Journal of
Applied Crystallography
ISSN 0021-8898

Received 19 January 2004
Accepted 14 May 2004

SNAP-1D: a computer program for qualitative and quantitative powder diffraction pattern analysis using the full pattern profile

Gordon Barr,* Christopher J. Gilmore and Jonathan Paisley

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

SNAP-1D is a computer program for the qualitative and quantitative analysis of powder diffraction data using the full measured data set. As measures of similarity between patterns, non-parametric statistical tests based on Spearman's correlation coefficient and the Kolmogorov–Smirnov test are used. Traditional correlation coefficients based on the Pearson formalism are also employed. This combination, suitably weighted, gives a reliable measure of qualitative pattern similarity. The method can be extended to the quantitative analysis of mixtures by using the above methods in conjunction with singular value decomposition techniques. A full description of the theory with suitable examples has been published elsewhere [Gilmore *et al.* (2004). *J. Appl. Cryst.* **37**, 231–242]; here the focus is on the computer software itself. The program is commercially available, and runs on PCs under the Windows 2000 and XP operating systems with modest hardware requirements. An easy to use graphical interface is supplied.

1. Introduction

Pattern-matching software in X-ray powder diffraction patterns has, until recently, relied on simplified patterns in which the full diffraction profile is reduced to a set of the strongest peaks, which are usually further reduced to a *d*-spacing (or 2θ value) and the corresponding intensity (the *d–I* system). This simplified approach to the analysis of powder diffraction patterns has advantages primarily in computer storage requirements and the speed of the associated search algorithms, especially when handling very large databases. *SNAP-1D*, in contrast, is a computer program that employs every measured data point for both qualitative pattern matching ('What is most like a given pattern?') and quantitative calculations ('What are the components of this mixture?'). The theory and several examples have been published (Gilmore *et al.*, 2004), but we present here a detailed description of the software and its options.

2. Importing and pre-processing data

On opening the program, the user can either select an existing database or create a new one into which a set of patterns is incorporated. Data import and pre-processing proceeds as follows.

(*a*) Data are imported either as ASCII *xy* data (2θ, intensity) with comma or tab delimiters, CIF format (Hall *et al.*, 1991), MDI ASCII or in Bruker raw data format. CIF files are a preferred option and the entries are scanned for unit-cell information, cell contents, formula *etc.*, which can be examined later in the program. A platform-independent binary format is also employed for this data, being used internally in the associated software. The ASCII format can also be used to import other data types, such as IR or Raman data, which can, with modification, be used with this software.

(*b*) The intensity data are normalized.

(*c*) The pattern is interpolated or extrapolated if necessary to give increments of 0.02° in 2θ. Neville's algorithm is used (Press *et al.*, 1992). It is important that all patterns have the same constant data step size.

(*d*) Background removal is optional. When requested, local *n*th-order polynomial functions are fitted to the data and then subtracted to remove the background. Three independent 2θ domains are usually defined, but this can be modified for difficult cases.

(*e*) Background removal is followed by the optional smoothing of the data using wavelets *via* the SURE (Stein's Unbiased Risk Estimate) thresholding procedure (Donoho & Johnstone, 1995).

(*f*) Peak positions are also optionally found using Savitsky–Golay filtering (Savitzky & Golay, 1964). Only two of the four matching techniques use the peak positions; if these tests are not used, peak positions are not needed.

Fig. 1 shows the Pattern Editor window for *SNAP-1D* in which all these facilities are used. All the options described above are set in this window. Processing may be applied to all patterns in a database at once, or individually as required.

2.1. Qualitative pattern matching

The sample pattern to be matched against the database is selected, pre-processed as necessary and then compared automatically in turn to each of the database patterns, data point by data point. For each sample pattern, a comparison is made as follows.

(i) The intersecting 2θ range of the two data sets is calculated, and each of the pattern-matching tests is performed using only that region.

(ii) A minimum intensity is set, below which profile data are set to zero. This eliminates noise and does not reduce the discriminating

DOI: 10.1107/S0021889804011847

power of the method. By default, this is set to $0.1I_{\max}$, where $I_{\max}$ is the maximum measured intensity.

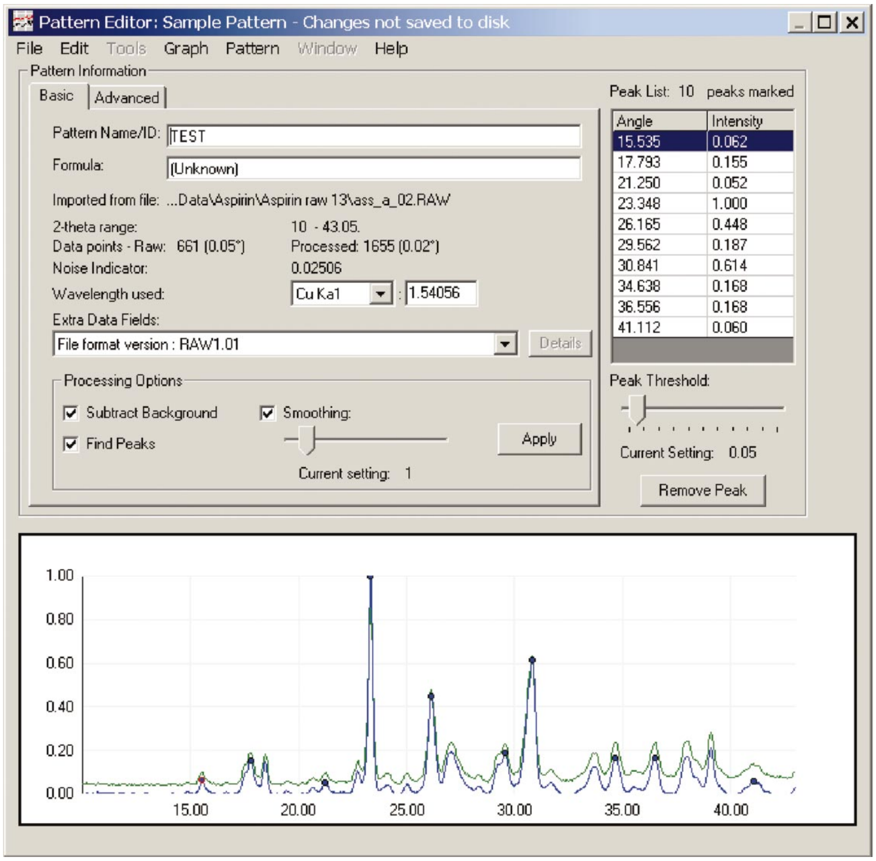
(iii) The full profiles of the patterns are compared on a point-by-point basis using the non-parametric Spearman rank-order coefficient test (Spearman, 1904; Conover, 1998). A score of 1.0 represents a perfect match, 0.0 a zero match, and −1.0 an anti-correlation (which is highly unusual).

(iv) A parametric Pearson equivalent of the Spearman test is then applied, as in (iii) above, again to all intersecting data points.

(v) If any peaks have been marked in either the sample or a particular database pattern and have the same value of $2\theta_{\max}$ within a user-specified tolerance, the correlation between the two peaks and its associated probability is calculated using the Kolmogorov–Smirnov (KS) test (Smirnov, 1939; Steck & Smirnov, 1969; Conover, 1998). The range of each peak to be tested is taken to be the intersection of the two peak ranges, calculated by tracing their shoulders until either the intensity falls below a set threshold, or the intensity of either starts to increase. The pattern with the greater number of peaks is taken as a reference. The KS test is then performed on each of these peaks and an associated probability, p_i, is returned for each. This has a value of 1.0 when the peaks are identical and zero when a peak is matched against no peak. The overall KS value, $p_{\rm KS}$, is

$$p_{\rm KS} = \sum_{i=1}^{m} p_i \Big/ m \qquad (1)$$

for m peaks in the reference sample; $p_{\rm KS}$ takes the values $0 \le p_{\rm KS} \le 1.0$.



Pattern Information

Pattern Name/ID: TEST
Formula: (Unknown)
Imported from file: ...Data\Aspirin\Aspirin raw 13\ass_a_02.RAW
2-theta range: 10 - 43.05.
Data points - Raw: 661 (0.05°) Processed: 1655 (0.02°)
Noise Indicator: 0.02506
Wavelength used: Cu Ka1 : 1.54056
Extra Data Fields: File format version : RAW1.01

Processing Options
☑ Subtract Background
☑ Find Peaks
☑ Smoothing:
Current setting: 1

Peak List: 10 peaks marked

Angle	Intensity
15.535	0.062
17.793	0.155
21.250	0.052
23.348	1.000
26.165	0.448
29.562	0.187
30.841	0.614
34.638	0.168
36.556	0.168
41.112	0.060

Peak Threshold:
Current Setting: 0.05

Figure 1
The Pattern Editor window in *SNAP-1D*. The options to subtract the background, find the peaks, set the peak level and smooth the data using wavelets are all set here. If CIF or raw files are used as the data source, extra data fields can be examined. The Advanced tab allows the input of unit-cell dimensions and contents for quantitative analysis to obtain the weight percentage. Multiple excluded regions in 2θ can also be defined here.

(vi) The parametric equivalent of equation (1) is also computed in the same way except that the Pearson correlation coefficient is used instead of the non-parametric KS test.

(vii) Finally, a rank value, $r_{\rm w}$, comprising a weighted mean of each of the available statistics, is calculated for each sample. These weights are user-definable and default to equal weighting for the Spearman and Pearson tests and zero for the KS test and its parametric equivalent.

(viii) An optimal shift in 2θ between patterns is often required, arising from equipment settings, sample preparation and data collection protocols. *SNAP-1D* provides three possible corrections, although these by no means encompasses all the possible correction geometries that can arise. These take the form

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

which corrects for varying sample heights in reflection mode, or

$$\Delta(2\theta) = a_0 + a_1 \sin\theta, \qquad (3)$$

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

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

which provides transparency and thick-specimen error corrections. The parameters a_0 and a_1 are constants that can be determined by maximizing the pattern–pattern correlation. It is difficult to obtain analytic expressions for the derivatives $\partial a_0/\partial r_{\rm w}$ and $\partial a_1/\partial r_{\rm w}$ for use in the optimization, so we use the downhill simplex method (Nelder & Mead, 1965), which does not require the calculation of derivatives.

(ix) It is possible to define multiple 2θ regions that are excluded from the calculations.

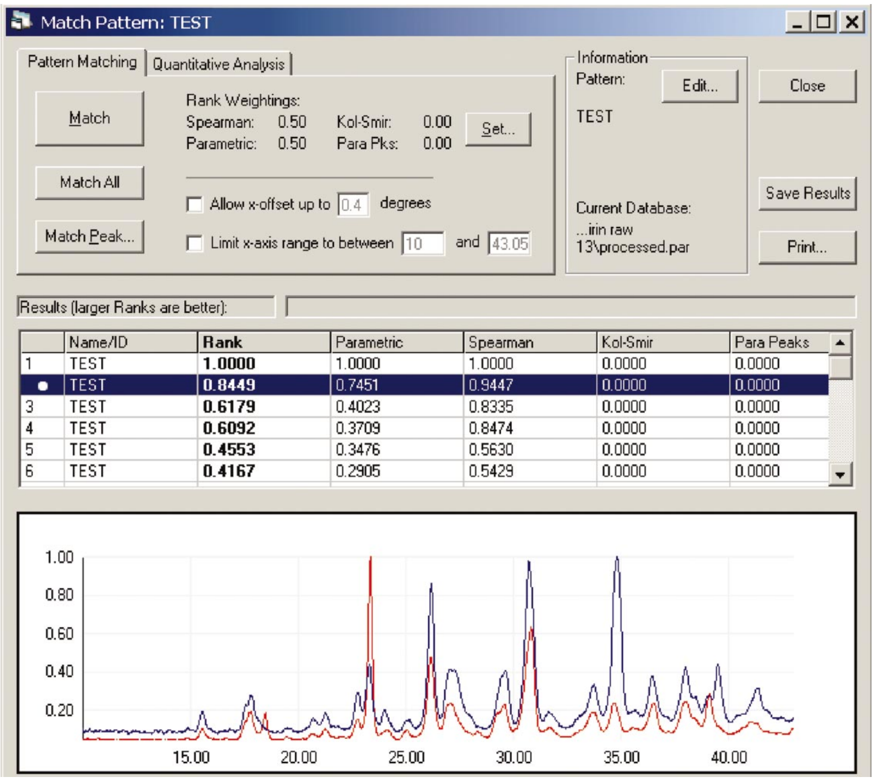
Fig. 2 shows a typical window display for qualitative pattern matching.

2.2. Generating a correlation matrix

Instead of selecting a single pattern and matching it against every entry in the database, it is possible to match every pattern against every other. If there are n patterns, this generates a symmetric ($n \times n$) correlation matrix, which can be exported to other statistics packages, *e.g.* for principal-component analysis, cluster analysis, *etc.* The use of the correlation matrix forms the basis of the *PolySNAP* computer program, which is discussed elsewhere (Barr *et al.*, 2004).

3. Quantitative analysis

If patterns corresponding to all pure phases in the mixture are present in an associated database, quantitative analysis can be carried out. The method used is an alternative to Rietveld refinement (*e.g.* Hill & Howard, 1987) and other methods. The Rietveld approach requires crystal structures to be known for all individual phases in the mixture; this approach does not require knowledge of the atomic coordinates in the unit cell or data of great accuracy; it is, however, less accurate.



Rank Weightings:
Spearman: 0.50 Kol-Smir: 0.00
Parametric: 0.50 Para Pks: 0.00

□ Allow x-offset up to 0.4 degrees

□ Limit x-axis range to between 10 and 43.05

Pattern: TEST

Current Database:
...irin raw
13\processed.par

Results (larger Ranks are better):

	Name/ID	Rank	Parametric	Spearman	Kol-Smir	Para Peaks
1	TEST	**1.0000**	1.0000	1.0000	0.0000	0.0000
●	TEST	**0.8449**	0.7451	0.9447	0.0000	0.0000
3	TEST	**0.6179**	0.4023	0.8335	0.0000	0.0000
4	TEST	**0.6092**	0.3709	0.8474	0.0000	0.0000
5	TEST	**0.4553**	0.3476	0.5630	0.0000	0.0000
6	TEST	**0.4167**	0.2905	0.5429	0.0000	0.0000

Figure 2
The Qualitative Analysis window. The patterns are sorted in descending r_w value and listed in the column labelled Rank. Patterns 1 and 2 are superimposed in the graphics pane. The individual correlation coefficients are in the next four columns. Only the Spearman and Pearson coefficients were calculated for this data set. The calculation of optimal 2θ offsets can be initiated here, and the maximum value specified. The 2θ ranges can also be set. The Quantitative Analysis tab opens the window shown in Fig. 3.

The method has been fully described by Gilmore *et al.* (2004) and employs full-matrix least squares with every measured data point, with singular value decomposition (SVD) (Press *et al.*, 1992) for the matrix inversion procedure. A brief summary, however, may be useful.

Assume we have a sample pattern, S, which is considered to be a mixture of up to N components. S comprises m data points, $S_1, S_2, \ldots, S_m$. The N patterns can be considered to make up fractions $p_1, p_2, \ldots, p_N$ of the sample pattern. The required equation to solve for p_i takes the form

$$\begin{pmatrix} x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mN} \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_N \end{pmatrix} = \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_N \end{pmatrix} \tag{5}$$

where x_{ij} is the ith measured data point for the jth pattern. Writing equation (5) in matrix form,

$$\mathbf{x} \cdot \mathbf{p} = \mathbf{S}. \tag{6}$$

The SVD methods allow $\mathbf{x}$ to be decomposed into three smaller matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$, and gives the solution

$$\mathbf{p} = \mathbf{V} \cdot \mathrm{diag}\,(1/w_j) \cdot \mathbf{U}^T \cdot \mathbf{S}. \tag{7}$$

$\mathbf{W}$ is a diagonal matrix with positive or zero elements. We accept the top min(15,N) values of p components of the mixture ranked on r_w. We also examine the elements of $\mathbf{W}$ and exclude any contributors with small values, and build a new matrix $\mathbf{p}$, thus repeating the entire procedure. Finally, the top j patterns (where j is an integer, $1 \le j \le 15$) are processed *via* the matrix decomposition once more. The results returned are the fractions of each pattern included in the mixture pattern. These are scaled to a percentage, and the number of possible phases is limited to j. The composition is normally displayed as a scale percentage, *i.e.* the percentage of the mixture pattern accounted for by each individual phase. If the unit-cell dimensions and contents for each component are available, the program converts this scale percentage to a weight percentage (Leroux *et al.*, 1953). The estimated error is also reported for each component. Additional feedback on the reliability of the results is given by these estimated errors, and by how good the matching results of the Spearman, parametric and KS tests are for each phase. Occasionally, if an incorrect pattern has been suggested by the program, this may be indicated by abnormally low values of the Spearman, Pearson and KS tests, and such patterns can be marked as ignored during subsequent runs of the procedure.

One drawback with the SVD procedure is that, because of its power and stability, it is almost always possible to decompose a matrix. This can mean that in some situations, for example, if the actual phases contained within a mixture are not present in the database, the method will give an incorrect solution. In these cases there are several signs available to warn the user to be cautious of an answer, such as abnormally high error values on the reported fractions, and/or large residuals when comparing the mixture and simulated patterns.

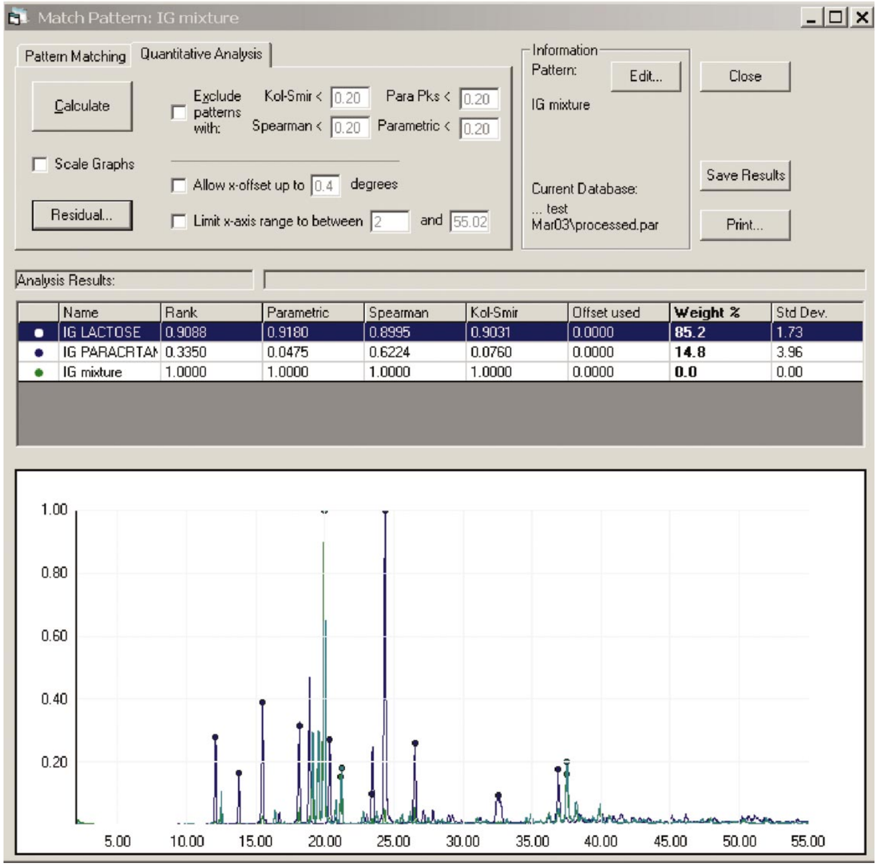
Other options for quantitative analysis are also available, as follow.

(*a*) Offsets. A 2θ offset can be refined to optimize r_w as described in §2.

(*b*) Residuals. To see if the suggested results are correct, or if they include a pattern not present in the mixture, or if they miss a phase that should be present, the Residual window constructs a calculated pattern made up from the individual patterns suggested as mixture components, in the proportions calculated. A difference plot between this and the sample pattern is available. The simulated mixture pattern can be saved as an ASCII text file, as can the difference plot. Fig. 4 shows the Residual window corresponding to Fig. 3.

(*c*) Automatic missing-phase detection. The program examines the results from the analysis and, using an algorithm based on the calculated error and the residual, can suggest if the resulting composition does not account sufficiently for all of the unknown pattern intensity. This can occur, for example, when not all of the phases present in a mixture pattern were in the pure-phase database. The quantitative analysis is then re-run to include a simulated pattern of the missing fraction as a known phase.

(*d*) Pattern exclusion. It can be useful to narrow down the number of patterns to be considered as components of the mixture. This is done by excluding patterns that are below user-set thresholds on any of the correlation coefficients included in the quantitative calculation. Generally, the best approach is to perform an initial standard analysis with defaults, and see if any poorly matching patterns have been included. The results from this will then give a feel for suitable cut-off values, and the analysis can be re-run.



Match Pattern: IG mixture

□ Exclude patterns with: Kol-Smir < 0.20 Para Pks < 0.20 Spearman < 0.20 Parametric < 0.20

□ Scale Graphs

□ Allow x-offset up to 0.4 degrees

□ Limit x-axis range to between 2 and 55.02

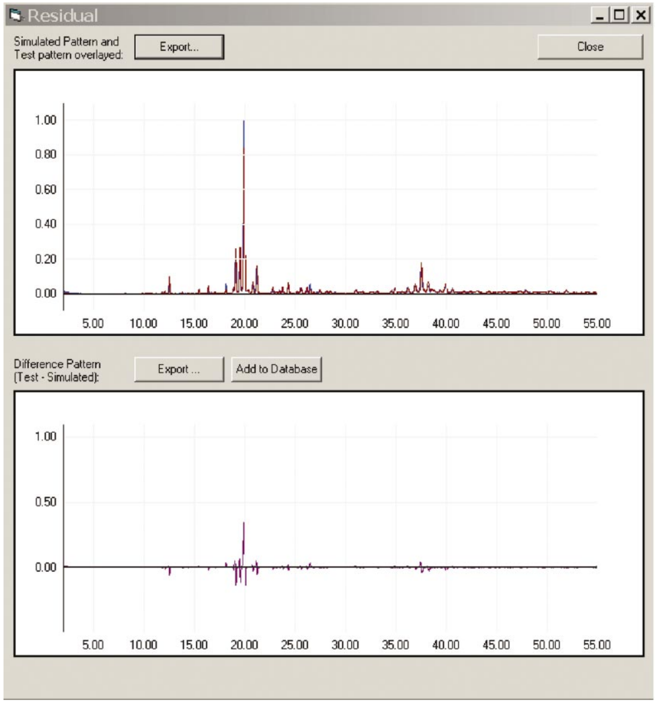
Information
Pattern: IG mixture

Current Database:
... test
Mar03\processed.par

Analysis Results:

	Name	Rank	Parametric	Spearman	Kol-Smir	Offset used	**Weight %**	Std Dev.
○	IG LACTOSE	0.9088	0.9180	0.8995	0.9031	0.0000	**85.2**	1.73
●	IG PARACRTAI	0.3350	0.0475	0.6224	0.0760	0.0000	**14.8**	3.96
●	IG mixture	1.0000	1.0000	1.0000	1.0000	0.0000	**0.0**	0.00

Figure 3
The Quantitative Analysis window. The mixture comprises lactose as entry 1 and paracetamol as entry 2. The weight percentages are 85.2 and 14.8%, respectively, with estimated errors of 1.7 and 4.0%. Just as in the Qualitative Analysis window, the calculation of optimal 2θ offsets can be initiated here, the maximum shift specified, and the 2θ ranges set.



Residual

Figure 4
The residuals following the quantitative analysis displayed in Fig. 3. The component patterns are superimposed in the upper pane to give a resultant, while the residual intensity is plotted in the lower pane. Both of these can be exported as ASCII files and re-imported into *SNAP-1D* or other software.

(*e*) Limiting the 2θ range. It is also possible to limit the analysis to subsets of the 2θ range of the unknown pattern. This can be useful if a particular feature of the pattern is causing problems, *e.g.* the presence of standards.

(*f*) Ignoring patterns. If a particular pattern included in the list of suggested results is known to be incorrect, it can be excluded from the calculation. It is possible to mark multiple patterns in this way. One can also ignore all patterns except those in a selected list if knowledge of the component phases is available.

A typical output display window is shown in Fig. 3.

4. Program details

The program is written in a mixture of C++ and Visual Basic. It runs on a PC using the Windows XP SP1 or Windows 2000 SP2 operating system or better. A minimal system requires a P4 processor (or AMD equivalent) operating at above 1 GHz, and 128 MByte of memory. Graphics and disk-storage needs are modest. There is a complete on-line and printed manual, tutorial and test data.

Up to 1000 patterns can be imported. In general, the manipulation of 100 patterns takes a matter of seconds, and matching 1000 patterns takes less than 1 min on a PC with a 2.0 GHz processor and 256 MByte of memory. These timings increase by a factor of ten if optimal 2θ shifts are calculated.

The program is available commercially from Bruker-AXS.

The authors would like to thank the Ford Motor Company, Detroit, for funding this work, and especially Charlotte Lowe-Ma whose input and support have been invaluable.

References

Barr, G., Dong, W. & Gilmore, C. J. (2004). *J. Appl. Cryst.* **37**, 243–252.
Barr, G., Gilmore, C. J. & Paisley, J. (2003). *SNAP-1D: Systematic Non-Parametric Analysis of Patterns – a Computer Program to Perform Full-Profile Qualitative and Quantitative Analysis of Powder Diffraction Patterns*, University of Glasgow. (See also http://www.chem.gla.ac.uk/staff/chris/snap.html.)
Conover, W. J. (1998). *Practical Nonparametric Statistics.* 3rd ed. New York: John Wiley.
Donoho, D. L. & Johnstone, I. M. (1995). *J. Am. Stat. Assoc.* **90**, 1200–1224.
Gilmore, C. J., Barr, G. & Paisley, J. (2004). *J. Appl. Cryst.* **37**, 231–242.
Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.
Hill, R. J. & Howard, C. J. (1987). *J. Appl. Cryst.* **20**, 467–474.
Leroux, J., Lennox, D. H. & Kay, K. (1953). *Anal. Chem.* **25**, 740–743.
Nelder, J. A. & Mead, R. (1965). *Comput. J.* **7**, 308–313.
Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C.* Cambridge University Press.
Savitzky, A. & Golay, M. J. E. (1964). *Anal. Chem.* **36**, 1627–1639.
Smirnov, N. V. (1939). *Bull. Moscow Univ.* **2**, 3–16.
Spearman, C. (1904). *Am. J. Psychol.* **15**, 72–101.
Steck, G. P. & Smirnov, G. N. (1969). *Annal. Math. Stat.* **40**, 1449–1466.