

PolySNAP: a computer program for analysing high-throughput powder diffraction data

Gordon Barr,* Wei Dong and Christopher J. Gilmore

Department of Chemistry, University of Glasgow, Glasgow G12 8QQ, Scotland. Correspondence e-mail: gbarr@chem.gla.ac.uk

In high-throughput crystallography experiments, it is possible to measure over 1000 powder diffraction patterns on a series of related compounds, often polymorphs or salts, in less than one week. The analysis of these patterns poses a difficult statistical problem. A computer program is presented that can analyse such data, automatically sort the patterns into related clusters or classes, characterize each cluster and identify any unusual samples containing, for example, unknown or unexpected polymorphs. Mixtures may be analysed quantitatively if a database of pure phases is available. A key component of the method is a set of visualization tools based on dendrograms and pie charts, as well as principal-component analysis and metric multidimensional scaling as a source of three-dimensional score plots. The procedures have been incorporated into the computer program *PolySNAP*, which is available commercially from Bruker-AXS.

© 2004 International Union of Crystallography
Printed in Great Britain – all rights reserved

1. Introduction

The *PolySNAP* computer program is designed to process and analyse high-throughput powder diffraction data (Barr, Dong & Gilmore, 2004; Storey *et al.*, 2004). It generates an ($n \times n$) correlation matrix by applying the Pearson and Spearman correlation coefficients to match the full profiles of each of the n patterns in turn with every other pattern in a database (Gilmore *et al.*, 2004), and then uses this matrix to partition the patterns into related clusters or sets. No reference sample database is required, but if one is available the partitions reflect the composition of each sample relative to the database, and mixtures are quantitatively estimated. All the calculations, including quantitative analysis, use the full measured data profile. Coupled with the calculations is a set of visualization tools that exploit the power of computer graphics and which are based on classification procedures (see, for example, Gordon, 1999). Those of special use include dendrograms, metric multidimensional scaling, three-dimensional principal-component analysis and scree plots. The theory has been described elsewhere (Barr, Dong & Gilmore, 2004); here we describe the operation of the associated computer software, *PolySNAP*, in detail.

2. Program operation

This section describes each step of the program operation and user interface in detail. It should be noted that some parts of *PolySNAP* are built upon the *SNAP-ID* software (Barr, Gilmore & Paisley, 2004), and all the options present in *SNAP-ID* are available in *PolySNAP*. A flowchart of the program is shown in Fig. 1.

2.1. Data input modes

The diffraction data are input into the software in two possible ways.

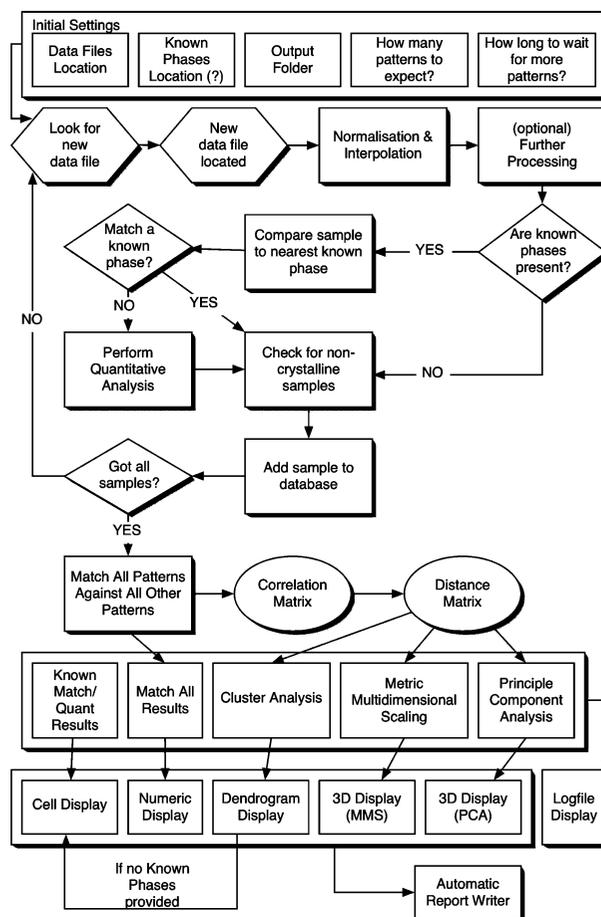


Figure 1
The *PolySNAP* flowchart from raw data to the final automatically generated report.

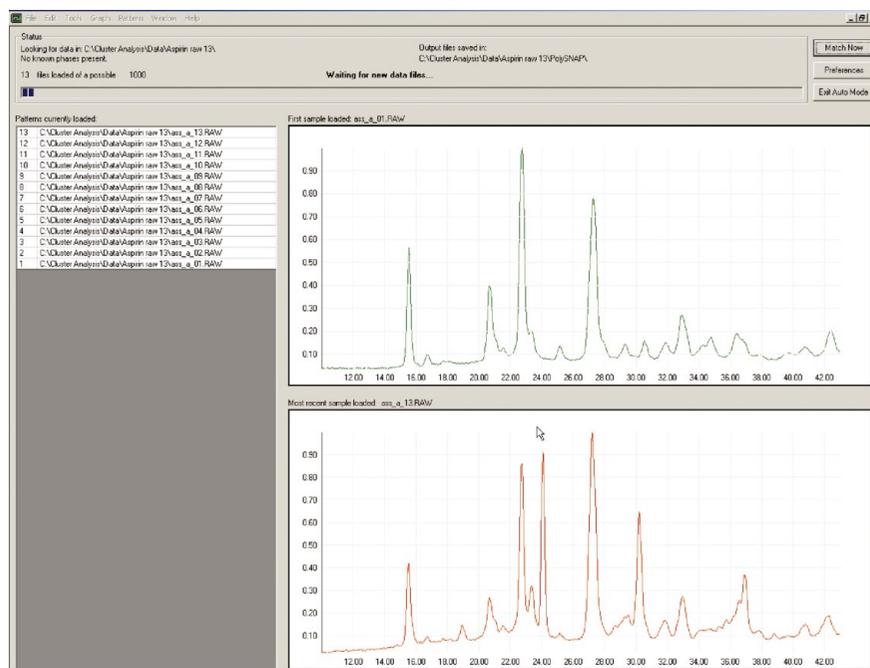


Figure 2

The initial pattern-import window. The top graphics pane contains the diffraction pattern for the first input pattern (or another selected from the left-hand column) and the lower pane contains the current imported pattern. *PolySNAP* either looks in a specified directory and loads all the pattern files in it, or it waits for patterns to arrive as they are collected on a high-throughput data collection system

(a) The program is given the name of a directory or folder in which a full set of patterns is located. *PolySNAP* then imports them all, one at a time, processing as it proceeds.

(b) The program accesses a folder in active use by the data acquisition hardware. Data sets are deposited here as they are collected. *PolySNAP* duly processes them until either the expected number of data sets has arrived or until a time-out occurs due, for example, to equipment malfunction.

Fig. 2 shows the typical user interface during data import. The top graphics pane contains the diffraction pattern for the first input pattern (or another selected from the left-hand column) and the lower pane contains the most recently imported pattern. This enables the user to monitor the data import process.

2.2. Data pre-processing

The data are pre-processed. This procedure has been described in full by Gilmore *et al.* (2004). In summary:

(i) Data are imported either as ASCII xy data (2θ , intensity), in powder CIF format (Hall *et al.*, 1991), MDI ASCII format, or in Bruker raw data format.

(ii) The intensity data are normalized.

(iii) Each pattern is interpolated or extrapolated if necessary to give standard increments of 0.02° in 2θ .

(iv) Background removal is optional. When requested, local n th-order polynomial functions are fitted to the data and then subtracted to remove the background.

(v) The data are optionally smoothed using wavelets.

(vi) Peak positions are optionally found using Savitsky–Golay filtering.

(vii) Multiple regions that the user wishes to exclude from the matching can be masked. This is often done when internal reference standards are present in the sample.

2.3. Quantitative analysis

At this point, *PolySNAP* looks for a database of pure phases as a user input option. If the database is present, then the program operates in quantitative mode: each input sample is checked against the database of known phases and if the pattern correlations (based on all the measured data points and using the Pearson and Spearman correlation coefficients) do not indicate a good match, then quantitative analysis is carried out using all the measured data points with singular value decomposition as the tool of matrix inversion to ensure computational stability. The percentage composition of the sample is determined. If no such database is present, then this step is bypassed.

2.4. Non-crystalline samples

The input sample is checked for crystallinity as follows.

(a) The total background for each pattern is estimated and its intensity integrated.

(b) The non-background intensity is estimated.

(c) The diffraction peaks, if any, are located.

(d) The ratio of non-background to background intensity is determined. If this ratio falls below a user-set limit (with a default of 3%) and

if, additionally, there are less than n peaks identified (with a default value of 3), the sample is flagged as non-crystalline.

Amorphous samples are treated separately during the remainder of the *PolySNAP* procedures.

2.5. The correlation, similarity and distance matrices

When pattern import is complete, each of the n patterns is matched against every other pattern present (including itself) using a mixture of Spearman, Pearson, Kolmogorov–Smirnov and peak correlation coefficients (Gilmore *et al.*, 2004). The weighted mean of these coefficients is used as an overall measure of correlation. The weights are user adjustable, and by default take the values 0.5, 0.5, 0.0, 0.0 for the Spearman, Pearson, Kolmogorov–Smirnov and peak correlation coefficients, respectively, *i.e.* only the non-peak specific tests are used. In this way, the program generates a symmetric ($n \times n$) correlation matrix, ρ , with a unit diagonal. Two symmetric ($n \times n$) matrices are generated from ρ , as follows.

(i) The distance matrix, \mathbf{d} , with elements

$$d_{ij} = 0.5(1.0 - \rho_{ij}); \quad 0.0 \leq d_{ij} \leq 1.0. \quad (1)$$

A high correlation implies a short distance, and *vice versa*.

(ii) The similarity matrix \mathbf{s} :

$$s_{ij} = 1.0 - d_{ij}/(d_{ij})_{\max}; \quad 0.0 \leq s_{ij} \leq 1.0. \quad (2)$$

Patterns with a high correlation will have high similarity coefficients.

If amorphous samples are present, they can either be discarded at this point or the distance matrix modified so that each amorphous sample is given a distance and dissimilarity of 1.0 from every other sample, and a correlation coefficient of zero. This automatically excludes the samples from the clustering until the last amalgamation steps, and also limits their effect on the estimation of the number of clusters present in the data (§2.6.1).

An optimal shift in 2θ between patterns is often required, arising from equipment settings, especially due to variation in the sample height, and data collection protocols. We use the forms

$$\Delta(2\theta) = a_0 + a_1 \cos \theta, \quad (3)$$

which corrects for varying sample heights in reflection mode, or

$$\Delta(2\theta) = a_0 + a_1 \sin \theta, \quad (4)$$

which corrects for transparency errors or, for example, transmission geometry with constant specimen–detector distance, and

$$\Delta(2\theta) = a_0 + a_1 \sin 2\theta, \quad (5)$$

which provides transparency and thick-specimen error corrections. The parameters a_0 and a_1 are constants, refined automatically to maximize pattern correlations using the downhill simplex method of Nelder & Mead (1965). The user can define maximum values of a_0 and a_1 . The resulting correlation matrix is examined for stability for subsequent eigenanalysis and cluster analysis calculations using singular value decomposition.

2.6. Classification and visualization of the patterns

Following matrix generation, classification analysis is carried out, resulting in a tabbed graphics pane with six user-selectable windows present, which summarize the results.

2.6.1. The dendrogram. The first of these is the dendrogram. A typical dendrogram is shown in Fig. 3(a). In this example, the data are partitioned into four clusters, containing 12, 6, 6 and 2 patterns, respectively. During the generation of a dendrogram, each of the diffraction patterns begins at the bottom of this plot in a separate class, and these amalgamate in stepwise fashion and become linked by horizontal tie bars as the algorithm proceeds. The height of the tie bar represents the similarity between the samples as measured by the relevant distance statistic. There are numerous ways in which to generate a dendrogram, which are differentiated by the way in which the distance matrix is modified as hierarchical clusters are generated. *PolySNAP* offers the following choices.

- (a) Single link.
- (b) Complete link.
- (c) Average link.
- (d) Weighted average link.
- (e) Centroid.
- (f) Group-average link.

As described by Barr, Dong & Gilmore (2004), *PolySNAP* attempts to choose the optimal clustering method. However, in general, the group-average link method works best with most diffraction data. Each cluster is given a unique colour that is used in other graphical representations. The horizontal yellow bar is the cut line. This is established by estimating how many clusters are present in the data (Barr, Dong & Gilmore, 2004). We use the following.

- (i) Eigenanalysis of the correlation matrix, both in its original form and suitably normalized, and the relevant matrix from metric multidimensional scaling routines.
- (ii) The Calinški & Harabasz (1974) test.
- (iii) A variant of the Goodman & Kruskal (1954) γ test as described by Gordon (1999).
- (iv) The C test (Milligan & Cooper, 1985).

To reduce the bias towards a given dendrogram classification scheme, these tests are carried out on four different clustering methods, namely the single link, the group-average link, the sum of squares and the complete link methods, to generate 12 semi-independent estimates of the number of clusters, plus three from eigenanalysis, giving 15 in total.

A composite algorithm combines these estimates. The maximum and minimum values of the number of clusters from the eigenanalysis results define a primary search range (c_{\max} and c_{\min} , respectively). The Calinški and Harabasz test, the Goodman and Kruskal γ test, and the C statistic are then used in the incremented range $\max(c_{\min} - 3, 0) \leq c \leq \min(c_{\max} + 3, 0)$ to find local optimum points. The resulting estimates are averaged, outliers removed, and a weighted mean value taken, which is used as the final estimate of the number of clusters. The upper and lower limits from the 15 estimates are also displayed on the dendrogram as a dashed horizontal line. The cut line can be moved from its initial position by the user to alter the number of clusters, creating more by lowering it, or less on raising it. Modifying the cut level on the dendrogram has consequences for the colours used in the PCA and MMDs plots. If the dendrogram is modified in this way, the user is asked if they wish to retain the change and map the new colour scheme onto these plots. It is possible to undo such an operation using a rollback mechanism.

Any pattern can be selected and displayed in the lower graphics pane. Multiple patterns can be superimposed using the standard Ctrl/Shift keys in the Windows operating system keyboard environment.

It must be emphasized that estimating the number of clusters is a difficult procedure (Graham & Hell, 1985) and that the MMDs and PCA plots described in the next sections also need to be consulted.

When there are more than 100 patterns it can be difficult to read a dendrogram and so a simplified form can be displayed in which each cluster has only three sample patterns displayed: the most representative (see §2.6.3) and the two samples least well joined to the cluster. It is a simple matter to toggle between the two representations. Amorphous samples are placed at the far right-hand end of the dendrogram with tie bars at the zero similarity level.

2.6.2. Well displays, pie charts and stacks. High-throughput powder diffraction experiments use flat plates with wells containing the samples arranged as a two-dimensional grid. It is therefore useful to display the contents of each well in the same format. *PolySNAP* does this by using the results of the dendrogram calculation where each sample is assigned to a cluster, and an associated representative colour is generated. These colours are used to display the samples in well format, with a key at the left-hand side of the pane. Fig. 3(b) shows a typical well display. Any well or combination of wells can be selected and the corresponding pattern(s) displayed in the bottom graphics pane, as in Fig. 3(b). Amorphous samples are duly identified.

If quantitative analysis has been carried out, then the colours are chosen to represent pure phases (displayed in the key on the left of the dendrogram pane). These colours map onto the wells as before, but if mixtures have been identified, pie charts are produced to show the relative proportions of the components of the mixture. Fig. 3(c) shows a typical pie-chart display. As an alternative to this, the data can be represented as stacks, as in Fig. 3(d).

2.6.3. Metric multidimensional scaling. Given an ($n \times n$) distance matrix \mathbf{d}^{obs} , metric multidimensional scaling (MMS; Gower, 1966) seeks to generate an Euclidean distance matrix, \mathbf{d}^{calc} , the elements of which closely approximate those of \mathbf{d}^{obs} . We use a three-dimensional Euclidean distance matrix, \mathbf{X} , to represent the data and to generate \mathbf{d}^{calc} . This may not always be appropriate and, as a check, the program computes a correlation coefficient (based on both the Pearson and the Spearman coefficients) between \mathbf{d}^{obs} and \mathbf{d}^{calc} , which is displayed in the MMDs graphics pane. Normally, one sees values >0.95 , but low values can indicate an inability to match the two matrices using three dimensions. This is rare, even with 1000 patterns, although there is a reduction in the overall correlation coefficient as n increases. From the MMDs calculation, we obtain a three-dimensional coordinate matrix, \mathbf{X} , which contains a set of coordinates (x_i, y_i, z_i) for each

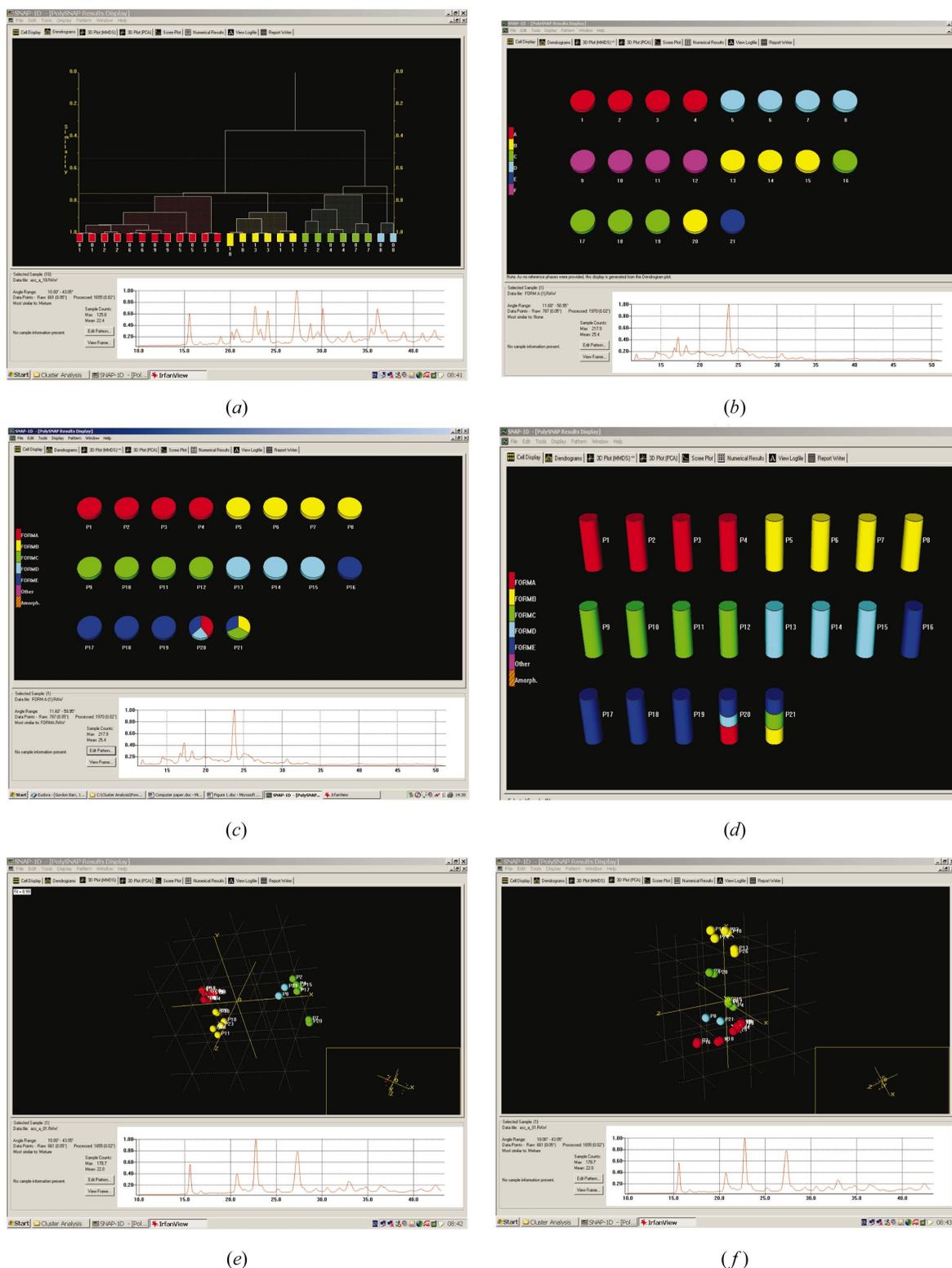


Figure 3

The main PolySNAP graphics and text panes selected by tabs at the top of the window. (a) The dendrogram. The horizontal yellow line is the cut line and defines the data clustering. The 26 patterns are partitioned into four clusters containing 12, 6, 6 and 2 patterns, respectively. One yellow box is vertically extended, having been so selected by the user, and the diffraction pattern corresponding to this is displayed in the lower graphics pane. (b) Pie charts when no database of pure forms is present. Each well is assigned a colour based on the clustering produced by the dendrogram. Mixtures are not identified. (c) The pie chart produced when PolySNAP operates in quantitative mode. The relative proportions are shown in the usual pie-chart way. (d) The use of stacks as an alternative to pie charts. (e) The metric multidimensional scaling (MDS) plot of the data. Each sphere represents a powder pattern and takes the colour assigned by the dendrogram. It is possible to rotate, zoom and pan, remove labels, and adjust foreground and background colours using a graphics toolbar. (f) The equivalent three-dimensional principal-component analysis (PCA) plot. The colour convention matches that of the MDS method. (g) The scree plot from the PCA analysis, used, in part, to define the number of clusters. (h) A dendrogram with the sample image displayed in the bottom right-hand corner. The image can be toggled on or off as desired. (i) The correlation matrix. Clicking on any entry in the matrix results in the display of the two relevant patterns as in this figure in the bottom graphics pane. (j) An automatically generated report on the calculations, which is stored in rich text format (RTF), which can be imported into most word processors. The figures are automatically generated and inserted in the report.

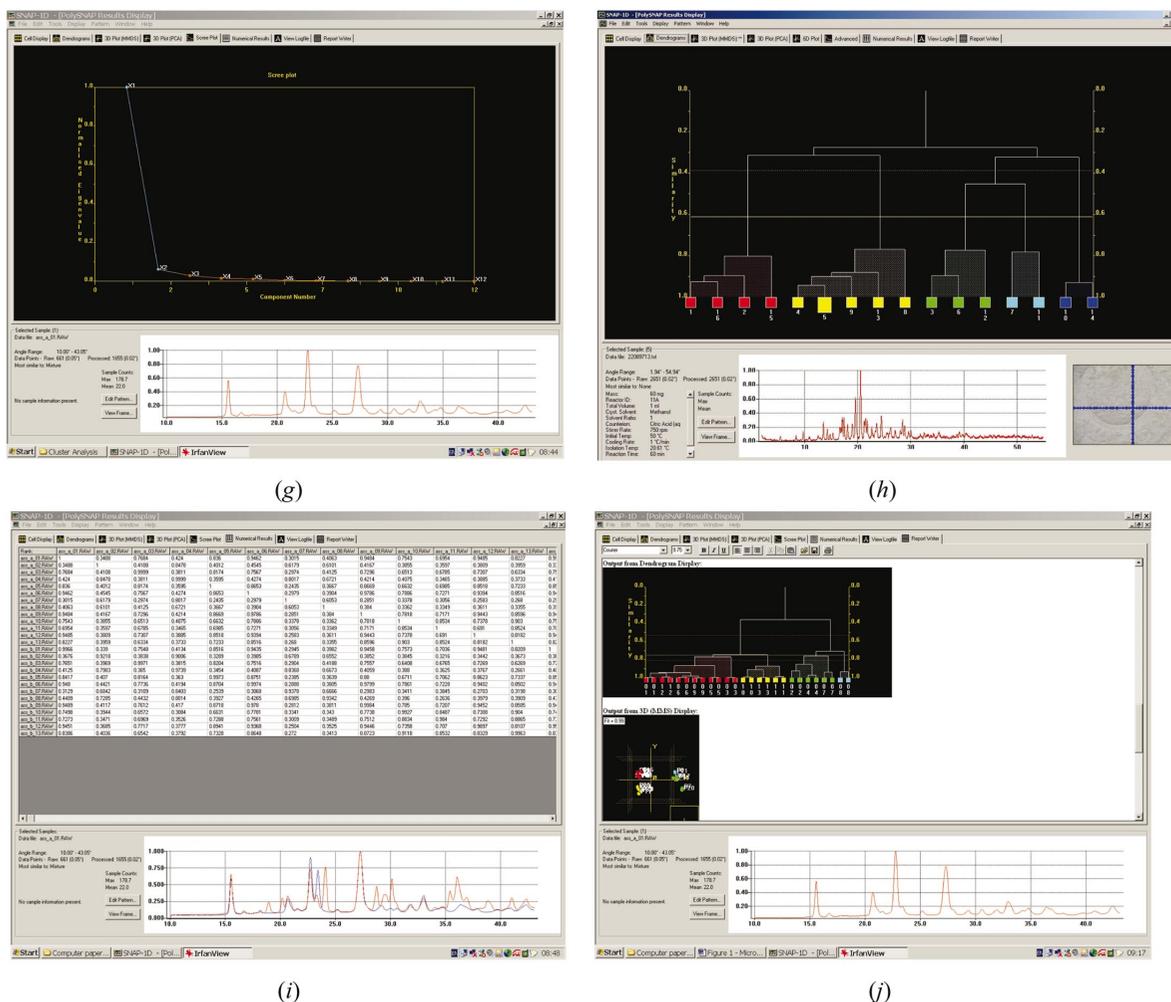


Figure 3 (continued)

sample, i , that can then be plotted in a standard three-dimensional plot with each sample represented as a sphere.

Fig. 3(e) shows a typical plot. It is possible to manipulate this in the usual way: it can be rotated; zoom and pan are possible, the sphere size can be altered; labels can be removed or displayed and display quality can be altered. It is possible to copy and paste from the graphics pane to the clipboard, either in total or using a selected region. The sphere colour is taken from the dendrogram.

The program seeks to identify the most representative sample for each cluster, defined as that sample which has the minimum mean distance from every other sample in the relevant group. This is flagged in the display. There is an optional graphics toolbar that be used to modify background and foreground colours; toggle the grid and labels on and off, and change fonts, *etc.*

2.6.4. Three-dimensional principal-component analysis score plots. A second procedure for displaying the data in three dimensions is provided *via* a three-dimensional score plot from the principal-component analysis of the correlation matrix. Score plots traditionally use two components with the data thus projected onto a plane, but we use three-dimensional plots in which three components are represented. Once the coordinate matrix has been generated from the eigenvectors of the correlation matrix, the display of the data is identical to that of the MMDS formalism and the same facilities are provided. Fig. 3(f) shows a typical three-dimensional

score plot for the same data as the MMDS plot in Fig. 3(e). Note that the two methods have arbitrary origins and orientations.

The program also seeks to identify which of the PCA or MMDS methods give the best representation of the data by identifying which \mathbf{d}^{calc} matrix has the highest correlation with \mathbf{d}^{obs} . The best representation is flagged using asterisks on the MMDS or PCA tabs in the main display window.

2.6.5. The scree plot. The final graphics pane is a simple scree plot resulting from the eigenanalysis of the correlation matrix. It gives a simple graphical representation of one of the primary estimates of cluster numbers. The colour of the plot is modified: it changes from blue to red once 95% of the variability of the data is accounted for. It can be a useful tool: a plot which descends steeply probably has well defined clusters; one which is slow to reach the 95% limit may not be so well behaved. A typical scree plot is shown in Fig. 3(g). This shows an encouragingly steep rate of initial descent.

2.7. The sample image

The Bruker Discover D8-GADDS diffractometer system provides images of each well plate; other hardware may do so as well. If these images are stored in the data directory, they can be displayed along with the diffraction data. Fig. 3(h) shows a typical image displayed at the bottom right-hand corner of the pane.

2.8. Output

Four main text output files are generated, as follows.

(a) A log file which is viewable, but non-editable within the program, that summarizes the input data, the results of pattern matching and classification. All input data choices and alterations to the program defaults are logged and time/date stamped; all re-runs of the software on the same data append the new output to the old file, thus providing an audit trail of all the procedures that have been followed with a given data set. A record of the user and computer identification is kept. This represents partial compliance with 21 CFR Part 11 requirements.

(b) The correlation matrix can be displayed in table format. Fig. 3(i) shows a typical small matrix. Clicking on any entry in the matrix results in the display of the two relevant patterns as in this figure in the bottom graphics pane.

(c) An error log that monitors all warnings and errors. This has the same properties as the log file (a).

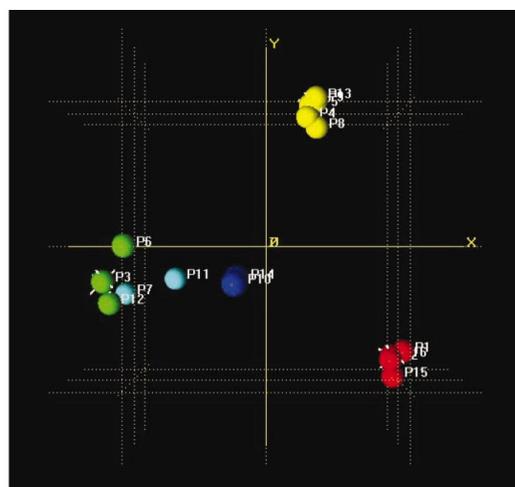
(d) An automatically generated report in rich text format (RTF), which can be easily imported into most word processors. Diagrams are optionally incorporated and a complete summary of the sample data, and the matching and classification calculations is presented. The user can subsequently edit this file within *PolySNAP* or elsewhere. Fig. 3(j) shows a typical text pane editor that *PolySNAP* uses for this file.

2.9. Incorporation of sample preparation details into the three-dimensional plots

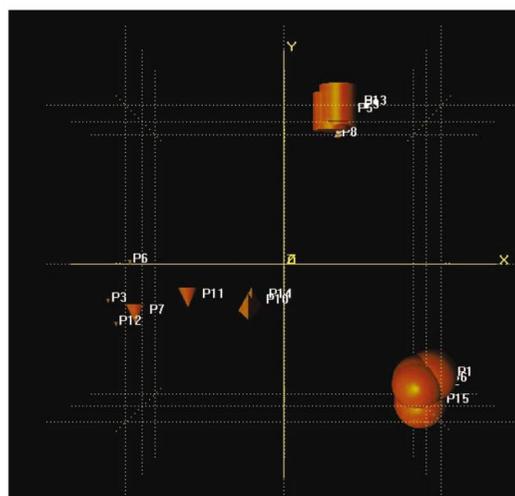
PolySNAP provides the facility to allow the MMDS or three-dimensional PCA score plot to be augmented by up to three additional dimensions. These are used to represent information recorded about each sample regarding its method of preparation. Available information that can be incorporated is by default: sample mass, total volume, counterion, stirrer rate, sample presentation, solvent, anti-solvent, initial temperature, isolation temperature, cooling rate, heating rate, reaction time and antisolvent volume. Normally, the three-dimensional plots represent each sample by a sphere of fixed size and by colour dictated by the dendrogram; to incorporate the additional information, the shape, size and colour of these spheres are modified to represent the relevant property. Up to three properties can, therefore, be displayed simultaneously. In this way it is possible to determine if there is a relationship between a particular combination of sample preparation conditions and the resulting clusters.

The sample preparation information may be read from either the ASCII or Bruker raw data file header, as semicolon-delimited fields, or directly from a specified text file. The labels, format and order of the fields may be flexibly changed by means of editing a configuration file.

Dialog boxes provide the facility for mapping the conditions onto the sphere properties. Fig. 4 shows some examples of this process for artificial data. (The only real data sets we have available are commercially sensitive.) Fig. 4(a) shows the standard MMDS plot for a data set of 21 samples. Fig. 4(b) shows the equivalent plot modified to incorporate solvent type and reaction time. The shape of each point represents the three different solvents used and the size of each point represents reaction time: the larger the shape, the longer the time. In Fig. 4(c), colour represents reaction time and shape represents the solvent. It can be seen that the cluster positions map the sample preparation data.



(a)



(b)



(c)

Figure 4

The inclusion of sample preparation data into the MMDS plot. These are synthetically generated data. Part (a) shows the standard MMDS plot for a data set comprising 21 samples. Part (b) shows the equivalent plot modified to display some of the variables associated with sample preparation. The shape of each point represents the different solvents used and the size of each point represents reaction time: the larger the shape, the longer the time. (c) In this plot, colour now represents reaction time and shape represents the solvent. It can be seen that the cluster positions map the sample preparation data.

3. Program details

The program is written in a variety of languages: the user interface is written in Visual Basic, the pattern-matching code in C++, the graphics in C++ and OpenGL, while the cluster-analysis code is written in Fortran 95. The software runs on PCs using the Windows 2000 or XP operating systems. Graphics cards with OpenGL optimization are recommended. The graphics demands can be considerable when a large number of patterns are being displayed. Computer times are highly variable. Parts of the calculation associated with estimating cluster number are of order $O(n^3)$; pattern matching is of the order $O(n^2)$ and there is a factor of ten increase in this part of the calculation if optimal non-linear shifts are to be estimated; the remainder of the cluster analysis runs as n^2 . Typically, on a 2.4 GHz Intel Xeon-based computer with 512 MByte of RAM running Windows 2000 (SP4), the program processes 50 patterns in 10 s, 96 patterns in 20 s, 400 in 5 min and 1000 in 1 h 5 min.

We are currently exploring methods of incorporating Raman, DSC and melting-point data into the software, as well as exploiting multiprocessor computing environments. Initial results are highly encouraging. We are also investigating other aspects of classification, including silhouettes, minimum spanning trees and fuzzy sets (Barr, Dong, Gilmore & Faber, 2004).

There is an extensive manual in Adobe PDF format and a tutorial (Barr *et al.*, 2003). The software is available commercially from Bruker-AXS.

We wish to thank Bob Docherty, Chris Dallman, Richard Storey, Neil Feeder and Paul Higginson of Pharmaceutical Sciences, Pfizer Global R & D, UK.

References

- Barr, G., Dong, W. & Gilmore, C. J. (2003). *PolySNAP: a Computer Program for the Analysis of High-Throughput Powder Diffraction Data*, University of Glasgow. (See also <http://www.chem.gla.ac.uk/staff/chris/snap.html>.)
- Barr, G., Dong, W. & Gilmore, C. J. (2004). *J. Appl. Cryst.* **37**, 243–252.
- Barr, G., Dong, W., Gilmore, C. J. & Faber, J. (2004). *J. Appl. Cryst.* **37**, 635–642.
- Barr, G., Gilmore, C. J. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 665–668.
- Calinški, T. & Harabasz, J. (1974). *Commun. Stat.* **3**, 1–27.
- Gilmore, C. J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231–242.
- Goodman, L. A. & Kruskal, W. H. (1954). *J. Am. Stats. Assoc.* **49**, 732–764.
- Gordon, A. D. (1999). *Classification*. 2nd ed. Boca Raton: Chapman and Hall/CRC.
- Gower, J. C. (1966). *Biometrika*, **53**, 325–328.
- Graham, R. L. & Hell, P. (1985). *Annal. Hist. Comput.* **7**, 43–57.
- Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* **A47**, 655–685.
- Milligan, G. W. & Cooper, M. C. (1985). *Psychometrika*, **50**, 159–179.
- Nelder, J. A. & Mead, R. (1965). *Comput. J.* **7**, 308–313.
- Storey, R., Docherty, R., Higginson, P., Dallman, C., Gilmore, C., Barr, G. & Dong, W. (2004). *Crystallogr. Rev.* **10**, 45–56.