pure phases in the mixture must be present in the current working database.**

3.5.1 Standard Quantitative Analysis:

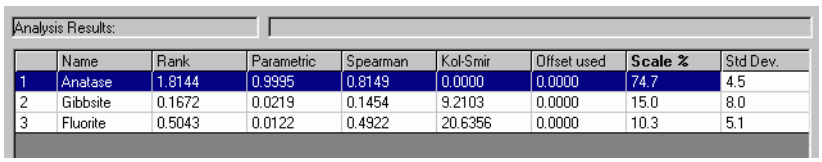
Click the *Calculate* button (or press F3). Analysis of the selected unknown pattern will commence. Prior to analysing, a pattern matching calculation is performed.

The progress and status bars indicate this:

Before showing that analysis is taking place:

Note that this process may take some time, depending on the number of patterns included in the calculation. The progress bar is **not updated** during the calculation of relative proportions, so patience is required as the calculations are very processor intensive, and may take several minutes. The screen is not updated or redrawn while this is happening.

Once analysis has completed, the results will be displayed in the results pane:



	Name	Rank	Parametric	Spearman	Kol-Smir	Offset used	Scale %	Std Dev.
1	Anatase	1.8144	0.9995	0.8149	0.0000	0.0000	74.7	4.5
2	Gibbsite	0.1672	0.0219	0.1454	9.2103	0.0000	15.0	8.0
3	Fluorite	0.5043	0.0122	0.4922	20.6356	0.0000	10.3	5.1

Note that the headers have changed. Although the first 6 columns are the same as before, there are 3 new columns:

1. Offset used:

   The offset given to a particular pattern that gave the best match; see section 4.4.2.1

2. Scale % or Weight %:

   Displays what proportion of the mixture is accounted for by an individual database pattern. This is normally displayed as a scale percentage *i.e.* what percentage of the mixture *pattern* does each individual phase constitute?

   If all the required advanced pattern information has been entered (see Section 3.3.11, Additional Pattern Information), the program converts this scale percentage to an actual percentage of the mixture by weight. In this case, the column heading is changed to Weight %.

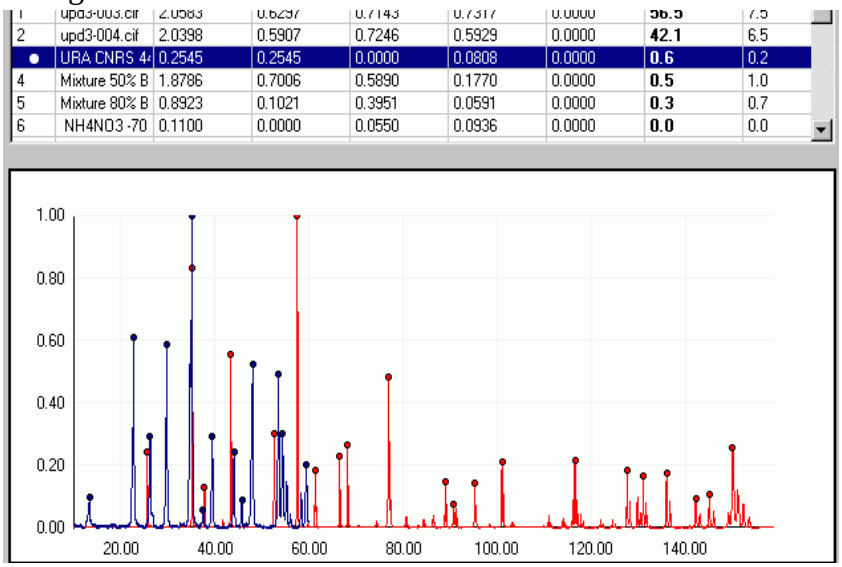
   Note that this information must have been entered for ALL phases in the mixture to be able to calculate a weight fraction.

3. Std Dev.

   The standard deviation of the percentage composition calculation reported for each pattern.

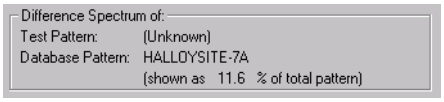
By default, the maximum number of patterns that will be reported as included in a mixture is 6, although this is easily changed. (see section 4.6, Matching Options)

As before, patterns can be displayed over the unknown pattern by clicking once in the leftmost column:



This can provide valuable visual feedback for the user to check if the suggested constituents of the unknown mixture are indeed correct.

To view a phase pattern on top of the mixture pattern scaled to the percentage result suggested, check the *Scale Graphs* checkbox, and then click on the pattern to display it in the usual way. If the *Scale Graphs* option is checked, and the *Show Difference Graph* option is selected, the individual phase pattern will be subtracted from the mixture pattern scaled to the relevant %. The amount of scaling performed is shown in the difference graph window:

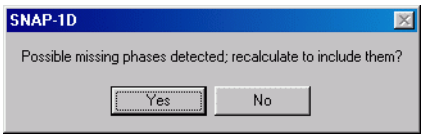


Additional feedback on the results is given by the calculated error on the suggested percentage, and by how good the matching results of the Spearman, Parametric and Kol-Smir tests were for each phase.

Occasionally if an incorrect pattern has been suggested by the program, this may be shown by extremely low values of the Spearman, Parametric and Kol-Smir tests, and such patterns can be marked to be ignored during subsequent runs - (see 3.5.2.2, Exclude patterns...).

3.5.1.1 Automatic Missing Phase Detection

The program examines the results from the analysis, and based on the calculated error and residual trace can suggest if the result does not account sufficiently for all of the unknown pattern intensity:



This would occur in a case where not all of the phases present in a mixture pattern were in the database being used.

Selecting *Yes* in the above dialog causes the program to simulate a pattern corresponding to the intensity unaccounted for by the current results. (This simulated pattern is equivalent to the pattern obtained by saving the positive intensity trace from the Residual window; see Section 3.5.2.6, Residual Window for more details.)

The calculation is then re-run to include this simulated missing phase, and results re-calculated:

	Name	Rank	Parametric	Spearman	Kol-Smir	Offset used	Scale %	Std Dev.
1	Mixture.TXT	**2.0000**	1.0000	1.0000	1.0000	0.0000	0.0	0.0
2	Phase 2.TXT	**1.1987**	0.9619	0.2368	0.6668	0.0000	44.9	0.8
3	Simulated Miss	**0.8548**	0.2642	0.5905	0.2242	0.0000	53.8	3.0
4	Phase 1.TXT	**0.0000**	0.0000	0.0000	0.0866	0.0000	1.3	0.5

Note that results cannot be calculated in terms of weight fraction when simulated missing phases are included.

After doing an analysis, and choosing to include a missing phase, when the user closes the match window, the program checks to see if the simulated phase should be retained in the current database or discarded.

If the pattern is retained, it shows up in the database window with a name of 'Simulated Missing Phase', although this can of course be altered in the Edit window later.

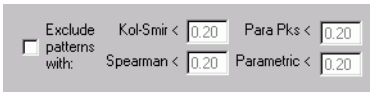
The settings used by the program to determine if a missing phase may be present can be controlled by the user; see section 4.3.10 on page 144.

3.5.2 Advanced Analysis

3.5.2.1 Refine linear/non-linear shift

This option works as described in Section 3.4.8.1, Refine linear/non-linear offset.

3.5.2.2 Exclude patterns...



Selecting the *Exclude Patterns with:* checkbox can be extremely useful to narrow down the number of patterns to be considered as components of the unknown pattern. It does this by excluding patterns that are below user-set thresholds on any of the matching tests to be included in the quantitative calculation.

The default values may be easily altered by changing the relevant values in the text boxes. The best values to use can change depending on the type of data being considered.

Generally, the best approach is to perform a standard analysis with defaults to begin with, and see if any poorly matching patterns have been included. The results from this will then give a feel for what

values the cut-offs should be set at. This option also has the side benefit of speeding up the process as less patterns are used.

To use this option, turn on the checkbox, enter the desired values, and click the *Calculate* button as normal.

3.5.2.3 Limit x-axis range...

Turning on the *Limit x-axis range* checkbox has the effect of limiting the analysis calculation to a subset of the entire x-axis range of the unknown pattern. This can be useful if a particular feature of the pattern is causing problems for the calculation.

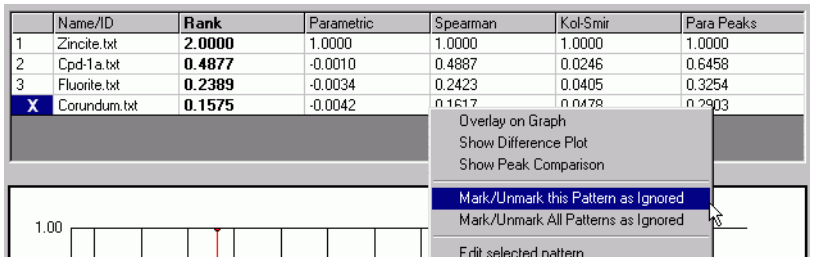
The default values in the textboxes are the full range of the unknown pattern, although these can be changed by either:

1. Typing in the desired start and end angles of the required range into the textboxes, or:

2. Zooming in to the desired range by clicking and dragging a zoom-rectangle on the graph pane. The start and end ranges of the x-range of the zoomed pattern will automatically be placed in the textboxes.

To use this feature, turn the checkbox on, enter the desired start- and end- details, and click the *Calculate* button.

3.5.2.4 Ignore selected pattern



If a particular pattern included in the list of suggested results is known to be incorrect, it can be excluded from the calculation, by clicking once on it in the results pane to select it, and then by either right-clicking to bring up the results-pane pop-up menu, using the *Mark/Unmark Pattern as Ignored* option in the *Pattern* menu.

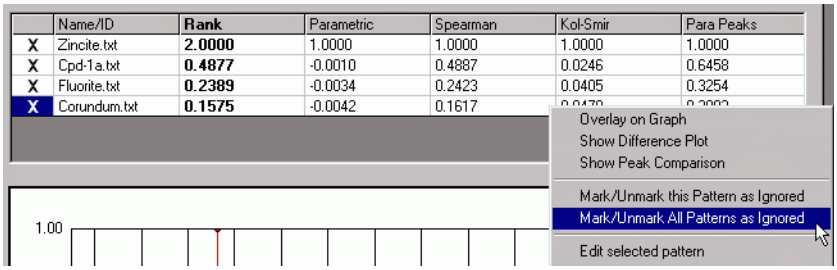
Select *Mark/Unmark this Pattern as Ignored*. A large, bold **X** will appear next to the pattern name in the first column, to indicate it is being ignored for the purposes of the calculation.

Once a pattern has been selected as ignored, click the *Calculate* button again to re-run the calculation, which will be performed without any the 'Ignored' patterns.

If the pattern was incorrectly marked as ignored, it may be unmarked, by simply repeating the above procedure, and selecting *Mark/ Unmark this Pattern as Ignored* once more. The large **X** will disappear.

It is possible to mark multiple patterns to be ignored by repeating this process with different selected patterns.

3.5.2.5 Ignore All patterns except.



In situations where the particular constituent patterns of an unknown mixture are known, but only the relative percentages of each are unknown, it is useful to be able to only include the correct components in the calculation. This not only speeds up the process but can generally improve the accuracy of the results.

The simplest way to achieve this is to first mark all of the patterns in the database to be ignored, and then un-mark the individual patterns that are known to be present in the mixture.

To achieve this, right-click on the results pane to bring up the pop-up menu, or use the *Pattern* menu: select *Mark/Unmark all Patterns as Ignored* from the options presented. A black bold **X** will appear in the first column of the results pane next to every pattern shown. *(N.B.* if any patterns have previously been marked as ignored, these will be un-marked by this process).

Next, individually unmark the patterns that are to be included by selecting them singly, and selecting *Mark/Unmark this Pattern as Ignored* from the pop-up menu. Then select *Calculate* as normal.

To unmark all of the marked patterns simultaneously, select *Mark/ Unmark all Patterns as Ignored* again from the pop-up menu. All patterns that were marked as ignored will become unmarked, and *vice-versa*.

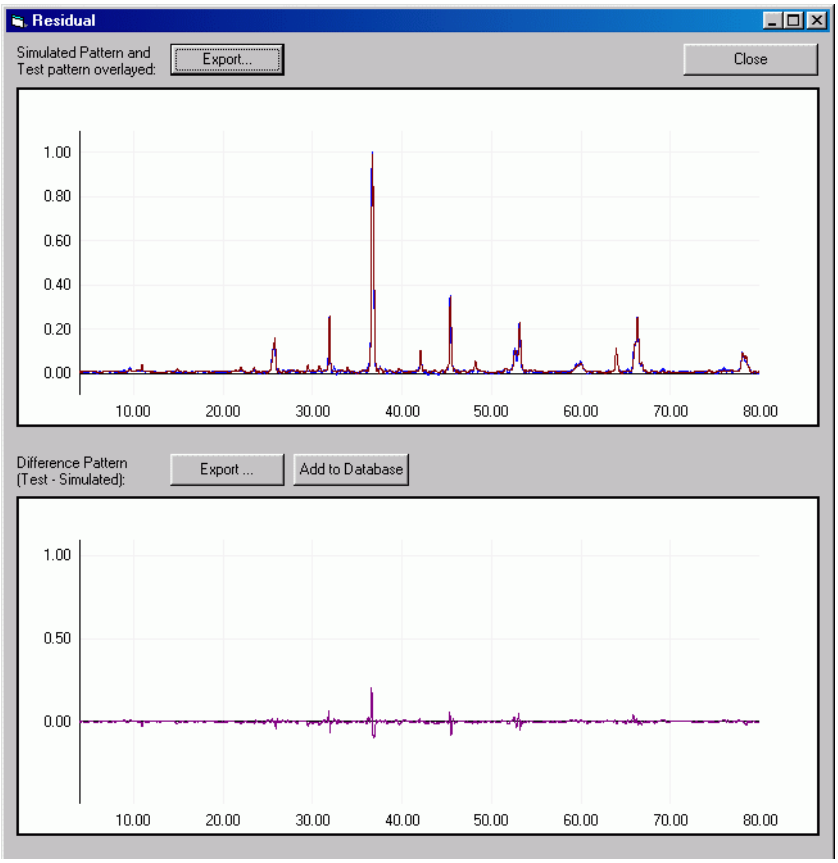
N.B. If the patterns that have been marked and unmarked become confused, it is possible to reset all of the patterns in the database to

the default state of all being included in a calculation, by closing the match window and then re-opening it.

3.5.2.6 Residual Window

To see if the suggested results are correct, or if they include a pattern not present in the mixture, or if they miss out a pattern that should be there, the residual window constructs a calculated pattern made up from the various individual patterns suggested as mixture components, in the proportions calculated.

Clicking the *Residual* button brings up the following window:



The upper window overlays the original mixture pattern (in blue) with the calculated one using the suggested results (in red/brown).

The difference between the two profiles is shown in the bottom window. Large negative peaks in this correspond to incorrect peaks, and would suggest one or more of the suggested constituent phases are incorrect. Positive peaks here would represent peaks in the original pattern not accounted for in the suggested results.

The simulated mixture pattern can be saved to an ASCII text file using the *Export Simulated* button. The difference plot can be saved to an ASCII file using the *Export Difference* button.

In the latter option, two different formats are available from the pop-up list at the bottom of the Windows Save dialog. The first (*Full residual*) saves the difference profile as it is shown in the window, with some peaks negative and some positive. The second *(Positive residual*) ignores any negative peaks and saves only positive ones in order to make a pattern comprising only intensity not accounted for by the current pure phases.

Exported difference profiles can be re-imported and treated as new patterns, and further quantitative analysis can be performed using them if required. This can be done manually or by using the *Add to Database* option.

3.5.3 Other Options

Edit Selected Pattern

Having performed a match or analysis procedure, it may be that one pattern listed in the results table is of particular interest. It is possible to view the Pattern Editor window for such a pattern directly, by first clicking once on the pattern of interest in the list, and either selecting *Edit* from the *Pattern* menu, or *Edit Selected Pattern* from the pop-up menu.

Either option brings up the editor window allowing detailed information about a pattern to be viewed or edited.

Match Against Selected Pattern

Having performed a match or analysis procedure, it may be that one pattern listed in the results table is of particular interest, and the user my wish to perform an additional match or analysis with that pattern.

It is possible to open another Match window for such a pattern directly, by first clicking once on the pattern of interest in the list, and either selecting *Match* from the *Pattern* menu, or *Match Against Selected Pattern* from the pop-up menu.

Either option brings up a new Match window. The original Match window is still open and may be viewed as desired. As many Match windows as required may be open at the same time.

3.6 References

1. Spearman, C. (1904). 'The proof and measurement of association between two things'. *American Journal of Psychology, 15*, 72-101.

2. Smirnov, N.V. (1939). 'Estimate of deviation between empirical distribution functions in two independent samples.' *Bulletin Moscow University, 2*(2) 3-16.

3.Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P. *Numerical Recipes in C: The Art of Scientific Computing*, Second Edition, Cambridge University Press, 1992.

CHAPTER 4

Using PolySNAP: Prescreening

4.1 Introduction

Pre-screening mode assumes you have a single new unknown sample file, which you wish to compare to a large number of existing patterns (e.g. > 10,000). On a reasonably modern PC, 10,000 patterns should taken around 10 minutes to process. This allows an effectively unlimited number of patterns to be narrowed down to a size suitable for full PolySNAP analysis.

The existing library patterns correspond to samples already collected; the user wishes to know which of the many library patterns the new sample is most similar to (if any).

PolySNAP is limited by default to performing cluster analysis on up to 1500 patterns per single dataset. It doesn't make sense however to perform a match everything-against-everything for the large library of data, so pre-screening it to identify the most relevant patterns is the obvious solution.

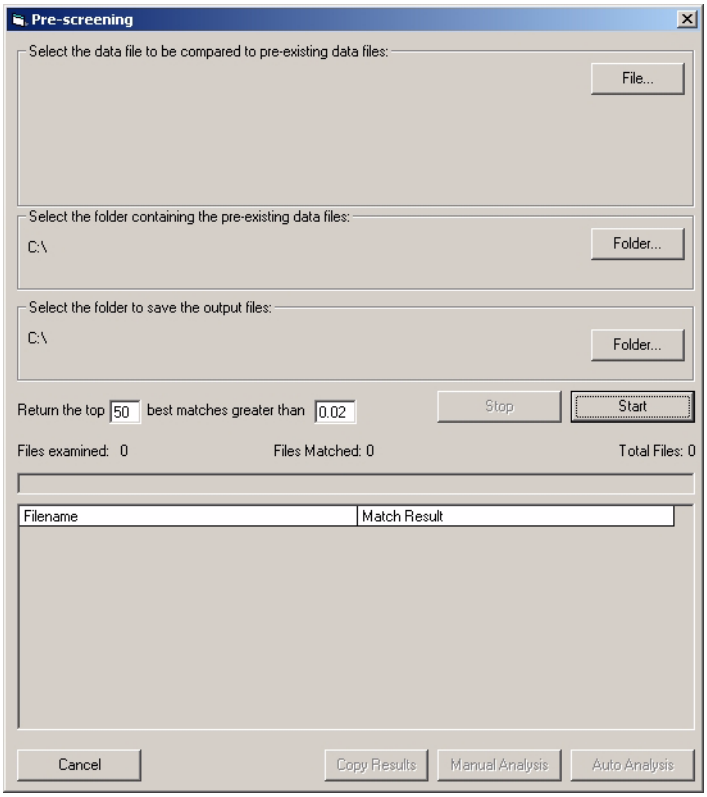
Once the most similar patterns have been identified in this manner, they can then be examined in PolySNAP, either in Manual mode, or Automatic mode, incorporating the normal cluster analysis and dendrogram display, etc.

4.2 Running Pre-screening Mode

Launch PolySNAP, and select File Menu > Automatic Mode > Pre-screen data…

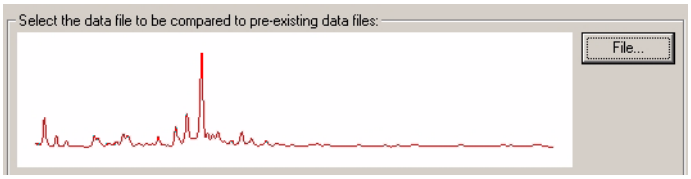
Alternatively, click the *Pre-screen Large Dataset* button on the Welcome window.

The pre-screening dialog box appears:

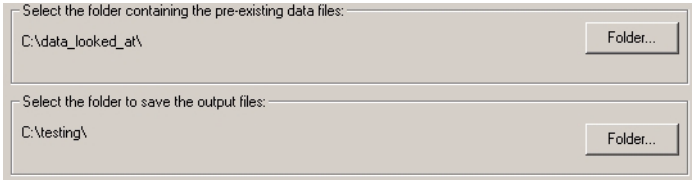


Select the location of the new, unknown pattern data file, by clicking on the File button.

This can be in any of the normal formats that PolySNAP imports; e.g. RAW or txt. A thumbnail view of the pattern is then displayed in the dialog box for reference.



Select the location of the folder containing the library of existing data files. The files can be in any mixture of formats that PolySNAP can read in. Subfolders of the selected folder will also be scanned for suitable patterns.



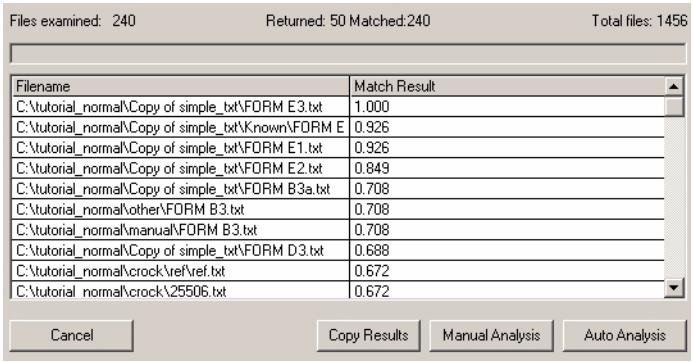
Select a folder location to store the results from the analysis. This should not be the same location as either input location. Two files are

output; a log file, and a SNAP Database file containing the returned best match patterns.

Decide how many of the top matches to the new sample you wish returned. This can be any number between 1 and 1500 (default of 50). To help not return poor matches, a default minimum rank threshold (default of 0.02)can also be set here. The matching rank is calculated using the match tests selected in the Options dialog box (default is 50-50 Spearman-Parametric full profile matching tests).

Click the button labelled *Start* to begin the analysis. The number of files in the library folder is then calculated; depending on the number of files and folders this may take some time. During this process, *Counting Files, Please Wait* is shown. To stop the filescan at any time, press the Escape key.

Once matching begins, the total files found, the number examined and the number matched is shown along with a progress bar. Matching continues until all the files found have been examined, or until the *Stop* button is clicked. *Stop* can be clicked at any time to end the process.



Once the matching is complete, or has been stopped by the user, the top *n* best matches to the sample pattern are displayed in the results grid.

The options then available are:

Copy Results from the table to the Clipboard, from where they could be pasted into *e.g.* a report or MS Excel:

Manual Analysis: This opens the identified patterns in a database window, from which the new sample pattern, labelled as 'New Sample, can be selected and matched against:

Auto Analysis: This brings up the PolySNAP Analyse Data window, with the database of best-fit patterns automatically selected. Cluster analysis, metric multidimensional scaling, etc. can then be performed and the results presented graphically using dendrograms and 3D plots as normal.

CHAPTER 5

Using PolySNAP: Quality Control

5.1 Introduction

This option is designed for situations where the stability of a material is being monitored over time; for example as part of a production line system, or for periodic checks for alignment or other issues compared to the last time a standard sample was collected.

A certain number of *Reference* patterns must be available; these are patterns that for the purposes of the analysis are considered to be a good representation of what is expected to be seen.

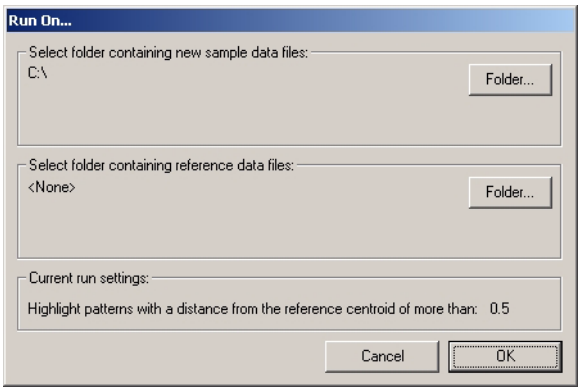
Various *Sample* patterns are then imported and compared to those reference patterns, and any that vary significantly from the ideal are noted and highlighted.

The results are displayed graphically with 'good' sample patterns shown within the surface of reference patterns, and 'bad' sample patterns appearing outside it. The tolerance for what is considered worth warning the user about can be adjusted to suit.

5.2 Worked Example

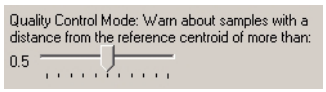
The best way to illustrate the value of this mode of operation is through a worked example. The data for this example can be found in the *quality* folder, within the *tutorial* folder installed along with the program.

Selecting *Quality Control* from the *Automatic* mode menu brings up the following control window:



The user simply identifies a folder containing the new sample patterns to be checked, and a folder containing some pre-existing reference patterns to compare them to. In this case we select *tutorial\quality\samples* and *tutorial\quality\references* accordingly.

The lower section shows the current setting of the *Quality Threshold* value in the program Options window, *Display & Advanced* pane:

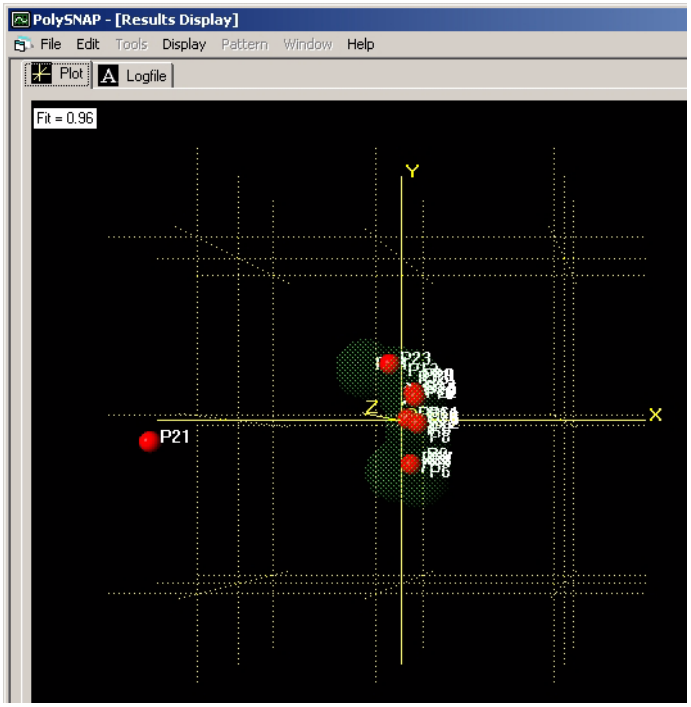


This acts a 'strictness' control. The lower the number (ranging from 0 to 1.0), the less far the sample patterns are allowed to vary from the centroid position of the reference patterns before they are highlighted as potentially a problem.

Clicking *OK* brings up the standard PolySNAP data loading window. During the analysis phase of the calculation, a warning dialog is displayed if any of the sample patterns are considered to be outwith the set threshold; in this case:

A simplified version of the standard PolySNAP results window is then displayed, showing just two tabs, the 3D Plot and the Logfile. By default the pattern information pane is hidden, but this can be

shown by selecting *Show Pattern Information* from the *Display* menu.



The main plot can be interacted with like the standard 3D plots. The red spheres are the new sample patterns, and the reference samples are represented by the shaded green area. Sample patterns within this shaded zone are considered to have passed the analysis, whereas samples outwith it, such as P21 in the above example, have failed.

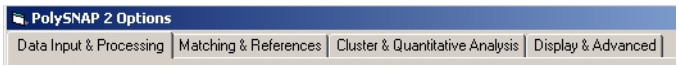
Clicking on the Logfile tab brings up a PolySNAP audit trail log, which notes details of the patterns imported, and highlights any thought to be suspect.

CHAPTER 6

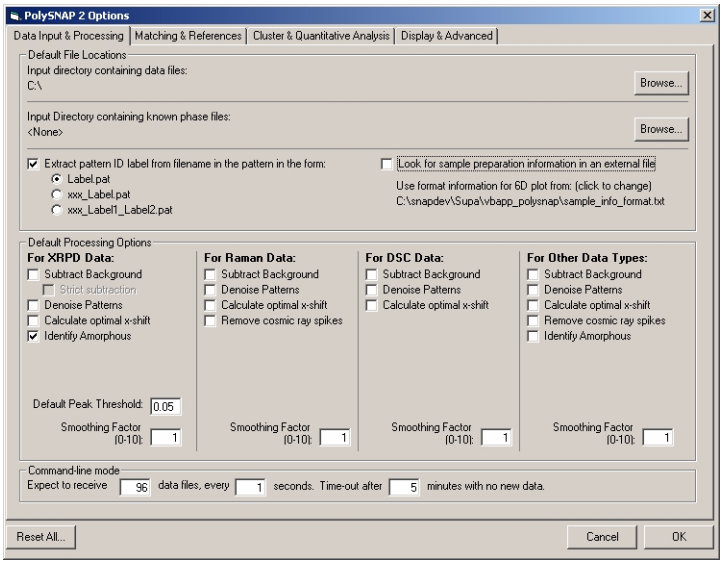
Program Options and Defaults

6.1 Accessing Options

The Options screen can be accessed at any time from the *Edit* menu. It is comprised of four tab sections which can be selected from the top of the window. These are each described in turn below.



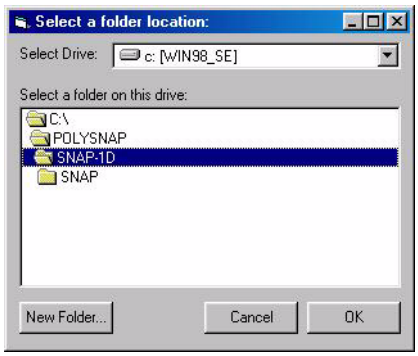
6.2 Data Input & Processing.



6.2.1 Default File Locations

If no working input directory has been specified *via* a command-line argument or other input method, the program uses the directories

specified here as its default. To alter any of them, click on the *Browse* button to bring up a directory selection window:



Selection of a new working directory is carried out by first selecting the correct drive letter from the pop-up menu at the top, and then navigating to the desired folder *via* the lower navigation box. To open a folder to see what other sub-directories it contains, double-click on it. When the desired folder has been located, double click on it to ensure it is selected, and click *OK*.

Extract label from filename in the pattern in the form - With this option on, sample ID information contained be parsed and used to identify the pattern on the Cell Display and Dendrogram Display. With the first option, *Label.pat*, the entire unique part of the filename is used (though only the final 7 characters are displayed; the full label can be seen in a tooltip by hovering the mouse over a sample). With the second option, *xxx_Label.pat*, only that portion of the filename after the final underscore character is used as a sample label. The third option, *xxx_Label1_Label2.pat* uses the information after the final two underscores as the label, and ignores any previous part of the filename.

Look for sample preparation info in external file: with this option on, the program looks for a file in the input directory called <input>.dat, where <input> is the name of the containing directory.

Use format information for 6D plot from: - clicking this option allows the user to select a different *sample_info_format.txt* file. The default format file is installed with the program in its Program Files folder. If an alternative location is selected here, it is used until a different location is selected.

For more information on the *sample_info_format* file, see Section 2.4.2.

6.2.2 Default Processing Options

Selecting items here cause them to be selected by default in the main Analyse Data dialog box. In addition, if running the program from

the command line, then the options selected here are used. Note that different defaults can be set up for each of the four main datatypes (PXRD, Raman, DSC, Other).

Some of the options are common to different datatypes; some are specific to a particular datatype. Some, like Background Subtraction, behave subtly differently depending which type of data is being processed.

6.2.2.1 Subtract Background

Selecting the first checkbox performs the standard background subtraction routines. This option is available for all types of data, but different algorithms will be applied depending on what type of dataset is present. The second checkbox, available only for PXRD patterns, *Strict subtraction*, applies an additional pass that is intended for use when there is a large amorphous contribution to the patterns that is required to be removed. It should be used with caution as the possibility of removing more pattern than is required is more likely.

6.2.2.2 Denoise Patterns and Smoothing Factor

Varying the number in this box controls the amount of noise removal performed by the wavelet smoothing routines. The larger the number, the smoother the resulting pattern. Be careful of oversmoothing and either losing valuable information, or introducing artifacts into the signal. The default value is 1. Available for all datasets.

6.2.2.3 Default Peak Intensity Threshold

For PXRD data only, this control changes the default peak threshold. Only peaks whose maximum intensity are above this cut-off value will be marked as peaks, and therefore included for comparison in the peak-based tests. The number of peaks marked can also be used as an indicator of crystallinity of a sample. The default value is 0.05; the acceptable range is from 0 to 1.0.

6.2.2.4 Calculate optimal x-shifts

This section controls how patterns will be shifted in an attempt to maximise the correlations between them. By default the program attempts to maximise the correlation between patterns by varying the values a_0 and a_1 in the equation $\Delta 2\theta = a_0 + a_1 \sin \theta$, although $\cos \theta$ or $\sin 2\theta$ may optionally be used instead by selecting the relevant item from the drop-down menu (accessed under the Matching & References section).

The values of a_0 and a_1 represent the maximum values that will be allowed by the program during the refinement.

6.2.2.5 Identify Amorphous

For PXRD only, attempt to identify which patterns are amorphous, using the criteria define in the Cluster & Quantitative Analysis section (see 5.4.1).

6.2.2.6 Remove Cosmic Ray Spikes

For Raman data only, check patterns for cosmic ray spike features (normally identifiable as very large intensities with very narrow pixel ranges) and remove them from the signal.

6.2.3 Command-line mode

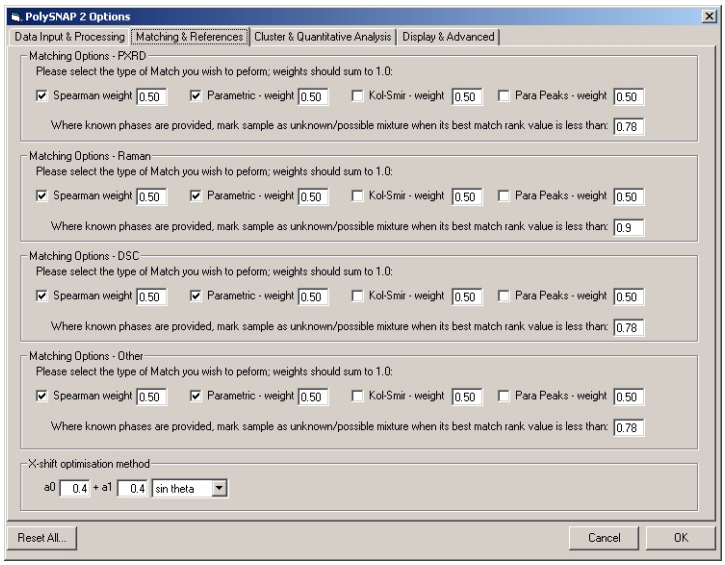
Expect to receive [ ] data files, every [ ] seconds. Time-out after [ ] minutes with no new data

If not set otherwise by means of a command line argument, the program looks here to see how many data files it should expect to receive before completing importing. The default is 96, any integer between 1 and 1500 is acceptable. Note that at least 4 patterns are required to be able to do any cluster analysis.

The latter number sets how often *PolySNAP* looks in the input directory specified above to see if any new data files have appeared. The default is every 0.05 seconds, and any numerical value between 0.01 and 60 (i.e. 1 minute) is acceptable. Note that between each check, the interface will be unresponsive, as the program is sleeping to conserve system resources. As a result, it make take up to the time set here for the program to respond to user input.

If no new data files have been found [x] minutes after the last file was imported, the program stops waiting for data even if the number of data files expected has not been met. The default is 5 minutes, a numerical value between 0.5 and 720 (12 hrs.) is required.

6.3 Matching & References



6.3.1 Matching Options

For each of the four main datatypes, individual settings can be applied as to the combination of different full-profile matching tests that are applied to them. Different combinations of tests may prove more effective with particular types of data.

The matching performed by PolySNAP takes all of the loaded sample patterns are compared to each other, and a correlation matrix produced. It is this correlation matrix that is analysed further by the cluster analysis and related routines in PolySNAP.

Therefore, matching is the key phase of the analysis, and a suitable selection of tests and weightings is often the deciding factor between a useful or less viable analysis. These settings should not normally need to be altered in day-to-day usage of the program.

Any combination and weighting of the four tests listed may be employed, however note that the combined weights of the tests to be used **must sum to 1.0** if sensible results are to be obtained. It is not recommended to mix Kol-Smir and Para Peaks with Spearman and Parametric.

The tests available are:

Spearman - with a weight of [ ] - this can have any value between 0 and 1.0. The default value is 0.5.

Parametric - with a weighting of [ ] - this can have any value between 0 and 1.0. The default value is 0.5.

Kol-Smir - with a weighting of [ ] - this can have any value between 0 and 1.0. The default value is 0.5.

Para Peaks - with a weighting of [ ] - this can have any value between 0 and 1.0. The default value is 0.5.

All of the above options can be selected by clicking on the check box beside each. By default only the Spearman and Parametric tests are selected with equal weightings.

Where known phases are provided, mark sample as unknown/ possible mixture when its match rank is less than [ ]

If known phases are provided, each sample is compared to the known phases on import. If the best match to a known phase is below this cut-off level, the sample is marked as being a possible unknown, and also then has quantitative analysis performed upon it. This option can have any value between 0 and 1.0; the default value is 0.78 for PXRD samples.

6.3.2 X-Shift Optimisation Method

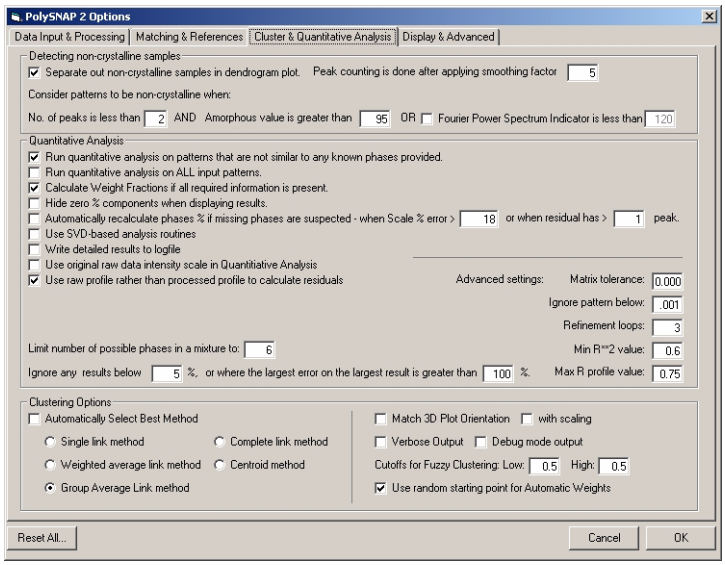
This option allows the program to calculate if a small x-axis shift of up to the selected amount will improve the matching results. This option is very slow, and can add considerably to the run time, so should only be turned on if needed.

By default, the program to attempt to find the coefficients a_0 and a_1 for the equation $a_0 + a_1 \sin \theta$, such that the matching result between two patterns is maximised. The user can also select to use $\cos \theta$ or $\sin 2\theta$ instead by means of the drop-down list. The a_0 term corrects for the zero point error. The $\cos \theta$ option corrects for varying sample heights in reflection mode, the $\sin \theta$ option corrects for transparency errors or, for example, transmission geometry with constant specimen-detector distance, or the $\sin 2\theta$ option provides transparency and thick-specimen error corrections.

Note that the calculated offsets can be investigated interactively through the Numerical Results pane (see Section 2.3.14).

The values in the text-boxes correspond to the maximum shift amount allowed in either the + or - direction, in terms of the coefficients a_0 and a_1.

6.4 Cluster & Quantitative Analysis



6.4.1 Detecting Non-crystalline Samples

Powder Patterns that are imported are examined and marked as possibly non-crystalline when the conditions below are met:

Separate out non-crystalline samples in dendrogram display

The final option controls how the program behaves when it encounters a pattern marked as possibly non-crystalline. With this option checked, such patterns are not included when constructing the dendrogram, and so stand out as individual patterns not connected to anything on the right hand side. With it turned off, amorphous samples are not distinguished from other samples in any way.

Patterns that are marked as possibly non-crystalline are highlighted as such in the results display for the cell layout; see Section 2.3.4 for more information.

Peak counting is done after applying smoothing factor [] - this sets how much the pattern is smoothed as part of the amorphous pattern identification calculation; the default is level 5; with very noisy data this should be reduced to 1 or 2 to prevent he occurrence of false positives.

No. of peaks is less than [ ] - this can be any value between 0 and 100. The default value is set at 2. This can be overridden by the program if less than two peaks are present, but they have high intensity values.

Amorphous value is greater than [ ] - this can have any value between 0 and 100. The default value is set at 95. The amorphous value used above is a measure of the percentage of a pattern that is considered to be amorphous as opposed to crystalline.

Fourier Power Spectrum Indicator is less than [] - as an additional check, the power spectrum of the Fourier transform of the pattern profile is examined to see how much signal is present at higher frequencies. If there is no signal about the default value of 120, the pattern is marked as amorphous if the peak setting requirements above are also met. This option is off by default.

6.4.2 Quantitative Analysis

6.4.2.1 General Options

Run quantitative analysis on patterns that are not similar to any known phases provided - This option can be activated or deactivated by clicking on the check box, and controls whether or not the program performs quantitative analysis on samples it considers to be possible mixtures.

Perform analysis on all files regardless of similarity - normally, the program only attempts quantitative analysis of samples that do not much sufficiently well to any of the known phases, and ignores samples that do match above a certain threshold. With this option on, it will attempt to perform analysis on all input samples. Note that this may result in spurious results, so use with caution.

Calculate Weight Fractions if all required information is present

This checkbox allows the program to automatically display analysis results as a weight fraction rather than proportional to scattering power, if all the necessary pattern information is available.

With it turned off, a scale percent is reported even if all pattern information is present. (See section 3.3.11 on page 108 for information on what details are required to calculate a weight fraction).

Hide zero % components in results

With a database containing a large number of patterns, performing an analysis usually results in a few patterns contributing towards a mixture, with the rest being ranked at 0.0%. This option hides any patterns that are not considered contributors to the mixture.

Recalculate for Missing Phases...

These controls set the values beyond which the program considers that missing phases may be present in a mixture (see section 3.5.1.1 on page 128 for more information).

The two criteria used are the size of the largest calculated error on the quantitative results, and if the residual intensity trace has any peaks above the default minimum peak height.

In addition, if the *Automatically Recalculate for Missing Phases* option is checked, the program will recalculate to include missing phases automatically, without consulting the user first. The default value for this is to consult the user.

Use SVD-based analysis routines - by default, stepwise regression is used to calculate the proportions of pure phases in a mixture; selecting this option returns to the alternative method used in earlier versions of PolySNAP. This may be useful to recreate results obtained from a previous version.

Write Detailed Results to Logfile

For debugging purposes, detailed output from the stepwise regression calculation can be added to the logfile.

Use Original Raw Data Intensity Scale in QA

For stepwise regression, it may give more accurate results to use the original intensity values stored in the imported pattern files, rather than the scaled versions used within the matching sections of PolySNAP.

Use raw rather than processed profile for residuals...

This checkbox determines whether the original pattern profiles, or the processed profiles (*e.g.* with background subtracted, *etc.*) are used to construct the simulated pattern in the Residual window.

Limit number of possible phases in a mixture to [ ]

This can have any integer value between 1 and 15. The default value is 6; if the initial number of results returned is greater than the value set here, only the top x results are reported (where x is the value entered here). The results which are reported can be further filtered by means of the following two options:

Ignore any results below [ ] percent - can have any value between 0 and 100. The default value is 5%. This sets the smallest percentage value which will be returned as significant; anything below this cutoff level will be ignored.

or where the error on the largest result is greater than [ ] percent - can have any value between 0 and 1000. The default value is 100. This sets the level for the error value calculated for each phase. If the error on a particular phase exceeds the amount entered here, it is not considered as a possible phase in the mixture pattern.

6.4.2.2 Advanced Options

Matrix tolerance

The second controls what proportion of the patterns that do not contribute significantly to the overall pattern are ignored. This should be a number between 1 and 0. The default is $1x10^{-5}$.

Ignore pattern below...

This (default is 0.001) controls how much of each database pattern is treated as background upon inclusion in the calculation. This is done to simplify the problem and reduce the effect of background noise.

Refinement loops

The Refinement loops field controls how many loops of refinement the program performs. This should be an integer between 1 and 3. If it is 1, all patterns are included. If it is 2, all patterns are used, then only the top 15 are included the second time around. If it is 3, then all patterns, then the top 15, then the top N, where N is the number of components selected above, are included

6.4.3 Clustering Options

The program can use up to five different clustering methods to attempt to explain the data provided. By default, the Group Average Link method is selected.

Optionally, the program can use all of the methods, and then apply various internal tests to automatically select the method which produces the most self-consistent results. The disadvantage to this is that program run-time may be dramatically increased as a result.

The overall method employed is agglomerative hierarchical clustering, using a distance matrix derived from the PolySNAP match results correlation matrix. Initially all the patterns are assigned to individual clusters which are then joined one pattern at a time. When two clusters are so joined, a measure of the distance between them is needed, and each clustering method has its own way of doing this.

In general, when two clusters *i, j* are combined, a new distance between the cluster and an existing cluster *k* is calculated as:

$$d_{k(i,j)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma \left| d_{ki} - d_{kj} \right|$$

The parameters α_i, α_j, β, and γ are different for each clustering method used:

Single Link Method:

$$\alpha_i = \frac{1}{2} \qquad \beta = 0 \qquad \gamma = -\frac{1}{2}$$

Complete Link Method:

$$\alpha_i = -\frac{1}{2} \qquad \beta = 0 \qquad \gamma = \frac{1}{2}$$

Weighted Average Link Method

$$\alpha_i = \frac{1}{2} \qquad \beta = 0 \qquad \gamma = 0$$

Centroid Method:

$$\alpha_i = n_i(n_j + n_j) \qquad \beta = \frac{-n_i n_j}{(n_i + n_j)^2} \qquad \gamma = 0$$

Group Average Link Method:

$$\alpha_i = \frac{n_i}{(n_i + n_j)} \qquad \alpha_j = \frac{n_j}{(n_i + n_j)}$$

$$\beta = 0 \qquad \gamma = 0$$

Advanced Options

Match 3D Plot Orientation

By selecting this checkbox the MMDS and PCA plots are both rotated so as to be as closely in the same orientation to each other as possible. This is to make for easier comparison. By default this option is turned off.

With Scaling

Selecting this checkbox scales the plots up or down if necessary in order that they take up roughly the same volume of the cube. Again this option is available to make comparison easier. By default this option is turned off.

In addition to the output methods displayed graphically in the main PolySNAP output window, the clustering algorithms also all output textual information to the logfile for the current run of the program. This may be found in the program output folder with the name *SNAPlog.txt*. The detail level of the text output can be modified by the two checkbox options:

Verbose and *Debug Mode Output*

The *Verbose* and *Debug Mode* options control the level of output the cluster analysis section of the program writes to the logfile for each run. In normal program usage, both options are turned off. Occasionally it may be useful to see more detailed information on the clustering results, in which case Verbose option may be turned on.

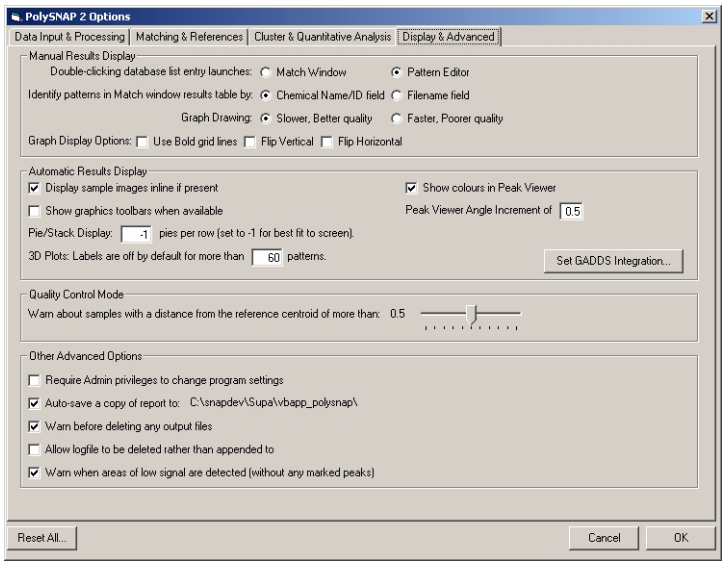
The Debug Mode option causes the program to output all of the working matrices generated during analysis to the logfile, to aid in debugging or solving very difficult problems. Note however that use of this option can cause the logfile to become very large - possibly up to tens of megabytes in size. It should be used with caution, as usage of this option may also slow the execution of the program, especially when large numbers of patterns are being analysed.

The two cut-offs for fuzzy clustering determine which patterns are displayed in the fuzzy clustering output. As the fuzzy clustering is only interested in outliers, only samples which score either below the low cut-off, or above the high cut-off value are shown in the output.

Use random starting point for Automatic Weights

This determines if a random starting point for automatic weights is used in order to optimally calculate the combined datasets when working with multiple data.

6.5 Display & Advanced



6.5.1 Manual Results Display

Double-clicking Database List Entry

The first option controls whether double-clicking a pattern entry in an open database window brings up the match window or the pattern editor window. By default this is set to being the *Pattern Editor.*

There is also a toggle to switch between displaying the Chemical name/ID or the Filename of patterns in the Match window and other results lists.

Graph Drawing

Decides whether the patterns are displayed in a high-quality, but slow to draw mode, or a faster but lesser quality mode. The slower option is the default.

The three *Graph Display* checkboxes control whether the patterns are displayed as read in from their original data files, or flipped in either the x or y axes. These options may be useful when examining different types of non-powder diffraction x-y data. *Use Bold Grid Lines* checkbox controls if a black grid shown behind patterns on the graph, or if light-grey lines are used. The latter is the default.

6.5.2 Automatic Results Display

Display sample images inline if present - with this option on, PolySNAP automatically resizes the graph pane in the Display

window to show a small preview of the sample images (if present). With this option off, the images may still be viewed using an external graphics program. Sample images are expected to be in the same input folder as the data files, with the same filename, differing only by having a *.jpg* file extension.

Show graphics toolbars when available - in the results display window, several of the graphics panes have toolbars to allow quick access to commonly used functions. This option sets if they are displayed by default or not.

Pie/Stack display: Show [ ] pies per row - For a display purposes the number of pies in a row can be altered to suit the users needs. *e.g.* for a run of 12 samples the number entered may be 6, which would create 2 rows of 6. The minimum value is 2 and the maximum value is 199. If the value is set to -1, the program makes its own judgement as to the best number to display for a given run to fit in the display window, and this behaviour is the default.

3D Plot label display: if the number of data files is greater than the value set here, then labels will not be displayed on 3D plots by default to reduce visual clutter. They can still be turned on manually at any time.

Show colours in Peak Viewer: with this checked, the Peak Viewer window accessed through *Tools* > *Show Peak Comparison* in the Results display highlights the different peak intensities using colours as well as numbers. The *Peak Viewer Angle Increment* setting controls the width of bin the x-axis is divided up into (default of 0.5 degrees). See Section 2.6.1.

Set GADDS Integration... - clicking this button brings up a dialog box in which to enter the location of the script directory for the Bruker GADDS Pilot software. Entering the correct location (by default, *C:\saxi\gadds\scripts*) and clicking *OK* copies a file, *DisplayFrame.slm*, to that directory. The presence of this file in the correct location permits the integration between the PolySNAP display window and the GADDS software (see Section 2.3.17).

6.5.3 Quality Control Mode

Quality Control Mode - see “Using PolySNAP: Quality Control” on page 135.

6.5.4 Other Advanced Options

Require Admin privileges to change program settings - with this option on, only a user that has Administrative privileges on the local

machine is permitted access to the program Options. Note that Admin privileges are required to toggle this option.

Auto-save copy of report to: <file location> - With this option checked, the a copy of the current report is automatically backed up to the designated location whenever a automatic report is generated by the user (Section 2.3.16). To change this location, with the checkbox selected, click once on the name of the file. A standard folder-selection dialog box will appear. The location displayed will update to reflect the new selection.

Warn before deleting any output files - with this option selected, the user is always warned when selecting an output folder that already contains files from a previous program run. The older files will be deleted if the user continues. If the related option, *Allow logfile to be deleted*, is off then logfiles and errorlogs are merely appended to on subsequent runs, rather than started afresh.

Warn when areas of low signal are detected - this option brings up a warning dialog box when the program detects large areas of a pattern that appears to have little or no signal in it. For example, if a pattern has been collected to 55 degrees, but has no peaks above 30, when this option is turned on the program will suggest masking the region above 30, which may reduce the noise and improve the results of the calculation.

6.6 Reset All

There is a *Reset All...* button located at the bottom of the options window. Clicking this button resets all of the program options to their default settings. This option cannot be undone

CHAPTER 7

PolySNAP 2 Tutorial

This tutorial has been designed to guide the user through a few examples using *PolySNAP 2* with typical data that might be encountered in general use. It is not intended as a replacement for the full program manual, but as a basic introduction to actually using the program. It should therefore be read in conjunction with the manual itself for a more detailed explanation where necessary.

The tutorial requires the user to have already installed *PolySNAP 2*, and be familiar with Windows-based interfaces. The data files used in the tutorial are installed along with the software, and can usually be found in the *tutorial* folder in *C:\Program Files\PolySNAP2*.

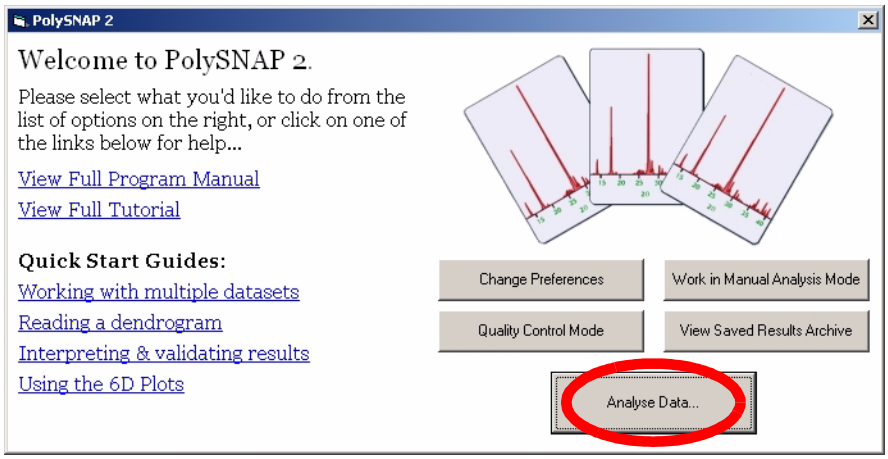
There are three main sections to the tutorial: getting used to working with a single dataset, working with multiple datasets, and working in manual mode with smaller datasets.

7.1 Automatic Analysis on a single dataset

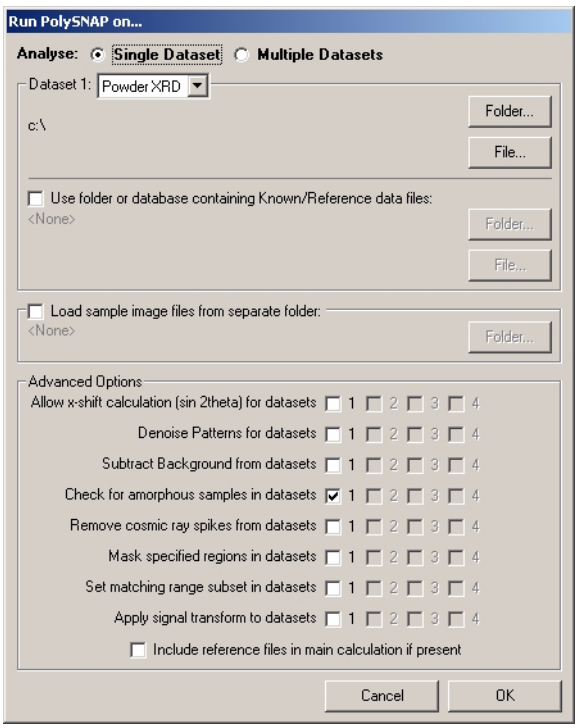
To gain experience with PolySNAP in automatic mode, a walkthrough of a simple run using 22 X-ray powder diffraction patterns is presented. The example assumes the program defaults are used.

To begin, launch PolySNAP 2 from either the icon on the Windows desktop or *via* the shortcut in the Windows *Start* menu.

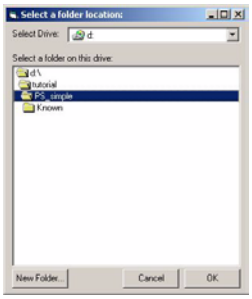
Once PolySNAP has been launched the user is presented with the *PolySNAP* welcome window.



From this, the first step is to define the input folder containing the necessary data. To do this click on the *Analyse Data* button. Alternatively, the same can be done from the *File* menu, by selecting *Automatic Analysis* and then the *Analyse Data...* option. The following dialog box should appear.



1. To specify the input folder containing the sample data files, click the *Folder* button in the top section of the dialog box. A folder-selection dialog box will appear.

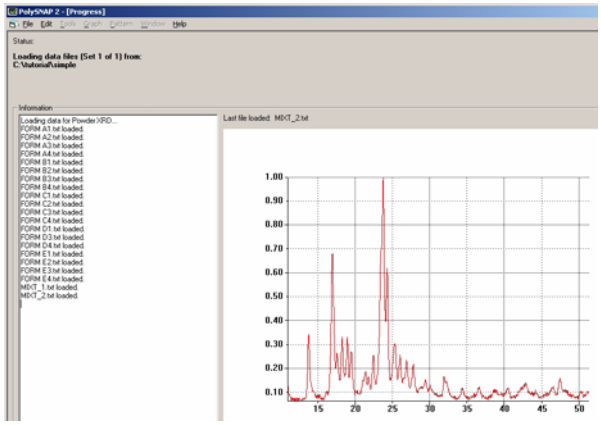


2. Select the drive containing the tutorial data from the pop-up menu at the top. In this case, select *C:*.
3. Navigate to the folder *C:\Program Files\PolySNAP2\tutorial\simple*. Ensure that the last folder of the desired path (*i.e. simple*) is selected by double clicking it (the folder icon should appear 'open').
4. Click *OK*. The selected path should be displayed in the upper portion of the dialog box. Check it is correct; if it is not, repeat the previous step.
5. Ensure that the other input sections has no defined folder, indicated by the word '<None>'. If this is not the case, uncheck the checkboxes labelled *Use Folder or Database containing Known/ Reference data files* and *Load sample image files from another folder*. The use of these settings will be explained later.
6. Leave the Advanced Options settings to their default values (all should be off except *Check for amorphous samples*).
7. Finally, click *OK* in the bottom of the window.

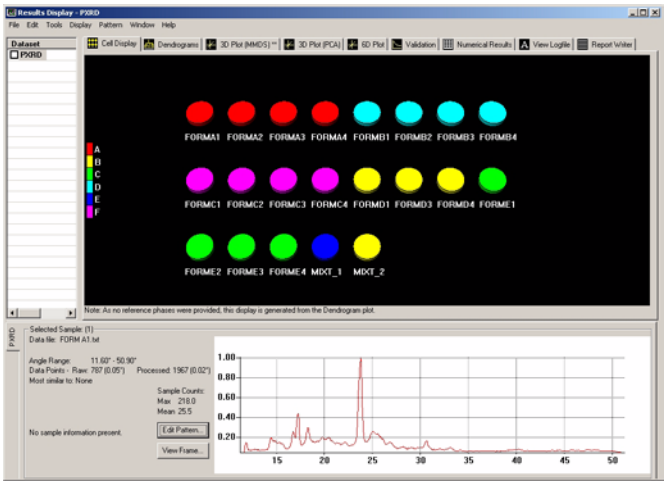
The pattern files from the specified input folder will now be loaded into *PolySNAP 2*.

If you get an error after clicking OK about no data files being found, double-check the data location specified is correct.

PolySNAP now proceeds by reading the sample files from the specified input folder, and loading these into the program.



Each pattern is loaded in turn, with the pattern profiles shown as they are read in. When all 22 patterns have been loaded, the program examines the input files, matches the data files against one another, and then performs cluster analysis. When complete the main results window will appear; this should only take a few seconds for this small datasest.

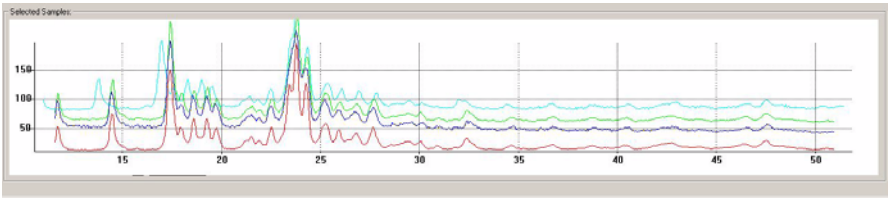


N.B. Depending on the size of your screen, the layout of the results window may not be exactly as shown in the screenshots.

There are now a number of ways to view and examine the *PolySNAP* results. The first of these - the default view displayed - is the *Cell Display*, which visually represents the contents of each sample. Each cell (shown as a disc) represents a different pattern from the input folder, with colour being used to denote the suggested grouping of compounds. In other words, similar samples are given the same colour. In the lower part of the window the sample information and pattern display of the selected cell is available. The first sample cell is selected automatically when the window is first opened.

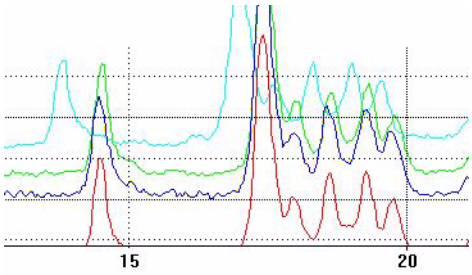
The program can be set to display a certain number of cells per row, so that the display represents, for example, the layout of a 96-well sample plate.

1. In the *Cell Display* click on cell *FORMD1*. The associated full pattern profile of pattern D1 is now displayed in the lower region of the window beside the sample information. In this particular case the information available is limited, so many of the fields are blank.
2. Hold down the *shift* key and click on cell *FORMD4*. The cells from D1 to D4 are now selected, and the sample information pane now disappears, leaving only an extended view of the profiles of patterns D1, D3 and D4 overlaid to allow a visual comparison. From this view they are all obviously the same compound.
3. From the *Display* menu select *Offset Overlaid Pattern Profiles* and click *on y-axis*. This will now display the multiple patterns with an offset along the y-axis of the plot, when any further cells are selected.
4. Hold down the *control* key and click on the other cell coloured yellow, *MIXT2*. The display now includes this new fourth pattern, and to allow an easy comparison are displayed overlaid with an offset along the y-axis.



5. Zoom into the area between around 15° and 20° by holding down the left mouse button and dragging a box over the desired area. When the mouse button is released, the graph region is redrawn to show just the selected area. Zooming in and out smoothly can also be accomplished by clicking in the display and then holding down the *control* key while moving the scroll wheel of the mouse (if available).
6. To move around the area and position the view more accurately, hold down the *Alt* key and *left mouse button* simultaneously while dragging with the mouse - the display updates in real time.

These zoom and movement functions are the same for all graphical displays within the *PolySNAP* program.



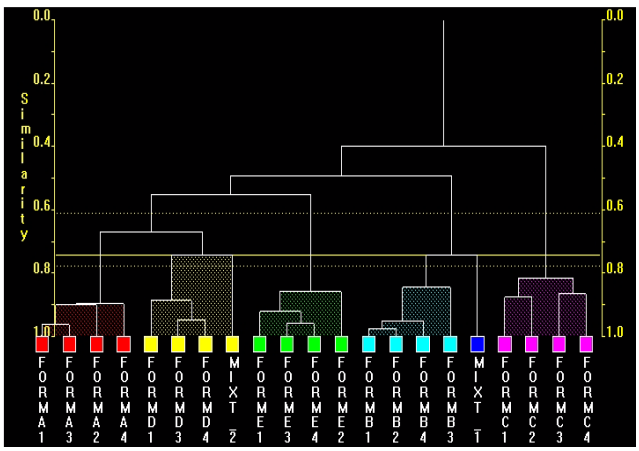
To check which of the plotted profiles correspond to which sample, 'hover' the mouse pointer over part of the plotted line; a tooltip appears with the sample filename in it.

This closer view of the overlaid patterns makes it easier to see that the top pattern (MIXT2) is noticeably different from the other patterns - note the extra peak around 14° for example - and, despite being coloured the same, may not actually belong to this group of compounds.

7. Reset the view with a *right click* of the mouse in the graph pane and selecting *Reset View* from the pop-up menu.

There are a series of tabs just below the menu bar, each of which display a different view of the data. Select the tab labelled *Dendrogram*.

The partitioning of the data into groups that were displayed in the coloured cells is carried out by the cluster analysis of the sample data. In *PolySNAP* there are five different methods of clustering available, each of which tend to give slightly different results. The program computes the best, in the sense of the most internally consistent, dendrogram method and displays the results from that (although the user can choose to view the results from the other methods and overrule the selected one if required).



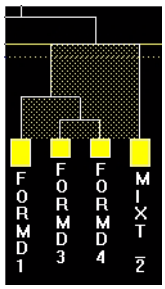
A dendrogram provides a visual display of the results from the hierarchical method of data classification using cluster analysis. The dendrogram itself takes the form of a tree-diagram in which each terminal branch (coloured box) is representative of a single pattern sample.

The higher up the similarity scale two samples are connected by a horizontal line (called a 'tie-bar'), the less similar they are. Hence samples FORMA1 and A2, with a similarity value of around 0.9 are very similar, whereas samples FORMA1 and FORMC1, which are only joined much further up the tree by a horizontal line with a similarity value of around 0.4, are quite different.

In this dendrogram there are 6 separate clusters, each distinguished by its own colour. These are the same colours displayed earlier in the cell display and throughout most of the other *PolySNAP* displays. The number of clusters are defined by the yellow cut-line which in this case was initially set to 0.708. The calculation of this level is *via* a number of statistics. The confidence levels on this choice of cut position are shown by the yellow dotted lines either side. Selecting a cut-line for a dendrogram is a difficult procedure, and the results must be treated with caution. The program-calculated level therefore should always be carefully examined by the user to see if looks sensible.

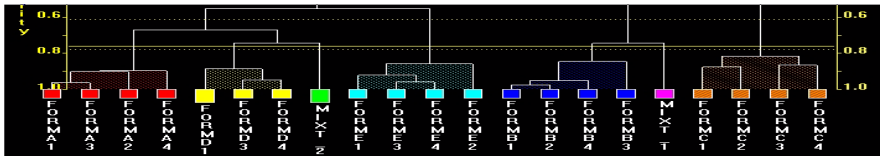
Adjusting the cut level upwards creates fewer separate clusters, and effectively reduces the discrimination between differences; adjusting the cut-level downwards creates more separate clusters.

1. In the *dendrogram* click on the yellow square cell *FORMD1*. The pattern profile of the sample is displayed.
2. Using the *control* key, also select cell *MIXT2*. These are the two patterns that appeared different from one another when they were overlaid in the cell display.

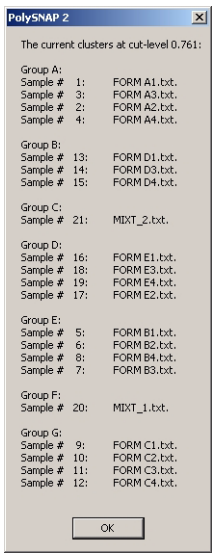


Looking at the position of the cut-line it is seen to be only very slightly above the similarity line between cell *MIXT2* and the other yellow cells. With the confidence levels indicated by the yellow lines being between 4 and 8 clusters, it is possible that the present level is not ideal.

3. To manually adjust the cut-line, either click in the Dendrogram area and use the *scroll wheel* or hold down *control* and the *left mouse button* while dragging up or down. Move the cut-line down slightly so that it is still above the lower confidence line but so there are now 7 different clusters. Notice that the assigned colours change, and that *MIXT2* is now in clusters of its own.



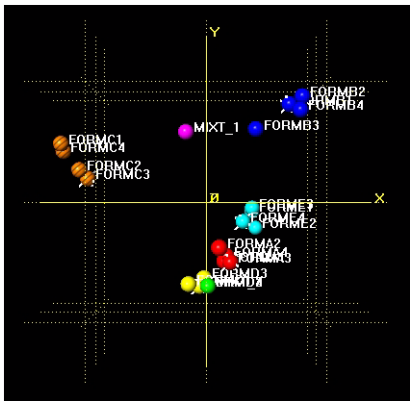
4. In the dendrogram area click the *right mouse button* and select *Save Modified Trees....*, to ensure that the changes are retained.
5. From the *Tools* menu select *List Pattern Cluster Members....* A dialog box will appear containing a list of the clusters.



This dialog box displays the different clusters with their sample numbers and actual file names. From the file names used in this demonstration example it can be seen that the samples are now all properly grouped, with the two mixtures separate from the rest. Notice in the dendrogram display that these two mixture patterns are quite dissimilar to anything else, having a low similarity connection value to their neighbours.

The differences between samples can also be viewed by making use of the distances derived from the correlation matrix to give a representation of the data in three dimensions. There are two methods used for calculating the resulting 3-D plots: these are *Metric Multi-Dimensional Scaling* (MMDS) and *Principle Component Analysis* (PCA). They both give different views of the samples because of the differences in calculation and will therefore give slightly different results. The control of each plot is the same so only the MMDS will be described here in detail.

Select the tab *3D Plot (MMDS)* (or PCA if you wish).



The initial view shows only the x and y axis, while the z axis lies in projection. Each point represents a sample. The position on the plot is taken from the MMDS calculation. The colour of each sample is taken from the dendrogram display to allow easy comparison of the results from these different methods. Allowing the mouse to hover over a sample displays the sample label in a tooltip popup.

Samples that are similar are plotted close to one another, so are seen to clump together in groups. Note that different coloured samples can also be close, this shows that they also have similarities. This can be seen by the yellow group being very close to the green sample, which is pattern MIXT2. When comparing with the dendrogram display, the MIXT2 sample is only separated from the yellow group by the current level of the cut-line.

Also notice the number plotted at the top-left of the display, in this case, 0.96. This is a correlation coefficient called the goodness of fit, that measures the quality of the 3D representation to the original data. The closer to 1.0 it is, the greater the reliability of the results. This value tends to decrease when larger data sets are used.

Use the following methods for exploring the 3-D plot.

Action	Control
Rotate the 3-D plot	Drag while holding shift key and left mouse button
Move plot laterally	Drag while holding alt key and left mouse button
Alter size of spheres	Drag up or down while holding control key and left mouse button
Zoom on centre	Click in area, hold shift key and use mouse scroll wheel

Zoom on area	Hold left mouse button and draw box over area
Select a sample	Click on sphere
Select multiple cells	Hold control key and select additional cell
Alter rendering quality	Press F2 and adjust scale
	(lower values are better for slower graphics cards)
Centre view	Right click mouse and select *Centre Selection*
Reset view	Right click mouse and select *Reset View*

The 3D plot is useful to spot patterns that are quite different from the others, as they tend to stand out on their own and are not easily grouped. Also, cases where the colours (from the dendrogram) and the positions (from the MMDS) of the samples appear to contradict each other are the samples that should be looked at manually in more detail.

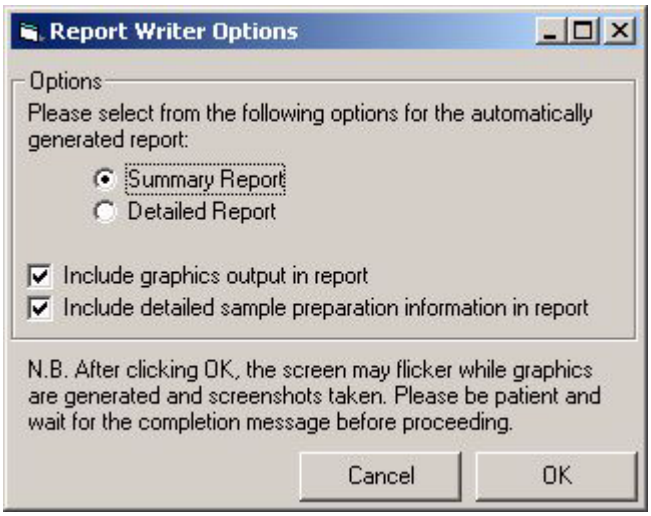
Finally, notice that some of the spheres have white 'spikes' protruding from them; these show which samples are the Most Representative Patterns of that cluster. Clicking on one brings up a dialog box showing the mean pattern-pattern distance for that cluster. The smaller the distance, the tighter the cluster, and the more similar the samples are within it.

7.1.1 Creating a Report describing the results

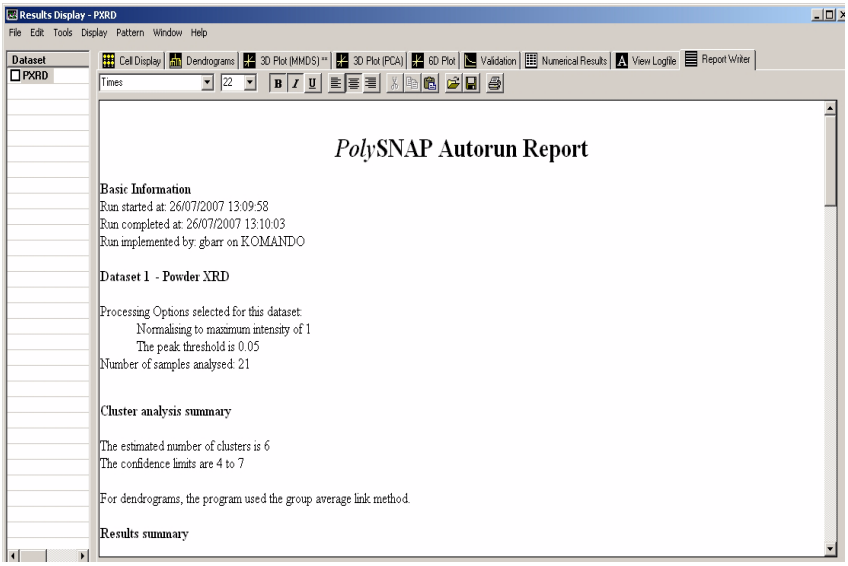
A basic report of the calculated results can be automatically generated using PolySNAP. This contains the information gained from the data and optionally some of the graphical views of the results that have been obtained.

1. Select the *Report Writer* tab.

2. From the *Tools* menu select *Generate Report...*. A dialog box will appear.



3. Select *Summary Report*. Deselect the checkbox labelled *Include detailed sample preparation information*, and then click *OK*. The report will now be generated and screenshots taken. This may take a few moments.
4. A Dialog box will appear confirming that "Report generation is now complete." Click *OK*.



The report begins by detailing when the calculations were done, who was the user and on which computer. It then continues with the settings used for the analysis and a summary of the results, including the cell display and dendrogram. If the *Detailed Report* option had been chosen, a more detailed report is presented, including output from all of the main graphic displays. The report may be edited, or additional information can also be added to the report manually, as in

a standard word processor, by using the simple formatting tools supplied.

For example, to add one of the profile plots to the report, switch to the Cell Display, and select say Form A1. In the profile pane at the bottom, right-click in the graph area and choose *Copy* from the pop-up menu. Then select the Report Writer tab, click in the report at the point where you want the image to be inserted, right-click and select *Paste* from the pop-up menu.

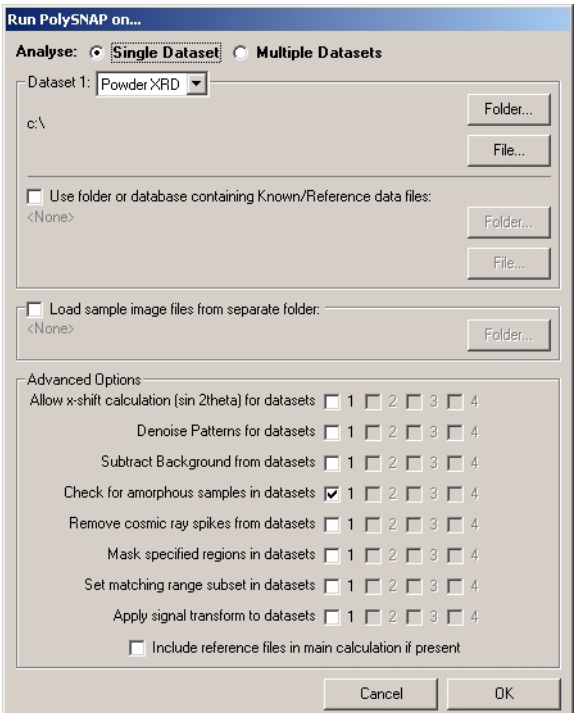
5. To save the report, click the *right mouse* button and select *Save as....*
6. A dialog box will appear. Use this to navigate to a suitable folder where the data can be saved and select a file name. *e.g. C:\PolySNAP Report.rtf.*

The report is now saved as an RTF file and can be opened and edited in any standard word processing package.

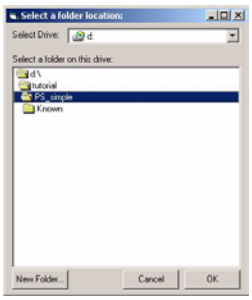
To end this session select *Close Window* from the *File* menu. If prompted to save changes to the report, select *No* because this has already been done. A second window will open asking if the results from the current run are to be saved. This would save the current results as a *PolySNAP* archive file (*.psnaparchive*) that could be accessed later, but this is not necessary at the moment. Select *No.*

Instead we will now run this dataset a second time, but this time there will be extra information and options used to give a more advanced understanding of the dataset. To begin, click on the white PolySNAP window background to bring up the welcome screen, and click *Analyse Data*, or select *Analyse Data... f*rom the *File -> Automatic Mode* menu.

The following dialog box should appear.



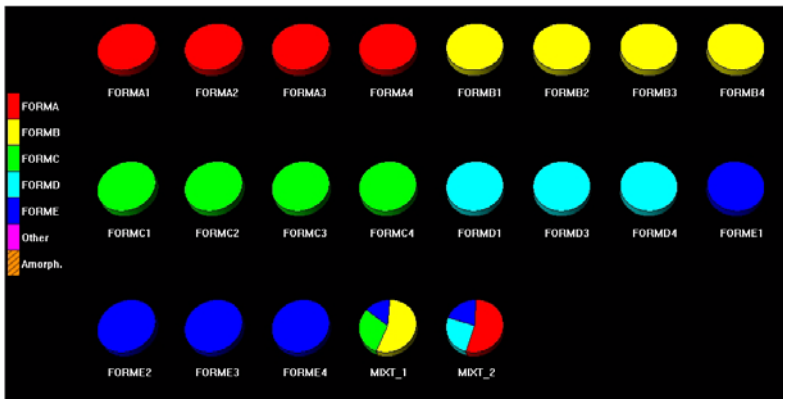
1. To define the input folder containing the sample data files, click the *Folder* button in the top section of the dialog box. A folder-selection dialog box will appear.



2. Select the drive containing the tutorial data from the pop-up menu at the top. In this case, select *C:*.
3. Navigate to the folder *C:\Program Files\PolySNAP2\tutorial\simple*. Ensure that the last folder of the desired path (*i.e. simple*) is selected by double clicking it (the folder icon should appear 'open').
4. Click *OK*. The selected path should be displayed in the upper portion of the dialog box. Check it is correct; if it is not, repeat the previous step.
5. Click on the checkbox next to the option *Use folder or database containing Known/Reference data files.* The *Folder* button can then be selected. Navigate to the *C:\Program Files\PolySNAP2\tutorial\simple\Known* and double-click to select this folder.

6. The *Load sample image files from another folder* section should still be empty and unchecked.
7. Finally, click *OK* in the bottom of the window.

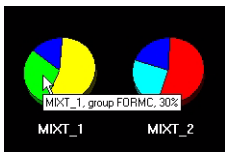
After the analysis has been performed the results open with the new cell display:



By selecting an extra option and including known reference files three important changes have been made to the cell display.

The reference files that were provided to the program are known patterns of pure phases that the samples in the dataset are to be compared to. Because of this the cells in the cell display are now coloured according to the known phases that they match and not according to the results from the dendrogram.

The presence of reference files has a more profound affect on the mixture samples.



Instead of being assigned to a particular cluster they are now analysed using the known phases as a reference, and how much of each known, pure phase present in each mixture is calculated. The cell is now displayed as a pie chart, divided up and coloured according to the percentage of the mixture that each known phase is thought to account for. Holding the mouse cursor over a slice of the pie chart will open a tooltip box displaying the phase, and percentage that the slice represents.

Finally, there are two extra keys that have been added onto the cell display legend.

Each known phase has its own entry, however there is now also a key called *Other*. Any pattern that did not match any of the known phases provided above a set similarity cut-off would be assigned here.

NB. - the Other group is not a group of similar patterns, but a collection of patterns that do not match the standard reference phases. However none of the samples in our dataset are in this group as they all match one of the phases, apart from the mixtures.

Switch to the *Logfile* tab. Scroll down through the information to the section headed *Quantitative Analysis*. The numeric output from the analysis can be found here.

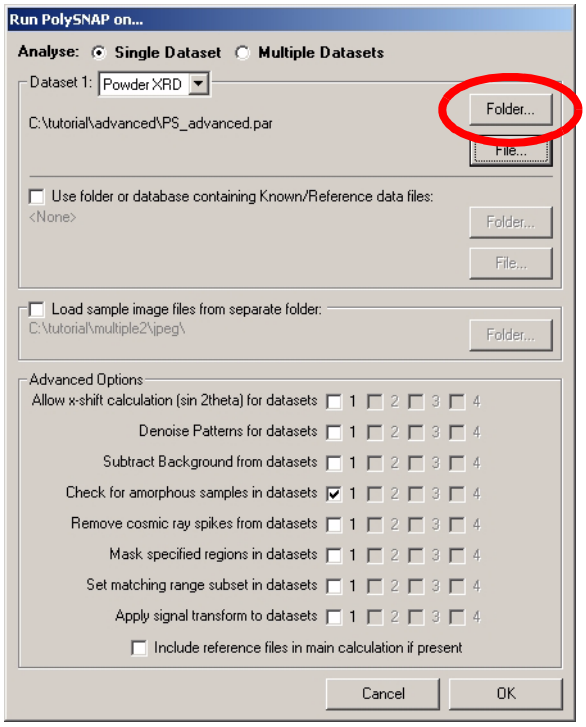
You can now close the Results window, as the first section of the tutorial is now complete.

7.2 Amorphous and Pattern Shifts

This part will cover some of the more advanced options available that include pattern matching while allowing for a 2θ-shift, and automatic identification of non-crystalline samples.

7.2.1 Loading Pattern Data from a Database

From the *File* menu, select *Automatic Analysis* and the *Analyse Data...* option. The following dialog box will appear.

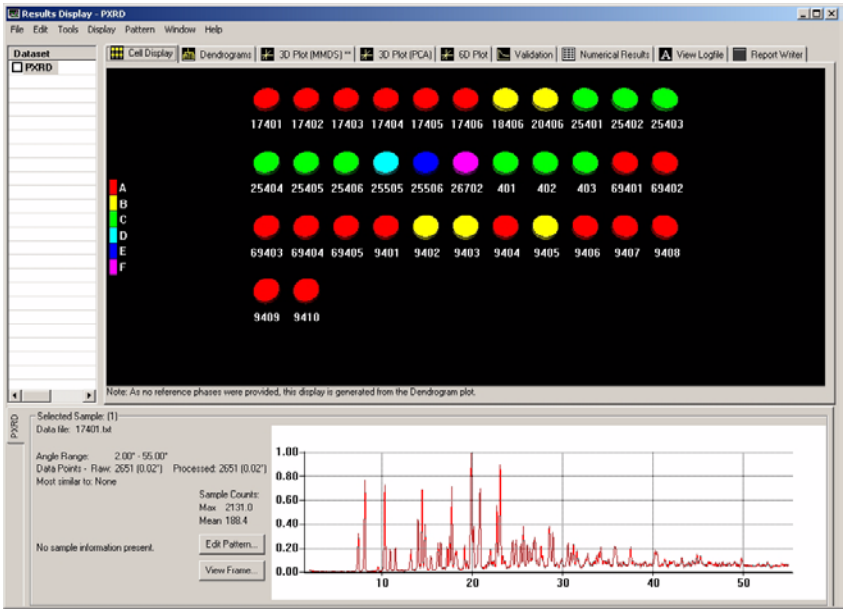


1. To define the input database containing the sample data, click the *Folder* button. A file-selection dialog box will appear.
2. Select the drive containing the tutorial data, in this case *C:*.
3. Navigate to the folder *C:\Program Files\PolySNAP2\tutorial\advanced*.
4. Ensure that no known phases are used by making sure the relevant checkbox is turned off.
5. Leave the Advanced Options settings at their default values.
6. Click *OK*.

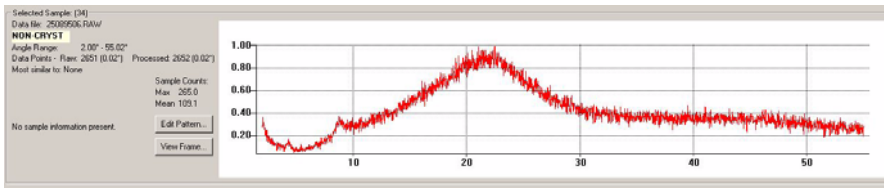
The pattern files in the specified folder are now loaded into PolySNAP, where it checks for amorphous samples, matches the patterns with one another, and performs a cluster analysis. This may take a little time to complete, but once finished the results window will appear with an initial view of the *Cell Display.*

7.2.2 Analysis of the Results

The 35 samples contained within the database are now presented in the *Cell Display.*

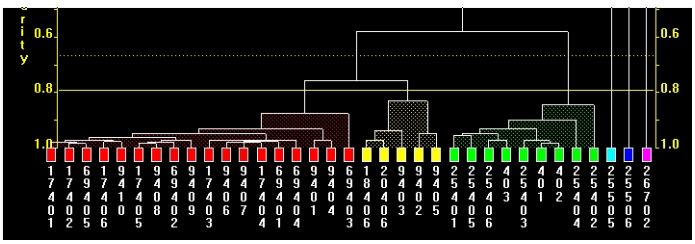


The program is using the last few digits of the filename of each sample to label them. Selecting pattern number 25505, the pattern information display shows that it has been labelled as a non-crystalline sample, which seems reasonable given its profile.

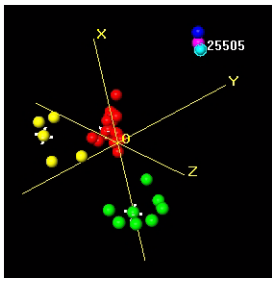


The other two patterns, 25506 and 26702 are similarly labelled for the same reason. Identification of such amorphous samples is done on the basis of checking to see if any signal (corresponding to peaks) would be left after subtraction of the entire amorphous hump. The method tends to err on the side of caution.

Looking at the *Dendrogram* tab:

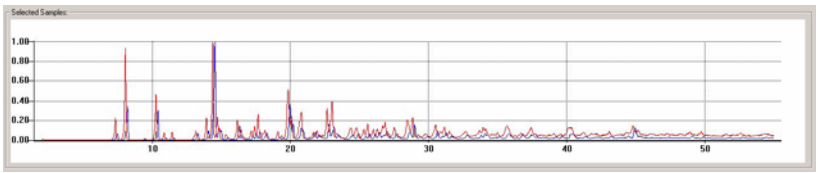


It is seen that three non-crystalline labelled patterns are placed on the far right of the diagram with a zero similarity to the rest. This is deliberate, in order to remove them from the main clusters. Compare this to the 3D (MMDS) plot:



Again, the three non-crystalline samples are quite separate from the rest of the patterns. Both the dendrogram and 3D plots suggest a loose grouping of the rest of the patterns, suggesting there may be some differences between them.

Using the *control* key, click two patterns which are on opposite sides of the dendrogram, for example patterns 17401 and 25402. Examining their profiles, it appears that in addition to some preferred orientation issues, there seems to be a noticable 2θ-shift between the otherwise relatively similar profiles:



Generate a full detailed report of the data including the graphics by selecting the *Generate Report* option from the *Tools* menu.

This completes the initial analysis of the data, but the 2θ-shift in some of the samples could be examined in more detail. From the *File* menu select *Close Window,* and do not save the results.

7.2.3 Reprocessing the Data allowing for an x-shift

When collecting powder diffraction data from a diffractometer the sample or instrument alignment can result in linear or non-linear shifts along the x-axis of the resulting pattern. This can especially be a problem if the sample height varies from sample to sample, giving rise to systematic errors in the pattern matching unless it is accounted for. However, to allow for this is a time consuming process and should therefore not be used unless such a shift is suspected - it is switched off by default.

A general expression for the shift is:

$$\Delta(2\theta) = a_0 + a_1 \sin\theta \qquad (7.1)$$

where the a_0 coefficient corresponds to a linear (zero-point) shift described earlier, and the a_1 coefficient a non-linear component [Zevin & Kimmel, 1995]. The requirement then is to find values of a_0 and a_1 that results in a maximum matching correlation result between two patterns.

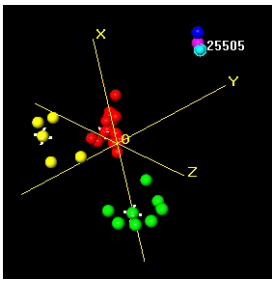
The same data as in the previous run will now be examined again. Unlike the last time, the program will vary the x-offset parameters to attempt to maximise the match result.

Define the input database, known phases and output folder as before:

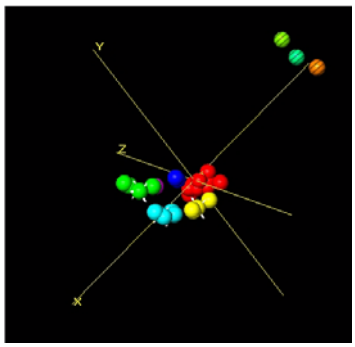
1. From the *File* menu, select *Automatic Analysis* and the *Analyse Data...* option. Define the same input data folder as used before in the previous run of PolySNAP.
   C:\Program Files\PolySNAP2\tutorial\advanced
2. Select *None* for the known phases directory.
3. In the Advanced Options area, turn on the *Allow x-shift calculation (sin theta)* checkbox.
4. In the *Run PolySNAP on...* dialog box click *OK*.

The same data as before is now run allowing x-shifts on the patterns. The time required to process this will take longer than a normal run of PolySNAP.

Now look at the 3D MMDS plot. Previously, it looked like this:

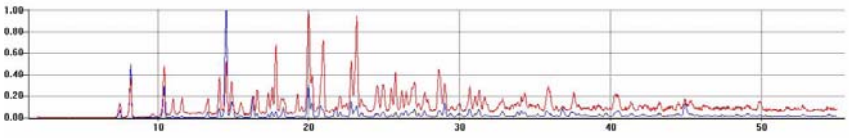


Now, with the option to calculate the best-offset value for each pattern turned on, it looks like this:

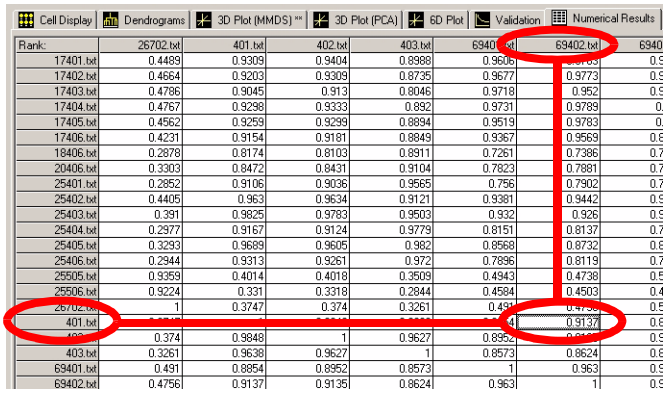


The three non-crystalline patterns are still quite separate, but the rest of the patterns have condensed together as a result of allowing for the 2θ-shift, showing that a large part of the differences between the pattern profiles was due to variation in sample heights during data collection. The program still separates them out within this grouping due to the preferred orientation problems.

Similarly with the dendrogram display, the similarity values between the patterns are much improved. The remaining differences appear to arise from preferred orientation effects, which are quite noticable in some cases. For example, overlay the profiles of samples 401 and 69402:



To see how much the patterns have been shifted as part of the calculation, switch to the Numerical Results tab. Locate 401.txt in the list on the left hand side, and read along to find where this row crosses the 69402.txt column:



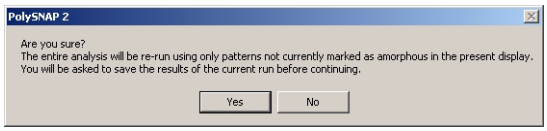
Rank:	26702.txt	401.txt	402.txt	403.txt	69401.txt	69402.txt	6940
17401.txt	0.4489	0.9309	0.9404	0.8988	0.9606		0.9
17402.txt	0.4664	0.9203	0.9309	0.8735	0.9677	0.9773	0.9
17403.txt	0.4786	0.9045	0.913	0.8046	0.9718	0.952	0.9
17404.txt	0.4767	0.9298	0.9333	0.892	0.9731	0.9789	0.
17405.txt	0.4562	0.9259	0.9299	0.8894	0.9519	0.9783	0.
17406.txt	0.4231	0.9154	0.9181	0.8849	0.9367	0.9569	0.9
18406.txt	0.2878	0.8174	0.8103	0.8911	0.7261	0.7386	0.7
20406.txt	0.3303	0.8472	0.8431	0.9104	0.7823	0.7881	0.7
25401.txt	0.2852	0.9106	0.9036	0.9565	0.756	0.7902	0.7
25402.txt	0.4405	0.963	0.9634	0.9121	0.9381	0.9442	0.9
25403.txt	0.391	0.9825	0.9783	0.9503	0.932	0.926	0.9
25404.txt	0.2977	0.9167	0.9124	0.9779	0.8151	0.8137	0.7
25405.txt	0.3293	0.9689	0.9605	0.982	0.8568	0.8732	0.9
25406.txt	0.2944	0.9313	0.9261	0.972	0.7896	0.8119	0.7
25505.txt	0.9359	0.4014	0.4018	0.3509	0.4943	0.4738	0.5
25506.txt	0.9224	0.331	0.3318	0.2844	0.4584	0.4503	0.4
26702.txt	1	0.3747	0.374	0.3261	0.491	0.4756	0.5
401.txt						0.9137	0.9
402.txt	0.374	0.9848	1	0.9627	0.8854	0.9135	0.9
403.txt	0.3261	0.9638	0.9627	1	0.8573	0.8624	0.9
69401.txt	0.491	0.8854	0.8952	0.8573	1	0.963	0.9
69402.txt	0.4756	0.9137	0.9135	0.8624	0.963	1	0.9

Clicking on the rank value shows the two profiles overlaid in the bottom pane; hovering the mouse over this value shoes a tooltip with the calculated values of a0 and a1 for this pair of patterns.

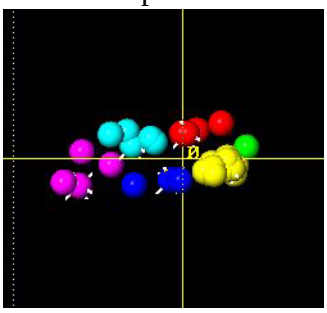
Create an auto-report on the run. This report will not have detailed any manual changes that might have been made to the dendrogram cut level. Therefore any information that has been changed can be included by adding them manually or copying from the *Logfile* pane, which details all the calculations and changes made during the PolySNAP session. Note that the use of an x-offset calculation has been noted in the output, as has the identification of the non-crystalline samples.

In cases where there are a larger proportion of amorphous samples in a dataset, it may be helpful to remove these samples and re-run the analysis without them to give a clearer idea of what is going on. This can be done automatically by selecting the *Tools* menu item *Rerun Analysis*, and from the sub-menu selecting *Ignoring All Non-crystalline Samples*.

A dialog warning box appears.



Select *Yes*. After being prompted to save the current results, a new analysis is performed, and a new results display window is opened as before. Note that this time, there are no amorphous samples visible in either the dendrogram or the 3D plots.



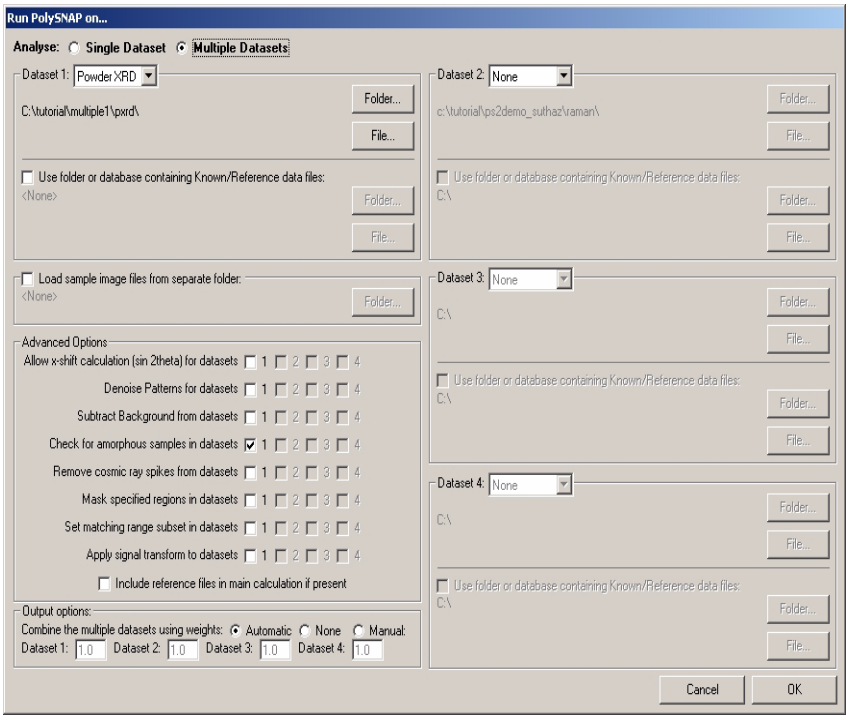
7.3 Multiple samples - Dataset 1

This tutorial involves the input and comparison of multiple datasets at the same time. This will demonstrate how combining results from different techniques can be useful and how looking at multiple datasets can yield extra information not available from just one.

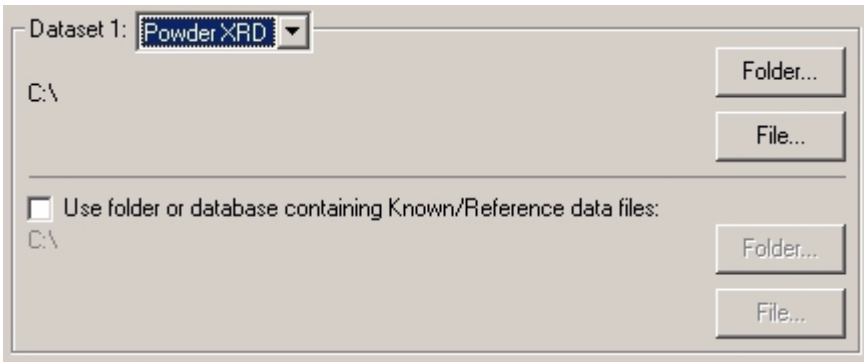
As before, first select *Analyse Data* from the welcome window:

This opens the data input window. When first opened there is only one section for inputting a single dataset. To input multiple datasets at the same time select the *Multiple Datasets* option from the top of the window:

This expands the data input window:

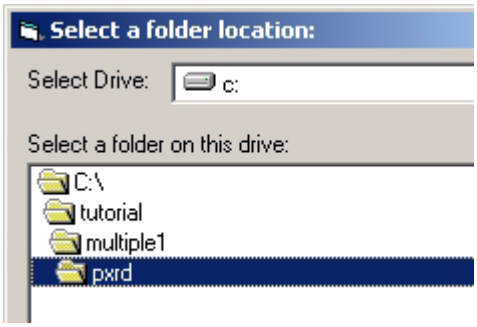


It now has four sections for entering up to four datasets; we will be using just two in this example. Each input section is identical:



There is a pull-down menu near the top that allows selection of the type of dataset that is going to be used. To the right there are buttons at the side that allow the location of the dataset to be specified. There is also the option to input known or reference files can be entered however these will not be used in this example.

For this example the first dataset is a dataset of powder X-ray diffraction data and the dataset is contained inside a folder. Ensure that the dataset type is set to *Powder XRD* and click on the *Folder...* button. This opens the following window:



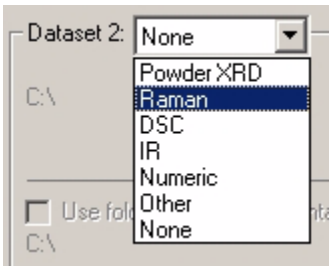
1. Select the drive containing the tutorial data from the pop-up menu at the top. In this case select *C:*.

2. Navigate to the folder *C:\Program Files\PolySNAP2\tutorial\multiple1* and open the folder called *pxrd.* This contains the Powder X-Ray diffraction data. Ensure that the last folder of the desired path (*i.e. pxrd)* is selected by double-clicking it. The folder icon should appear open, as in the screenshot above.

3. Click *OK*. The selected path (*C:\Program Files\PolySNAP2\tutorial\multiple1\pxrd*) should now be displayed in the upper portion of the section for Dataset 1. Check it is correct. If it is not, repeat the previous steps.

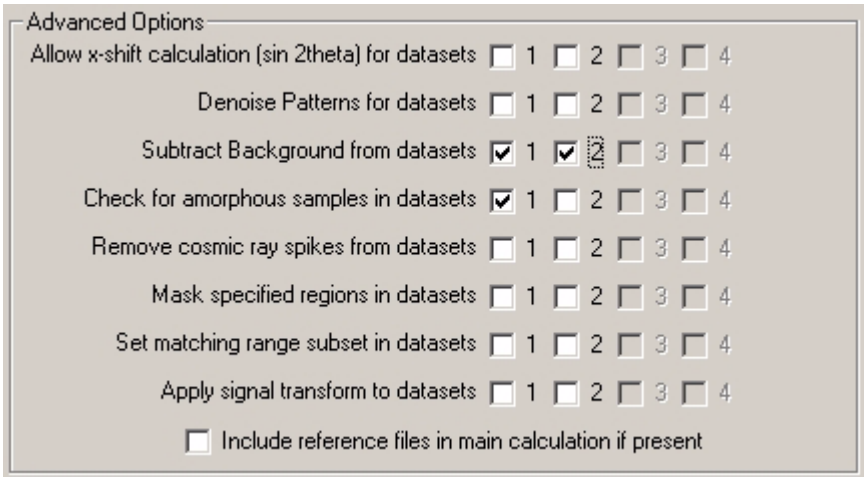
If we were to click *OK* in the main input window now, it would start an analysis run using the single dataset that has just been entered. However for this run two different datasets are going to be used.

The second dataset is a folder of Raman data. In the input section for *Dataset 2* select *Raman* from the pull-down menu.



Click on *Folder...* for *Dataset 2* and select the *Raman* folder located in same location as the *pxrd* folder and click *OK*. The filepath of the second dataset should now be displayed in the main window (*i.e. C:\Program Files\PolySNAP2\tutorial\multiple1\raman*). Now both datasets have been entered and both will be used in the analysis.

Below the input sections there is a section of advanced options. These are all the processing that can be applied to the data as it is being input and before it is fully analysed.

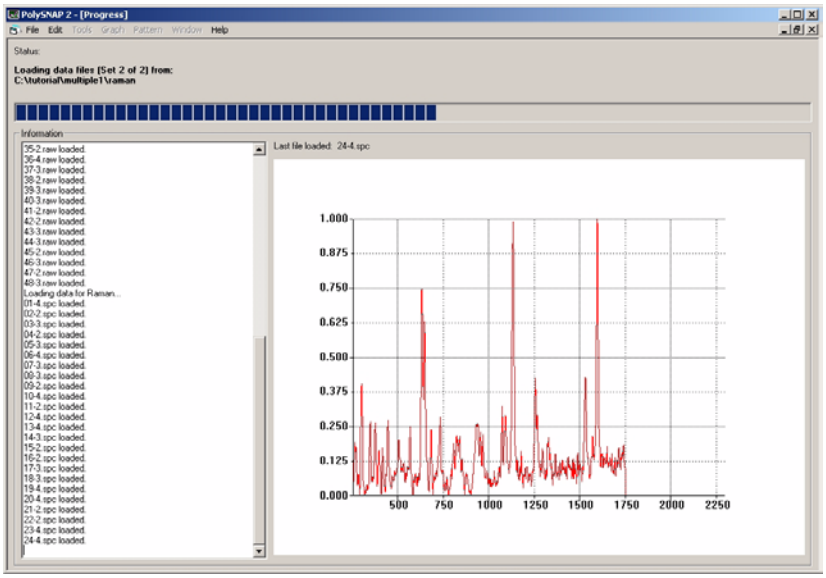


These options can be applied to individual datasets selectively. In this case only *Datasets 1* and *2* are being used, so only these two datasets have selectable checkboxes. The checkboxes for *Datasets 3* and *4*, which are both currently empty, are grey and unselectable.

Select to Subtract Background from both datasets 1 and 2.

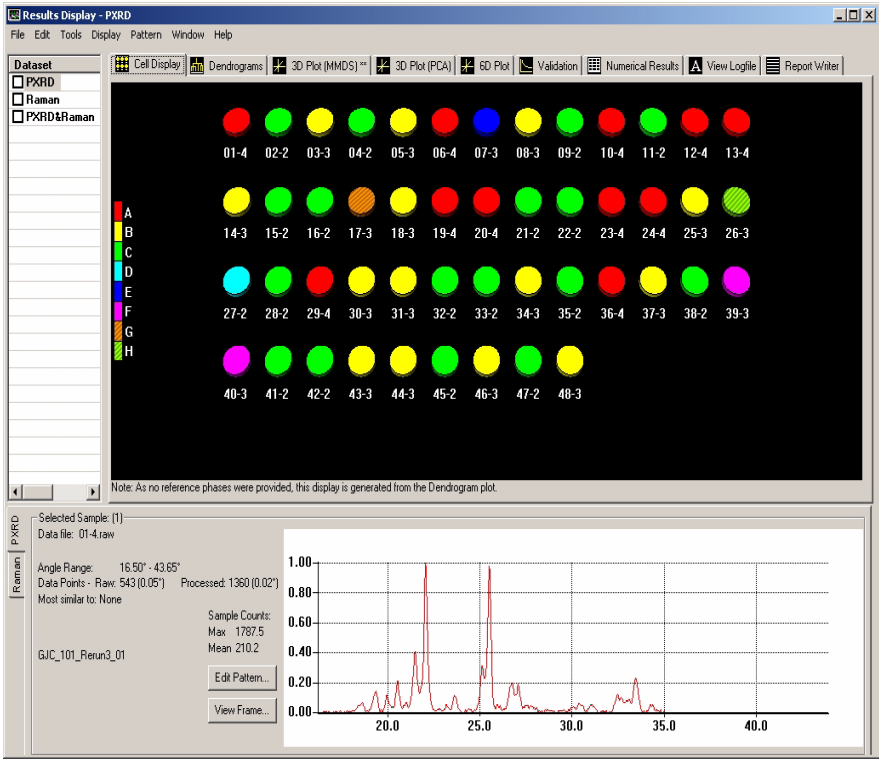
All of the relevant instructions have now been provided to the program. Click *OK* in the main input window to start the analysis.

A progress window will open as *PolySNAP* reads in the two datasets in turn, before performing analysis on both of them.

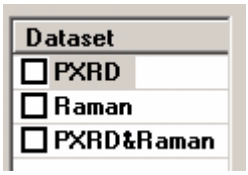


After analysis is performed on each dataset separately, a third analysis is carried out to create the combined results using

information from both datasets. Once this analysis is complete the progress window will close and the results display will appear:



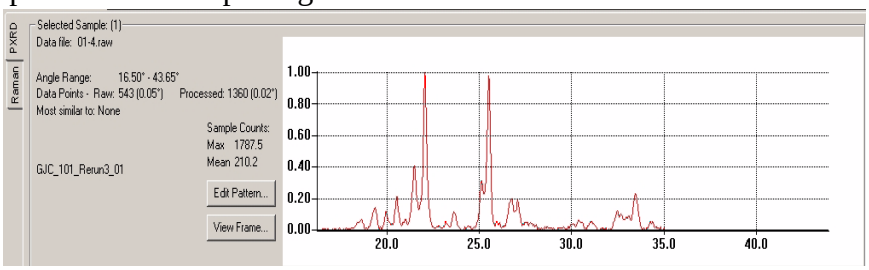
By default it will display the cell display of *Dataset 1* which is the Powder X-Ray dataset. There is a list of all the available datasets in a menu running along the left hand side of the display.



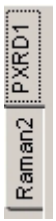
In this case there are three available datasets - the original datasets that were input (*PXRD* and *Raman*) and the combined dataset that was obtained using both the Power X-Ray and Raman data (*PXRD&Raman*).

Click on the **label name** *e.g.* **Raman** (not the checkbox) to switch to the results display for that dataset. The tab bar along the top can be used to switch between the display screens (Cell Display, Dendrogram, 3D plot *etc.*) within each datatype as before.

At the bottom of the results display is the display pane that shows the spectra for the sample patterns as well as any image files that were specified when inputting the data.

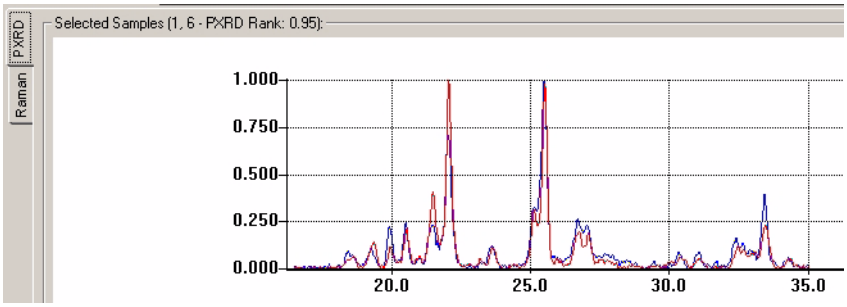


There is a tab bar running vertically along the side of this display that allows the spectra being displayed to be changed between showing either the original PXRD or Raman spectra for the currently selected pattern.



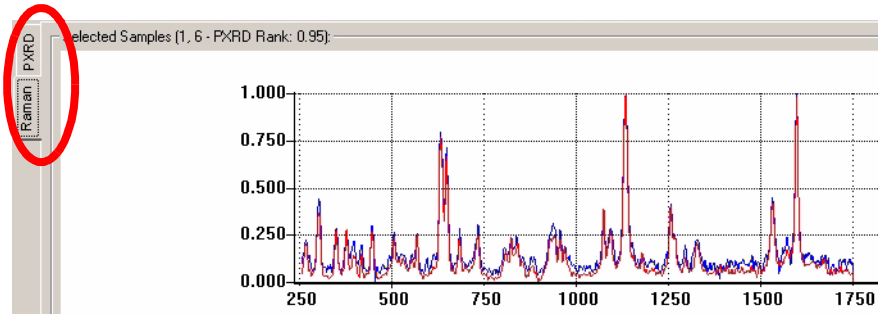
Spectra are only available for the two original datasets and these can be alternated between at anytime by using this tab. Changing the type of pattern shown has no effect on the main display.

Switch to the PXRD results display, by clicking on the label **PXRD** in the list in the top-left. Now, using the control-key, select two patterns of the same colour in the cell display - for example, *01-4* and *06-4*. The two PXRD profiles are displayed overlaid in the bottom portion of the window:



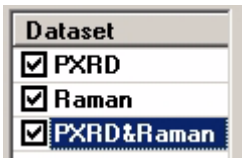
Based on the profiles, it looks reasonable that these have been clustered together. Now, switch to viewing the Raman profiles of the

selected samples, by choosing *Raman* from the vertical tab-bar on the lower left corner.

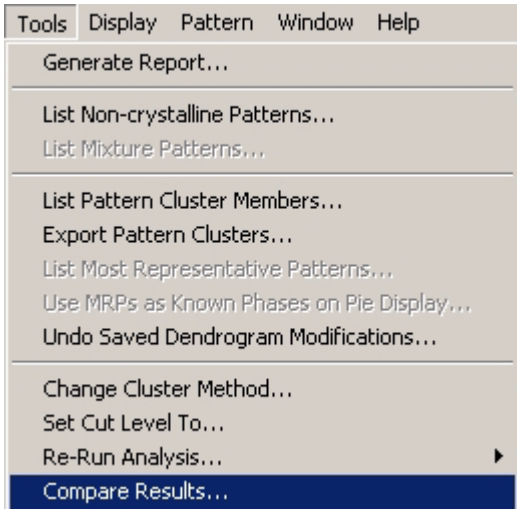


By comparing the profiles from the two separate methods, and seeing they are in agreement, we can have greater confidence that the analysis has been correct.

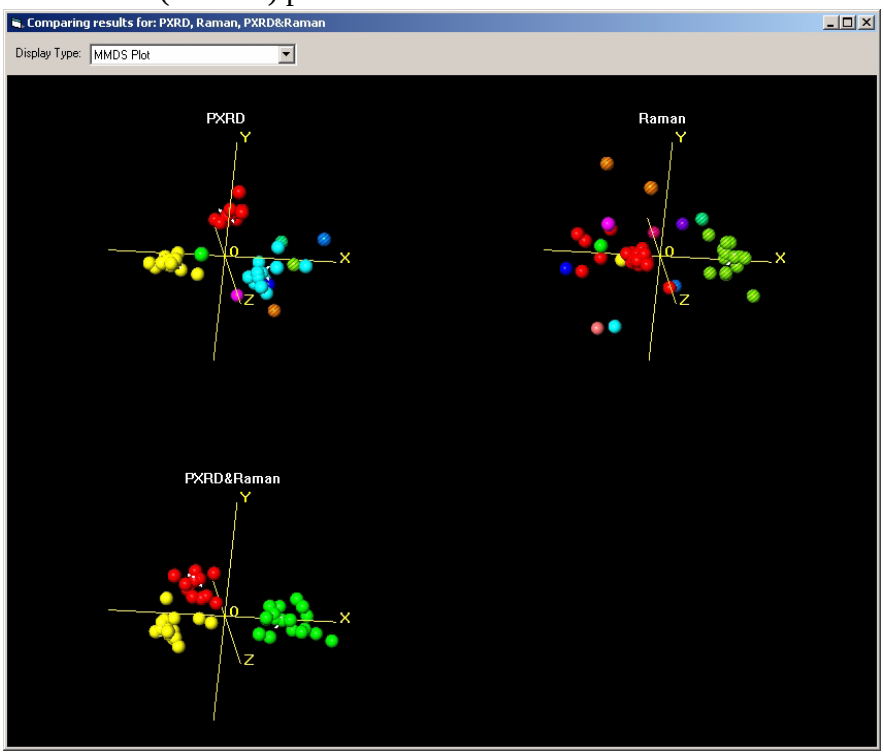
As well as switching between different displays, the results can also be compared directly in a new window. To do this select all of the checkboxes in the dataset list.



Select *Compare results...* from the *Tools* menu.

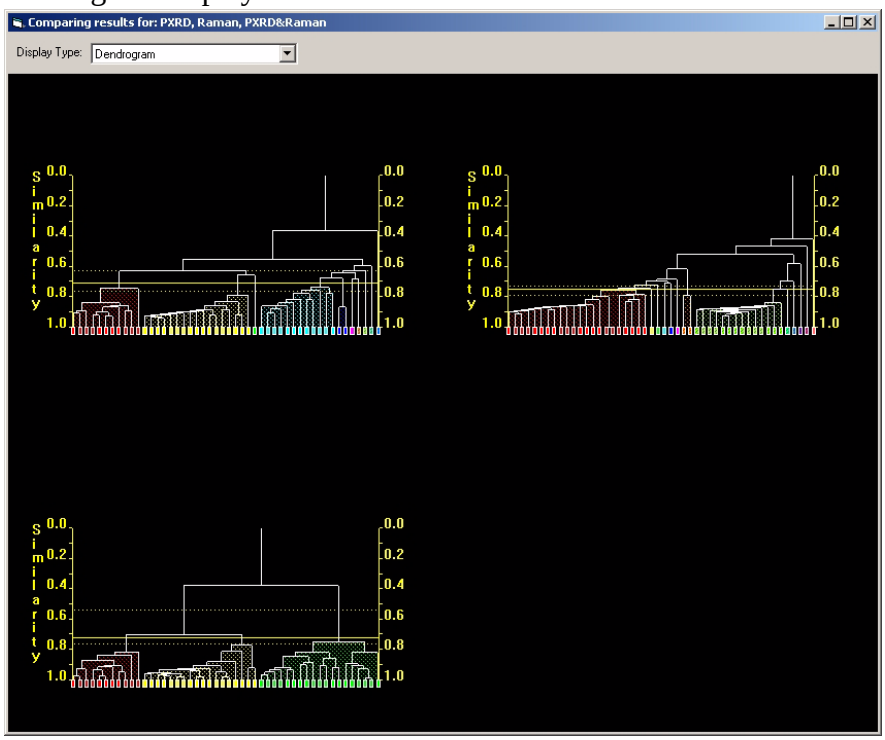


This opens a new window where the visual results from the three selected datasets are displayed next to each other. By default it opens with the 3D (MMDS) plot.



In this way the features of the different 3D plots can be directly compared, as the display can be rotated and zoomed into like a normal 3D plot. Here, we can easily see that the clusters for the combined PXRD&Raman dataset seem to be better than for either of the individual original datasets.

There is a pull-down menu in the top-left of the window that allows the 3D plot (MMDS) to be changed either to the 3D plot (PCA) or the dendrogram display.

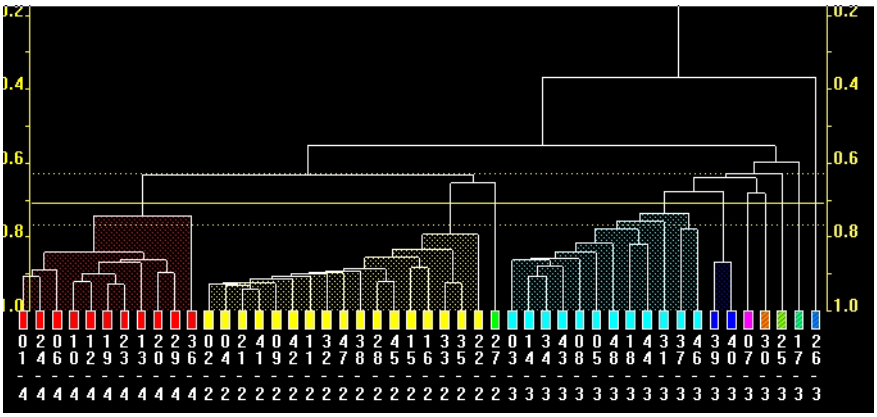


Now, let's look in more detail at the dendrograms of each of the datasets. Close the *Compare Results* window, and switch to the PXRD dataset view, then select the *Dendrogram* tab.

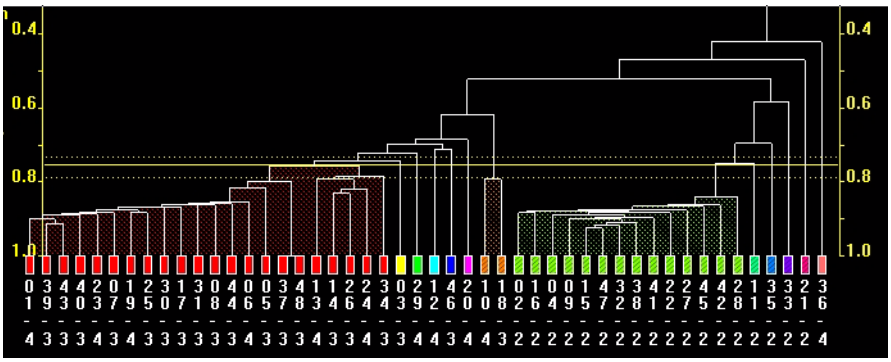
This dataset consists of three forms of a particular compound. The last character of the pattern labels show if a given sample should be Form 2, 3 or 4.

From the initial default cut-level of the PXRD dendrogram, it can be seen that using PXRD data alone has done quite a good job of partitioning the data into clusters corresponding with the different forms; with the only problem being the Form 3 samples have been

split into different places on the dendrogram (blue cluster, and group of patterns at the far right):

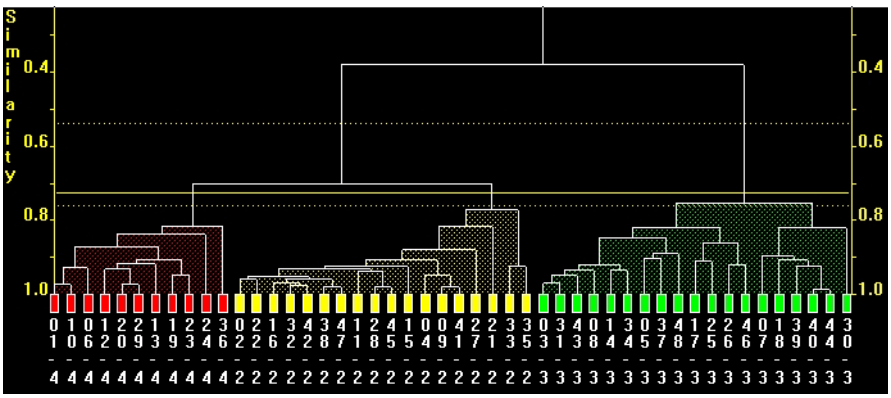


Now, switch to looking at the Raman results, by clicking on *Raman* in the list on the upper left. The display will update to show the dendrogram based on purely the Raman spectra provided:



Clearly, Raman by itself has not done as good a job as the PXRD in this example; the red cluster has forms 3 and 4 all mixed up, with the Raman unable to distinguish between them.

Now switch to the dendrogram for the combined PXRD&Raman results:



This has identified the 'correct' cluster for every pattern in the dataset, showing the value of combining results from multiple techniques.

7.4 Multiple samples - Dataset 2

Close any open results windows, and click in the white background of the main PolySNAP window to bring up the welcome screen. Select *Analyse Data*, and select the following options:

Dataset 1: Powder XRD

Data folder: *C:\Program Files\PolySNAP2\tutorial\multiple2\pxrd*

Background subtraction on.

Set Matching Range Subset to 5 - 30 degrees.

Dataset 2: Raman

Data folder: *C:\Program Files\PolySNAP2\tutorial\multiple2\raman*

Background subtraction on.

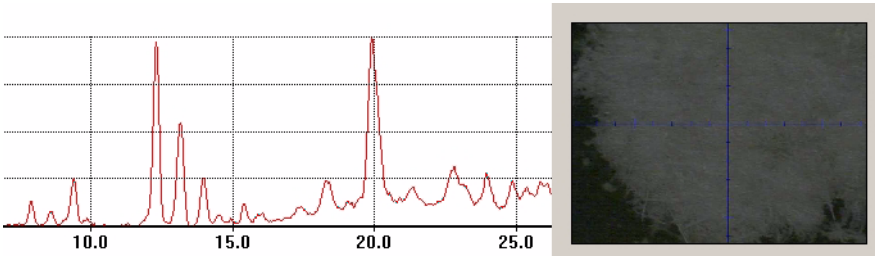
Set Matching Range Subset to 250 - 1750.

Finally there is an option that allows image files of the samples to be included and displayed along with the results.

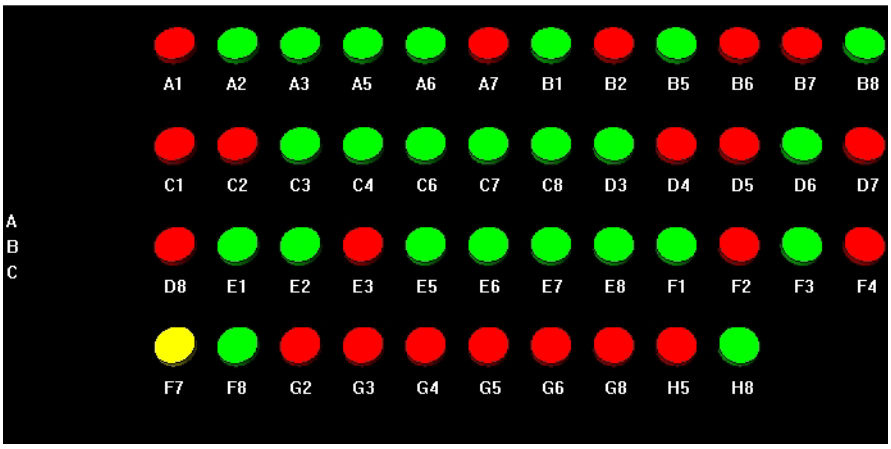
Click on the checkbox to activate this option then click on the *Folder...* button. Select the *jpeg* folder from the same location as before and click *OK*. The filepath (*C:\Program Files\PolySNAP2\tutorial\multiple2\jpeg*) should now be displayed.

Click OK to begin the analysis.

The results window appears, including in the bottom-right, a thumbnail view of the sample image corresponding to the currently selected sample in the display. To see this image full-size in a separate window, click on the thumbnail view.

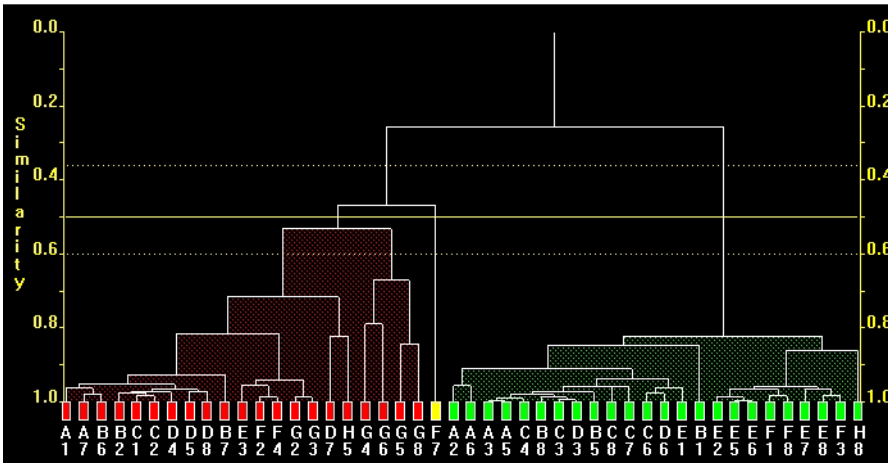


Look at the cell display for the combined results (*PXRD&Raman*).



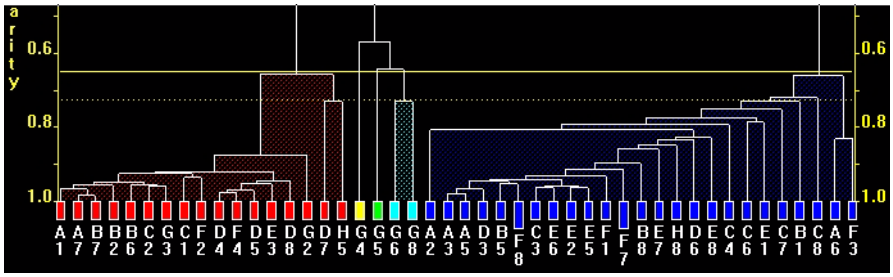
Most of the patterns have either been grouped into the red cluster or the green cluster; interestingly the yellow pattern *F7* has been assigned to a cluster of its own.

Switching to the dendrogram it can be seen that *F7* is quite isolated and highlighted as markedly unusual compared to everything else:

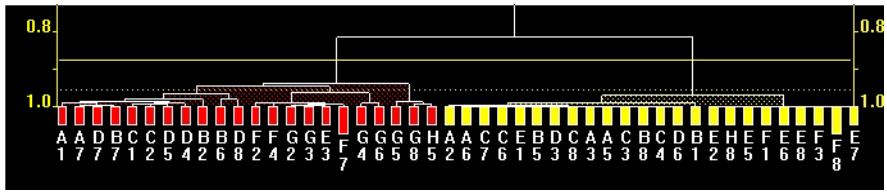


There is a very high tie-bar between it and the any other pattern. The combined results have identified this as pattern of interest. The results from the dendrogram are very similar to those 3D (MMDS) plot where *F7* is plotted noticably far from the other samples.

To understand this, switch to the dendrogram of *Dataset 1 - PXRD*. In the X-Ray data results there appears to be nothing unusual about pattern *F7*. It is in the same cluster as a similar sample, *F8*.

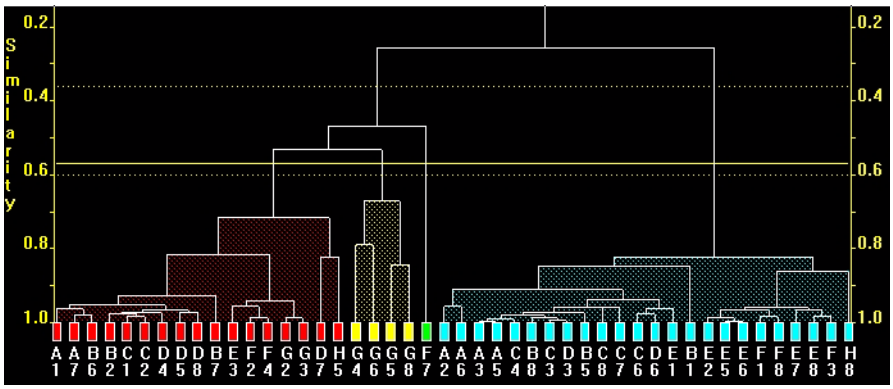


Switch to the raman dataset dendrogram.

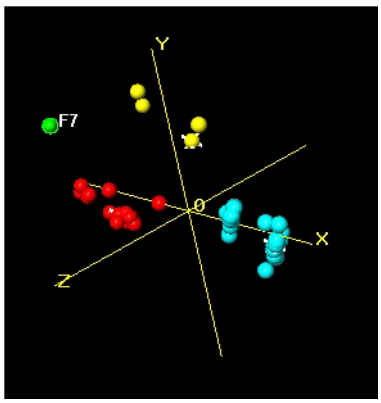


According to the Raman data, these two samples are in completely different clusters, contradicting the PXRD results.

Switch back to the dendrogram for the Combined results. Lower the cut-level to around 0.6, separating out G4, G5, G6 and G8.



Switch to the 3D MMDS plot, and see that the cut-level chosen describes the clusters shown there as well.



We now have a result that well describes the dataset; two large clusters of pure material A and B, four mixtures of A&B (the yellow cluster), and the one odd outlier.

Looking at each individual dataset display alone there would be no indication that there is anything unusual about *F7*. However by combining the results, it becomes clear there is a contradiction between the two methods, and this results in this particular pattern being highlighted for further investigation by the user.

This shows the value of having different techniques available to examine a sample, and the worth of combining the results from each. With a single dataset this unusual sample that needs to be brought to the user's attention can end up hidden amongst the rest of the results. It is only when the results are combined that the discrepancy becomes clear, drawing the attention of the user to that sample and allowing it to be investigated in more detail.

7.5 6-Dimensional Plots

There are a number of methods for analysing groups of patterns, some of which were detailed in the previous examples. However, these methods only concerned the powder diffraction pattern itself and not any other associated data that may be available. For example, this additional information can include sample preparation details such as the solvent used, concentration, pH, volume, temperature, *etc.*, and up to three of these can be incorporated into the 6D-plot at one time.

To begin, start with only the empty PolySNAP window containing the menu bar.

From the *File* menu, select *Automatic Analysis* and the *Analyse Data...* option.

Input Folder: *C:\Program Files\PolySNAP 2\tutorial\6D-plot*

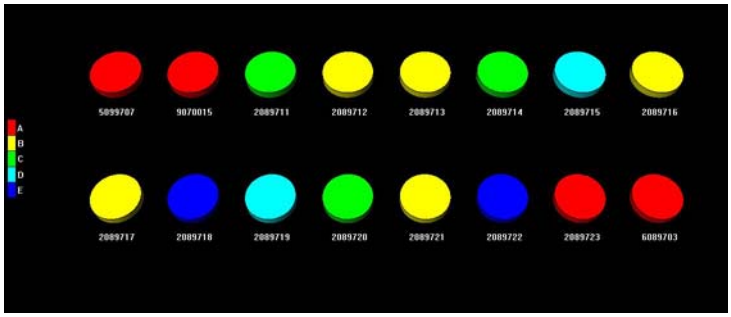
Known Data Files: *<None>*

Datatype: Powder XRD

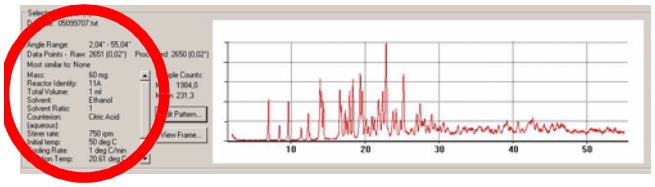
Click *OK*. The pattern files in the specified folder are now loaded into PolySNAP, which carries out pattern matching and cluster analysis of the data as before. Once complete, the results window will appear with a view of the *Cell Display.*

7.5.1 Analysing Results using 6-D Plots

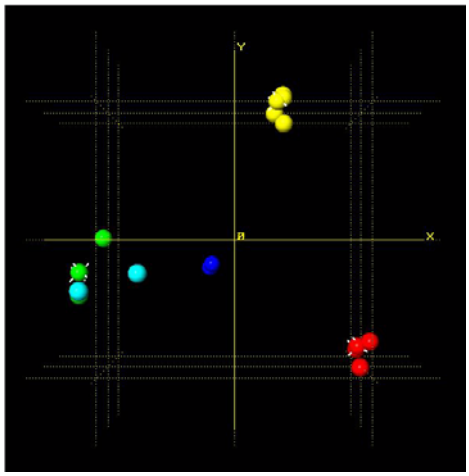
The 16 patterns samples contained within the input folder are presented in the *Cell Display*, where it can be seen that there seem to be five different types of pattern.



When an individual pattern is selected it can now be seen that not only is there a view of the pattern profile, but there is also textual pattern information (on the left).



The sample information shown is embedded in the pattern data file, Select the tab labelled *3D Plot (MMDS)*. The following display appears:



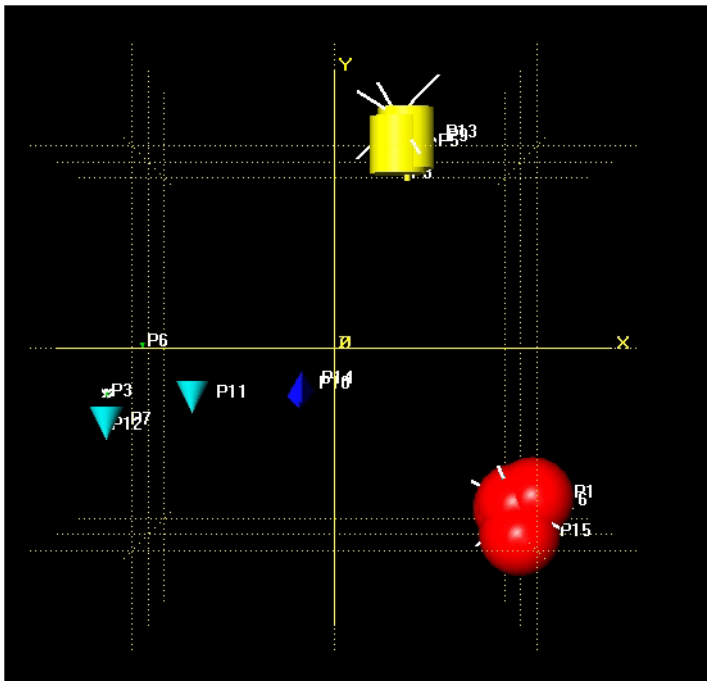
This view displays several distinct main clusters with the cells spread out along the x-axis. By selecting the cells individually, and examining the sample information pane, it can be seen that there appears to be a correlation between the solvent used in preparing the samples, and the different groupings of patterns. For example, all of the red samples are ones that were prepared with ethanol as solvent.

This information can be viewed more easily in a graphical manner, by enhancing the 3D-plot by plotting additional information on it, in the form of extra dimensions. For example, we can plot the different types of solvent used in terms of different shapes, and different reaction times in terms of different sizes of each plotted point.

1. From the main toolbar select *6D Plot...*. At the top of the frame several options are presented.

2. Set the *Plot* option to *MMDS*.
3. Make sure the *Ball Size* option is checked, and select to plot *Reaction Time*.
4. For the *Ball Colour* option, select *Dendrogram Colours*.
5. Ensure *Ball Shape* is selected and set to *Solvent*.

6. Click *Apply.*

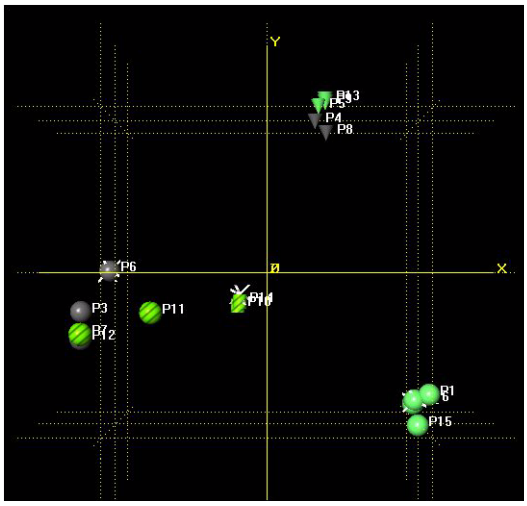


It can be seen that while the plotted points are in the same positions and colours as before, there are now several different sizes and shapes on the display. Note that the different shapes are clustered together - spheres are next to spheres, cylinders are next to cylinders, *etc*. Select one of the sphere-shaped patterns by clicking on it. Note that the solvent used in all of the patterns plotted as red spheres is ethanol; whereas all the patterns plotted as cylinders, for example, used methanol. While this information was available before, it is easier to spot trends in the data when viewed in this manner.

From the different sizes of shapes plotted, it can be seen that there were three different reaction times used in this set of experiments. This may be seen more clearly by replotting the display, using colour instead of size to represent the different amounts.

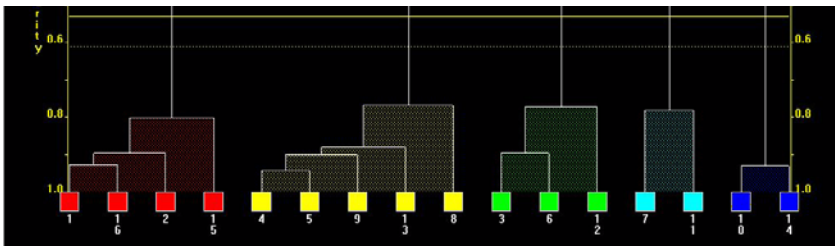
Deselect the *Ball Size* option, turn on *Ball Colour* and set it to plot *Reaction Time*, and ensure *Ball Shape* is still set to *Solvent*. Click *Apply.*

The new plot now looks like this:

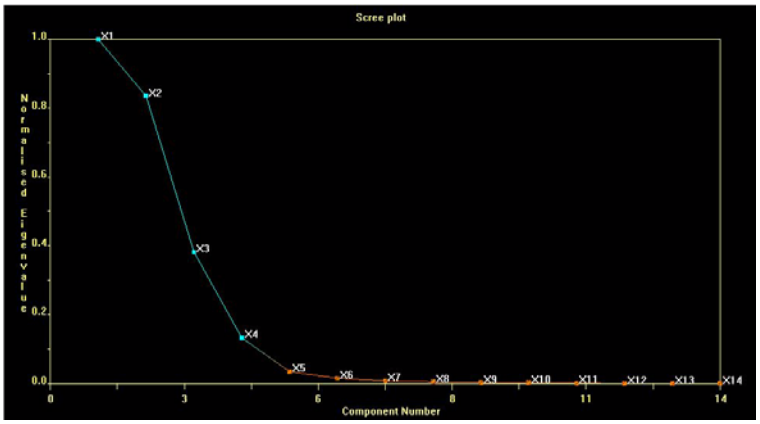


The different colours - green, grey and striped, represent the differing reaction times used. It is interesting to note that the two patterns 4 and 8 come up as separate colours - and hence times - from the rest of the patterns they are similar to.

Switch to the *Dendrogram* tab. From the *Display* menu, uncheck the option *Show Well Identity If Available*. Notice that pattern 8, in yellow, seems more separated than any of the rest of the yellow patterns.



We need to consider if the cut-line is positioned correctly. Switch to the *Validation* tab and check that *Scree Plot* is selected in the pop-up menu in the upper left corner.



The scree plot is one of the many methods used for estimating the cut-level of the dendrogram, and can be a useful visual aid for the user. It is derived from the eigenvalues of the pattern correlation matrix. To interpret the scree plot the gradient between the values (number of clusters) need to be analysed. Where the eigenvalues are very different from one another there will be a steep gradient between clusters. This is one of the methods used to select the number of clusters.

For this dataset the number of clusters suggested from this is between 4 and 5, because it is here that the plotted line changes colour, and the gradient starts to level out. 5 clusters corresponds to the current cut level; 4 would correspond to raising the cut line such that the green and blue patterns were considered part of a single cluster.

In cases such as this where the graphic displays may not give a clear-cut solution, it can often be useful to examine the detailed statistical output from the analysis. This is shown in the *Logfile* tab. Switch to it, and scroll to the very bottom of the output. Scroll back up slowly, until you see the output from the cluster analysis. One section of it reads as follows:

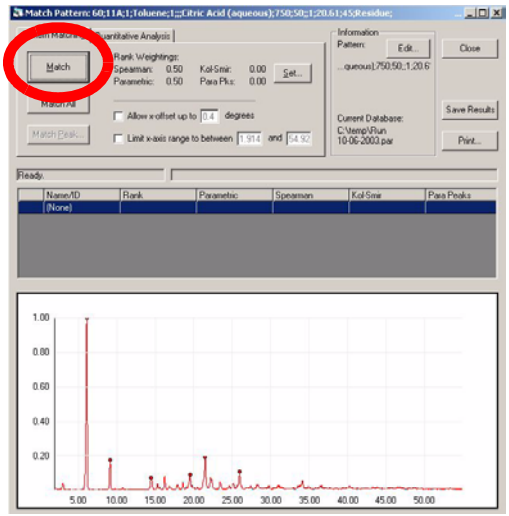
```
Summary of the estimates of cluster numbers:


From principal components analysis (non transformed matrix): 5
From principal components analysis (transformed matrix):     4
From multidimensional metric scaling:                        4
From the gamma statistic using single linkage:               5
From the Calinski-Harabasz statistic using single linkage:   5
From the C-statistic using single linkage:                   5
From the gamma statistic using group averages:               5
From the Calinski-Harabasz statistic using group averages:   5
From the C-statistic using group averages:                   5
From the gamma statistic using the Ward method:              5
From the Calinski-Harabasz statistic using the Ward method:  5
From the C-statistic using the Ward method:                  5
From the gamma statistic using complete linkage:             5
From the Calinski-Harabasz statistic using complete linkage: 5
From the C-statistic using complete linkage:                 5


Maximum estimate is 5
Minimum estimate is 4
Combined weighted estimate of the number of clusters is 5
The median value is 5
```

This suggests that 5 is indeed the correct number of clusters.

Go to the Dendrogram view, and select patterns 8 and another of the yellow patterns. Consider their overlaid profiles - the similarity between them suggests that the program was correct in its initial placement of the cut line, and the reason 8 is separated somewhat from the rest appears to be down to an x-shift.

7.5.2 Manual Matching

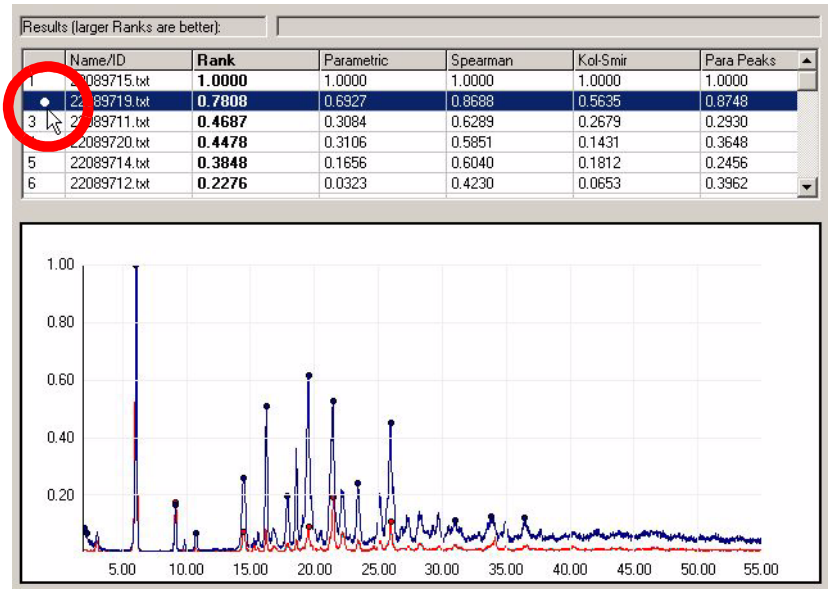
Select pattern 7. To get a better idea of what is going on with particular patterns, sometimes it is useful to compare them manually to the rest of the patterns. From the *Pattern* menu, select *Manual Match/Quantitative*, and from the resulting sub-menu, select *against Sample database*. A new window will appear.



This window, the manual match window, has the selected pattern shown in the bottom half. The upper half has the controls for the manual matching. Click the *Match* button in the top left, and the selected pattern is compared to each of the other patterns in the database in turn.

The results are then displayed in a table in the centre part of the screen. The most similar pattern to the selected pattern is that pattern itself, which appears at the top of the list, with perfect correlation scores of 1.0 in each of the 4 tests displayed. The results are sorted in descending order of Rank value, which a user-controllable combination of some of the individual matching tests. By default, it is the average of the Spearman and Parametric tests.

Click on the number **2** in the first column of the second row. The corresponding profile for that pattern is shown overlaid on the original to allow visual comparison.

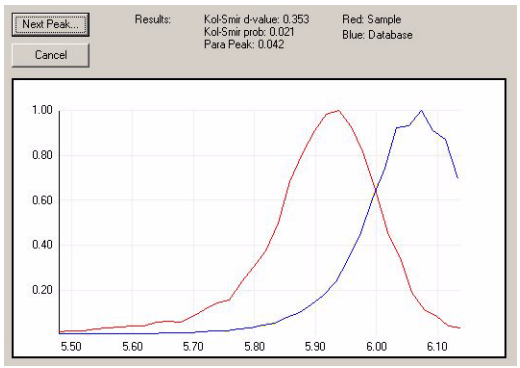


This pattern is the most similar to the selected profile, but has quite a few obvious differences. This is reflected in the relatively low scores that the pattern gets in each of the matching tests. This pattern corresponds to pattern 11 on the dendrogram earlier, which that method also identified as being most similar. It seems there are preferred orientation problems here.

Click again at the same place in the first column to make the overlaid pattern disappear. Now click on the number **3** to see what the next most similar pattern is. This corresponds to pattern 3 on the dendrogram display, part of the green cluster.

Notice that the numerical results are much lower than for the previous pattern, which supports the choice of the program to place these patterns in separate clusters.

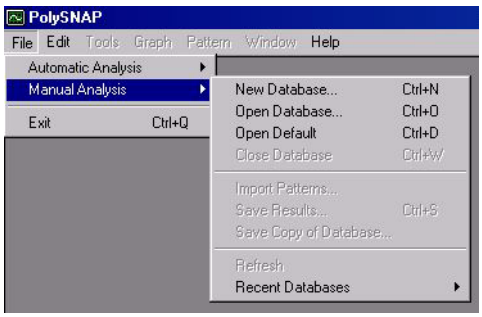
To examine the differences between them on an individual peak-by-peak basis, click the *Match Peaks* button.



A small window appears showing the first two coinciding peaks, and individual matching tests results between them. Clicking the *Next Peak* button repeatedly moves through the marked peaks on the selected patterns. Click *Cancel* to dismiss the window.

7.6 Manual Matching Options

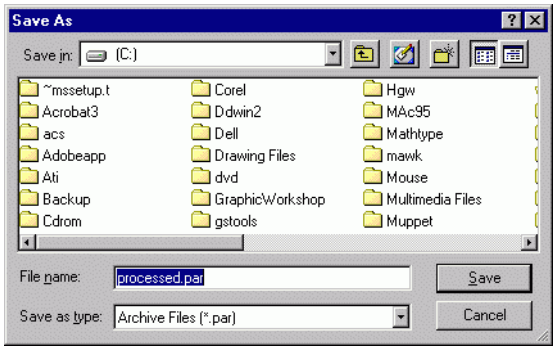
The rest of the examples in this tutorial demonstrate the manual analysis section of PolySNAP. This mode can be accessed from the *Manual Analysis* section of the *File* menu.



The first thing to do is to create a new database of patterns with which to work.

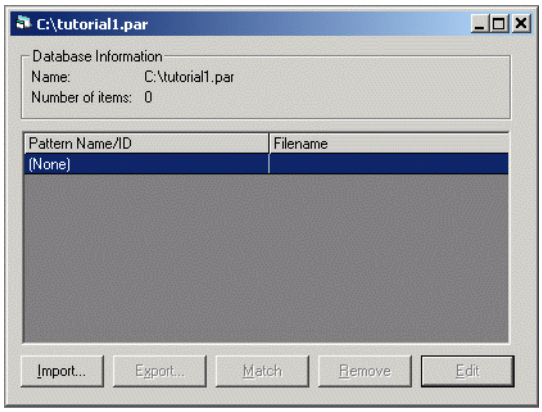
We first need to create a new database, which will initially be empty We will then load some powder patterns into it, and work with them.

Choose *New Database* from the *Manual Analysis* section of the *File* menu. A standard Windows file dialog box appears, and you are invited to choose a name and location for the new database.



Enter *tutorial1* as the database name, select a location (*C:/* for example), and click *Save*. The dialog box disappears, and a new window opens inside the PolySNAP workspace.

This is our new empty database; its filename is displayed in the window title bar. Like any other window, it can be moved around the screen by clicking and dragging on the title bar, and can be re-sized by dragging at the edges of the window.



The filename and your selected location appear in the Database Information section of this new window.

The number of patterns in this database is also displayed - as expected, it is currently zero.

Several buttons are arranged along the bottom of this window. Most of these are unavailable as there are no data files in the database. The only one that is available for use is the *Import* button.

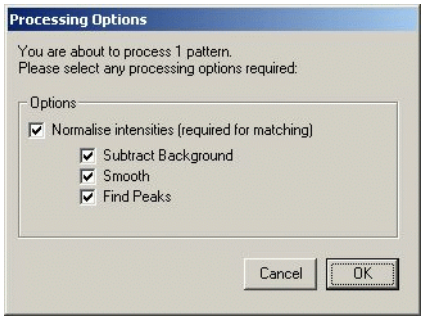
We now need to bring some pattern data files into the database.

Click the *Import* button. We need to navigate to where the data is stored. Locate the folder

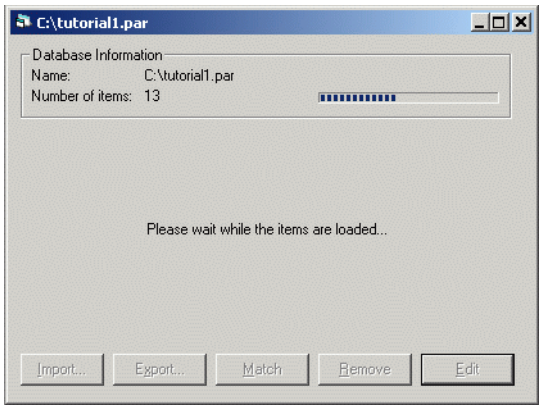
C:\Program Files\PolySNAP2\tutorial\manual

Once in the correct location, a list of different files in the folder that the program recognises is listed.

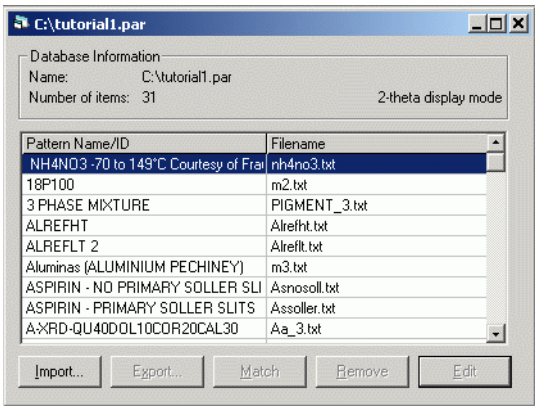
We want to open all of the files in this particular folder, so hit *control-A* on the keyboard to select them all, and then click *Open*. A new window appears with the pattern processing options.



These will be discussed in more detail later, so leave them with their default values for the moment and click OK. A progress bar appears at the top of the database window to allow progress to be monitored. As there are only a few patterns in the folder, this process should only take a couple of seconds.



Once importing has finished, the patterns will be listed in the main part of the database window:



If the window is too small to see all of the patterns, the scroll bar on the right hand side can be used to view the rest, or you can drag the window border to enlarge the window.

Each entry in the list represents one pattern file that has been imported.

Initially, the patterns are listed in the order they were imported. The list can be sorted by either *Pattern Name/ID*, or *Filename*, by clicking on the headers at the top of the list.

If some of the chemical names here look very similar to filenames this is because the ASCII files we have imported do not contain any chemical name information, so the filename is used instead. The actual chemical name, if known, can be added manually later.

You can select a particular pattern in the list by clicking once on it. When you do so, the other buttons along the bottom of the window become activated.

The *Import* button has already been used. The *Export* button lets you make a copy of a particular database entry to a separate file, in either ASCII text or PolySNAP pattern format.

Remove deletes the selected pattern or patterns from the database.

Experiment with these options by exporting a pattern to a separate ASCII text file under the name *exporttest.txt*. Make sure to note where you saved it! Then try importing the file you have just saved. Finally, select the *exporttest.txt* pattern in the list and delete it from the database using the *Remove* option.

Removing patterns from the database does not in any way affect the original data files which are left intact and unchanged throughout all

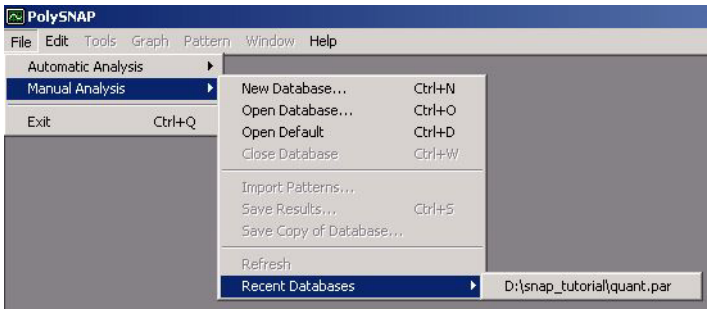
operations performed using PolySNAP - the program works on a copy of the data only.

The *Match* and *Edit* options will be dealt with in following sections, so leave them be for the moment.

Now quit the program by choosing *Exit* from the *File* menu. All windows will close. Note that any changes made to a database are automatically saved, so there is no need to manually save any changes before closing.

Relaunch the program again in the same manner as before. We want to return to the database we have just been working with. You could select *Open Database* from the *Manual Analysis* section of the *File* menu, navigate to where the database was saved, select it and click *Open*, but there is a much easier way.

The program keeps a record of the last four databases used in the *Recent Databases* submenu of the *Manual Analysis* section of the *File* menu. Open this now; one of the entries should be the

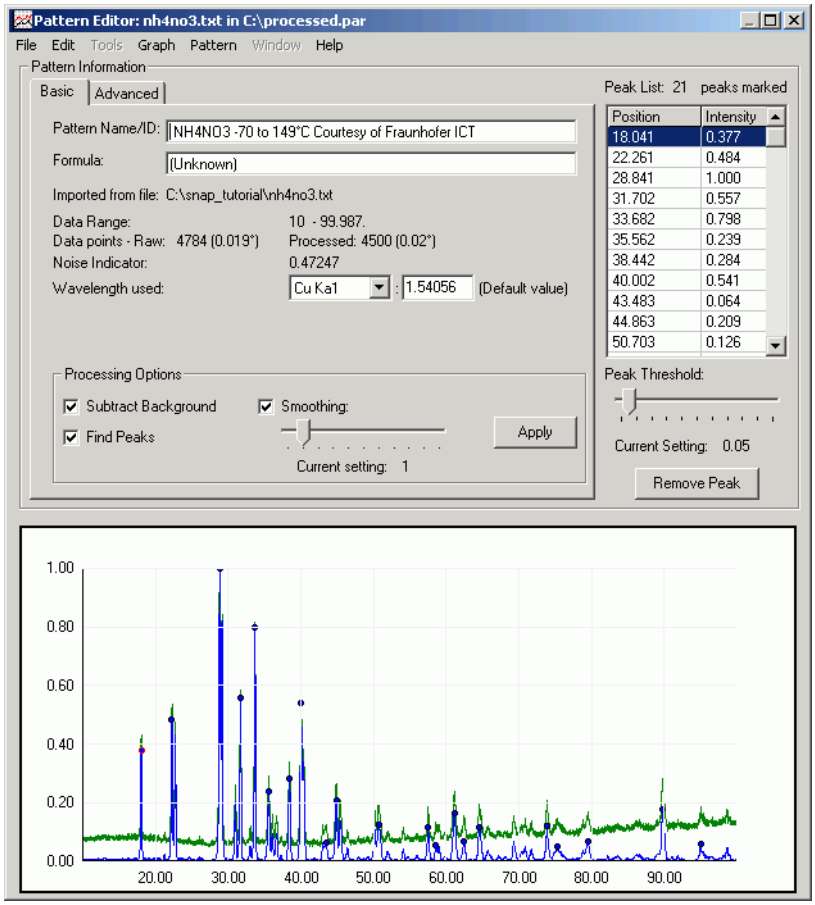


tutorial1.par database we were using earlier. Select it from the menu, and it should start to open automatically.

Next, we want to examine some of the patterns we have just imported in more detail.

7.6.1 Editing Patterns

Select the pattern *nh4no3.txt* from the list in the database by clicking once on it. Click the *Edit* button, and the Pattern Editor window appears.

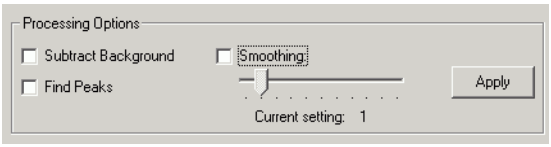


Several useful pieces of information about the pattern you have selected are displayed in this window, and a plot of the pattern itself is shown at the bottom.

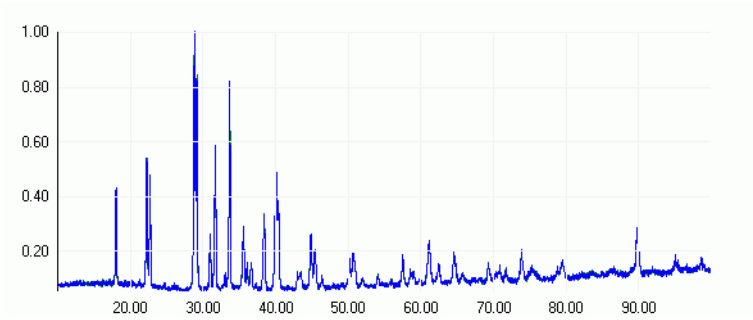
The chemical name or pattern ID can be changed here by editing the current name in the white text box.

Details such as the pattern filename, its start and end angles, and the number of data points are also displayed.

There are three check boxes labelled Processing Options just above the region where the pattern is displayed. To begin with, click once in each of the checkboxes, so that they are all turned off:



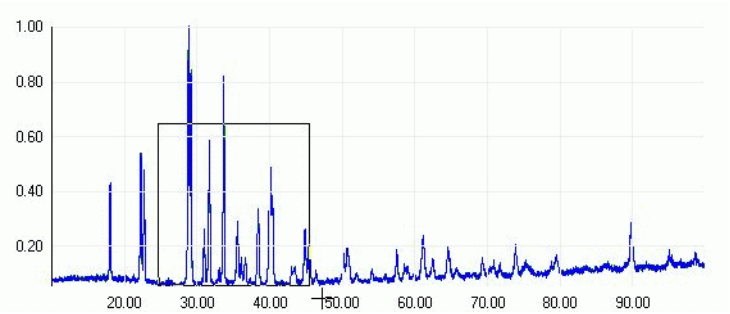
Then click *Apply.* The graphing region should now look like this:



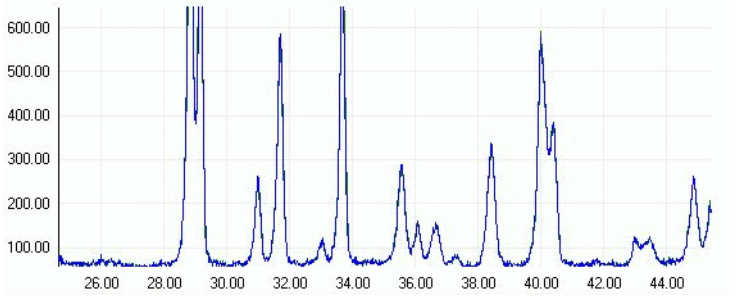
As you can see, the data has been scaled along the y-axis to run from 0.0 to 1.0. This operation is performed for all data being imported into PolySNAP in order to allow for suitable scaling between different data sets.

The raw data has also been interpolated from its original resolution, to the PolySNAP standard 0.02 degrees. Again, this is for consistency between patterns.

You can zoom into the pattern display, by clicking and dragging a rectangle on the region you wish to see more closely:

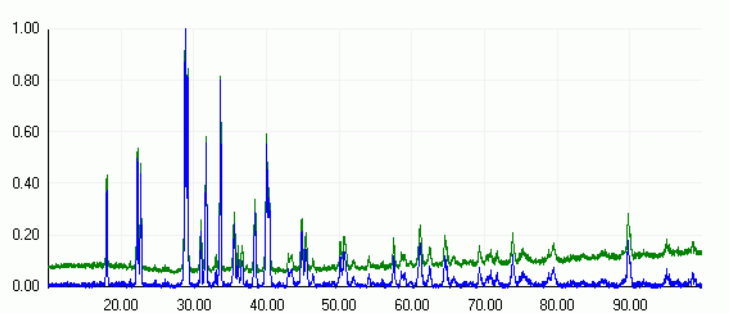


The view will then change to just the selected region:



To return to original view, right-click anywhere on the pattern display, and select *Reset View* from the resulting pop-up menu. Multiple zooms are possible by repeating the process.

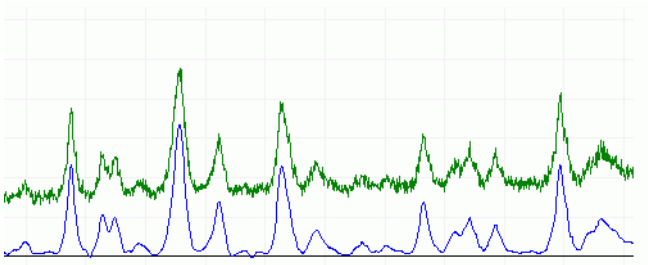
Now click the *Subtract Background* checkbox, and then click *Apply.* The pattern display should change:



The green line represents the raw pattern data. The blue line shows the same pattern after the background level has been subtracted. To see what the program considered to be background before subtractions, select *Show Background Curve* from the *Pattern* menu. A new window will open showing the subtracted background curve where the blue line now represents what is subtracted as background. Once you are finished looking, click *Close* to close this window and return to the standard view.

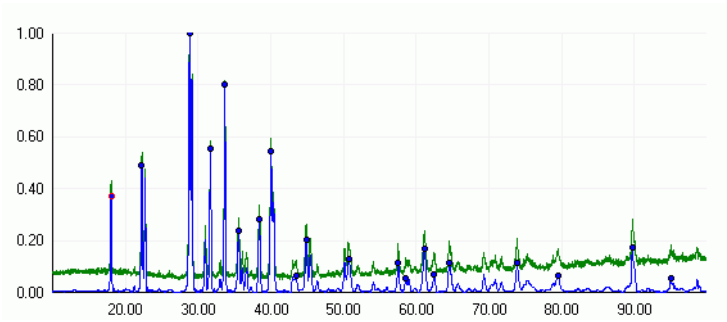
Now click the *Smoothing* checkbox, so both it and the *Background Subtraction* boxes are checked, and click *Apply* again.

The graph pane will be updated, and any noise in the pattern will be smoothed out (shown here zoomed in).



Finally, check the *Find Peaks* option, and click *Apply.*

The pattern display will update, and several small blue circles should appear on the top of the larger peaks:



These mark the location of what the program considers peak regions. There is a minimum peak height below which any peaks are ignored. This is set as a default to 0.05, which is why the smallest peaks are unmarked.

A list of the peaks the program has found is shown in the upper right corner of the pattern editor:

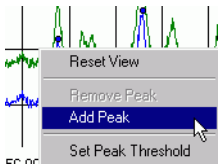
Peak List: 20 peaks marked

Angle	Intensity
18.041	0.372
22.261	0.479
28.841	1.000
31.702	0.554
33.682	0.794
35.562	0.237
38.422	0.284
40.022	0.536
43.463	0.064
44.863	0.205
50.703	0.123

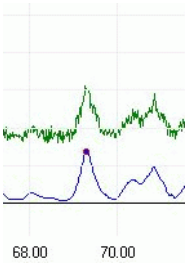
This lists the peaks found in order of increasing angle, and displays the corresponding intensity for each.

If the program has missed a peak you believe should be included, it is easy to add manually. Note that peak locations are added to the blue, processed pattern line, not the green, raw data line on the graph profile.

For example, if we wish to add the small peak that is located at around 54°, just right-click once at the point on the graph where you judge the peak maximum to be:

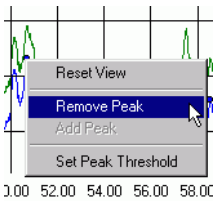


Then select *Add Peak* from the resulting pop-up menu. A round blue peak marker should appear at the top of the peak, and an entry for it should be added to the peak list:

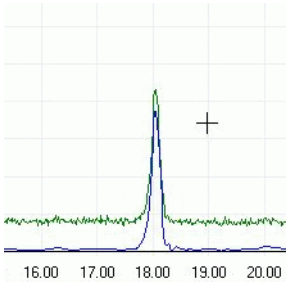


If this does not appear, try clicking as close as possible to the top of the peak. It may help to zoom in to the area of interest.

It is also possible to remove peaks you believe to be incorrect. For example, say we wish to delete the marker from the peak around 18°:

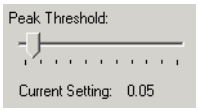


Right click as close to the peak marker as possible, and select *Remove Peak* from the pop-up menu that should appear. The peak marker, and its corresponding entry in the peak list, should vanish.



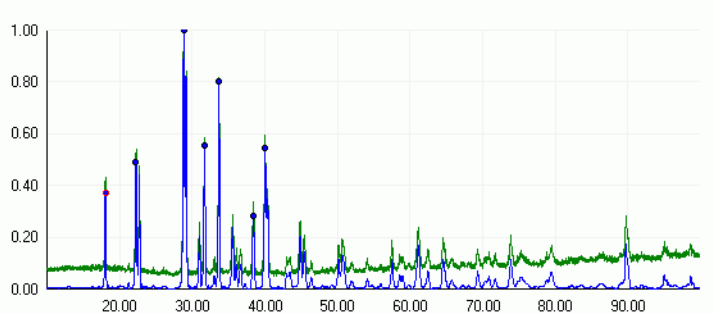
If the *Remove Peak* option is unavailable from the pop-up menu, you have not clicked close enough to the marker. Zoom in to make this easier, and try again.

Finally, it is possible to adjust the minimum peak height threshold to include less of the smaller peaks. Look at the slider just below the peak list - it should be set to 0.05.



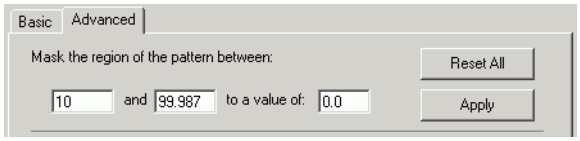
Drag the handle on the slider until it is set to about 0.25.

Several additional peaks that were marked before should now show up without peak markers, as they are now below the new minimum threshold of intensity.



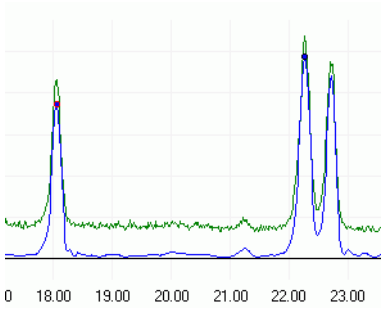
Finally, return the slider to 0.05. The peaks that previously became unmarked should show up again.

Now click on the *Advanced* tab in the top-left of the window.

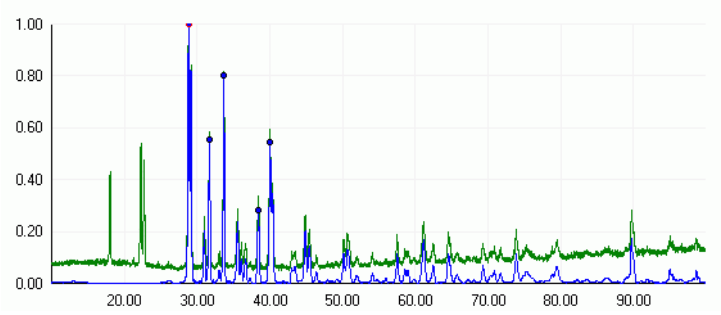


The top section revealed allows us to mask selected regions of the pattern to be ignored during matching or analysis processes. This may be useful where a particular peak is negatively affecting results. Say we wish to mask the peaks in the region of the pattern between 15° and 25°. Enter the start and end angles of the region to be masked in the relevant text boxes.

Examine the blue processed pattern in this region to determine its average background level – in this case, a value of zero should suffice, so enter 0.0 in the level text box, and click *Apply.*



The change in the pattern should be quite noticeable – while the green line of the raw data is untouched, the blue processed data line no longer has any peaks or features between the ranges we have entered:



Now repeat the process and mask the peak at approximately 90° using the same method.

Once finished, return the pattern to its initial state by clicking *Reset All*.

The rest of the *Advanced* tab options will be discussed in the Analysing Mixtures section later in this tutorial.

Choose *Close Window* from the *File* menu to dismiss the editor window. The program will check if you wish to retain any of the changes made in the editor. At present you do not want to keep any changes, so click *No*. You are returned to the main database window.

Feel free to look at some of the other patterns in the database in the editor window, and investigate the various processing options.

7.6.2 Matching Patterns

Assume that you have just obtained a new, unknown, powder pattern.

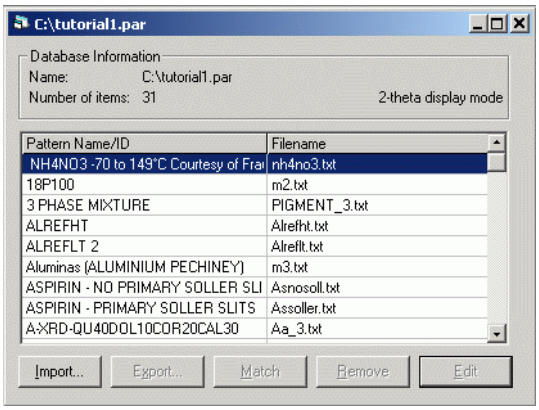
You want to find out what pattern in your existing database it is most similar to - and hence possibly identify what substance it may represent.

First, close any currently open database or editor windows.

Now, go to the *Manual Analysis* section of the *File* menu, and select *Open...*

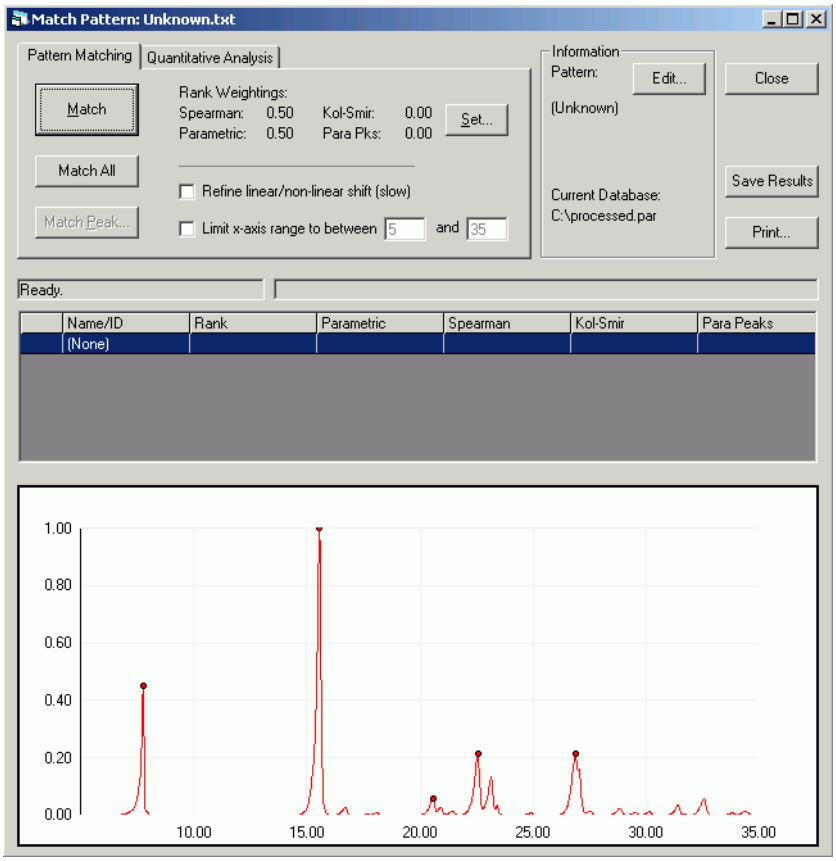
Open the pattern database you created earlier, *c:\tutorial1.par.*

A new database window will open, and the pattern data will be loaded into it.



The first step would normally be to import the unknown pattern into this existing database in order to examine it. In our case however, it is already loaded in the database.

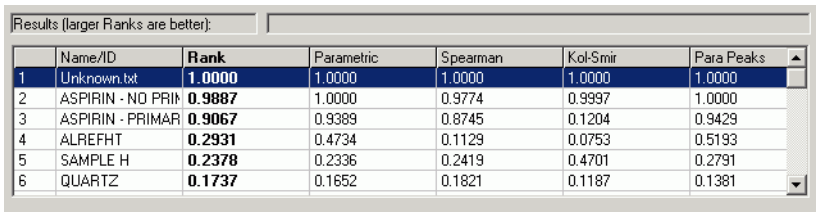
Locate the pattern *Unknown.txt* in the scrolling list of patterns in the database, click once to select it, and then click on the *Match* button. The main program match window will appear.



Our unknown pattern is displayed in the graph pane in the bottom half of the window. Note that the pattern name and the database we are using are listed at the top right. On the top left of the window are the matching and analysis controls contained within two tabs – *Pattern Matching* and *Quantitative Analysis*. The default tab is *Pattern Matching*, which is what we wish to do first.

Click on the *Match* button. A progress bar appears as the program runs through various tests in order to compare our selected pattern to every other pattern in the database. Once matching is complete, the

centre section of the window fills with the numerical results, sorted by the column in bold type, Rank.



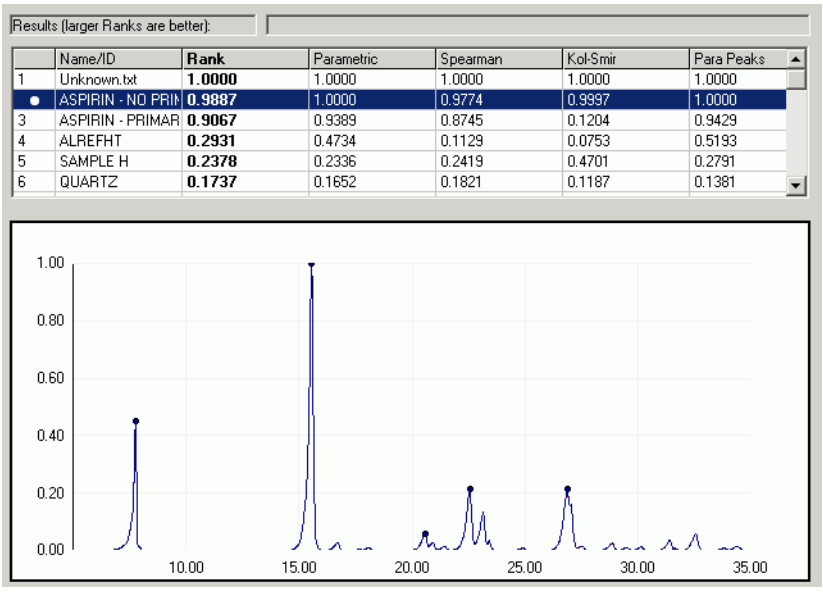
	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	Unknown.txt	**1.0000**	1.0000	1.0000	1.0000	1.0000
2	ASPIRIN - NO PRI	**0.9887**	1.0000	0.9774	0.9997	1.0000
3	ASPIRIN - PRIMAR	**0.9067**	0.9389	0.8745	0.1204	0.9429
4	ALREFHT	**0.2931**	0.4734	0.1129	0.0753	0.5193
5	SAMPLE H	**0.2378**	0.2336	0.2419	0.4701	0.2791
6	QUARTZ	**0.1737**	0.1652	0.1821	0.1187	0.1381

To sort the results by a different column, click once on that column's header. To change the sort order, *e.g.* to sort a column ascending instead of descending, or *vice versa*, click the header again. For the moment, click on the Rank header to re-sort by this column, and ensure the largest value (normally 1.0) is at the top of the list.

Look at the pattern associated with this value. This is what the program considers the 'best match' to our unknown. If it looks familiar, that is because it is – it is the *Unknown.txt* pattern itself. Because it is in the database, it is compared to itself. If this does not result in perfect match scores, there is a problem somewhere, so this is a useful check.

More interesting is the next entry down in the list. The next best match has a rank value of above 0.9, quite close to the perfect score of 1.0. The individual test scores reading along the row reflect this: all close to 1.0. The rank value is calculated by default by summing the Spearman and Parametric test scores and dividing by 2.0.

To see these scores reflected visually, we can overlay the best-match pattern with the unknown pattern. Click once in the left-most column of the pattern you wish to overlay (the one with '2' in it in this case).



	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	Unknown.txt	**1.0000**	1.0000	1.0000	1.0000	1.0000
●	ASPIRIN - NO PRI	**0.9887**	1.0000	0.9774	0.9997	1.0000
3	ASPIRIN - PRIMAR	**0.9067**	0.9389	0.8745	0.1204	0.9429
4	ALREFHT	**0.2931**	0.4734	0.1129	0.0753	0.5193
5	SAMPLE H	**0.2378**	0.2336	0.2419	0.4701	0.2791
6	QUARTZ	**0.1737**	0.1652	0.1821	0.1187	0.1381

The '2' is replaced by a coloured dot, and a pattern in this colour

appears in the graph pane. It is obvious the two patterns are very similar. Click on the coloured dot in the left-hand column once again to remove the extra graph.

This suggests our unknown sample is aspirin.

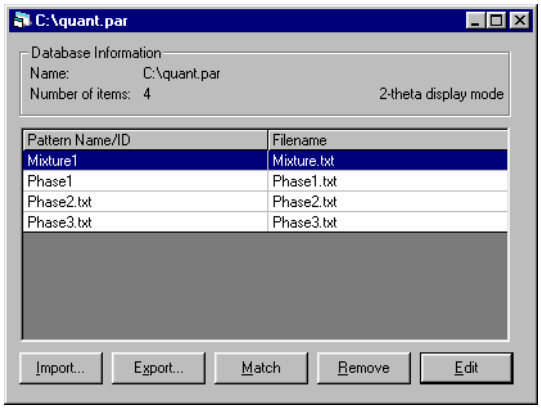
Experiment with overlaying different graphs on the unknown; this helps to get a feel for how the pattern scores correlate to similarity between patterns. Several patterns can all be overlaid at once on the graph; to clear them all in one go select *Clear Overlaid Graphs* from the *Graph* menu.

7.6.3 Analysing Mixtures Manually

Using PolySNAP, open the database file *quant.par* in the tutorial data folder. (Normally found in

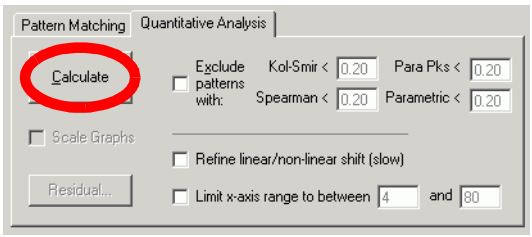
C:\Program Files\PolySNAP2\tutorial\manual)

It contains four patterns – one mixture, and the three pure component phases.

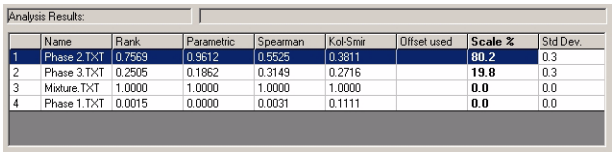


We will use the program to quantify the amounts of each phase in the mixture. Select the mixture from the list of patterns, and click *Match* to bring up the match window.

Go to the *Quantitative Analysis* tab, and click *Calculate*. After a few seconds when a progress bar is displayed, the program should display its answers.

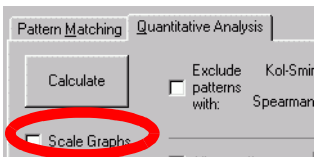


These are in a similar form as before, except that there are two additional columns – Scale % and Std Dev – the amount of each phase in the mixture, and the error on the calculation.

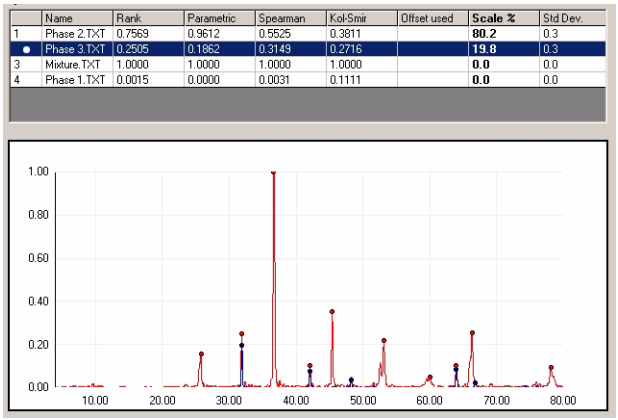


	Name	Rank	Parametric	Spearman	Kol-Smir	Offset used	Scale %	Std Dev.
1	Phase 2.TXT	0.7569	0.9612	0.5525	0.3811		**80.2**	0.3
2	Phase 3.TXT	0.2505	0.1862	0.3149	0.2716		**19.8**	0.3
3	Mixture.TXT	1.0000	1.0000	1.0000	1.0000		**0.0**	0.0
4	Phase 1.TXT	0.0015	0.0000	0.0031	0.1111		**0.0**	0.0

There are various ways available to see if the programs suggested percentages are sensible. First, check the *Scale Graphs* box on the upper left.



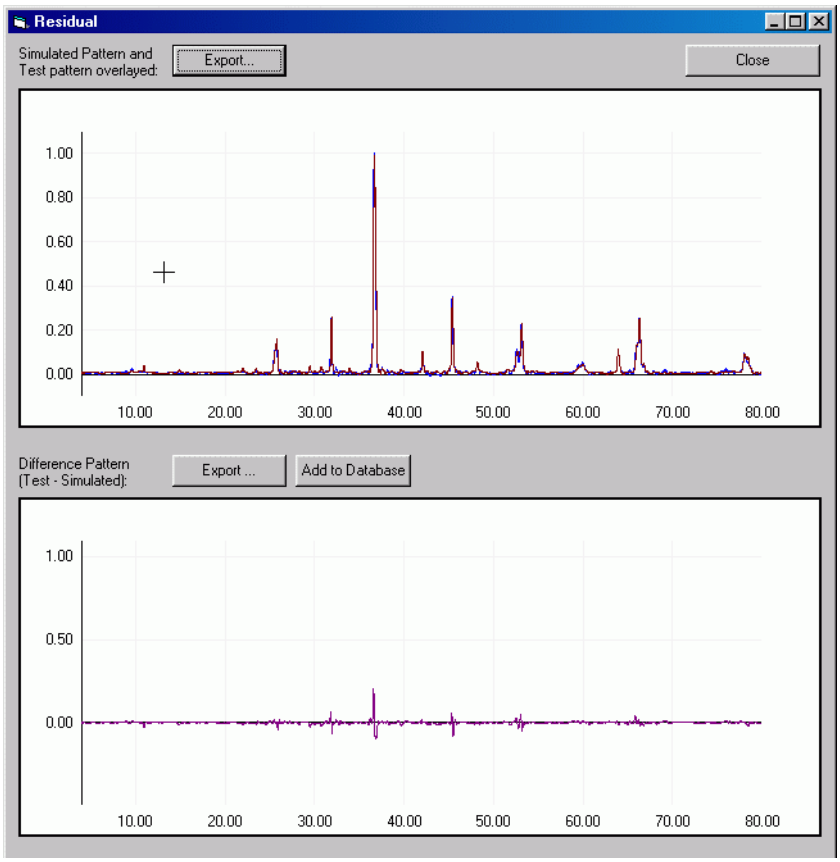
Then click on the left-most column of the results table for one of the phases.



	Name	Rank	Parametric	Spearman	Kol-Smir	Offset used	Scale %	Std Dev.
1	Phase 2.TXT	0.7569	0.9612	0.5525	0.3811		**80.2**	0.3
●	Phase 3.TXT	0.2505	0.1862	0.3149	0.2716		**19.8**	0.3
3	Mixture.TXT	1.0000	1.0000	1.0000	1.0000		**0.0**	0.0
4	Phase 1.TXT	0.0015	0.0000	0.0031	0.1111		**0.0**	0.0

As before, this overlays the selected pattern over the mixture, but is now scaled to the percentage intensity suggested by the programs analysis. The pattern should hopefully look sensible.

Another option is to click the *Residual* button, which brings up a window with two graph panes.

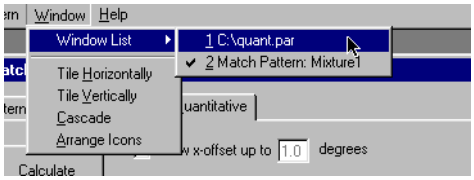


The upper one superimposes a simulated mixture pattern, made up of combining the pure phases in the amounts suggested by the program, on top of the original mixture. The bottom panel is then the difference between the two. This difference plot helps show up either missing phases or extra intensity which is not part of the mixture. If required, it can be output to a file to be imported as a new pattern at a later date. In this case, the small amount of residual intensity suggests the calculated answer is quite good.

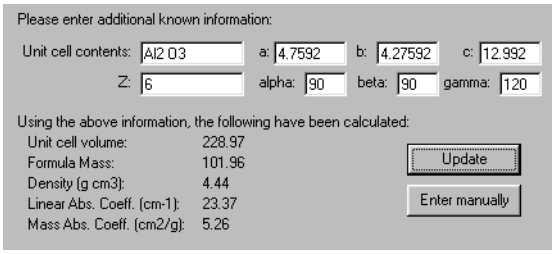
Click *Close* to dismiss the window.

So far, we have quantified the mixture in terms of how much of each of the pure phase *patterns* are required to make it up. A more useful number in the real world would be in terms of the weight of each phase used in the mixture.

In order to calculate this, we need to add some additional information to each of the pure phase patterns. To do this, first select the database *quant.par* from the list of open windows in the *Window* menu.



This brings it to the front. Now select the first pure phase pattern: Phase 1. Click *Edit* to bring up the pattern editor window, and click on the *Advanced* tab.



We need to enter information into the lower region of this window, in particular the chemical formula of the phase, its unit cell dimensions, and the number of formula units per unit cell.

Phase	**Formula**	**a**	**b**	**c**	**α**	**β**	**γ**	**Z**
Phase 1	Al2 O3	4.7592	4.7592	12.992	90	90	120	6
Phase 2	Ca F2	5.4649	5.4649	5.4649	90	90	90	4
Phase 3	Zn O	3.2501	3.2501	5.2071	90	90	120	2

Using the information in the table, enter the relevant details for this phase, and click *Update*. The fields showing molecular weight and absorption coefficients and so on for this phase should update. If not, be sure to enter the formula in the form

<Atomic Symbol><No. of atoms> <AtomicSymbol><No. of atoms> etc.

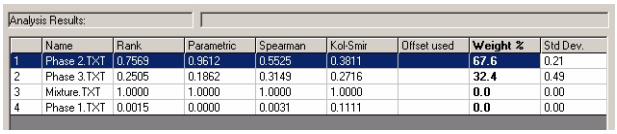
(The allowed formats for formula entry are discussed in more detail in section 3.3.11, Additional Pattern Information in the main program manual).

Repeat for the other two phases.

Click on the mixture pattern database entry to select it, before clicking the *Match* button to bring it up in the main match window.

Return to Match window, and in the *Quantitative Analysis* tab, click *Calculate* again.

This time, the results table should be slightly different – the results should be headed Weight % rather than Scale %. If not, go back and check that all of the three pure phases has had the extra information added successfully.


Analysis Results:

	Name	Rank	Parametric	Spearman	Kol-Smir	Offset used	Weight %	Std Dev.
1	Phase 2.TXT	0.7569	0.9612	0.5525	0.3811		**67.6**	0.21
2	Phase 3.TXT	0.2505	0.1862	0.3149	0.2716		**32.4**	0.49
3	Mixture.TXT	1.0000	1.0000	1.0000	1.0000		**0.0**	0.00
4	Phase 1.TXT	0.0015	0.0000	0.0031	0.1111		**0.0**	0.00

7.7 Conclusion

This completes the basic tutorial. There are many other features and options in the program that should allow much more complex problems to be examined. Each of these are described in the full program manual, which should be consulted for more information.

CHAPTER 8

PolySNAP Release Notes

Please report any bugs or problems encountered that are not already listed in the ‘Known Problems’ sections below via email to *snap@chem.gla.ac.uk*

Version 2.1

A new pre-processing feature has been added for situations when a large set of reference patterns is available, and needs to be narrowed down to find a smaller subset of the most similar to a particular sample pattern. This allows for databases of far more than 2,000 patterns to be scanned quickly and efficiently. This new mode is accessed via the *Pre-screen data...* option in the Automatic Mode main menu option, or via the Welcome window button.

A new display feature has been added allowing an easy visual comparison of the marked peak positions between selected patterns. This is accessed via the *View Peak Comparison...* option in the Tools menu. The colours and angle-step size can be altered via the *Display & Advanced* pane of the Options window.

Infra-Red (IR) and Near-Infra Red (NIR) data in Bruker Opus format can now be read in by PolySNAP.

A bug where a 'Type Mismatch' error dialog was displayed when the first pattern read in for a given dataset had problems has been fixed.

Version 2.0.1

A bug where DSC data was not being correctly interpolated to a constant step size was resolved.

A problem that could cause errors to be displayed if the last imported pattern did not have any marked peaks was fixed.

Version 2.0

Initial release of Version 2 incorporating multiple dataset combinatorial functionality and a re-designed Options screen.

Version 1.7.2 - Build 3-4-07

Data files load much more quickly using the default settings than in previous versions.

Updated copyright information in the Automatic Report Writer.

Fixed a problem where 'Rerun analysis…' options were not working in Default mode.

A bug where paths entered at the command line would not work correctly if the were enclosed in quotes has been rectified.

Fixed an issue in countries using the comma as a decimal point where shifted patterns were not being displayed correctly in the Numerical Results pane.

The 'Find' option for a selected Silhouette or Fuzzy Cluster is now F4 (previously F5).

A 6D-plot issue where the first two 'shapes' were shown as the same shape has been fixed.

Selecting patterns from the Key in the Cell Display, and then choosing 'Rerun Selected Samples' now works as expected.

Entering information in the Advanced tab of the Pattern Editor is now much easier, as tabbing between fields now works as expected.

The main Run On... dialog box has been slightly re-arranged.

A new program splash screen and window background are displayed.

The program version number is now displayed in program title bar.

The program, documentation and website now use the new Bruker logo.

The main clustering analysis code is now more efficient and uses a smaller memory footprint.

Version 1.7.1 - Build 1-3-07

Feature Additions

When x-shifts have been refined, clicking on an entry in the numerical results pane of the Results window now shows the relevant two patterns superimposed with the calculated optimal shift value.

A new feature highlights areas of a dataset where there is little or no signal or marked peaks, and whose removal may improve the results of the analysis (this notification can be deactivated by deselecting the '*Warn when areas of low signal are detected…*' checkbox is the Automatic: Advanced section of the program Options dialog box).

A new option has been added to force quantitative analysis to be attempted on all input samples regardless of how similar they are to the reference phases. This can be selected via the '*Perform analysis on all samples regardless of similarity*' checkbox in the Options: Automatic: Advanced / Quantitative options dialog box.

Bug Fixes

Fixed a bug where a dialog box would repeatedly appear and could not be dismissed when using the automatic report writer.

Quantitative results were not previously correctly showing up in View Results mode; now they are.

Version 1.7.0 - Build 14-2-06

A new program operating mode has been added - *Quality Control* mode. This allows the user to investigate if new samples are within a user-specified cut-off of similarity to known reference materials - *e.g.* for taking several samples over time to monitor production. Results are shown on a 3D plot, with the references shown as a 3D surface; if new samples are within this surface, they pass; if they fall outside the surface, they fail, and the user is warned.

A new option in the *Run On...* dialog box allows reference phases to be included in the main calculation - rather than keeping them separate as before, they will also now show up on the dendrogram, 3D plots *etc*.

Two new tools have been added to the *Validation* pane - *Parallel Coordinate Plots* and *Space Exploration*. These allow the user to check that the suggested clustering remains valid in higher-dimensional space. As a result of these additions, the *Tools* menu

option to *Export Results for ClusterVision* has been removed as it is no longer required.

A new option in the *Options* dialog to Use Sample Preparation information from an external file rather than the data file headers has been added. When it is turned on, PolySNAP will look for a *<foldername>.dat* file in the current data input folder; if that doesn't exist, PolySNAP checks as before in the sample_info_format file for a different file location; if that fails it then looks in the file headers.

If a database containing reference phase patterns with their unit cell contents known is provided, PolySNAP can now calculate the weight % for quantitative results as well as the scale %.

A new option allows the user to decide if the calculated simulated mixture patterns used for checking quantitative results output are constructed from the original raw profiles, or with processed (*e.g.* background-subtracted) profiles. The default is to use the original raw profiles.

A new *Mask Regions* dialog in Manual Mode (accessed through the *Tools* menu) allows easier selection of multiple pattern regions to be ignored during matching.

A new option has been added to the Matching Options dialog box, allowing the user to set how much the pattern is smoothed as part of the amorphous pattern identification calculation; the default is level 5; with very noisy data this should be reduced to 1 or 2 to prevent he occurrence of false positives.

To aid the clarity of interpretation, patterns on the dendrograms are no longer drawn with a white border round them for very large data sets.

ASCII data where the angle range runs from high to low can now be correctly imported.

It is now possible to use the control-key to select multiple items on the cell display key.

As a result of the use of a new, more flexible copy-protection system, the License menu item has been removed from the Help menu as it is no longer needed.

A copy of all of the PolySNAP tutorial data is now automatically installed along with the program itself in a subdirectory of the main program folder; normally *C:\Program Files\PolySNAP\tutorial*

Bug Fixes

A bug where the incorrect mu value for sulphur was being used in absorption coefficients calculations has been corrected.

Combining peak-based and profile-based tests when also allowing an x-shift no longer causes an incorrect correlation matrix to be generated.

An issue where the residual pattern was not calculated correctly after quantitative analysis when there were zeroes in the error column has been fixed.

A problem introduced in PolySNAP v1.6.4 where it was no longer possible to copy text from the numerical results pane has been resolved.

A problem where reported errors on quantitative results were off by one when patterns in the results list had been marked as ignored is now fixed.

A bug in the absorption coefficient calculation that could cause a crash under certain circumstances has been resolved.

It is now possible to correctly manually match a sample pattern against a reference database using the option in the auto-mode *Display* window.

Version 1.6.4 - Build 29-9-05

The PolySNAP manual and other documentation has been comprehensively reorganised.

A new PolySNAP logo and program splash screen has been introduced.

The user can now set the location of the *sample_info_format.txt* file (used to specify parameters for 6D plots) in the Advanced pane of the program *Options* window.

A bug that prevented more than 500 entries being displayed in the Numerical Results display has been fixed.

A bug that prevented dendrogram colours being overlaid on 6D plots has been resolved.

An issue where RAW v4 format files did not have their 'Comment' and 'Chemical Name' fields read correctly was fixed.

Calculations of mu-star for quantitative analysis are now correct in cases where different phases have different values of Z.

PolySNAP now supports basic read support of Bruker RAW files in the new 21 CFR Part 11 compliant format.

Version 1.6.3 - Build 4-7-05

The basic menu structure has been simplified to clarify program operation; the 'PolySNAP' sub menu is now named Automatic Analysis, and all manual options are in the new 'Manual Analysis' submenu.

The tabs in the program Options dialog have also been renamed to fit this new pattern.

The menu item 'Run Automatically' has been renamed 'Run on Defaults'.

The 'Run On' dialog box now includes details of which shift method is currently selected (sin theta, sin 2theta or cos theta), and also if the 'Strict' background correction option is turned on or not.

In unlocked demo copies, a new dialog box reminds users when the demo will expire and how to register their copy.

More detailed information about the checking for amorphous samples is now included in the logfile.

If no known phases were provided, the menu item to match against them is now correctly disabled.

The user is now correctly informed if the fuzzy clustering did not find any problems with a data set.

The requirement that a user have Administrative privileges to change program settings is now off by default.

An issue where the Strict background subtraction could be turned on without first selecting the 'normal' subtraction has been fixed.

A bug where re-running silhouettes caused the graph pane to disappear has been resolved.

Location of known phases are now reported correctly when they have come from a database file, not a folder.

A bug in manual quantitative analysis where the presence of an ignored pattern was preventing weight fraction calculation has been resolved.

The Average Link method has been removed from the Clustering options dialog box as it is the same as Group Average Link.

The program now checks if there are any other copies of the program already running & exits if so.

A bug where selecting patterns to overlay from the numerical results pane could cause a crash if shifts had been calculated has been resolved.

A new graphics toolbar option has been added allowing line sizes to be changed on the graphical plots.

Any changes made to the graphics colours are now reset when the program is closed and opened again.

Version 1.6.2 - Build 27-5-05

A bug where the program would refuse a valid registration code if it was entered after the initial time-limited demonstration period had expired has been corrected.

Version 1.6.1 - Build 27-4-05

The splash screen has been updated to replace the University of Glasgow logo with the new WestCHEM logo. Additionally, a list of relevant references for the software has been added to the About screen, which can be copied to the clipboard by right-clicking on them.

The calculated % crystallinity is now output to the logfile for all patterns, not just those considered to be amorphous.

Version 1.6.0 - Build 4-3-05

Feature Additions

The *Display* menu now features the option *Show Current Display in New Window*, which opens a copy of the current main graphics display (*e.g.* dendrogram or 3D plot) in a separate new window. This new display window can be resized, zoomed into *etc*., but changes made here are not reflected in the main display; it is only a snapshot of the current view which can be used to view several different displays at once for comparison purposes. This menu item can be used multiple times to open as many new display windows as are required.

A new option in the initial *Run On...* settings dialog allows for the selection of a restricted subset of the full angle range. With this option selected, only this range of the patterns will be used for the matching comparison.

A new sub-menu *Tools -> Re-run Analysis...* gives access to the existing *using Most Representative Patterns* option, as well as two new options: *Using Only Currently Selected Patterns*, which takes the current user selection of at least 4 samples on the display, and performs a new analysis on just those items, and *Ignoring All Non-Crystalline Patterns*, which performs an analysis of all samples in the

current run that were not flagged as possibly amorphous. Both the latter two options give the option to select a location to store the new output files in; the default location is a sub-folder of the current output folder.

The way samples are labelled in the display has been drastically changed. New options allow the format of a label in the filename to be set - either the full filename, or various subsets of it may be selected. If the label is longer than 7 characters, only the last 7 characters are displayed; the full label may be shown by hovering the mouse over the display until a tooltip appears. Labels are now always off by default on the 3D plots, but can be accessed at any time by the tooltip method described above. Finally, the menu item *Show Well Identity if Available* has been retitled *Show Sample Labels If Available*, to better reflect their new flexibility.

It is now possible to import data from ASCII text files with the .UXD extension.

Bug Fixes

When the system locale is set to use the comma as the decimal separator, ASCII files containing commas within the values for angle or intensity are now correctly read in.

A bug where the fact that an x-axis shift was calculated was not recorded in the automatic report writer has been fixed.

An issue where manually matching against a reference phase database gave incorrectly scaled Rank values has been resolved.

A bug that could cause a crash in certain situations where the View Mixtures option was selected has been fixed.

Version 1.5.4 - Build 10-1-05

When zoomed into the dendrogram display, and changing the selected pattern by means of the keyboard, it is now possible to navigate across the entire dendrogram, not just the currently visible portion.

An issue where nonsense results were generated in the quantitative analysis when a pattern start angle was less than 0 has been resolved.

Version 1.5.3 - Build 5-11-04

A bug where quantitative results were not being calculated correctly has been fixed.

An issue where the report-writer pane did not resize correctly when the Pattern information screen was hidden has been fixed.

Version 1.5.2 - Build 15-9-04

Feature Additions

An additional technique to help identify amorphous samples has been added. This uses a Fourier Transform Power Spectrum to determine if a given pattern has any useful structure. New options have been added to the Advanced Matching Options dialog to control this. As an experimental feature, it is turned off by default.

When using reference phases, the generated known phase database file is now created in the program output directory, and not the reference data directory as before. This approach means we need only read access to the data, and not read-write to be able to use known phases.

The *Hide Group* menu option has been renamed *Mask Group*, and is now available for the dendrogram display as well as the 3D plots.

When a particular cluster grouping is selected using the *Mask Group* option in the 3D plot, pressing the space bar toggles between hiding just that selected group, or hiding all of the other groups instead while showing only the selected group.

It is now possible to show the colours from the Dendrogram as part of a Modified 3D Plot (6D Plot), by selecting *Dendrogram Colours* from the *Ball Colour* drop-down list.

Selecting two patterns to be overlaid in the main results window now automatically shows the Rank correlation coefficient for the two patterns in the Pattern Information pane titlebar.

The *Toggle Mode* option in the right click menu on the pattern graph now displays all but the most recent pattern in the same colour, to allow easier comparisons between a particular pattern and multiple other samples.

An additional option in the 6D Plot allows the dendrogram colours to be used as the *Colour* field for a plot. This is accessible from the *Colour* drop-down menu item *Dendrogram Colours*.

The option to mask selected pattern regions from the *Run On...* dialog box has been turned from a button into a checkbox. Selecting

the checkbox brings up the settings dialog. This makes it easier to see if masking will be used for a given data set or not.

If appropriate, the installed serial number and the demo expiry date are now displayed on the program splash screen.

The report writer has been updated to display additional program references.

Bug Fixes

The program should now behave more reliably on international systems when the system locale decimal separator symbol is set to something other than "."

An issue where the use of the F5 key failed to locate patterns on the graphic displays after a zoom or rotate operation was resolved.

A bug where patterns of different data ranges were not masked correctly has been fixed.

Some issues that prevented the program being run in command line mode were resolved.

Some issues that prevented the program running correctly when the output folder was set to be on a remote network drive have been fixed.

A problem where the 6D plot did not fill the screen when the pattern information pane was hidden was fixed.

An issue where the labels in the cluster member listings were not output in the correct order in the logfile has been resolved.

Version 1.5.1 - Build 1-4-04

Feature Additions

The ability to display a colour-coded key for which patterns are which on the graph display has been added; it may be toggled by means of the item *Show Pattern Key on Graph* in the *Display* menu.

A new option has been added to the preferences to set the number of datafiles above which label display will be off by default on the 3D plots; this helps to prevent confusion when large numbers of patterns are being plotted. The default is 60.

The ability to import data from the MDI ASCII output format has been added.

The ability to hide or show the dendrogram axes has been added; it may be toggled using the right-click menu item *Show Axes*.

A new option in the dendrogram and 3D plots allows the user to locate a particular pattern by either pattern index, or pattern label. This option, accessed via the right-click menu item *Find Item...* brings up a tooltip to show where on the display the particular pattern is located. The located item can optionally be selected and centred in the display.

The new find pattern option is also integrated with the silhouette and fuzzy clustering results display. Clicking on a silhouette selects the corresponding patterns, after which repeatedly pressing the F5 key 3D plots are visible highlights each of them in turn. This allows outlier or other suspicious patterns to be easily located.

An option to change the equation used to try to refine pattern offsets has been added; as well as the default of $D2q = a_0 + a_1 \sin q$, it is now possible to select $D2q = a_0 + a_1 \cos q$ and $D2q = a_0 + a_1 \sin 2q$ as alternatives, giving more flexibility to deal with different data collection geometries.

Several program output files have been renamed to make them more self-descriptive and consistent.

References to the PolySNAP software have been added to the automatic report writer.

The *Advanced* tab has been renamed the *Validation* tab to better reflect its usefulness.

The default sphere size on the 3D plots is slightly smaller than in previous versions.

Bug Fixes

Several bugs relating to updating the 3D plot colours after the cut-level was altered have been resolved.

A bug where pattern selection was not possible on the 6D plot has been fixed.

A bug where known phases were not displayed on the Cell Display when Well IDs were showing has been corrected.

The program manual has been reorganised to give greater prominence to the PolySNAP sections, and the installation chapter has been rewritten to reflect the new installer mechanism.

Version 1.5 - Build 20-3-04

Feature Additions

PolySNAP can now work with up to 1500 patterns in a single run.

A new option has been added to the *Run On...* dialog allowing up to three independent regions of the pattern to be masked at the start of a standard program run. This makes this facility much more easily available than in previous versions.

Several new graphics options are available in the various 3D plots. These include the ability to render the spheres with either partial transparency or as dots, allowing any hidden features to be more easily spotted. Facilities to employ a clipping plane, as well as to hide or bring forward a selected colour group, make dealing with complex plots and large data sets more practical. The ability to manually drag pattern labels on 3D plots into different positions has also been added, allowing for the creation of more professional looking plots when creating diagrams for use elsewhere. These features are all accessed through the right-click pop-up menu.

A new tab, *6D Plot*, has been added to make access to the multi-dimensional plot easier. A more intuitive interface also acts as a key as to what dimension is representing what information field. Once created, a 6D plot now is not deleted automatically, so it is now possible to switch between the other displays and the 6D plot without having to redraw it each time. The *Tools* menu option previously used to access this feature has been removed.

Another new tab, *Advanced*, has been created and contains several options that may not be required in a straightforward program run, but that could be useful in trickier cases. The *Scree plot* has been moved here. New plots include *Minimum Spanning Trees*, *Silhouettes* and *Fuzzy Clustering*. MSTs give a different way of splitting up the patterns into different clusters, based on the maximum distances between samples. Silhouettes and Fuzzy Clustering can give information about particular members of clusters that should perhaps be looked at in more detail - borderline cases, or possible mixtures. New options in the clustering advanced options pane give upper and lower cut-off values for the fuzzy clustering. Full details on these new plots, their meaning, and how best to utilise the information they give, can be found in the full program manual.

A new *Tools* menu item, *List Most Representative Patterns…*now gives a list of the patterns considered to be the most representative in

each cluster. A list is displayed on the screen, and is also automatically added to the logfile. Note that the results depend on the current dendrogram cut-level, and also that different results will be obtained depending on if the MMDS or PCA displays are selected when this option is chosen.

A new *Tools* menu item, *Rerun Analysis using Most Representative Patterns…* takes the current list of MRPs, and uses it to create a known phases database using those patterns. The entire analysis is then re-run for the current dataset, but now additionally using the MRP known phases for comparison. The output from this re-run is saved inside a new folder called *MRP_rerun*, which is created inside the original output folder. This may help to identify possible mixture patterns. However, caution should be used when examining the results, as it is entirely possible one of the MRP patterns could be a mixture itself.

A new *Display* menu item, *Show Amorphous Samples on 3D Plots*, toggles whether or not samples flagged as amorphous are plotted. This can make quite a difference to plots in some cases - especially the PCA - as it often removes extreme outliers and therefore makes seeing the actual clustering simpler.

When several profiles are displayed on the graph plot, control-clicking to deselect one of them on the cell, dendrogram or 3D displays now correctly results in the relevant profile being removed from the graph plot.

A new setting in the program *Options* dialog box, *Extract Well ID from Filename* toggles if the last 3 or 4 characters of the pattern filename (after a separating underscore character) represent the Well ID of the pattern, and are parsed and used to identify the pattern on the Cell Display and Dendrogram Display.

The Dendrogram cut-level is now drawn as a solid yellow line; the upper and lower confidence lines are now dotted yellow lines to help increase their visibility.

The format of the report has been changed very slightly to output the list of clusters including a separation colon, in order to make turning it into a table in Word later more feasible.

If the user selects an output folder that already contains PolySNAP output files, they are now warned before the older files are deleted. An option to turn off this warning is available in the program options. Additionally, an option to start each logfile afresh rather than appending to any existing file is also available.

The *Tools* menu option *List Pattern Clusters* has been renamed to *List Pattern Cluster Members*.

Program options are now saved in a file, *snap.ini*, in the program directory (normally *C:\Program Files\SNAP*). This means settings are persistent across multiple users on the same machine, and can be more easily edited by hand if necessary allowing for simpler troubleshooting.

Entries in the error logfile are now time and date-stamped. Some list entries in the main logfile are now tidier.

Bug Fixes

An issue where an amorphous sample may sometimes appear by itself on the left-hand side of the dendrogram rather than on the right has been corrected.

Version 1.0.5 - Build 13-11-03

Feature Additions and Bug Fixes

Identified an issue that occurs on machines with Office 2000 Professional installed, when logged in as a non-administrative user, who has never run an Office program on that particular computer, and PolySNAP is run for the first time. The Office Installer pops up repeatedly at various points when running PolySNAP, and great persistence is required to dismiss it. This appears to be an issue with the MS Office program, and the current workaround is to launch any Office program, *e.g. Microsoft Word*, once and then exit it, before launching PolySNAP for the first time. If this is done once, the problem does not reoccur. For more information, see Microsoft Knowledge Base Article Q298385:

http://support.microsoft.com/default.aspx?scid=kb;EN-US;Q298385

An issue where the list of clusters obtained by selecting *List Pattern Clusters* from the *Tools* menu was not correctly updated after saving changes to the dendrogram was resolved.

An additional menu item, *Show Graphics Toolbars*, was added to the *Display* menu. This allows the toolbar visibility to be set persistently for an entire session.

The program installer now creates a shortcut to PolySNAP on the Windows Desktop which is available to all users, and not just the user who initially ran the installer.

An issue where the currently selected graph was reset to the initial view when switching to the report or logfile panes was fixed - the current view is now retained.

An issue where the colours used to display peak profiles in the *Peak Compare* window were reversed from those used to display the profiles in main graph window was resolved.

A display bug where the Cell Display was failed to draw correctly when the Display window is first created with some datasets was fixed.

Various small issues with the tutorial documentation were corrected, and some wording changed to make certain steps clearer.

Version 1.0.4 - Build 1-11-03

Feature Additions and Bug Fixes

The PolySNAP *Run On...* window has been redesigned to make it easier to use. Options are given to turn on processing controls such as background subtraction and allowing x-shift calculation on a per-run basis. It has also been made the first entry in the *PolySNAP* main menu.

The security feature that prevents users without Administrative privileges to access the program options can now be disabled if required. Admin privileges are required to toggle this particular option however.

A new preference option has been added to allow the location of any installed GADDS software to be specified. Upon specifying a location, the PolySNAP copies a script file to the GADDS script directory automatically. This script file, *DisplayFrame.slm*, allows for the basic integration between the Display window and GADDS software.

A bug where a manual match against a known phase database failed to work with certain patterns has been resolved [Bug ID: 000121].

A problem where turning off the option to separate amorphous samples caused the cluster analysis to fail was fixed [Bug ID: 000122].

Version 1.0.3 - Build 7-10-03

Feature Additions and Bug Fixes

A bug where the program sometimes crashed when attempting to view the numerical results pane after performing a match with an offset has been resolved.

Version 1.0.2 - Build 13-8-03

Feature Additions and Bug Fixes

The program has been updated to read in and display the revised, expanded format of sample preparation information. Fields that are empty of information are now no longer shown in the sample information pane.

It is now possible to view two patterns that have had a best x-offset calculated shifted by the correct amount. This is accessed by selecting the two patterns of interest in the *Numerical Results* display pane, when the *Display* menu item *Show Calculated x-shift* is selected. Additionally, the calculated offset values are shown in the tooltip text for that cell after clicking.

A new option, *Include Detailed Sample Information In Report*, lists each pattern by Well ID (if present), and the corresponding sample information fields, in the standard order and separated by semi-colons. When the saved report is opened in Word, the *Convert to Table* feature can then be used to turn this into a table for easier reading.

Fixed a problem that prevented 6D customisable plots from working in View Results mode, or when a database was used as the pattern source [Bug ID #000117].

A bug that could cause patterns that were not mixtures to appear as if they were after some settings has been changed has been fixed [Bug ID #000116].

An issue where using the *Reset All* option for preferences did not take effect for the matching settings has been fixed [Bug ID #000115].

PolySNAP no longer asks if you want to save the report when it has no actual content.

A problem where a predefined masked pattern region in a SNAP database would be lost when running PolySNAP on the data has been resolved.

A bug which prevented the Most Representative Pattern (MRP) from being highlighted on the 3D plots has been fixed.

The bug in which only the first page of the report or logfile would print should hopefully be resolved.

A serious memory leak that occurred when non-linear shifts were being calculated has been repaired [Bug ID #000118].

Version 1.0.1 - Build 7-7-03

Feature Additions and Bug Fixes

Added the facility to allow for a non-linear offset x-shift that varies with sin q. This is accessed through the Options -> Matching Options screen, and slows the matching operation down by around a factor of 4. Output from this is saved in the file *distances_offsets.txt* in the program output folder.

A new facility, accessed through an extra checkbox in the PolySNAP Options dialog that enables a 'strict' background subtraction on import has been added.

To speed up the analysis, the default clustering method is now Group Average Link (previously known as the Ward Method), instead of Auto-Select.

A new option in the *Display* menu to toggle between showing profiles with or without processing has been added.

If one of either of the PCA or MMDS plots are giving markedly better results than the other, then that display will have two stars (**) next to its name in the tab bar header to highlight that fact.

Fixed a bug that caused all but the first cell to be grey with a question mark under it after the dendrogram cut-level had been modified manually (Bug ID: 000054).

The default program wait between loading each pattern has been changed to 4 seconds.

Fixed an issue where the standard background subtraction checkbox setting was not being retained.

An issue where the program timeout was triggering far too soon when loading data from a network drive should now be resolved.

An irritating problem where spurious error messages about images were displayed when opening main results window has been fixed.

Fixed an issue that prevented Bruker Frame (*.*gfrm*) files from being viewed in the GADDS software despite it being installed correctly.

Fixed an issue where the amorphous indicator failed to behave as expected when any of the background subtracting options were turned on.

Fixed a bug where the well identities on the dendrogram/cell display were lost after changing dendrogram method.

Fixed a problem where DSC data was not importing correctly.

Version 1.0 - Build 30-4-03

Feature Additions and Bug Fixes

PolySNAP now pre-screens for patterns that have no signal, and does not include them in the analysis. The logfile is updated accordingly.

After selecting a particular pattern in the cell or dendrogram display with the mouse, the left and right arrow keys can be used to quickly select and scan through adjacent samples.

The mechanism used to highlight the currently selected sample on the 3D plots has been improved.

The algorithm to identify samples that are entirely amorphous has been enhanced, and now has a much higher success rate. See the program manual for a description of the modified method.

The algorithm to calculate components of a mixture has been improved to make it less likely spurious phases are included in the answer. See the program manual for a description of the modified method.

The dialog box that allows the user to list and examine all the samples marked by the program as amorphous now lists patterns by filename, rather than chemical ID field, as the latter is not always present. It also now optionally displays the amorphous-subtracted profile calculated by the program to allow the user a better understanding of why a particular sample may have been flagged.

The clustering method previously known as Ward is now referred to in the program and documentation by its correct name of the Group Average Link method. Additionally, the option to use the Sum of Squares method has been removed due to its poor performance.

Program instability and occasional crashes that occurred when a large number of samples were marked as 'amorphous' has been corrected.

An issue, also present in SNAP-1D, where the calculated error on quantitative analysis could produce unrealistic values when going from a scale to a weight fraction is now fixed.

A serious bug where changing the cluster method several times on a large dataset, and then returning to original could give different results in certain situations has been corrected.

Amorphous phases are now correctly labelled as such on the cell display.

An issue where the automatic report generation could cause the program to crash was resolved.

The 6D customisable plots are now more robust and reliable. Additionally, the initial 'large' ball size no longer dwarfs the entire display.

An issue where the Simplified Dendrogram option was not selecting the expected patterns for display has been corrected.

Known problems/bugs in this build

Opening a Frame file in GADDS is untested and requires the user to manually install a script file.

Printing from the report or logfile panes can be unreliable at times, depending on the current printer and operating system combination. The recommended workaround is to save a copy of the document, open it in Microsoft Word or similar, and print from that application.

Beta Version - Build 10-04-03

Feature Additions and Bug Fixes

The pattern profile graph has been replaced with a more flexible OpenGL based version. This allows for overlaid profiles to be offset in the x or y directions, and individual profiles can be identified by 'tooltips' containing the corresponding pattern filename. The limit on eight overlaid profiles has been lifted, and up to 1000 profiles can be shown at once on the same display. The graph can now be copied to the clipboard and pasted into the report or elsewhere. New options have been added to the *Display* menu to allow control of these features.

A basic automatic report can now be generated to order using the *Generate Report* option in the *Tools* menu. The use can select a detailed or summary report, and whether or not to include copies of the graphics output.

Added the ability to directly import all current types of Bruker RAW format files, including Diffrac-AT.

Code to change the behaviour of the program to prevent a non Administrator user from changing program settings has been added.

When importing data from a network drive, the program should recover transparently from most types of temporary network failure.

A goodness of fit indicator has been added to the 3D MMDS plot (shown in the upper left corner), to give an indication of the quality of the clustering results. Numbers closer to 1.0 are more reliable.

A new option in the preferences allows a backup copy of any report to be automatically saved to a pre-defined location whenever a report is saved manually.

Added the ability to save the report in ASCII text format.

Added the ability to open Bruker Frame files in the GADDS software, assuming it is installed on the current machine, and the relevant script file (*DisplayFrame.slm*) is installed in the GADDS software scripts directory.

If filenames include the well ID of a sample (e.g. 1A01) this may optionally be shown automatically in the cell and dendrogram displays.

If sample images are present in the data directory, they are by default

shown in thumbnail format next to the pattern profile when a corresponding sample is selected. Clicking on the thumbnail opens the full-size image in the system image viewer.

Fixed a problem where re-running the same datasets multiple times could cause a memory overflow, resulting in nonsense results being generated.

An issue where the Run On… dialog box did not remember the last number of files set to import was resolved.

An issue that prevented the program from running to completion when a database was selected as the data source in the *Run On...* window was fixed.

An issue where the busy cursor was not shown during Change Cluster Method operations was fixed.

The label to identify the 'Simplified' dendrogram has moved to the top-left to be more obvious.

The program installer now checks to ensure that the system is Windows 2000 or more recent, and refuses to install if this requirement is not met.

Known problems/bugs in this build

Opening a Frame file in GADDS is untested and requires the user to manually install a script file.

Printing from the report or logfile panes is not yet fully implemented. The recommended workaround is to save a copy of the document, open it in Microsoft Word or similar, and print from that application.

Crystallinity Problems - the program occasionally mis-identifies patterns as being non-crystalline. This is currently being investigated and a revised methodology is planned for the next release.

Mixture problems - the mixture identification routines are unfinished. Additionally, the facility to manually override a program suggested mixture component is not yet available.

Beta Version – Build 04-02-03

Feature Additions and Bug Fixes

An issue with the parsing of the sample info where the 11th data field was not being read in was corrected.

Added a 'None' button available in Options for the Known Phase Database location, and fixed issue where it didn't work. The default location for Known phases is now '<None>'.

Beta Version – Build 17-01-03

Feature Additions and Bug Fixes

The underlying SNAP-1D program is now upgraded to version 1.2. See the *SNAP-1D Release Notes* for more information on changes.

A new menu item in Pattern Menu, *Open Pattern Using EVA*, allows a selected pattern to be opened for viewing in EVA from PolySNAP. This only works for single RAW files in the standard data directory, and assumes EVA is set as the default editor for RAW files.

The *Toggle Mode* option in the right-click menu for the dendrogram display now switches between the standard dendrogram, and a 'simplified' dendrogram. The latter displays only the first, last and middle pattern in each cluster.

Upper and lower confidence bounds on the suggested number of clusters are now plotted as dotted lines on the dendrogram display.

A new option in the right-click menu of the numerical results pane brings up a window to show a difference plot for the selected pair of patterns.

A new button has been added in the *Options* window to allow the user return all preferences to their default values.

The program now checks that the weights entered by the user for the selected matching tests sum to 1.0, and warns when they do not.

The *Find Peaks* preferences option is now no longer available as a separate preference in PolySNAP, as there is no good reason for it to ever be turned off.

The selection of the default input, known and output directories for Automatic runs of PolySNAP has been simplified, and the user interface improved.

Within a given program instance, the *Run On...* dialog box now remembers the last location for each folder used each time it is opened. The first time it defaults to the default program locations set in the *Options* dialog box.

The interface to override the default amount of pattern smoothing has been simplified to being a value from 0 to 10. 1 is the default amount.

The contents of the PolySNAP data import window now resize correctly when displayed on screens larger than 1024x768.

The print mechanism for the report and logfile panes has been changed in an attempt to circumvent printing problems where spurious extra pages were being printed. There are now separate *Print Selection* and *Print* options in the right-click menus; the latter uses the experimental new printing mechanism, but requires to be tested on-site.

A problem on slower machines where the Pattern Editor window was opened, and the pattern display flickered badly, has been fixed.

An issue where the directory listing in the Select Folder screen was not updating after the New Folder button was used has been resolved.

Fixed a problem where the program would sometimes not exit properly after an automatic run had been cancelled by the user.

Experimenting with the changing the clustering method in the results screen no longer changes the default cluster method for the next run.

Fixed a problem where trying to plot too many profiles on the graph caused the same error message to be repeatedly displayed.

Pattern filenames instead of chemical names are now shown in results fields of the match window.

Fixed a problem where nonsense values were sometimes being written to the *distances_offsets.txt* file.

A problem with the parametric Pearson test behaviour, where results between -1.0 and 0.0 were not reported correctly, has been rectified.

The Match All progress bar is now only displayed with more than 50 patterns, or more than 15 when an x-offset calculation is being performed.

A new version of the program installer is now being used to prevent problems that kept occurring when trying to install on certain Windows 2000 systems.

Known problems/bugs in this build

[Bug ID: 000100] Amorphous phases are not correctly labelled as such on the cell display yet.

[Bug ID #000085] Customised 3D plots (aka “6D Plot”) are still largely untested due to lack of data containing varied sample information.

[Bug ID #000084] Graph Pane cannot be copied to the clipboard.

[Bug ID #000083] Button to open Frame files not working yet.

[Bug ID #000071] Amorphous Indicator calculation not always reliable.

[Bug ID #000064] Preference setting to Subtract Background is not always retained correctly.

[Bug ID #000056] Hardware graphics acceleration needs to be switched off on some PCs or graphics display strangely.

[Bug ID #000054] View Results option does not take any notice of known phases.

Beta Version – Build 17-12-02

Feature Additions and Bug Fixes

Match All code has now been optimised, resulting in roughly halving the time taken for that part of an average calculation. For example, performing Match All on 300 patterns took 3 minutes 20 seconds in the previous beta version, and now only takes 1 minute 20 seconds.

A new option to allow a variable 2q-offset to be calculated for patterns that may incorporate a shift has been added. This is accessed through the *Edit* menu -> *Options* -> *PolySNAP* tab -> *Matching* button -> *Allow x-offsets of up to...* checkbox. Note that the process greatly adds to the program run-time, and at present, only offsets for the Spearman and Parametric tests are calculated. A new results file in the program output folder, *distances_offsets.txt* is now created. This resolves [Bug ID/Feature Request 0000075].

New code has been added to improve system performance when the program is waiting for the input directory contents to change during an automatic run - the program now goes to sleep until it is sent a signal by the system to wake and continue processing. This new function does mean the interface is unresponsive during the sleep period however. The maximum sleep time has been changed to one minute as a result since system resources are not being used otherwise in that time.

It is now possible to change the number of files expected by the program once an auto-run importing has started; the display and internal counters are updated accordingly [Feature Request/Bug ID #000061].

An option is added in the *Display* menu to view the dendrogram-generated version of the cell display even when known phases are present. This resolves [Feature Request/Bug ID: 000086].

A new feature is added that displays the multiple-pattern overlay key in appropriate colours for the patterns, thus making it easier to see which profile corresponds to which filename; this resolves [Feature Request/Bug ID: 000069].

A new standardised function is used for all operations that involve writing to the program logfile. This is more robust, as it checks for the existence of the file, and creates it if necessary. It also uses more modern file-handling methods, and resolves [Feature Request/Bug ID 000096].

PolySNAP will no longer allow the user to set the program output directory to the same location as the input directory. Error checking of the input directories is also improved.

A problem where writing the cluster members list to the logfile with more than 26 groups caused strange ASCII characters to appear has been fixed, resolving [Bug ID 000046].

A problem where in some cases, switching between display modes, or clicking twice in succession on the Dendrogram tab, causes a saved, modified cut-level to be lost & replaced with the original calculated version should now be resolved [Bug ID #000048].

The known phases key on cell display no longer repeatedly uses the first file name for all entries; this key is limited to 8 characters, so the last 8 characters of the filename (minus the extension) are used, resolving [Bug ID: 000099].

Fixed [Bug ID: 000101] where the *Match Now* button on the import screen did not work correctly in some situations.

Some problems with parsing sample data should now have been resolved [Bug ID 000102].

Beta Version – Build 08-11-02

Feature Additions and Bug Fixes

The most representative pattern (MRP) for each cluster can be (optionally) shown in the 3D plot (it has spikes coming out of it); selecting it brings up a dialog box reporting the mean pattern-pattern distance for that cluster.

The colouring of the different clusters has been improved using shading, so larger numbers of separate clusters can be told apart more easily.

The quality of the plotting of the spheres in 3D plots can now be user-controlled (press F2 when 3D plot is shown).

Improved cluster analysis code automatically selects the best clustering method for the problem; this can be manually set or overridden by the user if required.

The option to re-run using a different clustering method is much faster than in previous versions.

Added an option to use an external graphics viewer (whatever the system default for JPEG format images is) to look at sample images, rather than the internal SNAP viewer.

[Bug IDs #000094 & #000095] Progress bars are now shown and correctly updated during long, computationally intensive pattern matching and cluster analysis calculations.

[Bug ID #000073] Fixed very annoying bug where changing the dendrogram cut-level manually caused colour assignations in Cell and 3D plots to become incorrectly sorted into ascending numerical order.

[Bug ID #000066] Fixed a bug where some RAW data files were not being fully imported correctly, and as a result data points were being lost (e.g. with pattern DOMAXL_A01.RAW).

[Bug ID #000076] Fixed a problem parsing pat names of known phases in certain situations that caused more known phases than actually existed to appear in the cell display key.

[Bug ID #000070] We now display the short not full path name in multiple pattern overlay key.

[Bug ID #000065] Fixed parsing code of sample information to not expect units in the concentration field.

[Bug ID #000050] We now hide 'x pats loaded out of y' labels on import screen when doing View Results only.

[Bug ID #000074] On import screen, long file-paths no longer wrap and look messy; they are truncated if necessary, and the full path can be revealed via a tooltip.

[Bug ID #000077] The icon for 3D (PCA) plot is now no longer incorrect in display screen toolbar - was showing icon for 2D plot.

[Bug ID #000052] Groups written to log file now do not use incorrect carriage return character and thus display all on one line.

[Bug ID #000060] Online run with GADDS should not now cause *targets.txt* file to be mistakenly imported as a pattern.

[Bug ID #000081] The graphics configuration preference files (*.cfg) are checked for consistency before saved settings are read; this should prevent problems where dendrogram lines appear to vanish due to corrupted preferences values.

[Bug ID #000072] The estimated number of clusters and dendrogram cut-level should now be correct and internally self-consistent.

[Bug ID #000097] Cell display no longer sometimes incorrectly shows results as if known phases present even when they are not.

[Bug ID #000098] Re-running the clustering method no longer causes tree3d.dat and score.dat files to be deleted incorrectly.

[Bug ID #000055] Hourglass cursor no longer often appears randomly and for no good reason when left hovering over the graphics displays.

A problem where program could hang when the return key was pressed to dismiss the ‘Save Modified Dendrogram?’ dialog was shown has been fixed.

Beta Version – Build 17-10-02

Modified test version.

Fixed bug where program would not run correctly given data on a network drive.

Added an option to allow a re-run of the cluster analysis using a different clustering method, *via* an option in the *Tools* menu of the display screen.

Beta Version – Build 14-10-02

Initial test version.