Journal of
Applied Crystallography

ISSN 0021-8898

Received 26 May 2005
Accepted 4 July 2005

dSNAP: a computer program to cluster and classify Cambridge Structural Database searches

Gordon Barr, Wei Dong, Christopher J. Gilmore,* Andrew Parkin and Chick C. Wilson

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

A computer program that automatically classifies and clusters structural fragments extracted from mining the Cambridge Structural Database is described. The methodology is based on cluster analysis and multivariate data processing of distance matrix information describing the extracted fragments. Coupled with the calculations is a set of visualization tools that enable the user to view and verify the proposed classification scheme, and further explore it in varying levels of detail. Two examples are presented: the first is based on a simple difluoroalkene fragment and the second, more complex, on a chiral vicinal dialcohol, R_1(OH)CHCH(OH)R_2.

1. Introduction: the problem

The continued increase in the number of entries in the Cambridge Structural Database (CSD; Allen, 2002) poses problems: it is an enormously powerful resource, but with more than 325000 entries, extracting meaningful chemical information can be daunting. There are numerous search algorithms and data-mining strategies in the literature (see, for example, Orpen, 2002; Allen & Motherwell, 2002; Taylor, 2002), but there is a need for a freely available, general purpose program for classifying the results of database searches. In previous papers (Barr *et al.*, 2004*a*,*b*,*c*,*d*), we have shown how pattern matching coupled with cluster analysis and multivariate statistical methods can be used to classify large numbers of powder patterns, for example to identify salts and polymorphs, or to analyze round-robin type data where there are multiple data sets relating to the same sample. Here we show how a subset of these techniques with appropriate extensions are incorporated into a new computer program, *dSNAP*, that can assist the structural chemist in extracting information from CSD searches.

2. The method

In this section we briefly describe the statistical techniques used. They are an extension and development of ideas originally proposed by Allen & Taylor (1991) and Taylor & Allen (1994), which are described more fully by Barr *et al.* (2004*a*,*b*,*c*,*d*).

Nomenclature: in this paper the terms 'structure' and 'hit' refer to an individual crystal structure determination mined from the CSD containing the defined search query and labelled by an individual CSD reference code (refcode), while the term 'fragment' refers to the part of the structure matching the search query. Thus a structure can contain one or more fragments; it is the geometry of the fragment that is defined and used to generate the input data matrix. In the case where more than one fragment is observed in a structure, these fragments frequently lie in chemically distinct environments and thus may exhibit significantly different geometries, and they are treated independently. In such circumstances the software will suffix the individual refcode with an integer consistent with the order in which the fragment is output; this allows simple reference back to the fragments when studying the structures within the *ConQuest* search program (Bruno *et al.*, 2002).

The procedure operates as follows.

(i) The search fragment or fragments (it is possible to use more than one) are defined and a scan of the CSD is initiated using *ConQuest* (Bruno *et al.*, 2002). For every fragment, all the interatomic distances (bonded and non-bonded, and not just those involving chemical bonds or nearest-neighbour contacts) and angles are defined using the three-dimensional options of *ConQuest*. The number of geometric parameters is thus equal to $(l/2)(l-1) + (l/2)(l-1)(l-2) = l/2(l-1)^2$, where l is the number of atoms. There is obviously redundancy here, but clustering algorithms are known to work effectively with such data, and this use of distances and angles together has proved to be optimal in practice. It is possible to include torsion angles as well, but, to date, there seems to be little benefit in doing so: the geometry is sufficiently characterized by the distances and angles. The search results are exported to *Vista* (CCDC, 1994) or Microsoft *Excel* along with the relevant *ConQuest* file. We now have a data matrix, $\mathbf{x}$, comprising n rows (where n is the number of fragments) and m columns (where m is the number of geometric parameters).

(ii) The data matrix is input into *dSNAP* along with the corresponding *ConQuest* file. Both files must be present. The data matrix is converted to a symmetric $(n \times n)$ Minkowski distance matrix, $\mathbf{d}^{\mathrm{s}}$, *via* the standard formula:

$$d_{ij}^{\mathrm{s}} = \left(\sum_{k=1}^{m} w_k \left| x_{ik} - x_{jk} \right|^{\lambda} \right)^{1/\lambda}. \qquad (1)$$

λ is a user-selectable parameter in *dSNAP* with a default value of 2 corresponding to a Euclidean distance matrix. A value of 1 is sometimes used and this corresponds to the traditional city-block distance. The superscript 's' for the $\mathbf{d}$ matrix implies that we are working in the subject or stimulus space to distinguish it from the related variables space which is discussed in §2.1. The parameters w are weights; each of the columns, corresponding to distances *etc.* in the original data matrix, can optionally be given a weight corre-

sponding to its importance. Weights are user-selectable parameters with the default of unity. The $\mathbf{d}^s$ matrix [equation (1)] is standardized by dividing each variable by its sample range giving $0 \leq d_{ij}^s \leq 1.0$ and $d_{ii} = 0.0$.

(iii) Metric multidimensional scaling (MMDS) is used to generate a three-dimensional Euclidean space in which each point represents one fragment in the data set. Given $\mathbf{d}^s$, a matrix $\mathbf{A}(n \times n)$ is constructed as

$$\mathbf{A} = -\frac{1}{2}\left(\mathbf{I}_n - \frac{1}{n}\mathbf{i}_n\mathbf{i}_n'\right)\mathbf{D}^s\left(\mathbf{I}_n - \frac{1}{n}\mathbf{i}_n\mathbf{i}_n'\right), \tag{2}$$

where $\mathbf{I}_n$ is an $(n \times n)$ identity matrix, $\mathbf{i}_n$ is an $(n \times 1)$ vector of unities and $\mathbf{D}^s$ is a matrix of squared distances. The eigenvectors of $\mathbf{A}$, v_1, $v_2 \ldots v_n$, form a vector $\mathbf{V}$, and the corresponding eigenvalues λ_1, $\lambda_2 \ldots \lambda_n$ give a second vector $\mathbf{\Lambda}$. A total of p eigenvalues are positive and the remaining $(n - p)$ are set to zero. A set of coordinates in p dimensions can be defined *via* the matrix $\mathbf{X}(n \times p)$,

$$\mathbf{X} = \mathbf{V}\mathbf{\Lambda}^{1/2}. \tag{3}$$

We now set $p = 3$ and work in three dimensions; the $\mathbf{X}$ matrix can be used to plot each data set fragment as a single point in a three-dimensional plot. This assumes of course that we can reduce the dimensionality of the problem in this way and still retain the essential features of the data. *dSNAP* offers two checks of this assumption.

(*a*) We compute a distance matrix from $\mathbf{X}(n \times 3)$ and compare it element by element with the observed matrix $\mathbf{d}^s$ using a mean of the Pearson and Spearman correlation coefficients. There are occasions where the underlying dimensionality of the data is one or two, and in these circumstances the data project onto a plane or a line in an obvious way without problems.

(*b*) An additional check on the integrity of clusters in higher dimensions is also supplied *via* an interface to the *CrystalVision* program of Wegman (Wegman, 2005; Wilhelm *et al.*, 1993) which can be downloaded free of charge from http://www.galaxy.gmu.edu (follow the links to the ftp server at this URL and select the software option). This program provides a continuous random sequence of projections from n dimensions into two (or more) with an interactive environment for exploring multivariate data. In the interface offered by *dSNAP* we have generated output for a maximum of six dimensions to be explored and it is possible to investigate the viability of the clusters using a set of 15 two-dimensional scatter plots or using six parallel coordinate plots.

(iv) Using $\mathbf{d}^s$, we also carry out agglomerative, hierarchical cluster analysis to put the fragments into classes or groups. The results are presented as a dendrogram (see, for example, Everitt *et al.*, 2001) examples of which are shown in Fig. 3(*b*). Each fragment begins at the bottom of the plot as a separate class, and these amalgamate in stepwise fashion linked by horizontal tie bars. The number of clusters defines the cut level which is represented by the solid horizontal line in the figure. Two estimates are generated: one from eigenvalue analysis carried out on the $\mathbf{A}$ matrix from the MMDS calculation, and the other from eigenanalysis of the correlation matrix, $\boldsymbol{\rho}^s$, corresponding to $\mathbf{d}^s$:

$$\boldsymbol{\rho}^s = 2\mathbf{d}^s - \mathbf{I} \tag{4}$$

(where $\mathbf{I}$ is the identity matrix). In each case the eigenvalues of the relevant matrix are sorted in descending order, and when 95% of the data variability has been accounted for, the number of eigenvalues is selected and this is used to define the number of clusters. The two values are averaged.

The program also supplies options for the generation of five possible dendrogram types: single link, complete link, weighted average link, centroid and group average link. The latter is used as a default and seems to give the most consistent results.

(v) For each cluster with three or more members, the most representative sample (MRS) is identified and highlighted in the MMDS plot. The MRS is defined as that member that has the minimum average distance from every other member of the cluster, *i.e.* for cluster $\mathbf{J}$ containing m patterns, the most representative sample, i, is defined as that which gives

$$\min\left[\sum_{\substack{j=1\\ i,j\in J}}^{m} d(i,j)/m\right]. \tag{5}$$

(vi) Semi-independent cluster validation tools are also employed using silhouettes (Rousseeuw, 1987; Barr *et al.*, 2004*c*). The concept is simple. Define the dissimilarity coefficient δ_{ij},

$$\delta_{ij} = d_{ij}/d_{ij}^{\max}. \tag{6}$$

If the fragment i belongs to cluster C_r which contains n_r structures, define

$$a_i = \sum_{\substack{j\in C_r\\ j\neq i}} \delta_{ij}/(n_r - 1) \tag{7}$$

and

$$b_i = \min_{s\neq r}\left(\sum_{j\in C_s} \delta_{ij}/n_s\right). \tag{8}$$

The silhouette, h_i, for fragment i is then

$$h_i = \frac{b_i - a_i}{\max(a_i,\, b_i)}. \tag{9}$$

Silhouette values are assigned to all members of a cluster, and give an estimate of membership. Clearly $-1 \leq h_i \leq 1.0$. Each cluster is displayed as a histogram frequency plotted against silhouette values and outliers are clearly identified. Clusters should have tight silhouettes with few or no outliers and values greater than 0. Ideal clusters have $h_i \geq 0.5\ \forall\ i$.

(vii) Scree plots are also used as a validation tool. Principal-components analysis of the $\mathbf{A}$ matrix yields a set of sorted eigenvalues and these are presented in *dSNAP* in the form of a scree plot. The graph should exhibit a steep descent without highly variable gradient changes, and can be used to check both the quality of the cluster analysis, and the input matrix.

(viii) In the MMDS window each sphere represents a crystal structure fragment, and in the dendrogram each node is also a fragment. It is possible to view the associated structures in either *Mercury* or *ConQuest* by selecting one or more node/sphere and using the 'View Selected Hits' option in the 'Tools' menu, which launches the required program. This is shown in more detail in Fig. 3(*b*) and described in §3.

2.1. The variables space

So far we have been exploring the subject or stimulus space, but it is also possible with *dSNAP* to investigate the variables space, *i.e.* the underlying geometries, stored in the data matrix $\mathbf{x}$, which can be used to create a new distance matrix $\mathbf{d}^v$. This is carried out as follows.

Each column of $\mathbf{x}$ contains a distance or angle (or torsion angle) and there are m such columns; each column is subjected to linear regression with every other column in turn to generate an $(m \times m)$ symmetric correlation matrix $\boldsymbol{\rho}^v$ and the corresponding distance matrix $\mathbf{d}^v$ using equation (4). We can use these two matrices in the

same way as $\mathbf{d}^s$ to generate a dendrogram, an MMDS plot, silhouettes *etc*., and these results are viewed by accessing the 'Variables' tab on the left hand side of the main *dSNAP* window.

The results are, of course, interpreted differently; in the MMDS plot each sphere now represents a geometric feature such as a distance or angle, and variables that are clustered tightly together are those which are highly correlated. The same feature is exhibited by the dendrogram: entries with low tie bars are highly correlated.

2.2. Graphics facilities

Visualization plays a major role in the operation of *dSNAP*, and a number of useful facilities for manipulating and exploring the graphics windows are provided.

(*a*) A graphics toolbar is available that can be used to change the colours of the background, axes, labels *etc.* in the usual way.

(*b*) Any given structure or set of structures can be located *via* their refcode or sequence number in the $\mathbf{d}^s$ matrix.

(*c*) Multiple selection of structures is achieved by clicking on entries with the Ctrl key held down on the keyboard; individual structures can be de-selected in a similar manner. Alternatively, a continuous number of consecutively displayed structures may be selected by holding down the Shift key and clicking on the first and then the last structure in the desired range. The structures represented can be viewed using *ConQuest* or *Mercury* with a built-in interface invoked by the 'View Selected Hits' option on a pull-down menu. [Fig. 3(*b*) and §3 describe this in more detail.]

(*d*) In all graphics screens, any particular area of the view can be 'zoomed in' by dragging a rectangle over the relevant region.

(*e*) The plot position can be moved by holding down the Alt key and then moving the mouse in any direction as required.

(*f*) The drawing quality of the spheres can be altered if needed. With many points displayed, the graphics display will be much faster if the drawing quality is reduced. Higher quality is useful for publication purposes.

(*g*) In variables space there is an additional feature. In the 'Numerical Results' pane the correlation matrix and labels are shown with all correlations having $|\rho_{ij}^{v}| > 0.8$ highlighted. Selecting any matrix entry (except, of course, the unit diagonal) brings up a linear plot of the two variables in a separate window. This graphics pane also has an optional toolbar and can be modified and copied *etc.* Using the standard OpenGL options, the point sizes can be modified. By positioning the cursor on any point, the corresponding refcode is displayed. Fig. 4(*d*) shows an example; this will be discussed further in §3.

2.3. Visualizing large clusters

When there are several hundred or more structures to display, the graphics panes can become very crowded and facilities are provided to navigate them under these circumstances.

(*i*) Show Grid: the grid can be hidden or displayed.

(*j*) Navigation: it is possible to show each cluster individually and navigate through them with the cursor keys.

(*k*) Hide Group: this option is useful in a crowded display with overlapping clusters overlapping. A 'Hide Group Option', temporarily removes a given group from the display.

(*l*) Popup Group: this option is used to bring the whole of a selected cluster to the front of the display.

(*m*) Transparency options: all the groups can be displayed as transparent spheres. This is useful in the case of overlapping clusters. To complement this an opaque zone can be defined or only a single group can be rendered transparent. Alternatively all spheres can be rendered as dot surfaces. The options of variable sphere size still apply.

(*n*) Labels can be turned on or off. It is possible to drag labels.

(*o*) When there are several large clusters, it is often useful to select one cluster and repeat the cluster analysis on this cluster only and facilities are provided to do this. The effect of this is to re-normalize the distance matrix, and it enables the user to explore more subtle differences in the structures.

(*p*) The dendrogram can be simplified by displaying only the first, middle and last member of each cluster. The associated MMDS plot also has this feature.

(*q*) The spheres representing the structures in the three-dimensional plots (or parameters in the variables space) can be enlarged or shrunk to suit any required 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.

2.4. Incorporation of new structures

It often occurs that the user has determined the crystal structures of one or more new compounds and wishes to compare fragments from these with existing CSD entries. *dSNAP* provides such facilities. On starting the program a dialog box appears in which the user specifies the input and output file locations. There is also an entry for the location of additional data from new structures. This must be in the same form as the data file derived from the database search. The new structures are inserted at the top of the list in the $\mathbf{x}$ and $\mathbf{d}^s$ matrices and thus can be easily located in any graphics pane as described in §2.2.

2.5. Local chemical symmetry: pre-processing the data matrix

There is one further topic of great importance that needs to be discussed: pre-processing the data matrix if local chemical symmetry is present in the search fragment. Consider the fragment shown in Fig. 1(*a*) in which a transition metal (TM) is octahedrally bound to six N atoms. The structure has local four- (and two-, three-) fold symmetry.

Because of the way in which the nitrogen atoms have been defined, there are a large number of possible ways that these atoms can be numbered and the data entered into the CSD are not consistent in this respect. This is inevitable, since there are no constraints on the ways in which authors themselves number their atoms or present their results. However, we can apply some structural rules to order the fragment geometries appropriately. In this case the rules to be applied are such that N2 has the shortest N—TM1 bond; N2—TM1—N3 is the largest N2—TM1—N angle; N2—TM1—N4 is the smallest N2—TM1—N angle; N4—TM1—N5 is the largest N4—TM1—N angle; N2—TM1—N6 is shorter than N2—TM1—N7. Thus N3 is

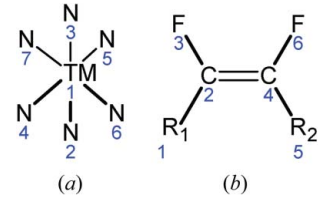


Figure 1
Examples of search fragments that have local chemical symmetry: (*a*) a transition metal octahedrally bound to six N atoms; (*b*) a difluoroalkene. These geometries and the constraints they generate are discussed in §2.4.

trans to N2, N5 is *trans* to N4 and N7 is *trans* to N6. In this way the input **x** matrix is optimally constructed for clustering.

As an alternative example, consider the *cis*-difluoroalkene search fragment shown in Fig. 1(*b*). It has local twofold rotational symmetry around the centre of the C═C double bond. In this example, one can stipulate that the C(4)—F(6) distance must be greater than the C(2)—F(3) bond length; if this is not the case then the bond pairs F(6)—C(4) and *R*(5)—C(4) are exchanged with C(2)—F(3) and C(2)—*R*(1).

A user interface is supplied to assist with this, and is shown in Fig. 2. The user is presented with a two-dimensional diagram of the search fragment, and is asked if local symmetry exists (Fig. 3*a*). If so, a dialog box allows the user to define the rules that must be applied in the building of the $\mathbf{d}^{s}$ matrix. The interface uses colours and line types to assist this process. Once is this is done the user clicks the 'OK' button and cluster analysis proceeds.

3. Examples

3.1. A simple example: difluoroalkenes, $FR_1C{=}CR_2F$

The operation of *dSNAP* is best illustrated by example. We present two here. The first is a simple case involving the $FR_1C{=}CR_2F$ fragment. The initial search, using *ConQuest*, resulted in 58 fragments from 33 crystal structures. The data matrix contains 15 distances and 60 angles for each fragment. At this point the *dSNAP* program begins with the dialog boxes concerned with local symmetry (Fig. 2). Once these are defined (and often they can be bypassed) the main *dSNAP* window appears as shown in Fig. 3(*a*). (For the computation time to reach this point see §4.) Across the top of the display window are tabs for cell, dendrogram and MMDS graphics displays. There is also a validation tab which offers the options of showing scree plots or silhouettes, and two text panes: the 'Numerical Results' pane shows the correlation matrix $\boldsymbol{\rho}^{s}$, and the 'View Logfile' tab gives a detailed output of the calculations. Any of the graphics panes can be detached and displayed in another window if required, so that more than one pane is visible at a time. This is especially valuable if two computer monitors are available.

In the initial display the clusters are represented as cells in which each cluster has a unique colour which is maintained in all graphical outputs. The cell display gives a simple, general overview of the clustering. The second tab gives the associated dendrogram and the third the MMDS plot as shown in Figs. 3(*b*) and 3(*c*). To the former have been added figures taken from *Mercury* that show the MRS structures and it can be seen that the clustering has an underlying, if broad, rationale. The colours in the dendrogram are the same as those in the MMDS and cell plots to facilitate comparison between the methods.

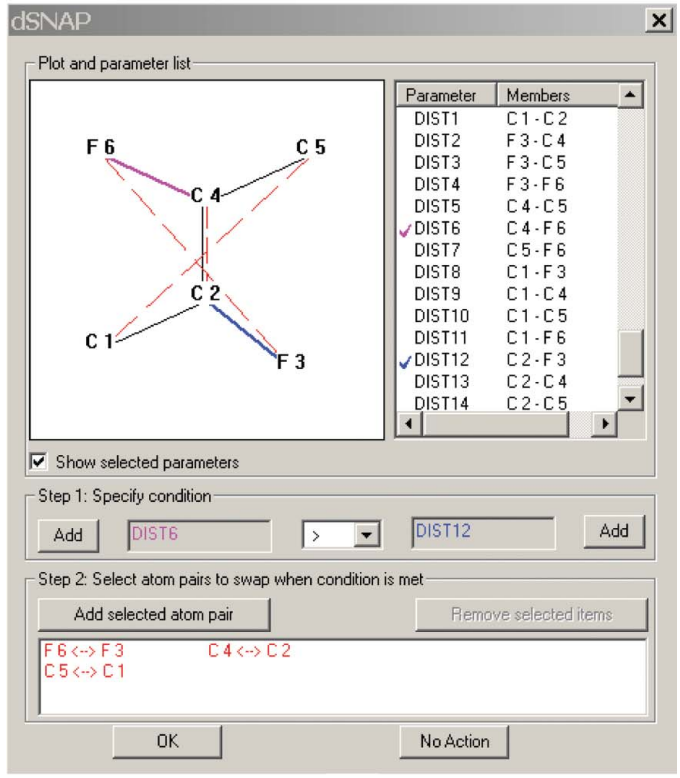
The program has generated five clusters (A–E) at a cut level calculated at 75% similarity. Running from left to right in Fig. 3(*b*): group A (red) contains all the *trans*-fluorine fragments, and groups B–E contain the *cis*-fluorine fragments. Group B (yellow) contains what might be considered the normal *cis*-fluorine structures, while groups C (green) and D (light blue) contain a total of four fragments (all from the structure with refcode XAFLIZ) from an aromatic carbon-cluster with F bound to the surface. These four fragments each contain a carbon atom which has four groups bound to it, one of which is a double bond, and are thus not chemically reasonable. Group E (dark blue) is a singleton cluster in which the search fragment is part of a cyclobutene ring and thus in a more strained chemical environment.

The corresponding MMDS plot confirms the appropriateness of the clustering: these five groups are clearly separated although the light blue and green coloured ones are, as expected, quite close. The large yellow group has a tightly bound cluster and some outliers. The validation tools are useful here. The silhouettes (Fig. 3*d*) for the red cluster are well defined with no outliers. The second cluster (yellow), however, has silhouettes < 0.5 and outliers at *ca* 1.0, which is indicative of a less than ideal cluster structure. (It should be noted that silhouettes are only computed for clusters with at least five members.) The scree plot is shown in Fig. 3(*e*), and has no sudden gradient changes with a steep fall to the 95% level which is indicative of good cluster definition.

The default classification is quite broad in this case, and we can increase the level of discrimination of *dSNAP* by manually lowering the dendrogram cut level as shown in Fig. 3(*f*). (It is possible to return to the original dendrogram cut level if the new clustering proves unsatisfactory.) We now have 16 clusters. The MMDS plot in Fig. 3(*g*) again confirms that, from the viewpoint of cluster analysis, this is an appropriate action. This is confirmed by the silhouettes in Fig. 3(*h*). Table 1 summarizes the characteristics of each cluster.

It is important to stress that *dSNAP* knows nothing about chemistry. It is, in some ways, that which makes it such a powerful tool: it is

(a)



(b)

Figure 2
The initial dialog boxes that question the user about local chemical symmetry. The interface allows the user to define the necessary rules to ensure optimal clustering.

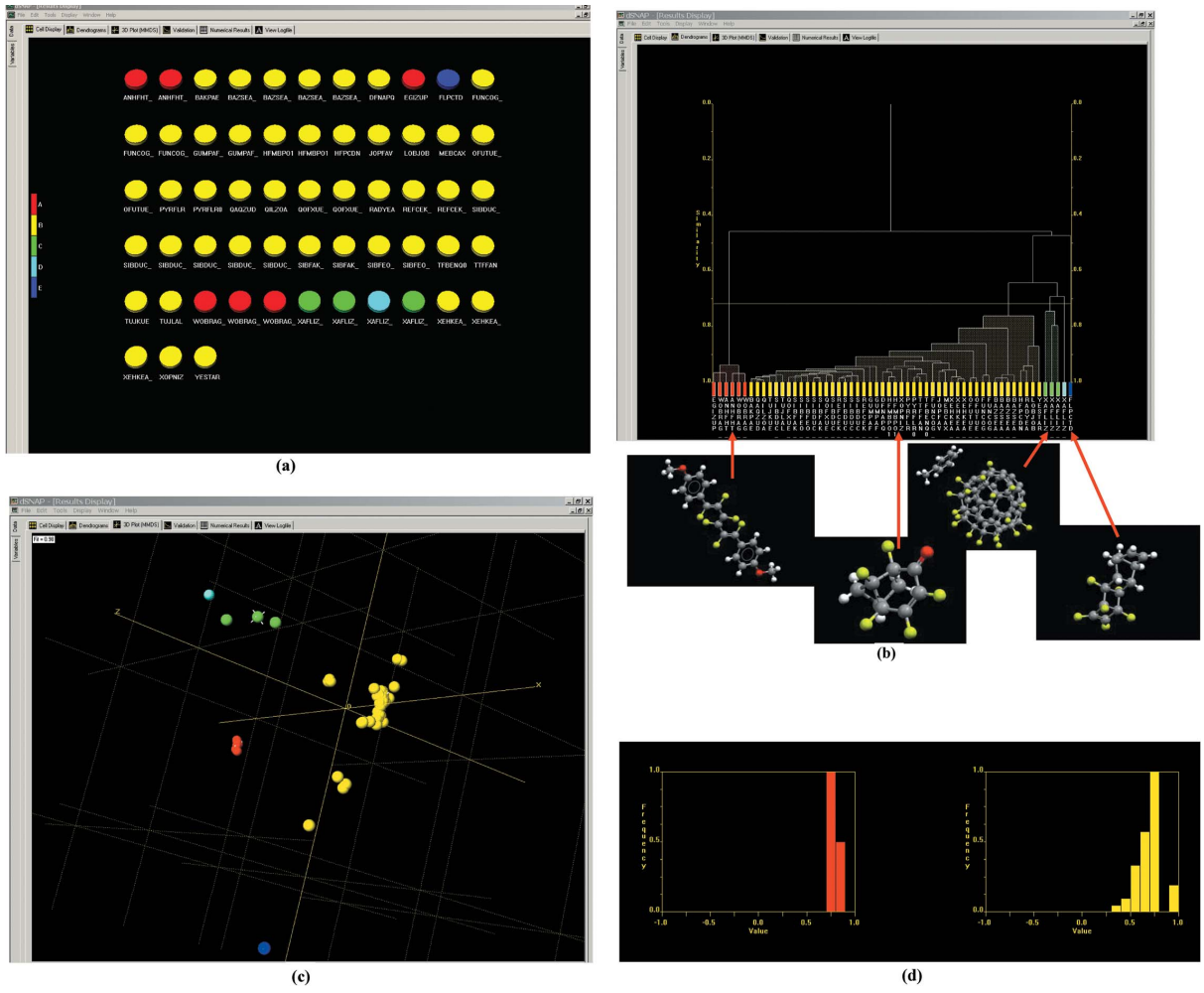


Figure 3
Results of a search based on the FR_1C═CR_2F fragment. The search resulted in 58 fragments from 33 hits. (*a*) The clusters represented as cells. Each cluster has a unique colour which is maintained in all graphical outputs unless altered by modifying the dendrogram cut level. (*b*) The default dendrogram. The *dSNAP* program has generated five clusters (A–E) at a cut level calculated at 75% similarity. The difference in the clustering of these groups can be related to differences in their structural chemistry, which is described in §3. Representative structures from four of these clusters, displayed by the *Mercury* program, are included. (*c*) The corresponding MMDS plot. The large yellow group has one tightly bound cluster with *ca* 8 outliers. (*d*) The silhouettes. The first (red) cluster is well defined with no outliers; the second cluster (yellow) has silhouettes <0.5 and outliers at *ca* 1.0. This is indicative of a less than ideal cluster structure. (*e*) The scree plot. There are no sudden gradient changes, which is indicative of good cluster definition. (*f*) The dendrogram level is lowered manually to produce 16 clusters. (*g*) The MMDS plot corresponding to (*f*). (*h*) Sample silhouettes corresponding to (*f*) and (*g*). The clusters are now well defined.

Table 1
Analysis of difluoroalkenes clustering according to the dendrogram in Fig. 3(*f*) with a rationale for the clustering.

Colour in Fig. 3(*f*,*g*)	Form	Description	No. of fragments
Red	*trans*	Sterically non-constrained; ideal geometry	3
Yellow	*trans*	Sterically constrained	3
Green	*cis*	All fragments of tetrafluoro-7,7,8,8-tetracyanoquinodimethane	21
Pale blue	*cis*	At least one three-coordinate carbon *R* group generally bound to an O atom	8
Dark blue	*cis*	Sterically constrained seven-membered ring	1
Pink	*cis*	Both *R* groups are four-coordinate	5
Orange, striped	*cis*	Resonant C═C bonds with partial double character	2
Yellow green, striped	*cis*	Sterically constrained B2	2
Green, striped	*cis*	Mislabeled bonds in the database	4
Blue, striped	*cis*	Bridged six-membered ring	3
Purple, striped	*cis*	Bridged six-membered ring	1
Next 4 entries	*cis*	Carbon cluster	4
Last entry	*cis*	Part of a cyclobutene ring	1

not biased in any way by chemical presupposition. However, it should be stressed that any explanation for differences in structure classification should be chemically reasonable; it is the responsibility of the user to verify the proposed clustering in terms of the underlying chemistry.

3.2. Variables space

We now explore the corresponding variables space by clicking the lower left hand 'Variables' tab in the main window. A new set of text and graphic panes appear that are very similar to the previous ones but contain quite different information.

(*a*) A well chart as before (Fig. 4*a*).

(*b*) A dendrogram with the default cut line (Fig. 4*b*). Each node is now a distance or angle entry in the original data matrix. As expected, the clusters are much more diffuse than in the subject space.

(*c*) The corresponding MMDS plot (Fig. 4*c*). Each sphere represents an angle or distance in the FR_1C═CR_2F fragment. Spheres which are close to each other are highly correlated variables. The

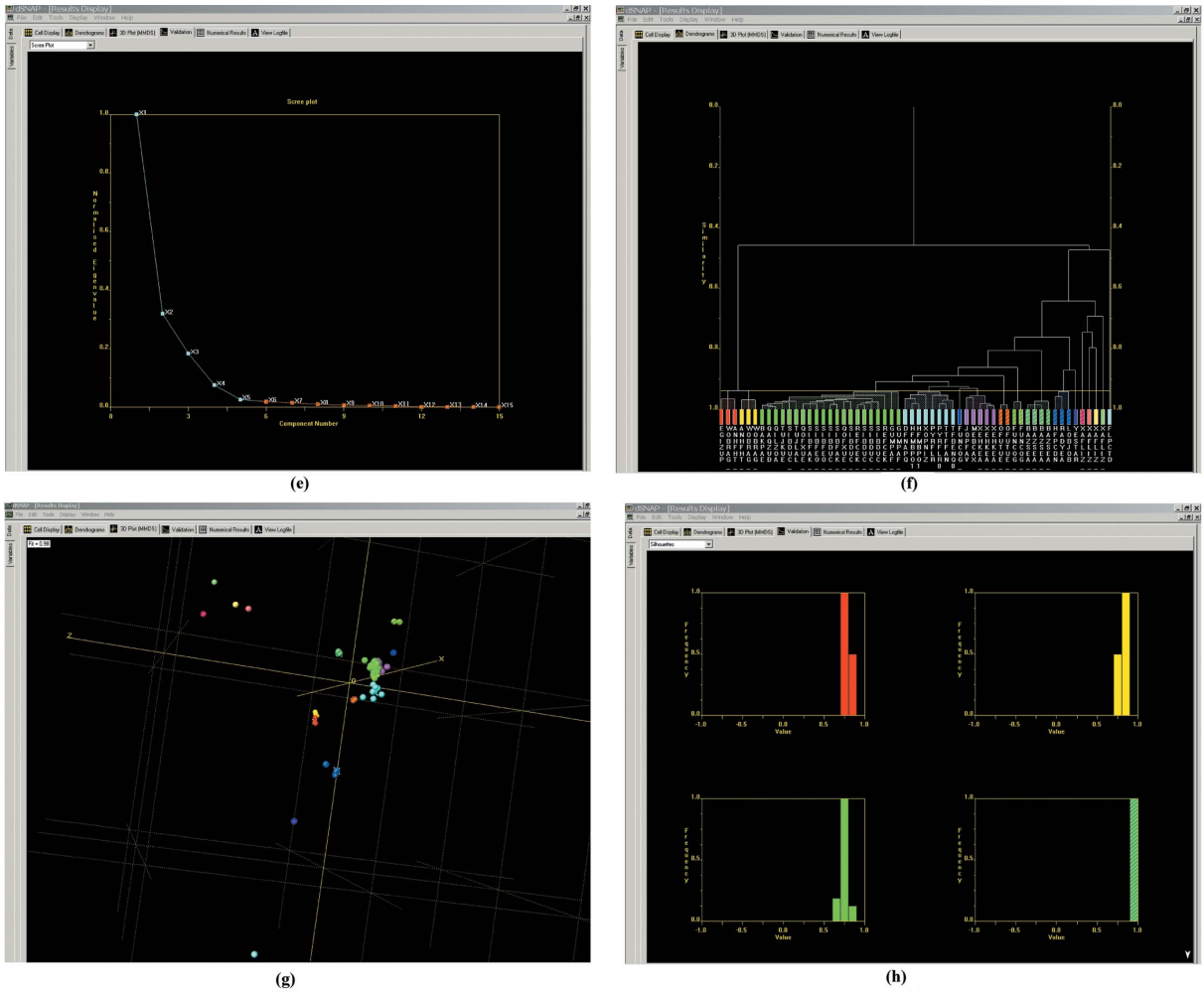


Figure 3 (continued)

easiest one to see in this context is the pair coloured purple at the bottom centre of the plot which correspond to angle 18 and angle 59 in the defined cluster. (It is possible to see which variables are represented by either switching on the 'Show Labels' option or placing the cursor over the sphere.)

(*d*) The variables correlation matrix. The scatter plot between any two variables is obtained by clicking the mouse on the relevant matrix entry. As an arbitrary example, the plot corresponding to angle 18 plotted against angle 59 is shown in Fig. 4(*d*) along with the correlation matrix. The correlation coefficient is 0.723. A straight line through the data can be plotted as an option. The labels associated with any points on the graph can be identified by placing the cursor over the relevant data point. Clicking the mouse on any point brings up a window with the corresponding structure displayed. Each point is given a colour taken from the current data space dendrogram.

3.3. A chiral, vicinal dialcohol, R_1(OH)CH—CH(OH)R_2

This search fragment, shown in Fig. 5(*a*), is a more complex example, and also shows the operation of the software when a large number of hits are found. The fragment not only exhibits free rotation around the central C—C bond, but also can contain up to two chiral centres. Because the geometry definition used as input to *dSNAP* includes no information from torsion angles, it is impossible to tell the hand of a particular molecule in a crystal structure. Since the search has not been restricted solely to crystals in chiral space groups containing molecules of known chirality, then this is entirely sensible. However, it is possible to tell the relationship between the two chiral centres and whether or not they have the same or a different hand. In general, for situations where chirality is important, the facilities of *ConQuest* need to be used to define a relevant search. The conformations that can be adopted by the search fragment are shown in Fig. 5(*b*). Of the staggered conformations shown here, forms 1 and 3 can only be adopted by structures with different-handed chiral centres, and conformers 2, 4 and 5 are structures with chiral centres of the same handedness. The same is true for eclipsed structures, with only conformers 1–1 and 1–3 possible for centres with different chiralities, and conformations 2–5, 2–4 and 4–5 possible for those with the same chirality. In theory then we might expect to see ten major clusters based on these conformers, with more structures observed in the sterically less hindered staggered conformations than in the eclipsed forms, although the difference between eclipsed structures and slightly staggered structures is somewhat arbitrary. It is not always easy to decide which are eclipsed structures and which are only slightly staggered structures. The input data have been ordered so that the O(3)—*R*(5) distance is always shorter than the O(6)—*R*(1) distance.

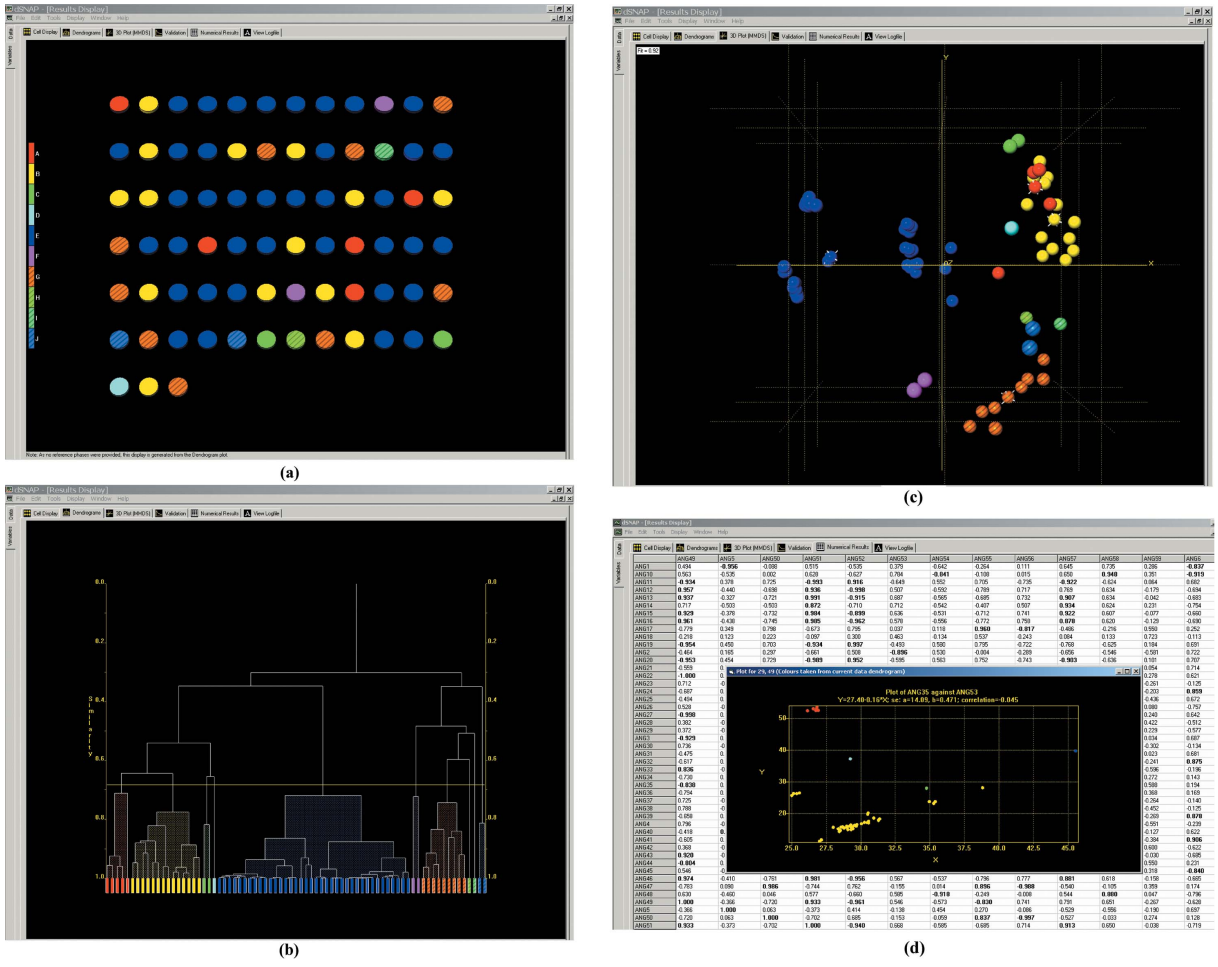


Figure 4
The variables space corresponding to Fig. 3 viewed by selecting the 'Variables' tab on the left-hand side of the screen. (*a*) The well chart. (*b*) The dendrogram with the software-selected cut level. (*c*) The MMDS plot. Each sphere represents an angle or distance in the $FR_1C{=}CR_2F$ fragment. Spheres which are close to each other are highly correlated variables. The easiest one to see in this context is the pair coloured purple at the bottom centre of the plot which correspond to angle 18 and angle 59 in the defined cluster. (It is possible to see which variables are represented by either switching on the 'Label Variables' option or placing the cursor over the sphere.) (*d*) The variables correlation matrix. Entries having $\left|\rho_{ij}^{v}\right| > 0.8$ are highlighted. A sample scatter plot (in this case angle 18 plotted against angle 59) obtained by selecting the relevant matrix entry is superimposed. The correlation coefficient is 0.723. The outliers can be identified by placing the cursor over the relevant data point. The colours of the points are taken from the dendrogram in (*b*). There is an option to display the straight line derived from linear regression.

A search of the CSD for the fragment in Fig. 5(*a*) matched 1356 fragments from 762 structures. Using an 89.5% similarity level 21 groups are observed, and the quality of the clustering can be seen by examining both the dendrogram, the reduced dendrogram and the MMDS plot in Fig. 6. All 21 groups can be explained chemically, and this information and the corresponding conformations are summarized in Table 2. For these calculations, a value of $\lambda = 1$ was used.

3.4. A data set showing no clustering: benzene as a solvate

Not all fragments give clusters using this method. This is not unexpected; if, for example, the structures show only a continual gradation of small differences then it is not possible to apply cluster analysis. As an example of this, a data set consisting solely of benzene molecules observed as a solvate in organic structures was compiled. The results are shown in Fig. 7. Note, in particular, the way the dendrogram gradually builds up from left to right in a stepwise fashion with no noticeable separation; this is typical of data that are not amenable to these clustering methods.

4. Program details

The *dSNAP* program is a highly modified and extended version of the computer program *PolySNAP* (Barr *et al.*, 2004*b*,*c*). The software runs on a PC under Windows 2000 or XP. The current limits are 4000 fragments (n) and 4000 structural parameters (m). The program is written in a variety of languages: the user interface is written in Visual Basic, the underlying code in C++, the graphics are also in C++ and OpenGL, while the cluster analysis code is written in Fortran 95. Graphics cards with OpenGL optimization are recommended. The graphics demands can be considerable when a large number of structures are being displayed. Computer times are highly variable. Parts of the calculation are of order n^3 in processor time. Typical times on a PC powered by a 2.8 GHz Intel Pentium processor with 1 Gb of RAM running Windows XP are 274 fragments in 5 s, 1478 fragments in 4 min 37 s, 1995 fragments in 16 min 30 s, and 2308 in 49 min.

There is an on-line manual in pdf format and a tutorial with sample data sets.

Table 2
Sub-groupings resulting from the cluster analysis of chiral alcohol structures R_1(OH)CHCH(OH)R_2, with a rationale for the clustering.

Group	Colour in Fig. 6	Form	Comment	*R*	O	H	No. of fragments
A	Red	*S*S*s1*	Staggered, part of five-membered ring containing C,N,O heteroatom	*cis*	*trans*	*cis*	22
B	Yellow	*S*S*s1*	Staggered, part of six-membered ring in chair conformation constrained by another ring at the other end of the molecule or staggered, part of six-membered ring in non-ideal twisted conformation	*cis*	*trans*	*cis*	18
C	Green	*S*S*s1*	Staggered, mostly part of six- or seven-membered ring or five-membered ring with Se heteroatom	*cis*	*trans*	*cis*	45
D	Pale blue	*R*S*s2*	Staggered, no obvious sub-structure	*trans*	*trans*	*trans*	80
E	Dark blue	*R*S*s2*	Singleton, outlier of group D; short C—C bonds *R*1—C2, C4—*R*5	*trans*	*trans*	*trans*	1
F	Pink	*S*S*s2*	Straight chains, no steric influences on chains	*trans*	*cis*	*cis*	88
G	Orange, striped	*S*S*s2*	Outlier of G, unusually short C—O bond of 1.36 Å	*trans*	*cis*	*cis*	1
H	Yellow green, striped	*R*S*s1*	Principally six-membered rings in chair conformation; three fragments in a straight chain, four fragments involved in five-membered rings with S or Se heteroatom; three fragments involving seven-membered rings	*cis*	*cis*	*cis*	252
I	Green, striped	*R*S*s1*	Straight chain	*cis*	*cis*	*cis*	6
J	Blue, striped	*R*S*s1*	Similar to J; straight chain, short O···O contact suggesting intramolecular hydrogen bond	*cis*	*cis*	*cis*	1
K	Purple, striped	*S*S*s3*	Varied: mostly part of six-membered rings or five-membered rings with S heteroatom, some straight chains; some short O(3)···O(6) distances	*cis*	*cis*	*trans*	564
L	Pink, striped	*R*S*e1*	The search fragment is part of a six-membered ring and all the fragments are almost exactly eclipsed	Eclipsing C	Eclipsing O	Eclipsing H	2
M	Terracotta orange	*S*S*e2*	Part of six-membered ring	Eclipsing C	Eclipsing H	Eclipsing O	1
N	Paler yellow	*S*S*s3*	Very sterically constrained structure, incorporating an intramolecular hydrogen bond, forcing the overall geometry away from a group K-type structure	*cis*	*cis*	*trans*	1
O	Paler green	*R*S*s1*	Part of a five-membered ring	*cis*	*cis*	*cis*	193
P	Paler blue	*R*S*e1*	Part of a five-membered ring; seven fragments eclipsed, rest semi-eclipsed/staggered	Eclipsing C	Eclipsing O	Eclipsing H	23
Q	Violet	*S*S*e2*	Part of a five-membered ring.	Eclipsing C	Eclipsing H	Eclipsing O	10
R	Paler pink	*S*S*s3*	Contains fragments of the five-membered ring versions of form 5; the search fragment lies opposite the 'flap' of the envelope structure	*cis*	*cis*	*trans*	41
S	Paler orange, striped	*S*S*s3*	As in R but the fragment lies next to the envelope flap, but is not part of it	*cis*	*cis*	*trans*	5
T	Paler green, striped	*S*S*s3*	Part of six-membered ring with very short (unrealistic) C—*R* bond	*cis*	*cis*	*trans*	1
U	Blue green, striped	*S*S*s3*	As T; probably a subset of K	*cis*	*cis*	*trans*	1

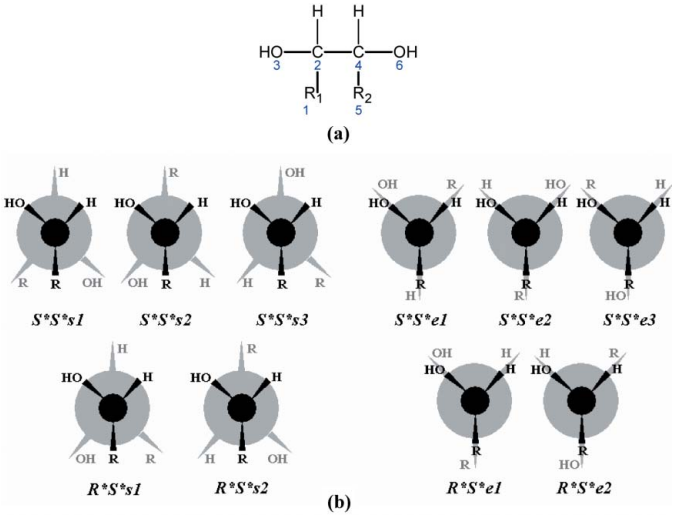


Figure 5
(*a*) The vicinal dialcohol; R_1(OH)CH—CH(OH)R_2 search fragment. (*b*) The possible conformers available to the chiral molecule. The standard notation for *R* and *S* forms is used for the hands of the molecules; a designator of either *e* or *s* denotes whether the form is eclipsed or staggered respectively; forms with similar designators are separately identified by an additional number. The chemically symmetrical nature of the search fragment means that the third *R*S** structure for both the staggered and eclipsed forms is identical to one of the other forms, and so is not explicitly given here. The eclipsed structures are named for their staggered equivalents.

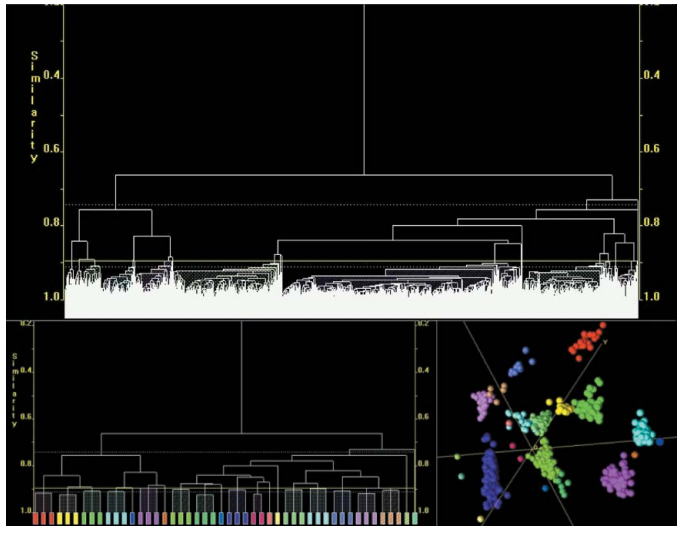


Figure 6
The dendrograms and MMDS plot for the fragment defined in Fig. 5(*a*). For clarity, the dendrograms and MMDS plot are shown as one figure. The upper dendrogram shows the full data set. This can be simplified by displaying only the first, middle and last member of each cluster. Such a dendrogram is shown in the lower left of the figure. The associated MMDS plot is on the right. There are 1356 fragments from 762 structures which have been clustered into 21 clusters. The MMDS plot corroborates this partition of the data with 21 well defined groups. This figure should be examined in conjunction with Table 2, which describes the geometric and chemical characteristics of the individual clusters.

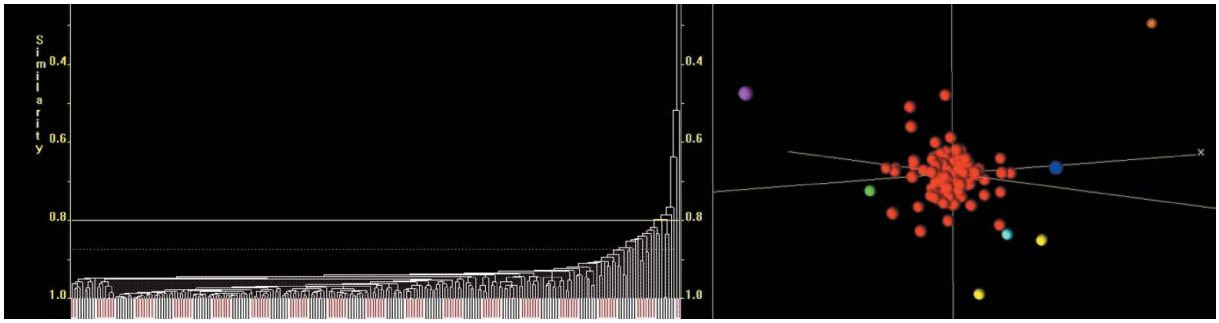


Figure 7
Dendrogram (left) and MMDS plot (right) of a benzene solvate data set showing no clustering structure. Although the dendrogram apparently shows some possible structure at a very low cut level, the tie bars rise in a continuous way that is indicative of a lack of structure as viewed in the context of cluster analysis. This is confirmed in the MMDS plot, with the vast majority of the structures forming a single large group distributed in a random fashion around the origin of the plot and with a few outliers.

Version 0.9 of the software is available free of charge for all crystallographers from Bruker-AXS (*via* http://www.bruker-axs.de). A licence key is required and this can be obtained by emailing dsnap@chem.gla.ac.uk.

We acknowledge support from Bruker AXS and the University of Glasgow.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Allen, F. H. & Motherwell, W. D. S. (2002). *Acta Cryst.* B**58**, 407–422.
Allen, F. H. & Taylor, R. (1991). *Acta Cryst.* B**47**, 404–412.
Barr, G., Dong, W. & Gilmore, C. J. (2004*a*). *J. Appl. Cryst.* **37**, 243–252.
Barr, G., Dong, W. & Gilmore, C. J. (2004*b*). *J. Appl. Cryst.* **37**, 658–664.
Barr, G., Dong, W. & Gilmore, C. J. (2004*c*). *J. Appl. Cryst.* **37**, 874–882.
Barr, G., Dong, W., Gilmore, C. J. & Faber, J. (2004*d*). *J. Appl. Cryst.* **37**, 635–642.
Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.
CCDC (1994). *Vista. A Program for Analysis and Display of Data Retrieved for the CSD*. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK.
Everitt, B. S., Landau, S. & Leese, M. (2001). *Cluster Analysis*, 4th edition. London: Arnold.
Orpen, A. G. (2002). *Acta Cryst.* B**58**, 398–406.
Rousseeuw, P. J. (1987). *J. Comput. Appl. Math.* **20**, 53–65.
Taylor, R. (2002). *Acta Cryst.* D**58**, 879–888.
Taylor, R. & Allen, F. (1994). *Structure Correlation*, Vol. 1, edited by H.-B. Bürgi & J. D. Dunitz, pp. 111–161. Weinheim: VCH.
Wegman, E. J. (2005). http://www.galaxy.gmu.edu/pub/.
Wilhelm, A. F. X., Wegman, E. J. & Syzmanzik, J. (1993). *Comput. Stat.* **14**, 109–146.