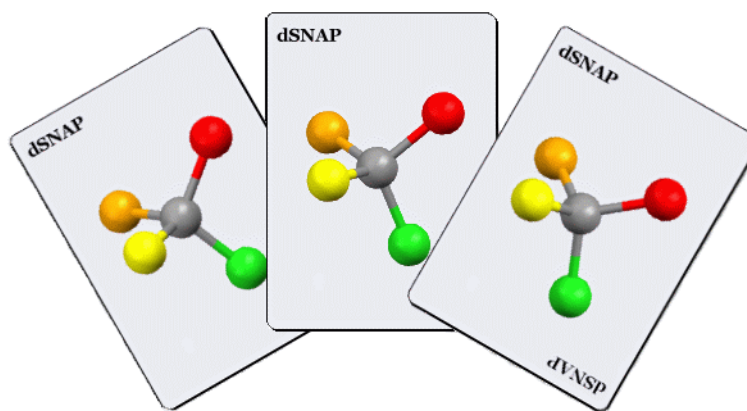




*d*SNAP

A computer program to cluster and classify
results from Cambridge Structural Database
searches



Program Manual

Credits

***d*SNAP:**

is a product principally of

Gordon Barr, Anna Collins, Chris Gilmore, Wei Dong, Andrew Parkin and Chick Wilson

of

*WestCHEM, Department of Chemistry
University of Glasgow
Glasgow
Scotland G12 8QQ
United Kingdom*

Any problems, comments or questions should be directed to:

Email: dsnapsnap@chem.gla.ac.uk

Documentation Production Notes

Manual written by A. Lamarque, A. Collins, A. Parkin and G. Barr.

*d*SNAP logo by R. Thatcher.

Created using Adobe Framemaker. Set in 12 point Times New Roman.

Last modified: 15/04/2009

Disclaimer

This manual, as well as the software described in it, is furnished under license and may be used or copied only in accordance with the terms of such license. The content of the manual is provided for informational use only, is subject to change without notice, and should not be construed as a commitment. We assume no responsibility for any errors or inaccuracies that may appear in this book. This software and printed materials are provided 'as is' without any warranty or condition of any kind, express or implied. You assume the entire risk as to the use and performance of the software or printed materials in terms of correctness, accuracy, reliability, currentness or otherwise.

Copyright © 1999 - 2009, The University of Glasgow, All Rights Reserved.

Contents

Introduction - - - - -	3
Installation and Registration - - - - -	9
Preparing Search Data for dSNAP - - - - -	21
Inputting Search Data into dSNAP - - - - -	27
Results Display Overview - - - - -	41
Data Space Results Display - - - - -	51
Variables Space Results Display - - - - -	83
Other Menu Items & Shortcuts - - - - -	91
Program Options & Default Settings - - - - -	97
Dealing with Structures with Local Symmetry - - - - -	105
Advanced Options - - - - -	117
References - - - - -	123

1.1 Program Overview

*d*SNAP is a computer program for automatically classifying and visualizing the results of database searches using the *Cambridge Structural Database*. This is done through cluster analysis and multivariate data processing of distance matrix information describing the extracted structures where an exhaustive table of geometries has been obtained.

The results of these calculations are displayed through a set of visualisation tools that allow the user to view and verify the proposed classification scheme, and explore it in varying levels of detail.

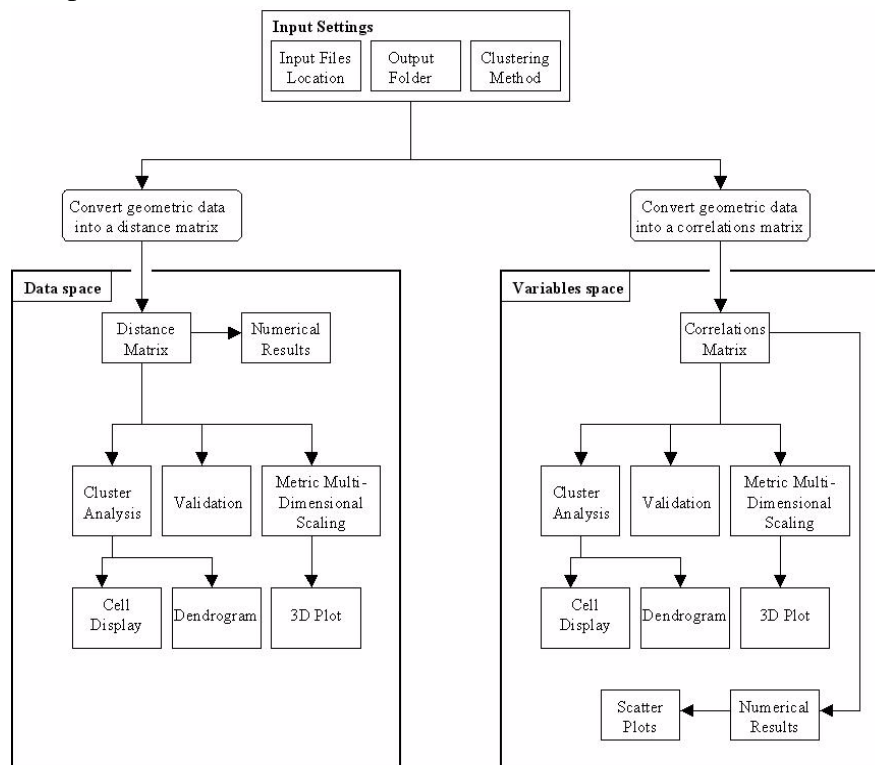
1.2 What *d*SNAP does

The main stages involved in a standard run are:

1. In *ConQuest*, define search fragment of interest.
2. In *ConQuest*, define a geometric variable in the fragment.
3. In *ConQuest*, perform a database search.
4. Export the relevant files from *ConQuest*.
5. Import and processing of data files into *d*SNAP.
6. Perform cluster analysis in *d*SNAP.
7. Output results to file and graphically to screen to allow interpretation.

8. Reclassification of results where necessary.

The process involved in a standard run of *d*SNAP is shown below:



1.3 Nomenclature

Some terms that are used in this manual and refer to use of *d*SNAP should be defined.

1.3.1 Fragments, Hits, Structures

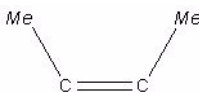
The terms *structure* and *hit* refer to an individual crystal structure determination mined from the *Cambridge Structural Database* (CSD), labelled by an individual unique CSD reference code (REFCODE). The term *fragment* refers to the part of the structure matching the search query.

In other words when the user is presented with a structure that has matched their database search, the fragment is only the section of the structure that matches the requirements they specified in the search.

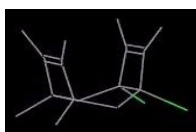
A single hit structure can contain one or more search fragments. When multiple fragments are observed for a hit, these fragments frequently lie in chemically distinct environments and thus may exhibit significantly different geometries. They are therefore treated independently.

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 correct fragment when studying the structures within the *ConQuest* search program.

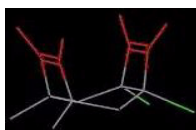
For example, a database search for the a carbon-carbon double bond with a methyl substituent on each carbon was performed.



This is the search fragment. One of the hit structures returned by the search had the refcode HAFTEN is shown below.



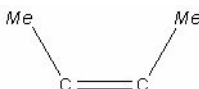
This hit has two instances of the search fragment, which are highlighted on the structure below in red.



These will be labelled separately (in this case HAFTEN_01 and HAFTEN_02) when used in *dSNAP*.

1.3.2 Variables

Variables are the geometric information in the structure of the fragment that have been defined and searched for. These include both angles between and distances between atoms. They are also termed parameters. Using the same search fragment as an example:



A variable would be the double-bond distance. Another would be the sp²-sp³ carbon distance (one for each side), or the sp³-sp²-sp² carbon angle (again, one for each side). All interatomic distances and angles are considered, so the distance between the two methyl groups is also a variable.

A single fragment will have many variables and the complete list of distances and angles completely defines a fragment.

In *d*SNAP a variable is named according to what type of variable it is and the atoms it is comprised of. For example *d_2_4* would be the distance between the second and fourth atoms and *a_1_3_4* would be the angle between the first, second and fourth atoms. In this system the atoms are numbered according to order they were drawn in the original *ConQuest* search.

1.3.3 Data and Variables space

*d*SNAP allows separate analysis on both the fragments and their underlying structural variables. These are both given their own section of the program and can thought of as being two separate domains: data space and variables space.

All of the analysis and calculations carried out in the *Data Space* section are concerned with the fragment structures themselves. These operate by collating all of the geometric data for a single fragment (for example collecting all of the distances and angles in HAFTEN_01) and treating these as a single entity to compare against the other fragments in the dataset. This enables the program to calculate similarities between fragments.

All of the analysis and calculations carried out in the *Variables Space* section are concerned instead with the geometric variables. Variables mode operates by grouping every instance of each variable in all of the fragments available and treating these as a single variable. This allows trends in geometric variables to be established.

1.4 Running *d*SNAP

1.4.1 Other software requirements

To run *d*SNAP, access to Cambridge Crystallographic Data Centre software is required. The Cambridge Structural Database (CSD) must be installed to be able to perform searches, and *Mercury* is helpful to view structures.

1.4.2 Limitations

The current limitations are 4000 fragments (*n*) and 4000 structural parameters (*m*) per fragment. Only 3-20 atoms of any fragment can be used in the analysis. Then number of parameters is given by:

$$l = \frac{l}{2}(l-1)^2$$

where *l* is the number of atoms in the fragment.

1.4.3 Run-times

Typical run-times are given below based on a PC powered by a 2.8 GHz Intel Pentium processor with 1 Gb of RAM running Windows XP. These times are measured from launching the program to the results display screen being displayed.

Number of fragments	Time
274	5 secs
1478	4 mins 37 secs
1995	16 mins 30 secs
2308	49 mins

These times are approximate, and may vary depending on the analysis options selected for a given program run and on the number of atoms per fragment.

2.1 Program Requirements

*d*SNAP is a highly modified and extended version of the computer program *PolySNAP*¹. The software runs on a PC under Windows 2000 or XP. The program is written in a variety of languages: the user interface is written in Visual Basic, the underlying code in C++, the graphics also in C++ and OpenGL while the cluster analysis code is written in FORTRAN 95.

*d*SNAP requires a modern, high-specification PC running Microsoft Windows 2000 SP4 or XP SP1 or later. Additionally, a monitor with minimum 1024 X 768 resolution at 32 bit colour depth is needed, though a larger size is recommended. Graphics cards with OpenGL optimisation are recommended. The graphics demands can be considerable when a large number of structures are being displayed.

2.2 Installation

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

Launch the installer program *Setup.exe* by double-clicking it.

Note that to install software on Windows 2000/XP systems, a system administrator password will be required.

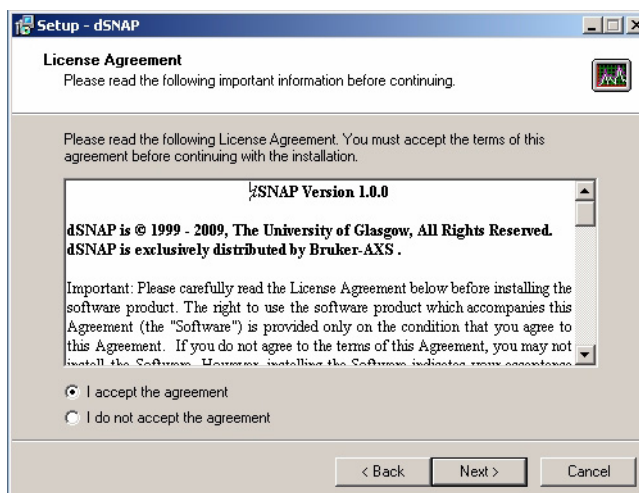
1. Barr, G., Dong, W. & Gilmore, C.J (2004). *J. Appl. Cryst.* **37**, 243-252.

Barr, G., Dong, W. & Gilmore, C.J (2004). *J. Appl. Cryst.* **37**, 635-642.

It will display a welcome window:

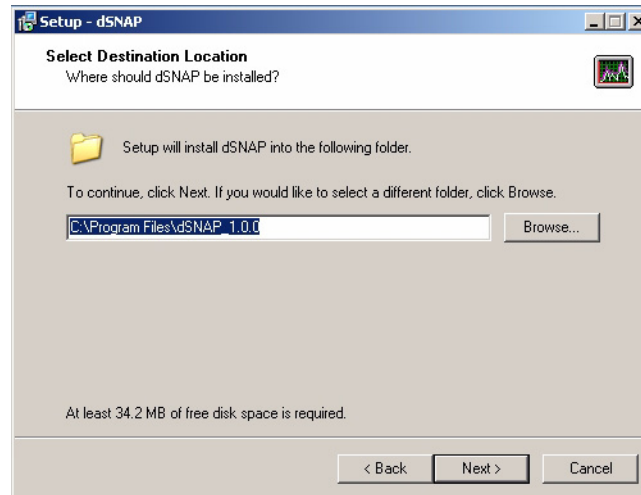


Click Next to begin the setup process. The screen details the licence agreement for installing and using *dSNAP*:



Read the details before choosing to accept or reject the terms of the agreement. Failure to accept the terms means that setup of the program cannot continue.

Once agreed to, the licence agreement is followed by a dialog box allowing the user to control where the program is installed:

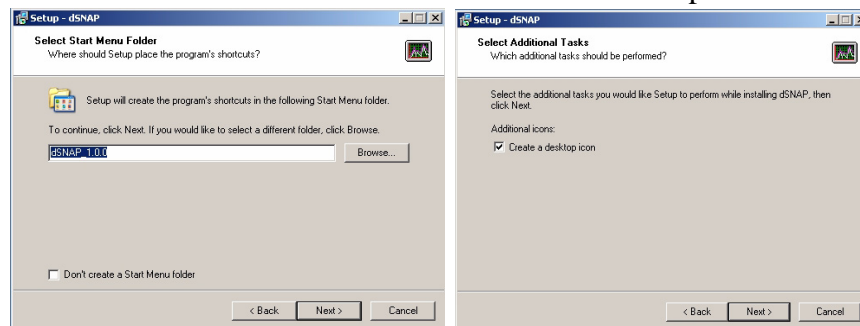


The default path,

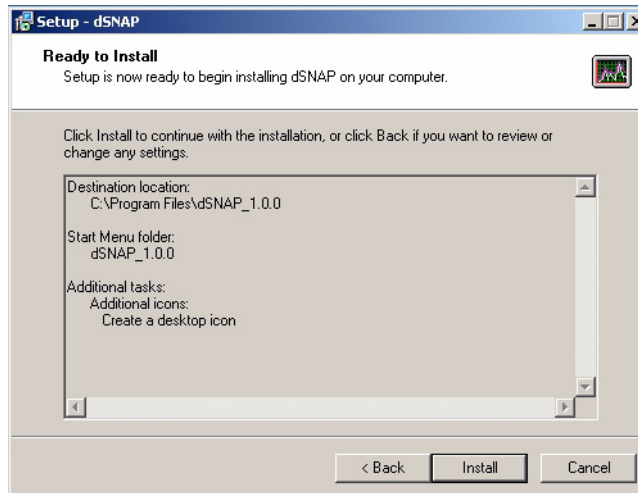
C:\Program Files\dSNAP

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

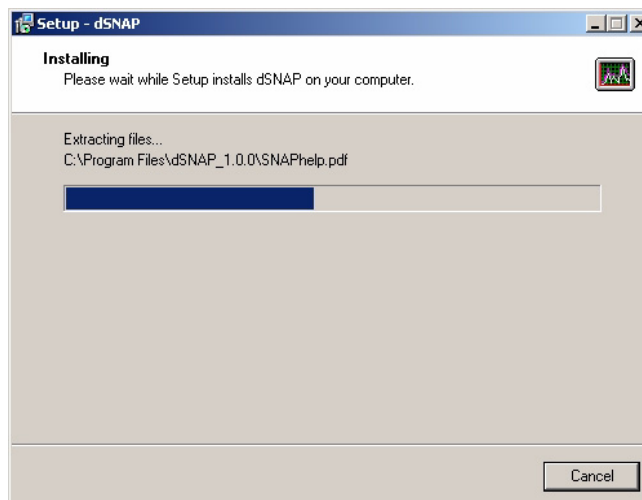
The next two screens control whether or not a shortcut to the software is added to the Start Menu and/or the Desktop:



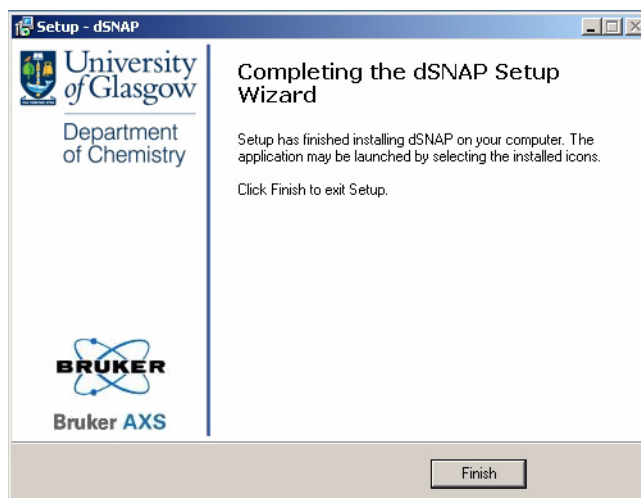
The final screen displays a summary of the selected options; click *Install* to start the installation process:



Installation should then proceed automatically. A progress bar will displayed as the installation is carried out:



The installer will display a message when it has completed installation:



Click *Finish* to close the installer.

2.3 Launching *dSNAP*

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

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



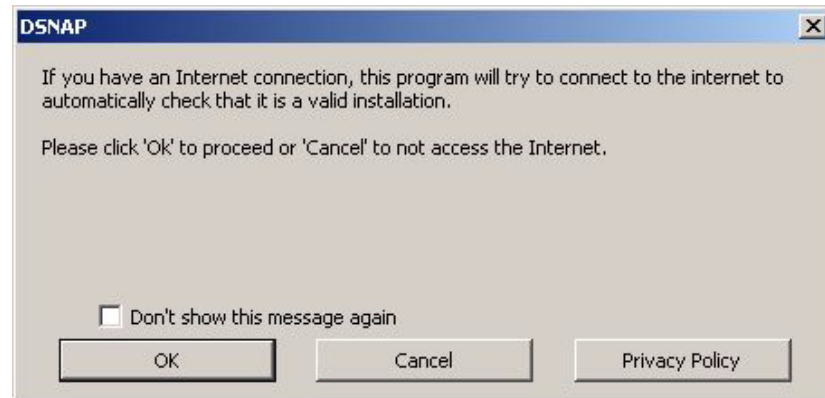
- From the Start Menu, select the *Programs* sub-menu, followed by the *dSNAP folder*, and then the *dSNAP* option. (It is usually installed at the very bottom of the list of programs).

2.4 Registering *dSNAP*

The copy of *dSNAP* that is now installed needs to be registered before it can be used.

Upon first launching *dSNAP*, a window will appear warning you it is

going to connect to the internet to perform a security check:



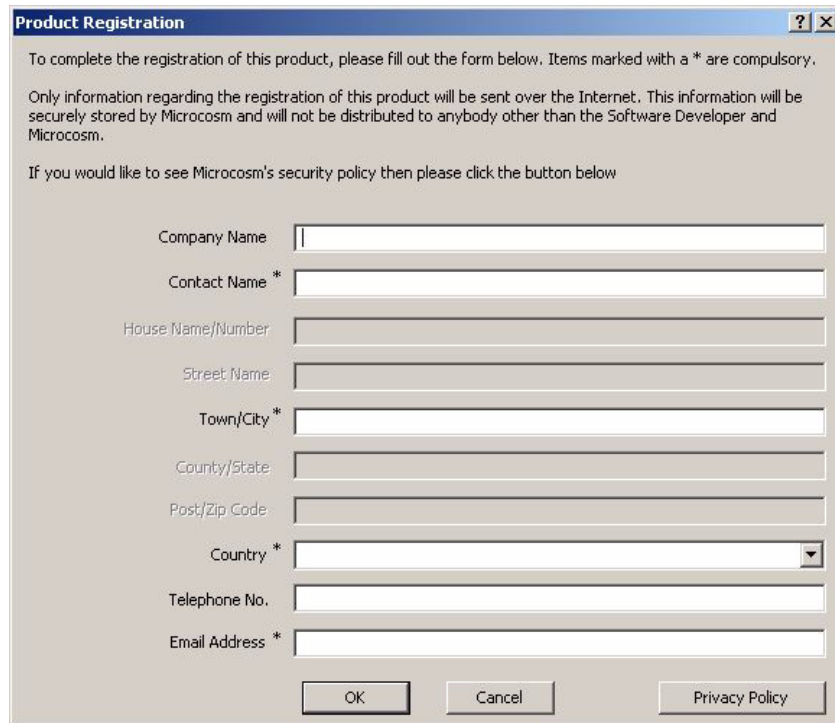
Click OK to continue [even if the computer you are installing *dSNAP* on is not currently networked].

The internet connection is then used to ensure that the copy of *dSNAP* being registered is a valid installation. A progress window will be displayed while this check is performed.



Once this is complete a form will open, asking for registration

details.



The image shows a 'Product Registration' dialog box with a title bar containing a question mark icon and a close button. The text inside the dialog reads: 'To complete the registration of this product, please fill out the form below. Items marked with a * are compulsory. Only information regarding the registration of this product will be sent over the Internet. This information will be securely stored by Microcosm and will not be distributed to anybody other than the Software Developer and Microcosm. If you would like to see Microcosm's security policy then please click the button below'. The form contains the following fields: 'Company Name' (text box), 'Contact Name *' (text box), 'House Name/Number' (text box), 'Street Name' (text box), 'Town/City *' (text box), 'County/State' (text box), 'Post/Zip Code' (text box), 'Country *' (dropdown menu), 'Telephone No.' (text box), and 'Email Address *' (text box). At the bottom of the dialog are three buttons: 'OK', 'Cancel', and 'Privacy Policy'.

Only your name, city, country and contact email address is required. Please fill them in, and click *OK*.

All details with an asterisk next to them are essential, and must be completed for registration to be successful. The authors of the program reserve the right not to issue a licence if the information given is incomplete or false. In particular your e-mail address must be valid.

NOTE: your email address and other information will not be passed to any third party and will not be used for any other purpose except to notify you of updates to the software.

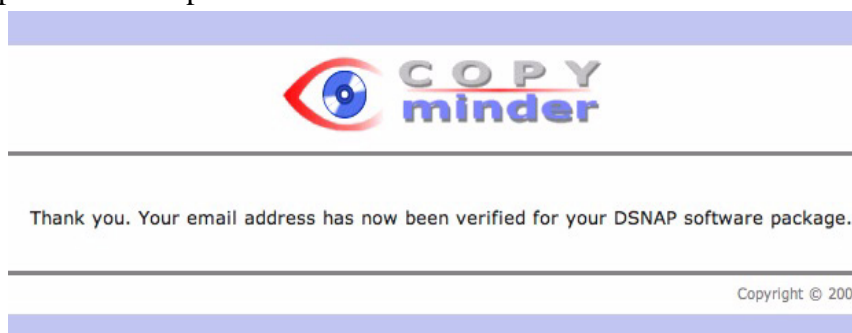
Click *OK* once all the relevant details have been entered.

Once the registration is processed, another window will appear tell-

ing you to check your email, and *d*SNAP will exit:



Once the confirmation email is received, please click the link contained within it, which will take you to the CopyMinder website to complete the installation. You should see the text below once the process is complete.



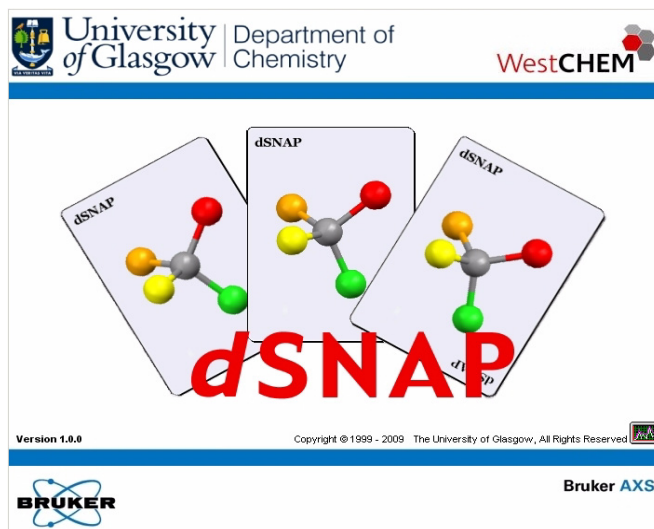
If you have any problems with the process, or are unable to connect to the internet, please contact us directly at

dsnapsnap@chem.gla.ac.uk

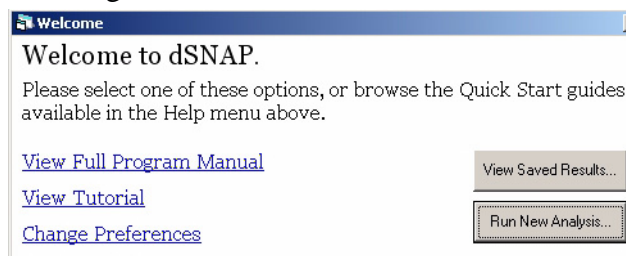
and we can process your registration manually.

Once this registration process is complete, you can now launch *d*SNAP again, and start using the now fully-registered program. The program logo screen will appear for a few seconds, before being

automatically dismissed:



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



You are now ready to begin working with the software.

Please note that a searchable PDF on-line version of this manual can be viewed at any time during the program by selecting *View Program Manual* from the *Help* menu or the welcome window.

2.5 Installation Troubleshooting

Problem

Program registration was not completed, because *dSNAP* was unable to connect to the internet.

Solution

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

www.copyminder.com/activate.php

to obtain an activation code manually.

Problem

An error is displayed saying *d*SNAP is unable to generate a matrix file, even when the input files are perfectly valid.

Solution

Try resetting the program preferences, by going to *Edit -> Options* and selecting *Reset All* in the bottom left corner. If the problem persists after this, please contact dsnapsnap@chem.gla.ac.uk

Problem

When *d*SNAP is launched, a Microsoft Office Installer window appears repeatedly and cannot easily be dismissed.

Solution

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

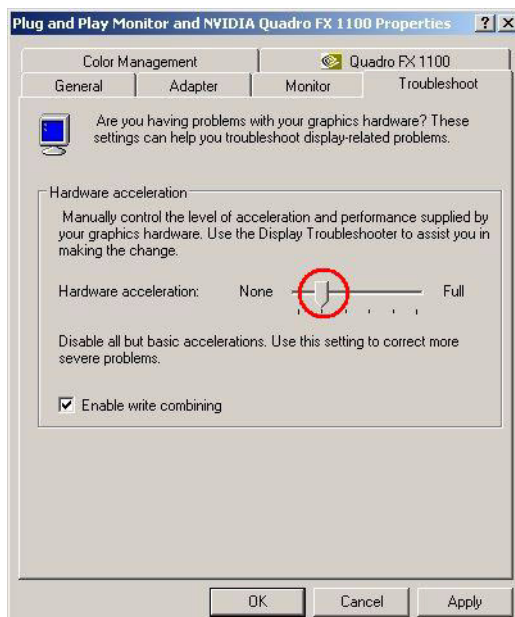
Problem

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

Solution

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

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



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

Problem

Suspicious behaviour is being flagged by Anti-Virus software during installation.

Solution

Anti-Virus software (such as Sophos) that look for 'suspicious' behaviour on your computer may flag up the *dSNAP* installer when it is running.

In particular, we have seen it complain and flag 'HIPSPROCMod004'

This is simply the Anti-Virus software being overly cautious; for more information, please contact us at **dsn@chem.gla.ac.uk**, or see the Sophos page here for details.

<http://www.sophos.com/security/analyses/suspicious-behavior-and-files/hipsprocmod004.html>

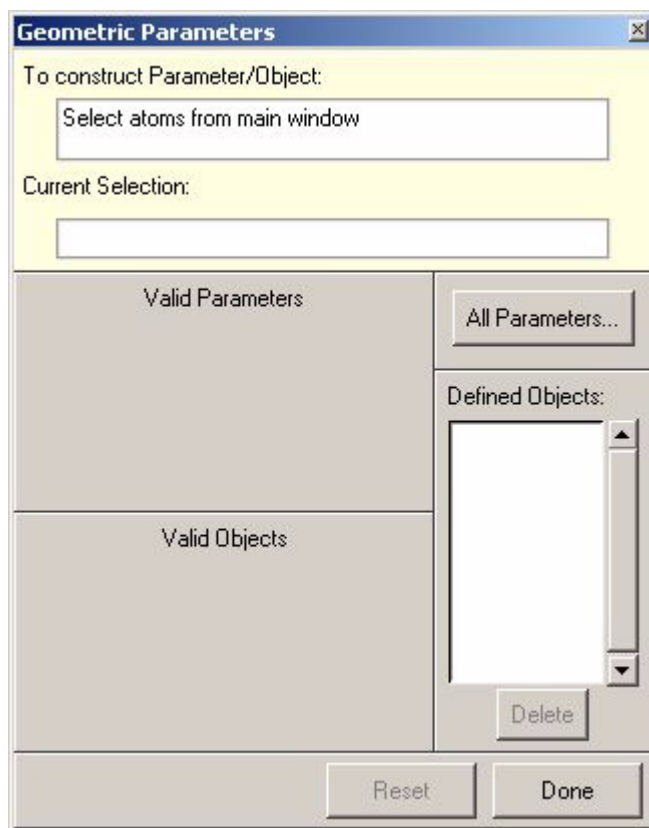
In particular, Sophos, like most anti-virus software companies, suggest disabling their software while you are performing any new software installations to prevent false-positives like this taking place.

3.1 Preparing data in *ConQuest*

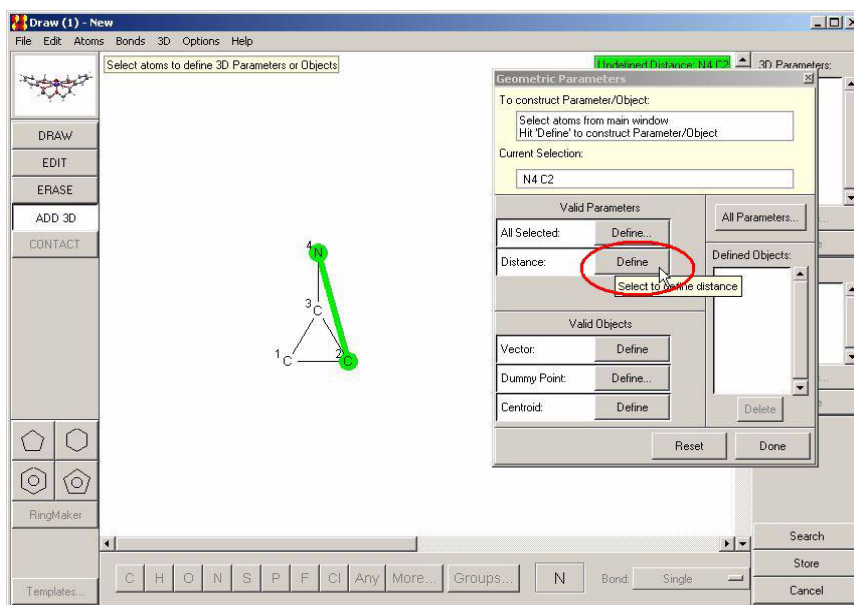
dSNAP performs analysis on the structural data obtained from the *Cambridge Structural Database* via the program *ConQuest*.

To extract this data properly the user must define at least one inter-atomic distance when drawing the desired structural fragment during the original *ConQuest* search.

To do this the user must be in the *Draw* section of the *Build Queries* screen of *ConQuest*. Once the desired fragment has been drawn select *Add 3D* from the left-hand side tab menu. A new window headed *Geometric Paramters* is opened:



To define the distance the user must select any two atoms in the structure by clicking on them, which highlights them in green and draws a green line between them. In the *Geometric Parameters* window there will be a *Distance* box which will now have a *Define* button available next to it.



Clicking on this button will define the distance and once this has been done the distance should now be assigned the name DIST1 in green. The *Geometric Parameters* window can now be closed and the search can be performed as normal.

N.B. There is a known issue with defining hydrogen atoms in the ConQuest search. If some hydrogen atoms are defined explicitly and others implicitly (as in Fig. 1, where the hydroxyl hydrogen has been defined explicitly and the alkyl hydrogen atoms implicitly), *d*SNAP will not run and an error message will be generated. However, *d*SNAP will run successfully in cases where all hydrogen atoms are defined explicitly (as in Fig. 2) or all are defined implicitly (as in Fig. 3).

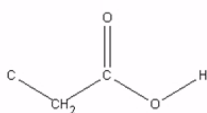


Fig. 1

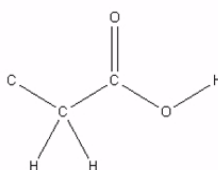


Fig. 2

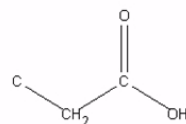


Fig. 3

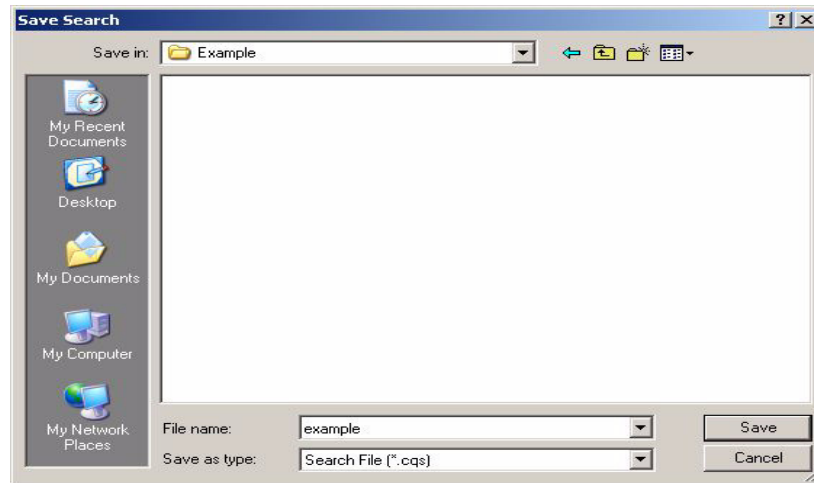
Additionally, fragments which have been defined with variable bond types cannot be analysed using *d*SNAP. We suggest repeating the search, defining the bond type as *Any*. The search can be run on the subset of data from the original ConQuest search which had the variable bond types to minimise the risk of generating additional, undesired hits.

3.2 Exporting Search Results from *ConQuest*

In order to analyse the results obtained from the CSD search, three separate files must be created to feed into the program. This is done in three steps.

3.2.1 Creating Search (.cqs) file

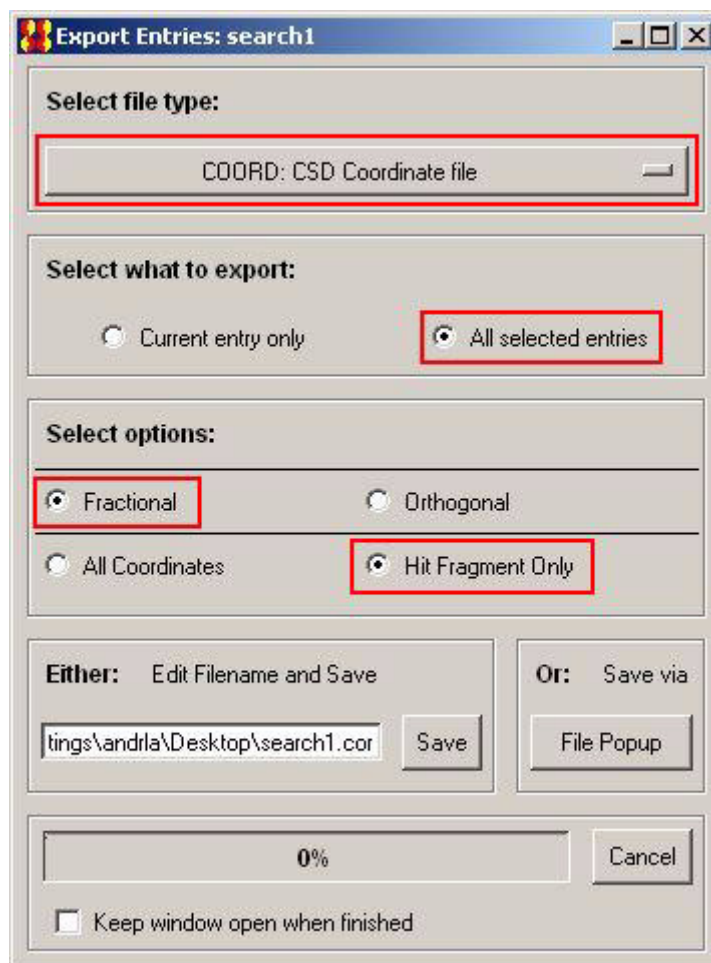
Once the search is complete select *Save Search As...* from the *File* menu.



Select a suitable destination folder in the browse window and select a suitable file name (*e.g.* example.cqs). This root filename must be used for all the files saved from this search.

3.2.2 Creating Coordinate (.cor) file

Select the *Export Entries as...* option from the *File* menu. A new window will appear with several options for saving your data.



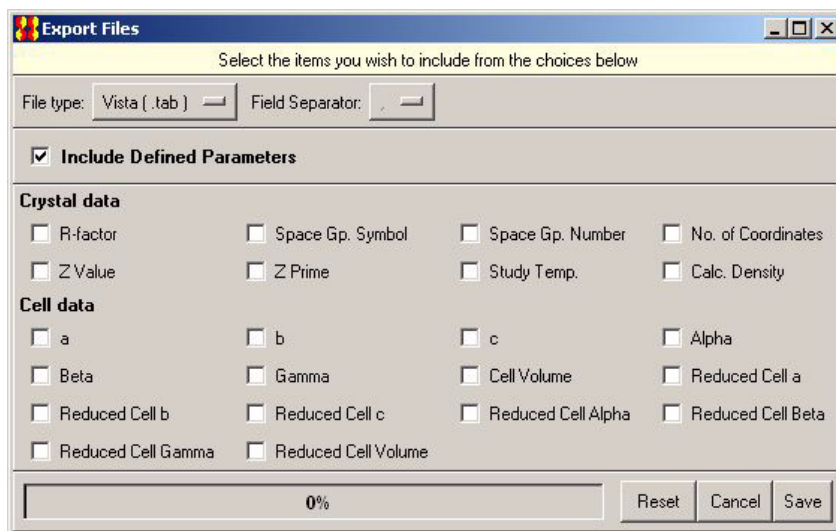
The correct settings needed for input into *d*SNAP are given below:

File type:	COORD: CSD Coordinate file
What to export:	All Selected entries
Select Options:	Fractional
	Hit Fragment Only

Once these settings have been selected click on the *File Popup* button at the bottom of the window. This opens a windows save box where the user must select the same destination folder and use the same base filename as before (e.g. example.cor).

3.2.3 Creating Parameters (.fgd) file

Back in the search results display select the *Export parameters and data...* option from the *File* menu. This will open a new window with options for exporting the data.



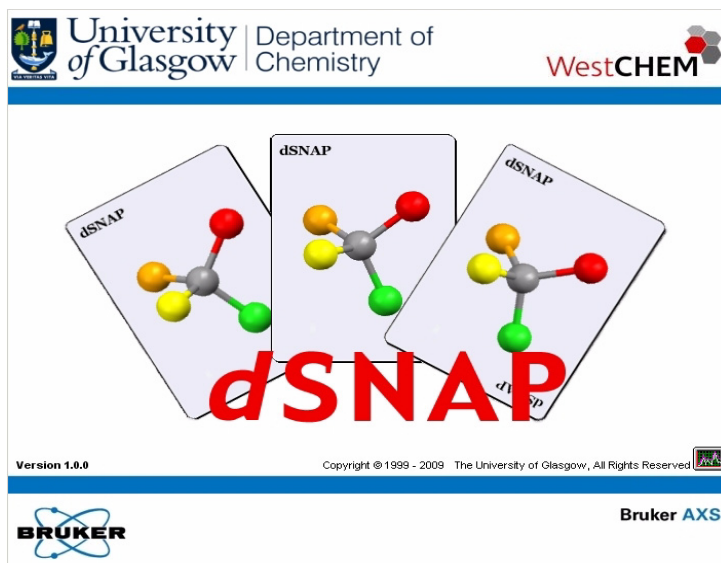
Export the data using all of the default options (i.e. the file type is Vista, *Include Defined Parameters* selected). To do this select *Save* from the bottom right of the window, select the same destination folder and use the same base filename as previously. This creates three files (e.g. example.fgd, example.fgn and example.tab) but only one (example.fgd) is required.

The selected folder now contains all the data files needed to run dSNAP. *ConQuest* can now be closed. This folder should contain five files, all bearing the same root filename.

4.1 Loading files into *dSNAP*

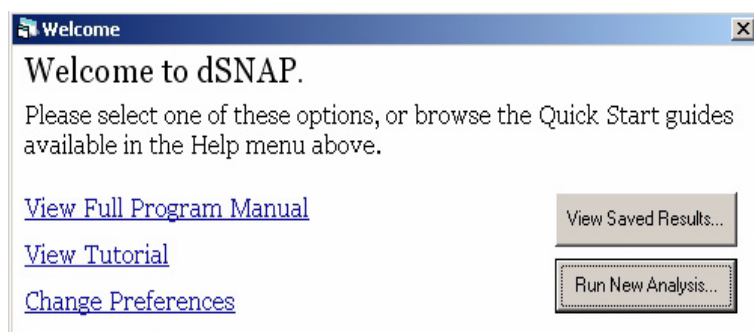
4.1.1 Opening *dSNAP*

Assuming a standard installation of the program, *dSNAP* can be opened either by double-clicking the desktop icon or by selecting *dSNAP* from its folder in the *Programs* section of the Windows *Start* menu. A splash screen will appear displaying information about *dSNAP*.



Once this has been displayed the main program window will open. The user should be met with a display screen with menus running along the top and featuring the *dSNAP* logo. A startup window welcoming the user to *dSNAP* will also be displayed.

4.1.2 dSNAP Welcome Window



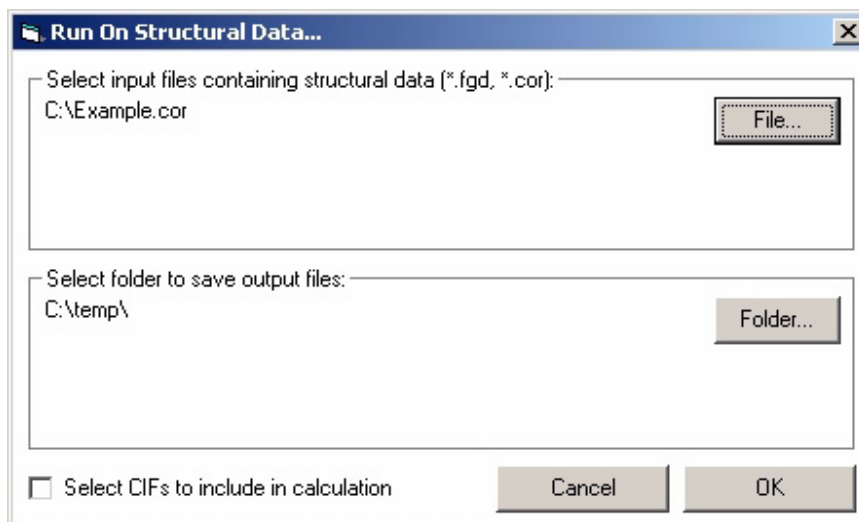
This window appears automatically when *dSNAP* is opened but can be accessed at any time by closing all results windows and clicking on the empty *dSNAP* display screen.

The two buttons allow the user to begin using *dSNAP*, either by fresh analysis of new data or by retrieving the results from a previous run.

There are links that open PDF versions of the manual and tutorial. The *Change Preferences* link allows the user to edit the program options

4.1.3 Inputting Data to dSNAP

In order to run a search through *dSNAP* either select *Run New Analysis* from the welcome window or select the *Run on...* option from the *File* menu. The following dialog box will open:



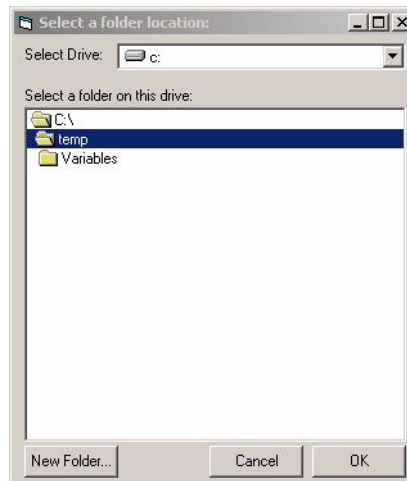
This allows the location of the required files to be selected on a run-by-run basis, without having to change the default settings in the program *Options* window. The initial paths listed in this window will correspond to the defaults specified in the *Options* window (see Section 9.1.1). The main two sections are as follows:

Select input file containing structural data (.fgd, *.cor)*

Here the user selects the relevant geometric data that was previously exported from *ConQuest*. The appropriate file is selected by clicking the *File...* button which opens a standard Windows file selection box. The user must navigate to the appropriate file and select it. Data files with the extension *.fgd* or *.cor* may be selected. The location displayed will update.

Select folder to save output files

The program has to be given a set location in which to save all of the output files created by an automatic run. This can be manually set to any folder as required by the user by clicking on *Folder...* and navigating in the following window.



To select an appropriate folder start by selecting a drive from the pull-down menu at the top of the display. Then select a suitable folder by double-clicking on it. If the selected folder contains more folders these will now be available to be selected. If required, new folders can be created by clicking the *New Folder...* button at the bottom of the screen.

A dialog box will open asking for a name for the newly created folder.



Once a name has been given the folder is created in the location of the last folder to be opened before clicking the *New Folder* button.

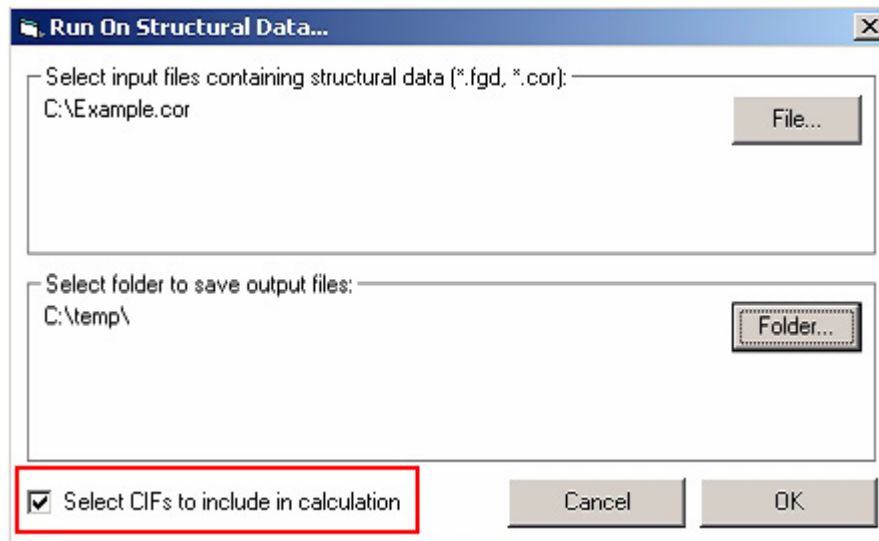
Clicking *OK* selects the last folder to be opened as the output folder. It is simple to see which folder this is as it will be the lowest folder in the display to be represented by an open folder icon - *i.e.* in the screenshot above, the folder *C:\temp* is selected.

Note that if an output folder is selected that already contains files from a previous *dSNAP* run, **those older files will be deleted** prior to the new run starting. Also note that the output folder cannot be a folder that contains either of the input files.

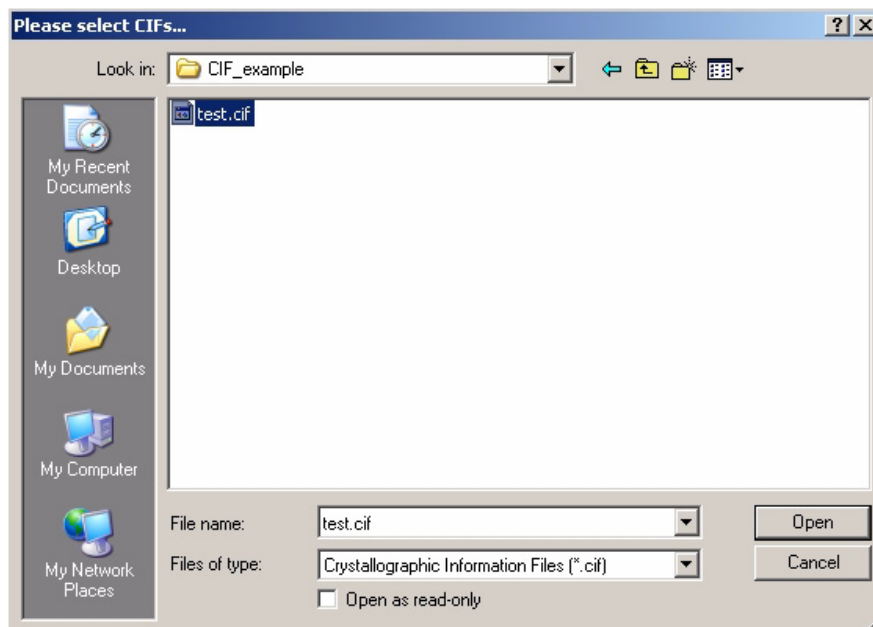
4.2 Including your own structures

When loading datasets into *dSNAP* it is also possible to include structures that do not yet appear in the database. This allows recently obtained structures to be directly compared with those that are already in the database by including them as crystallographic information files (*.cif*) in the analysis. This is demonstrated in the example shown below:

Once the file input window has opened click the *Select CIFs to include in calculation* checkbox at the bottom of the input window:

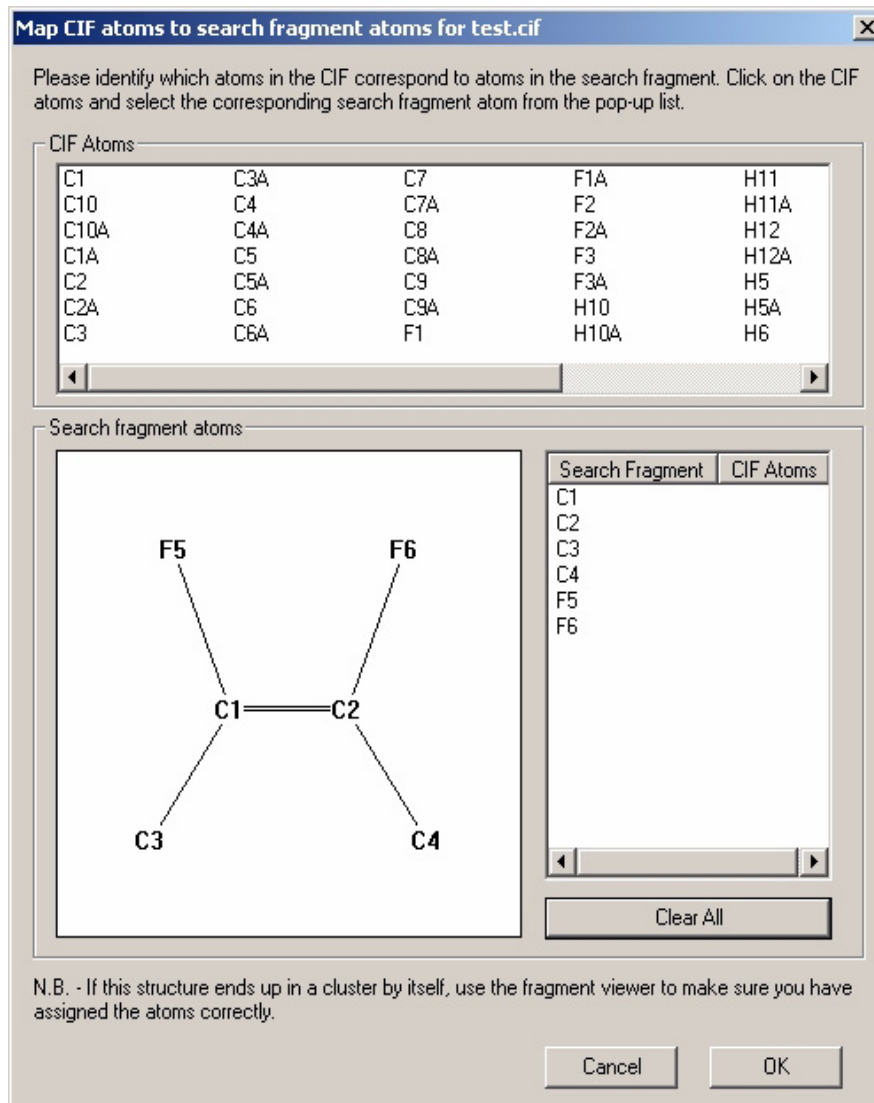


Once the input and output folders have been selected as normal the user clicks *OK* to proceed to an additional input window that asks for the location of the CIFs that are to be included:



Multiple CIFs can be selected at the same time. In the example shown above there is only one CIF being included. All CIFs must be selected at the same time so when including multiple CIFs they must all be located in the same folder.

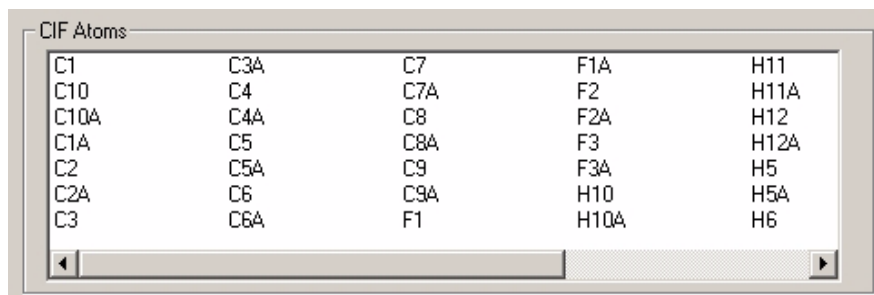
Clicking *OK* selects the specified CIF and opens a new window:



As the included CIF is separate to the entries returned from the database *d*SNAP has no information on how to correctly map the search fragment atoms to the corresponding atoms in the CIF. The user must define the correct way for this to be done by specifying the atoms in the CIF that are part of the search fragment, and then specifying where in the search fragment that particular atom belongs. Note that when multiple CIFs have been included, this process must be done for each CIF in turn. The name of the CIF currently being considered is displayed along the top of the window. In this example the CIF is named *test*:

Map CIF atoms to search fragment atoms for **test.cif**

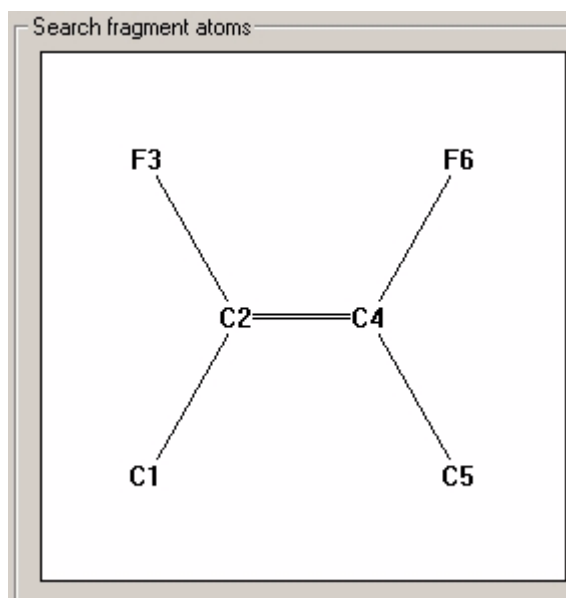
In order to define which CIF atoms are to be used the window is split into three sections. At the top there is a section named *CIF atoms* which displays a list of all the atoms that are present in the CIF:



C1	C3A	C7	F1A	H11
C10	C4	C7A	F2	H11A
C10A	C4A	C8	F2A	H12
C1A	C5	C8A	F3	H12A
C2	C5A	C9	F3A	H5
C2A	C6	C9A	H10	H5A
C3	C6A	F1	H10A	H6

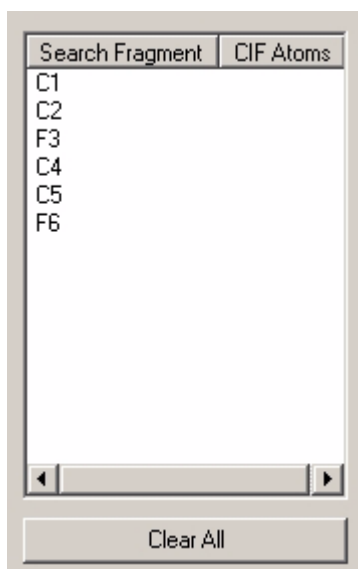
A scroll bar appears along the bottom when the number of atoms present becomes so great they can't all be displayed at the same time.

Below there is a section called *Search fragment atoms* which displays a 2D diagram of the search fragment, listing all of the atoms and their positions in the fragment:

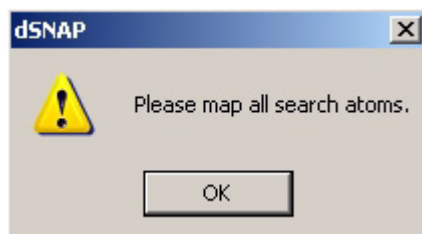


It is these atoms that need to be assigned a partner from the list of CIF atoms. It is important to note that although they take an identical numbering format to each other, the labels used in the diagram and the labels used for the list of CIF atoms do **not** correspond to each other. The diagram of the search fragment uses the *d*SNAP labelling system which is taken from the search query in *ConQuest*, and the labels given to the CIF atoms in the list are taken from the CIF itself. For example in this case there is a C1 in the list of CIF atoms, and a C1 in the list of CIF atoms, and a C1 on the diagram, but as will be seen, these are **not the same atom**.

Finally there is a section where all the fragment atoms are listed in a table:



The corresponding CIF atoms are then displayed in the table as they are defined. When assignment is complete all of these atoms must have a CIF atom assigned to them. If the user tries to proceed without defining all atoms a message is displayed:



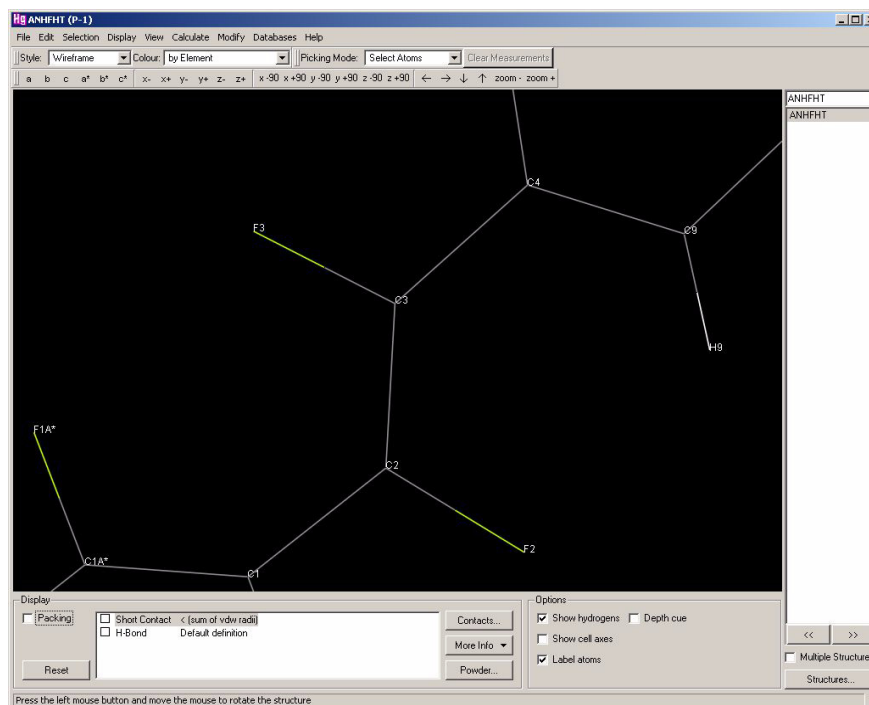
There is a *Clear All* button at the bottom of the section to reset all of the assignments that have currently been made.

In order to properly assign atoms from the CIF to the search fragment the user needs to know where in the structure in the CIF the atoms in the list appear. In most cases it will be necessary to open the CIF in a viewing program such as *Mercury* in order to access this information.

Once the CIF has been opened in *Mercury* the labels need to be displayed by right-clicking on the display and selecting *Show labels* from the *Labels* section of the right-click menu:



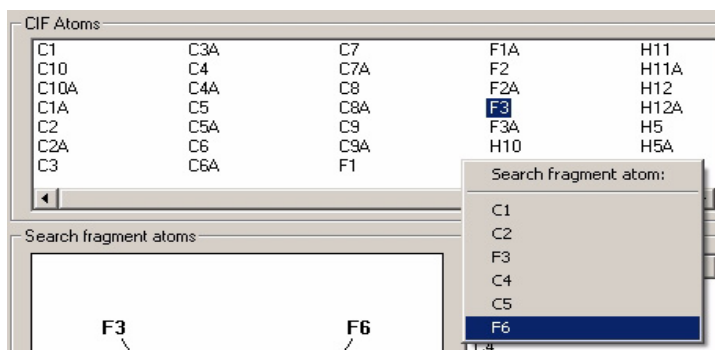
The user can then navigate to the relevant part of the structure. This displays the search fragment with all of the positions labelled using the CIF labels:



Using this the user can then map the correct CIF atoms onto the FGD search fragment diagram. For example, in this case F3 in the CIF

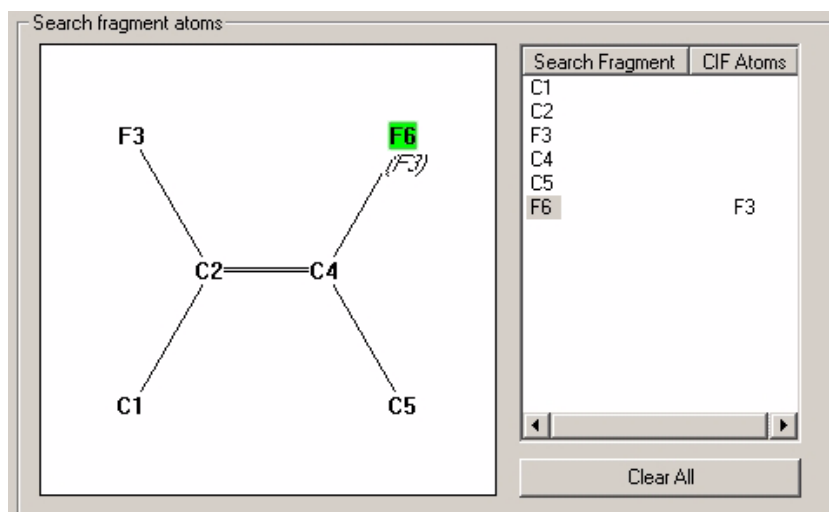
corresponds to F6 in the 2D FGD diagram. To assign this go back to the window in *dSNAP* and click on atom F3 in the *CIF atoms* section. By doing this we are telling *dSNAP* that the atom F3 in the CIF is one of the atoms in the search fragment.

A pull down menu then appears asking which search fragment atom F3 corresponds to:



In this case atom F3 in the CIF corresponds to atom F6 in the original search fragment. This is specified by scrolling down and clicking on F6 in the pull down menu.

Now that this atom has been fully assigned the other two displays change:



Atom F6 is now highlighted in green on the diagram to show it has been assigned, with the CIF atom that corresponds to it written in italics below. In the other section F3 is now listed as the corresponding CIF atom to search fragment atom F6.

This is repeated for all atoms on the diagram until the entire fragment has been assigned. Once this is done the user can proceed by clicking *OK*. You will be asked if you wish to select another fragment from the same CIF.

If multiple CIFs have been included then the same window will appear again for the second CIF and the entire process must be repeated, ensuring that the relevant CIF is opened in *Mercury* for reference. Only one CIF can be assigned at a time. If there was only one CIF included then the window closes and analysis begins.

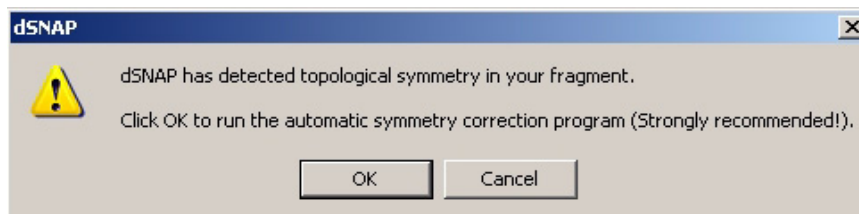
When the results display opens the included CIF can be easily identified as it is labelled with a + sign followed by the last 7 characters of the name of the parent CIF, so it is best to ensure that CIFs have short names that allow easy identification if multiple CIFs are included. Make sureAs the file used in this example was called *test* the fragment from the CIF is labelled as *+test*:



If the included CIF appears in a cluster by itself then this suggests that the atoms may have been assigned incorrectly and that the CIF atom assignment should be reperformed. The fragment viewer can also be used to help assess whether the atoms have been assigned properly or not. If the atoms have been assigned correctly then when the included fragment is opened in the viewer its structure should look sensible and broadly follow the same arrangement as fragments from other clusters. A structure that does not look sensible suggests the assignment is wrong.

4.3 The presence of symmetry in the fragment

If the presence of topological symmetry is detected, a dialog box will appear asking if the user wishes to apply a symmetry correction.

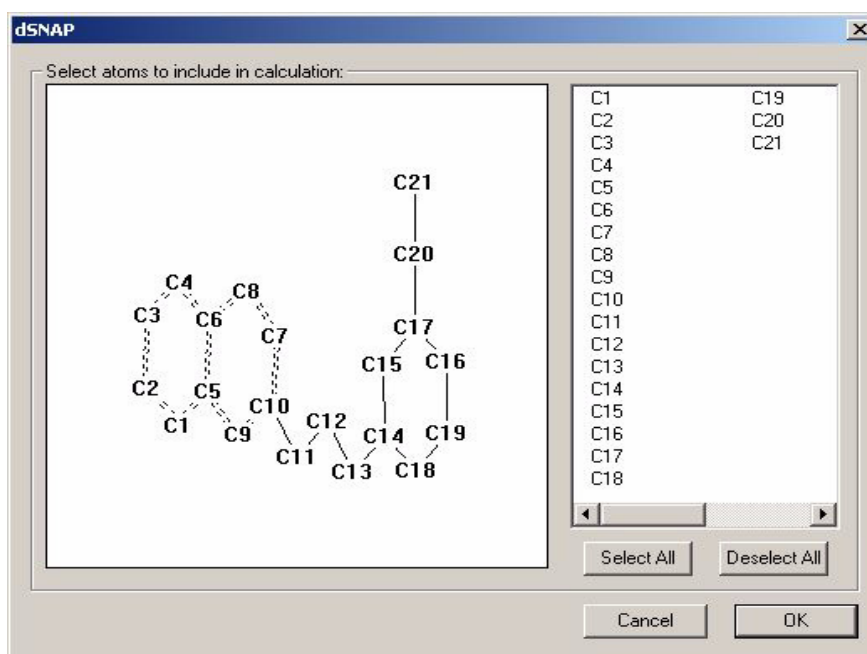


Clicking *OK* means that the symmetry correction is performed. Clicking *Cancel* will run the analysis but without applying a symmetry correction. Applying the correction is very strongly recommended.

For more information see the Chapter on symmetry.

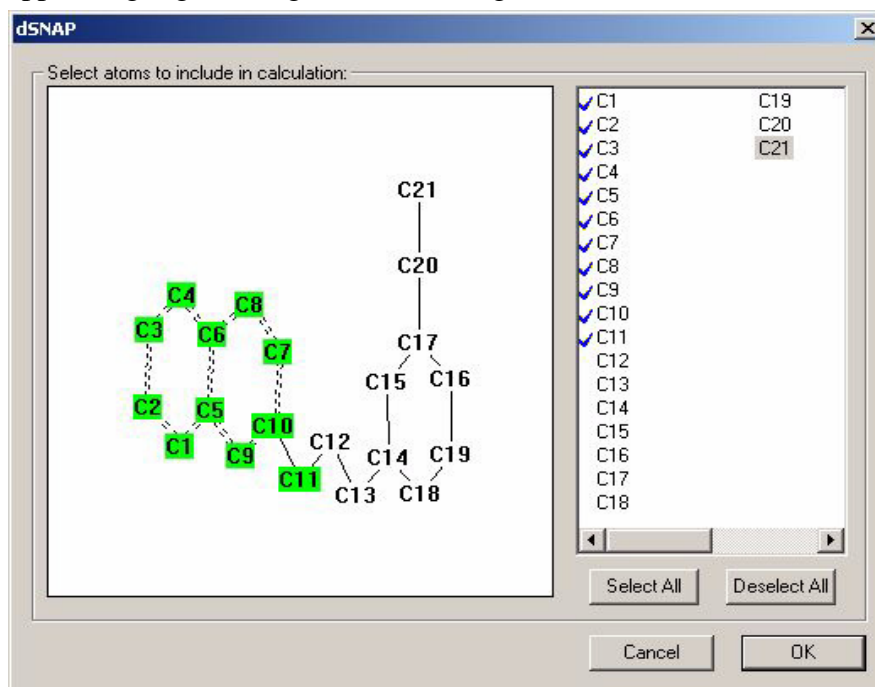
4.4 Selecting Atoms

If the number of atoms in the search fragment is greater than 20 the user will be given the opportunity to select which atoms, up to a maximum of 20, are to be used in the analysis. Before the analysis is performed a window will open with a diagram of the fragment.



All of the atoms are currently unselected. Atoms can be selected either by clicking on them in the list on the right-hand side or by

clicking on them in the diagram. Once atoms have been selected they appear highlighted in green on the diagram and are ticked on the list.

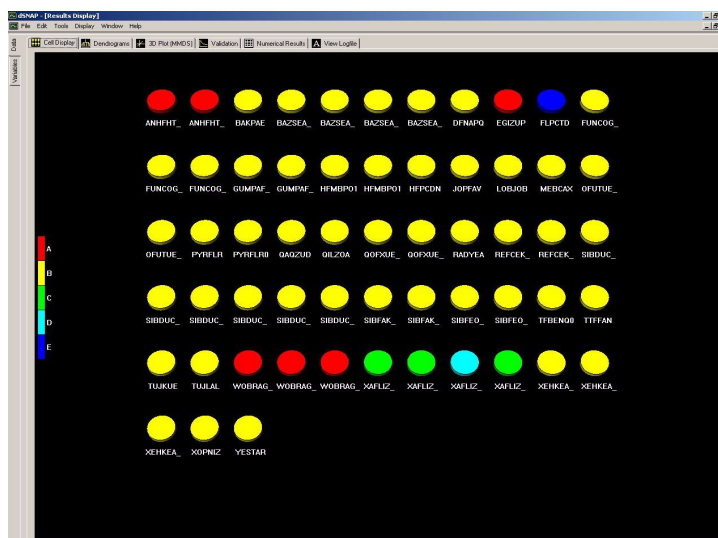


There are buttons that allow the user to *Select All* or *Deselect All* atoms. If the user tries to include too many atoms then analysis will not start and a dialog box will be displayed explaining that no more than 20 atoms can be specified. Once a suitable number of atoms have been selected clicking *OK* will start the analysis.

In cases where less than 20 atoms are present in the fragment this window will not appear by default and analysis will begin immediately. However a similar option is available for fragments with less than 20 atoms; see Chapter 12 on Advanced Options.

5.1 Results Display

Once the analysis is complete the display screen is loaded and results can then be examined in a number of ways.



The display area will initially open to present the *Cell Display* representation, but other display options can be accessed using the tabs running along the top of the window. The tab running vertically down the top left side allows access to two different sections of information, *Data* and *Variables*. *Data* space displays the relationships between whole fragments, and *Variables* space displays the relationships between the individual parameters of those fragments (*i.e.* angles and distances). By default the data section will be displayed first. The two spaces have highly similar results display options.

5.2 Display Options Overview

The various display modes from the tab bar are as follows:



Each tab opens a different results display pane:

Cell Display

The cell display provides a simple overview of the results by only displaying each fragment as a coloured disc, reflecting its cluster assignment. It is useful for obtaining an overview of the distribution of results. For more details see Section 6.1 and Section 7.2.

Dendrogram

The dendrogram is a chart that indicates the level of similarity between fragments. For more details see Section 6.2 and Section 7.3.

3D Plot (MMDS)

The 3D plot is a spatial representation of the level of similarity between fragments and is generated using a different method to that used in the dendrogram. For more details see Section 6.3 and Section 7.4.

Validation

The validation screen holds four displays, a *Scree plot*, *Silhouettes*, a *Parallel Coordinates plot* and a *Space Explorer plot*. All of these are used to assess the validity of the current classifications. For more details see Section 6.4 and Section 7.5.

Numerical Results

The numerical results pane displays the correlation matrix generated from the raw numeric data that was used in the cluster analysis. In Variables space it can be used for further analysis. For more details see Section 6.5 and Section 7.7.

View Logfile

A textual display that saves a record of the results of the file import, processing, clustering and any subsequent changes. For more details see Section 6.6 and Section 7.9.

5.3 General Display Features

The various graphical display screens all share many similarities in their options and controls, and are therefore initially described together here.

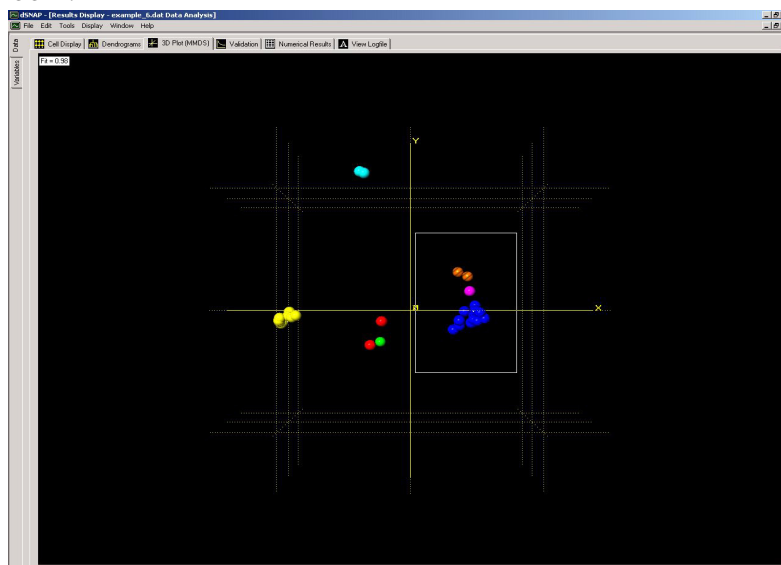
5.3.1 Display Controls Common to All Modes

There are two different modes for manipulating the display screens. Both perform the same functions, but the commands to perform these functions differ between the two. The *Standard* settings can be switched to the *PolySNAP-style* settings by de-selecting *Use Alternate Modifier keys to manipulate graphics displays* from the *Options* section of the *Edit* menu. By default, the *Standard* settings option is used.

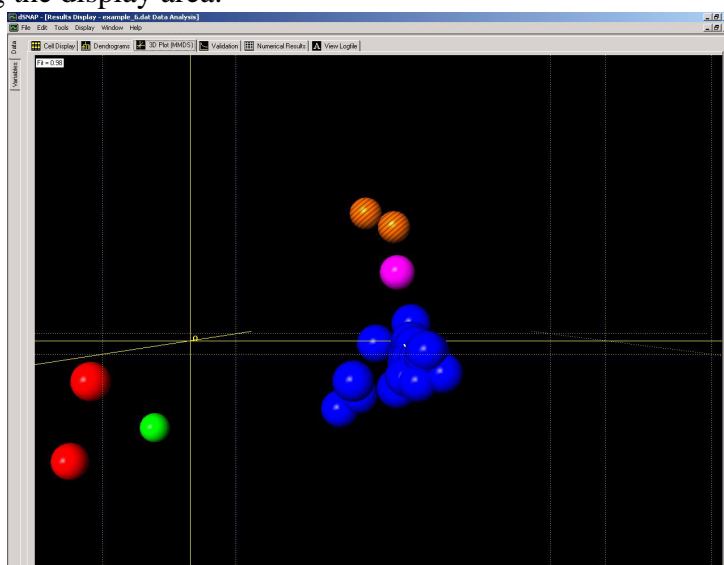
Below is a description of all the ways the user can interact with the display screens along with the relevant commands, in both modes.

Zoom

This allows the user to focus on an area of the graphical display. To zoom in on a region of a graphical display, with the *Shift* key and the left mouse button held down, drag a rectangle over region you wish to zoom:



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



Standard Command:	Hold down the <i>Shift</i> key and the left mouse button and draw rectangle
PolySNAP-style Command:	Hold down left mouse button and draw rectangle

Translate

This allows the user to move the contents of the display window, for example to move the contents up to see more results than will fit in the window by default. This is especially useful when the display has been zoomed.

Standard Command:	Hold down the <i>Ctrl</i> key and the left mouse button and drag display
PolySNAP-style Command:	Hold down the <i>Alt</i> key and the left mouse button and drag display

Rotate

The 3D displays, such as the *3D plot* or the *3D viewer*, can be rotated in 3D to show the display from all angles. These commands have no effect on 2D displays.

Standard Command:	Hold down the left mouse button and move mouse
PolySNAP-style Command:	Hold the <i>Shift</i> key and the left mouse button and move mouse

Change size

There are certain objects in some displays that can have their size altered to make the display easier to view. For example the data points on a plot can be made larger so they are easier to spot, or smaller if they are all grouped together in one place and are difficult to distinguish. This can be useful when there are lots of data being graphically represented, or when a display has been zoomed, or when producing diagrams for printing. As the objects been altered will vary depending on which visual display is being viewed, a fuller explanation of this feature can be found in the sections describing each display screen in turn.

Standard Command:	Hold down the <i>Alt</i> key and the left mouse button and move mouse
PolySNAP-style Command:	Hold down the <i>Ctrl</i> key and the left mouse button and move mouse

5.3.2 Additional General Options in the Graphics panes

These options can be accessed by right-clicking in the display area in any of the graphical displays. This brings up a pop-up menu where certain features can only be accessed for certain displays. However many features are common to all:

Reset View

This feature will return to the original view of the display if it has been moved or if zoom has been activated. This can also be done by using the *F5* function key.

Zoom In

Will zoom in on the centre area of the current display.

Zoom Out

Will zoom out from the centre of the current display.

Centre Selection

The currently selected item will be centred on the screen.

Deselect All

Any items that have been selected in the display will be deselected again with this option.

Find Item (Alt-F/Ctrl-F)

This features allows the user to search for a specific item, fragment or variable in a display. Clicking on this option, or using the shortcut-key after clicking in the display area, opens the following dialog box:



The item can be searched for either by using its unique label or by sequence number.

Note that when entering labels it is not case sensitive. Also, the search looks for the labels used in *d*SNAP, not the database refcodes. For example, if two instances of the hit fragment are found within the structure with the refcode HAFTEN, entering the refcode HAFTEN into a *Find Items* search will produce no results. The search will recognise only HAFTEN_01 or HAFTEN_02.

Objects Colour

This menu option is used on the *3D plot* to change the colour of a single sphere manually. It is described fully in Section 6.3. It has no function in other display screens.

Toggle Mode

This menu item is used to switch between different ways of presenting the data. This has different effects on different display screens and so is described in fuller detail in the sections describing those displays.

Show Axes/Grid

Certain display screens in *d*SNAP contain axes or a grid which can be displayed or hidden using this option. The displays where this option is applicable include the 3D plot, 3D fragment viewer, space explorer and scatter plots.

Show All Labels and Show Selected Labels

This option can be used either to hide or display the labels associated with the items. In particular this allows the user to bring up the labels if there are more items than the automatic cut-off for displaying labels by default.

Show Toolbar

The graphics toolbar that runs along the top of the display area can be toggled on or off at any time using this option. The contents of the toolbar are explained in Section 5.3.3.

Accelerate Wheel (Alt-A)

This option allows the user to increase the rate of movement that operating the mouse scroll wheel will achieve. This is useful for navigating several of the graphical displays where the user may frequently be required to zoom in on certain areas.

Print

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

Copy (Ctrl-C)

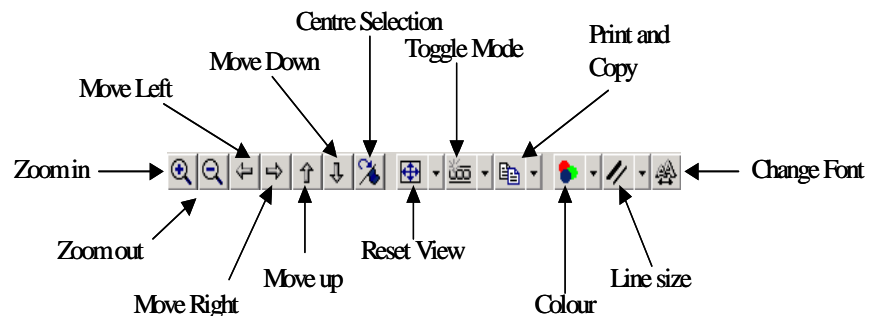
The whole of the current graphics display will be copied to the clipboard, and then can be pasted into any standard Windows program - for example *Microsoft Word*.

Copy Selection

This option allows smaller specific regions of the graphic display to be copied. Click on *Copy Selection*, and then drag a rectangle over the area to be copied; now only this area will be copied to the system clipboard.

5.3.3 The Toolbar

The basic functions of the toolbar are as detailed below:



Note that unless the default settings regarding the toolbar are altered by the user (Section 9.1.2) the toolbar will become hidden again once the user leaves the current display screen.

Many of the functions are exactly analogous to functions accessed *via* the right-hand menu.

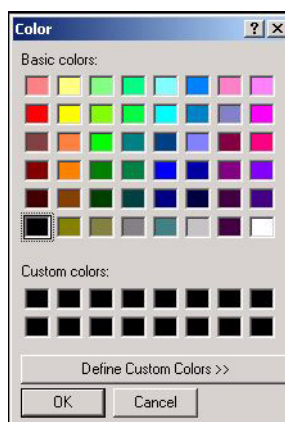
Clicking on the small arrow next to the *Reset View* button allows access to the options for both *Reset View* and *Deselect All*.

Clicking on the small arrow next to the *Toggle Mode* button allows access to the options *Toggle Mode*, *Toggle labels*, *Mask group* and *Accelerate wheel*. There are further options relevant to the *3D plot* available when using the toolbar on that display. The *Toggle Mode* feature itself operates slightly differently in each display mode and so is described in those specific sections.

Clicking on the small arrow next to the *Copy* button allows access to the options for the *Copy*, *Copy Selection...* and *To Printer...* which is analogous to *Print*.

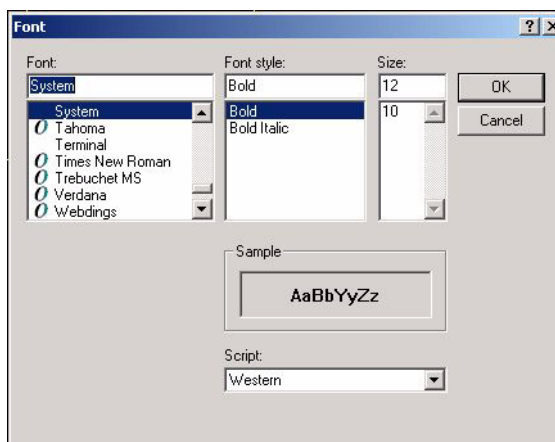
Clicking on the small arrow next to the *Colour* button allows access to the options *Background*, *Foreground* and, in the appropriate display modes, *Axes* and *Objects*. Note that the *Foreground* option refers to the lines on the dendrogram and the labels in the Scree plot.

Clicking on the first three options will open a colour palette from which a new colour can be chosen:



Clicking on the arrow next to the *Line Size* button opens a pull-down menu with a series of options for line size numbered 1 to 5, with 1 being the thinnest and 5 being the thickest. A dot appears next to the thickness currently selected. The line size options changes the thickness of lines on the graphical displays which may be useful when preparing images for publication. For suggestions on optimising the displays for print, see the Quick Guide *Preparing Graphics for Publication*, which is accessible from the *Help* menu.

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



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

Note that changes concerning colour and text settings take effect for all of the graphical display screens and remain in effect until the user closes the program. On reopening *dSNAP* all changes will be reset to the defaults.

There is one final toolbar option that is only available when viewing the *Space Explorer* and so is described in Section 6.4.4.

5.3.4 Selecting fragments on the graphics displays

In many cases it may be useful to select only certain items when performing certain functions, such as choosing fragments to visualise in a viewing program. There are several ways in which the user can select particular items from the displays.

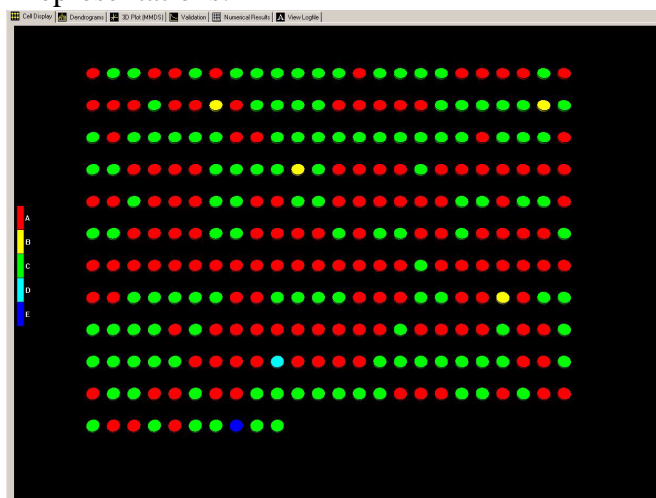
A single item can be selected by clicking on the graphical object that represents it. Multiple items can be selected at the same time by holding *Ctrl* while other items are selected. Also a range of items in sequence, for a example an entire cluster in the dendrogram display, can be selected by clicking on the first item in the series then holding down *Shift* and clicking on the last item in the series. Now all items in between will also be selected.

When a single item has been selected, the selection can be moved on to items on either side through the left and right arrow keys. This allows the user to scroll through related items easily. settings Note that this only works after a single item has been selected.

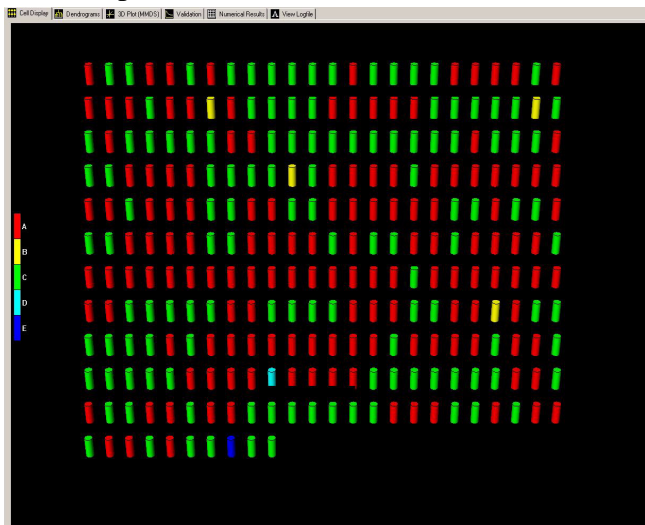
6.1 Cell display

This screen is the default graphics pane and arranges the fragments alphabetically in Refcode order, with fragments from cifs at the end. It is a simple overview of the data that displays the results by colour coding, with fragments belonging to the same cluster group having the same colour. This colour labelling is derived from the dendrogram. The Cell display is useful for easily identifying the number and variety of fragments in a single hit by displaying not only all the fragments in a hit together, but also if these hits have been assigned to the same or different clusters.

The cell display can be shown in either of two modes - as standard single cell representations:

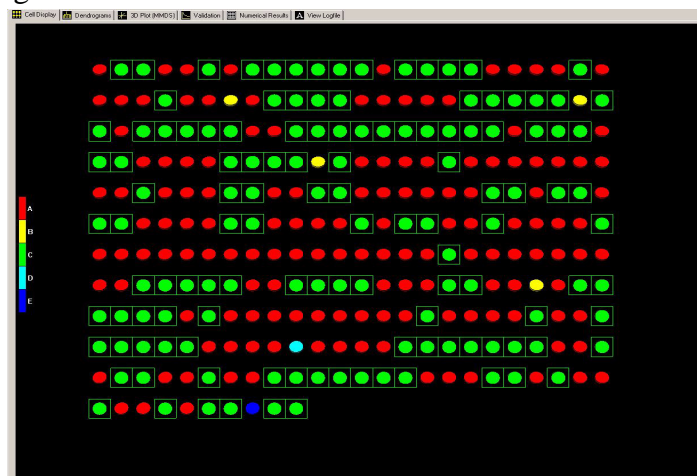


Or as column shaped 'stacks':



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

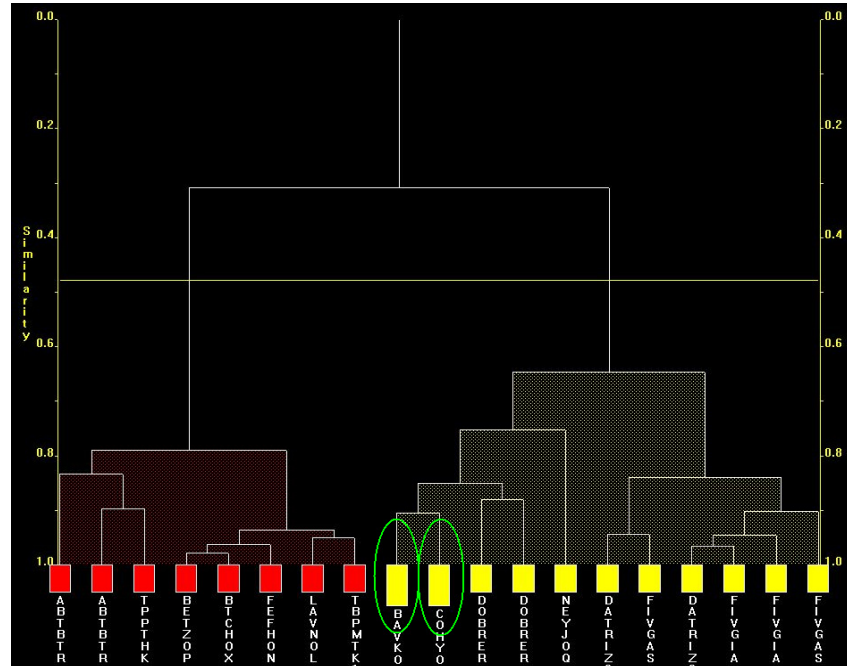
There is also a key on the left-hand side of the display with a legend to represent each cluster. By clicking on one of these the user can highlight every cell in that cluster which can be useful when dealing with larger data sets.



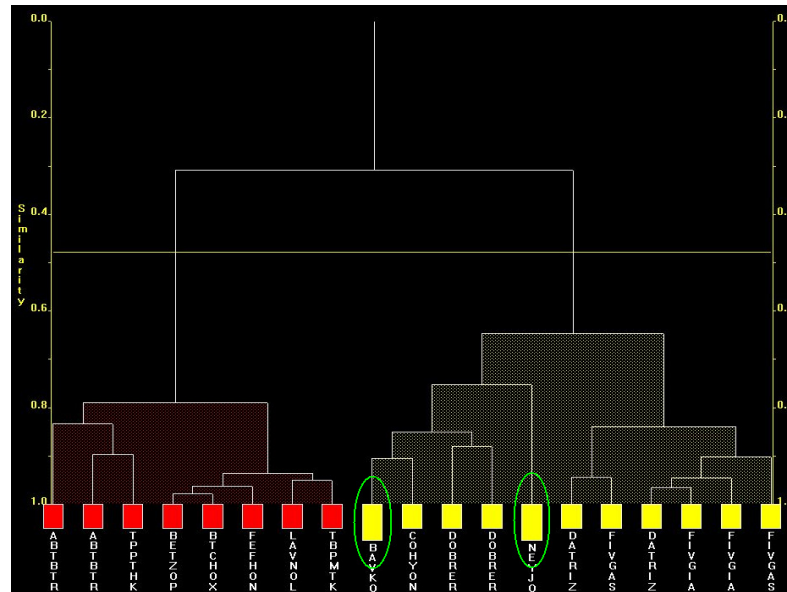
6.2 Dendrogram

The dendrogram display can be accessed by clicking on the Dendrogram tab along the top of the display window. It displays the results from the cluster analysis and is often the most important display in *dSNAP* as it visually presents the results in a hierarchical method of data classification. The dendrogram takes the form of a tree with each fragment represented by one of the boxes at the bottom of the screen.

A scale showing similarity is displayed on the vertical axis and the dendrogram's tree branches out according to the calculated similarity between the fragments that each branch connects to. For example, if two branches are joined by a horizontal bar (called a tie-bar) nearer the bottom of the dendrogram then the fragments associated with those branches can be considered to be highly similar and can be justifiably grouped together. However if two branches do not meet until nearer the top of the dendrogram then the associated fragments are less similar and more loosely related to each other. An example of this is given below:

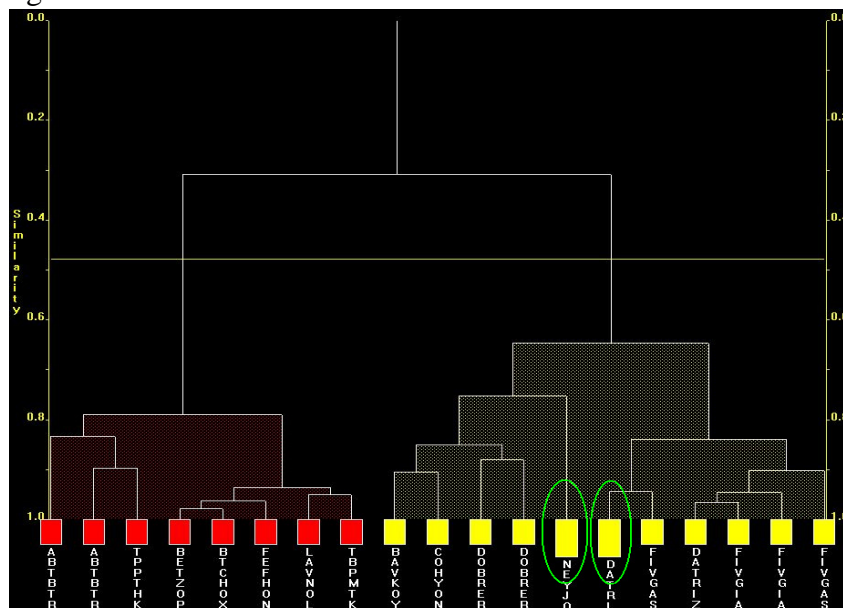


The two fragments that are selected (indicated by the larger boxes and circled in this picture) show two fragments with a tie bar low down in the dendrogram, and therefore they are very similar and belong to the same cluster.



Here the two highlighted boxes are joined by their tie-bar at a higher level of similarity, and both fragments belong to the same cluster, but are far less similar than the previous two.

It is important to remember that it is the tie-bar that indicates the level of similarity. The two fragments highlighted below appear next to each other on the dendrogram; however they are connected by a higher tie-bar and are in fact less similar to each other than the two fragments highlighted above, which are separated by three other fragments.



A manually adjustable cut-level, which decides how the dendrogram is split into separate clusters, is shown as a horizontal line which spans the width of the display.

Unlike the cell display screen, in the dendrogram the fragments are arranged according to cluster, with the most similar fragments appearing next to each other and all fragments in a given cluster identically coloured. This representation allows quick comparison of the different types of fragments and the level of similarity between fragments both within an individual cluster or in the dataset as a whole.

6.2.1 Navigating the dendrogram

The dendrogram can be navigated like all graphical displays, but there are several other controls that are specific to the dendrogram.

Standard commands:

Operation	(Click on dendrogram followed by...)
Adjust cut-level	Use mouse scroll wheel <i>or</i> Hold <i>Alt</i> and drag with left mouse button
Zoom in/out	Hold <i>Shift</i> and draw area to be zoomed <i>or</i> Hold <i>Ctrl</i> and use mouse scroll wheel
Translate horizontally	Hold <i>Shift</i> and use mouse scroll wheel
Move dendrogram	Hold <i>Ctrl</i> and drag with left mouse button

PolySNAP-style commands:

Operation	(Click on dendrogram followed by...)
Adjust cut-level	Use mouse scroll wheel <i>or</i> Hold <i>Ctrl</i> and drag with left mouse button
Zoom in/out	Draw area to be zoomed <i>or</i> Hold <i>Ctrl</i> and use mouse scroll wheel
Translate horizontally	Hold <i>shift</i> and use mouse scroll wheel
Move dendrogram	Hold <i>Alt</i> and drag with left mouse button

Remember that at any point the view of the dendrogram can be returned to its original state by right-clicking in the screen and selecting the *Reset View* menu item.

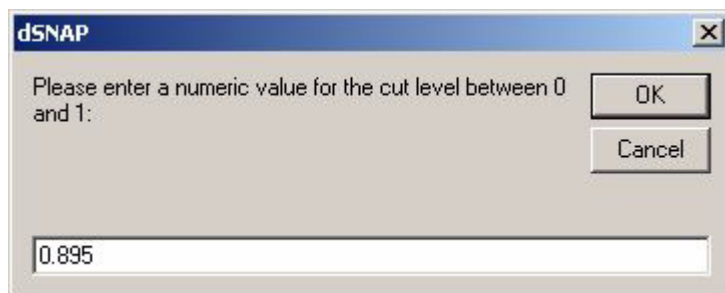
6.2.2 Modifying the dendrogram

-Changing the cut-level

If the initial program calculated cut-level is not considered to be suitable, the user may choose to adjust it. This may frequently be necessary, as the choice of cut-level will depend on the kind of problem being examined. The user is encouraged to investigate the effects of varying the cut-level for all analyses. As mentioned above,

this may be done in two ways, either by scrolling the mouse wheel or holding *Alt* and holding the left mouse button while dragging the mouse up or down.

The cut-level can also be accurately set to a specific numerical value by selecting *Set Cut-Level to...* from the *Tools* menu. This opens a pop-up box where the user can enter a new value between 1.0 and 0.0.



This value corresponds to the vertical similarity scale on the dendrogram. Once the new value has been entered the new cut-level can be set by clicking on *OK*.

The dendrogram cut-level can be returned to its original position at any time by selecting *Undo Saved Dendrogram Modifications...* from the *Tools* menu.

There are three other options that affect the appearance of the dendrogram and that can be accessed by right-clicking on the dendrogram display.

-Mask group

When toggled on this option will hide all of the fragments except for the fragments belonging to the cluster selected. If fragments from several clusters are highlighted when this option is selected then this mode is activated for the selected cluster at the left-most of the screen. The dendrogram branches only acknowledge the remaining fragments.

-Toggle Mode

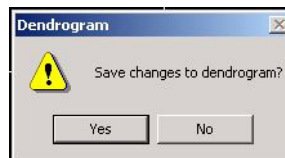
When selected Toggle Mode will only show the first, middle and last fragments of the cluster. This simplified view can be useful when dealing with a congested dendrogram of many different clusters. Note that the dendrogram cut-level cannot be adjusted in this mode.

-Show Axes

Selecting this menu item allows the user to hide or display the yellow similarity axes that run along either side of the dendrogram. This does not affect any of the other dendrogram functions.

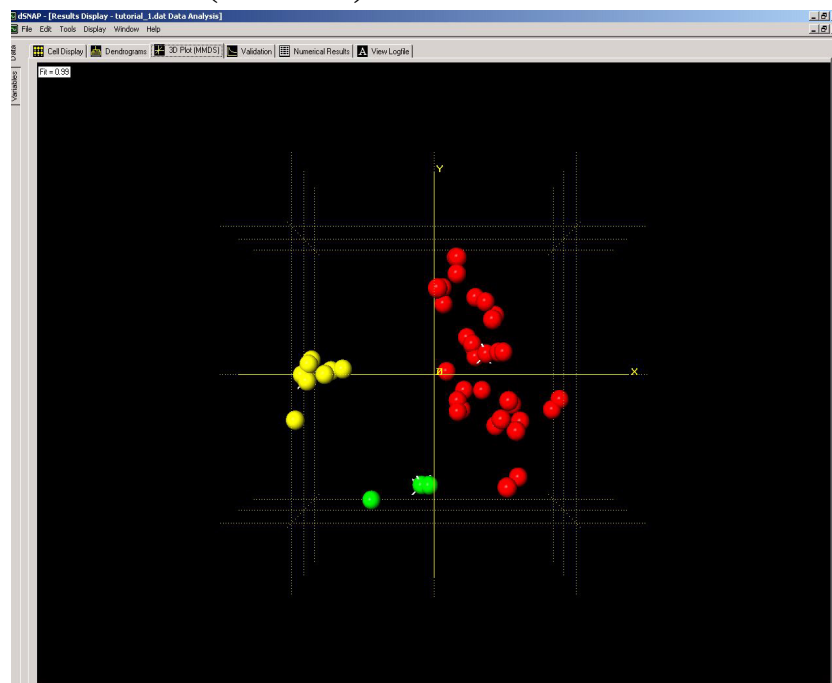
6.2.3 Saving Dendrogram modifications

Whenever the user attempts to leave the dendrogram screen after carrying out modifications a dialog box will ask if the user wishes to save any changes.



If the user selects *Yes* then the changes are retained, though still reversible using the *Tools* menu, and the Cell display, 3D plot and Logfile are all updated accordingly. If the user selects *No* then the changes are discarded, and the next time the dendrogram is opened it will have reverted to the last saved version.

6.3 3D Plot (MMDS)



This screen shows the plotted results of performing metric multidimensional scaling upon the distance matrix. The fragments are shown as a series of spheres arranged in 3D space. The similarity is now represented by the spatial distance between spheres in the plot. The closer two spheres are together, the more similar the

fragments are. Strongly related fragments tend to group together in the 3D plot and form spatial clusters. Therefore this screen gives a second way to visually interpret the level of similarity between fragments both within a cluster and within the dataset as a whole.

As with the cells in the cell display, by default the 3D spheres will take the same colour coding for the fragments as defined by the dendrogram. This is to allow for easy comparison of the two methods. Should they be consistent with each other then identically coloured spheres should cluster together. If the dendrogram cut-level is adjusted, the colour of the spheres in the 3D plot are updated accordingly.

A small numerical label in the top-left corner of the display gives an indication as to the goodness of fit of these results. This is computed as a correlation coefficient between the observed distance matrix and the derived calculated one. Numbers close to 1.0 suggest that it is a good fit, and low numbers suggest that caution may be required, or that the program had trouble adequately partitioning the data. In general high scores (over 0.9) can be expected although as the number of samples increases, the average GOF score, in general, will decrease. There is no set value for when a plot becomes unreliable. It is at the discretion of the user to decide when to exercise caution.

6.3.1 Display controls

There are several additional display controls that are specific to manipulation of the 3D plot.

Standard commands:

Operation	(Click on 3D plot followed by...)
Zoom in/zoom out	Hold <i>Shift</i> and draw area to be zoomed <i>or</i> Hold <i>Ctrl</i> and scroll mouse wheel
Plot rotation	Hold left mouse button and move mouse as required
Plot translation	Hold <i>Ctrl</i> then hold left mouse button and move mouse as required
Change sphere size	Hold <i>Alt</i> then hold left mouse button and move mouse up/down

PolySNAP-style commands:

Operation	(Click on 3D plot followed by...)
Zoom in/zoom out	Draw area to be zoomed <i>or</i> Hold <i>Ctrl</i> and scroll mouse wheel
Plot rotation	Hold <i>Shift</i> then hold left mouse button and move mouse as required
Plot translation	Hold <i>Alt</i> then hold left mouse button and move mouse as required
Change sphere size	Hold <i>Ctrl</i> then hold left mouse button and move mouse up/down

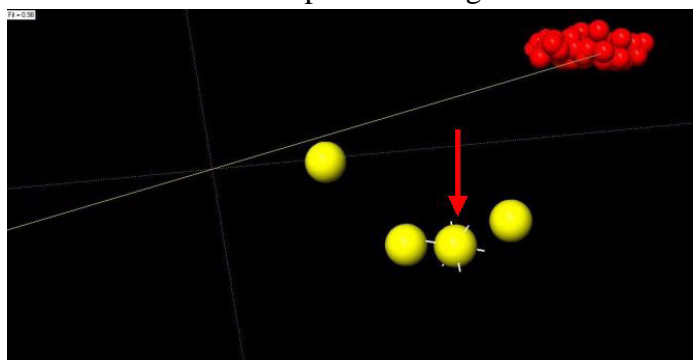
Furthermore there are a few more pop-up menu options that are specific to the 3D axes and grid plot:

Show Grid/Axes

The user can decide to have the each of the 3D grid and axes hidden or visible as they view the plot. As a default both will be displayed.

Show MRM marks

The most representative member of a cluster is the member with the minimum mean distance to all other members in that cluster. For this reason they can only be calculated for clusters with at least three members. The most representative member in each cluster can be highlighted if required with this option. It appears on the display as a normal member with several 'spikes' coming out of it:

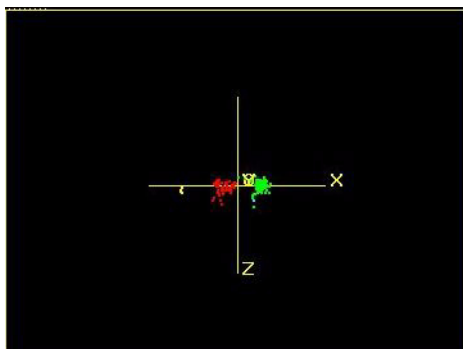


These spikes can be hidden or shown by means of this menu option. Clicking on a MRM sphere opens a dialog box containing information about the mean member-member distance for that particular cluster. The smaller the distance, the tighter the cluster:



Show Top view

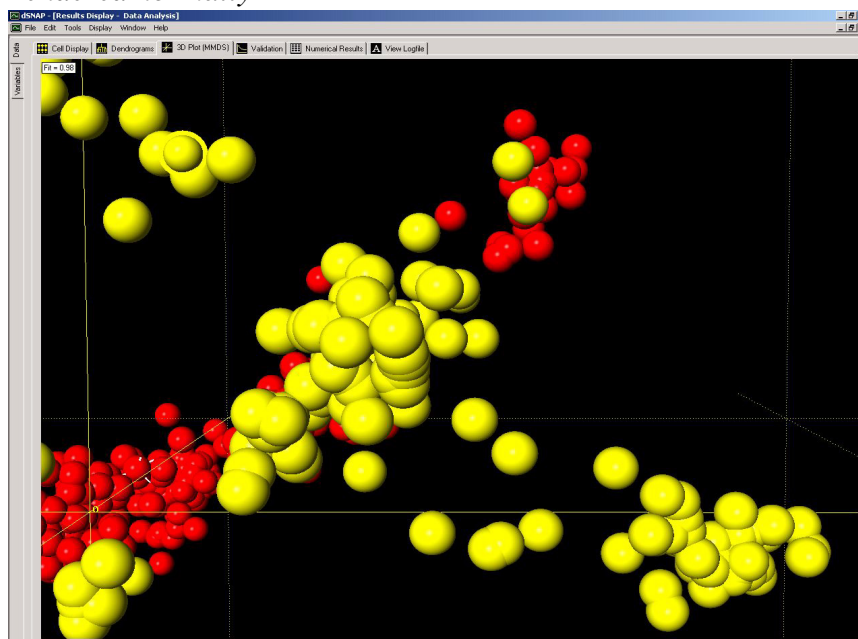
This option brings up a small simplified overview of the plot in the lower right hand corner. It can be useful for orientating yourself when zoomed into the display.



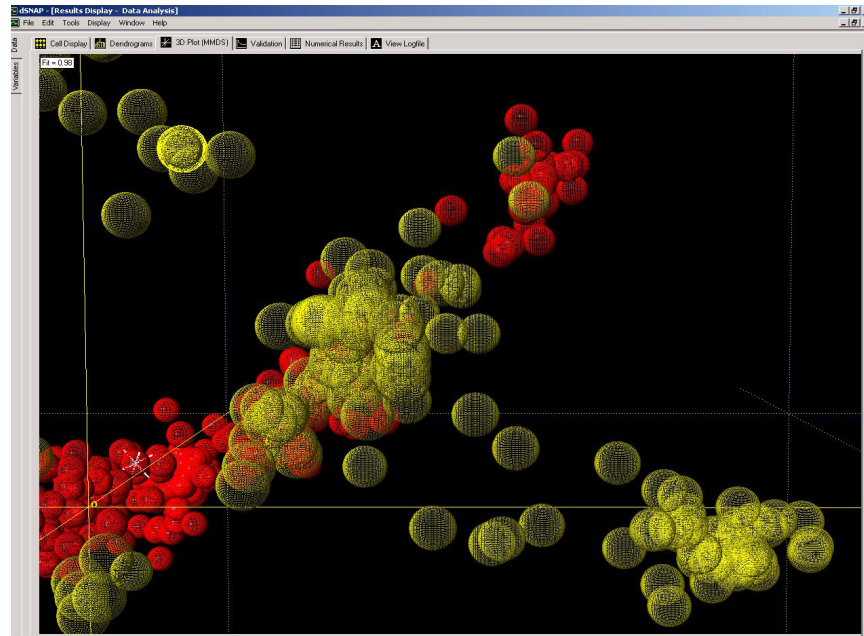
Render as dots and Transparent

Both of these menu items alter the way in which the spheres are plotted as shown in the diagrams below. This can be useful if, for example a single fragment of one colour is hidden within a larger group of other fragments.

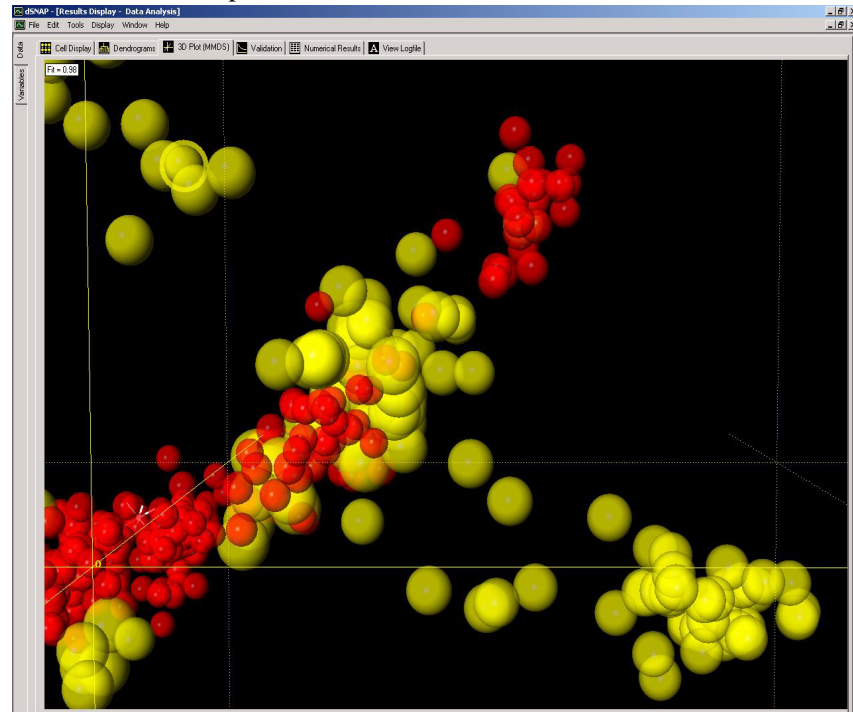
Rendered normally



Rendered as dots

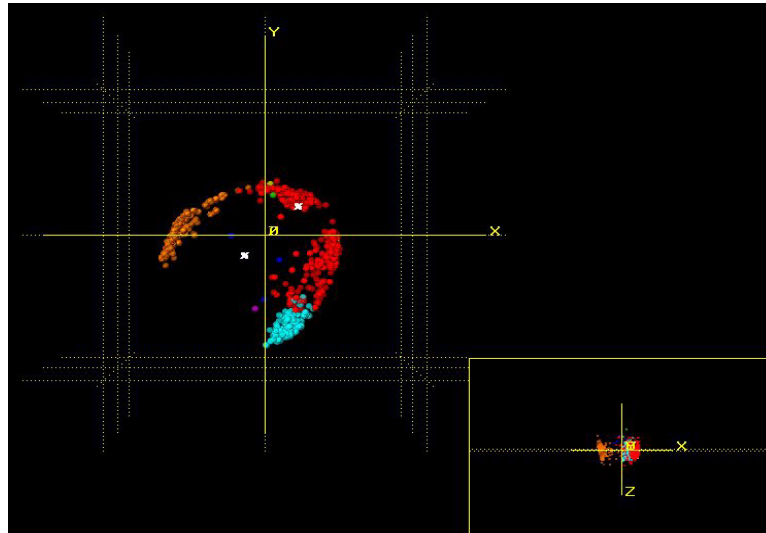


Rendered as transparent



Opaque zone

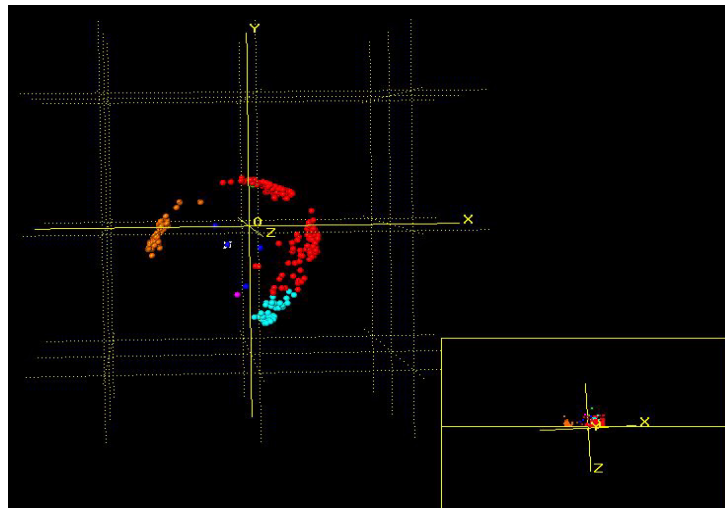
This option is available when *Transparent* mode has been activated. It creates a zone where all spheres are rendered as opaque again. This can be seen most clearly by activating the *Top View*.



The location of the opaque zone is shown as a dotted yellow line dissecting the top view display and is not adjustable. The 3D plot must be manoeuvred to move spheres in or out of the zone as required.

Clip plane

Selecting this option creates an invisible plane in the plot. Again this is best visualised using the Top View mode.

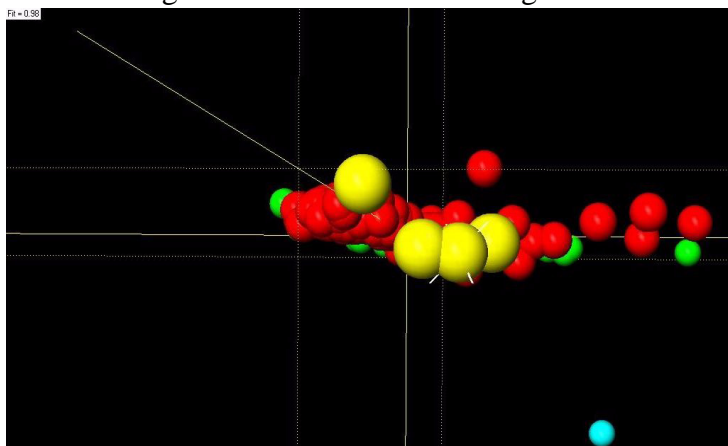


The clip plane is shown as a solid yellow line dissecting the top view display. Any spheres above this plane are displayed on the plot, while any underneath it are hidden.

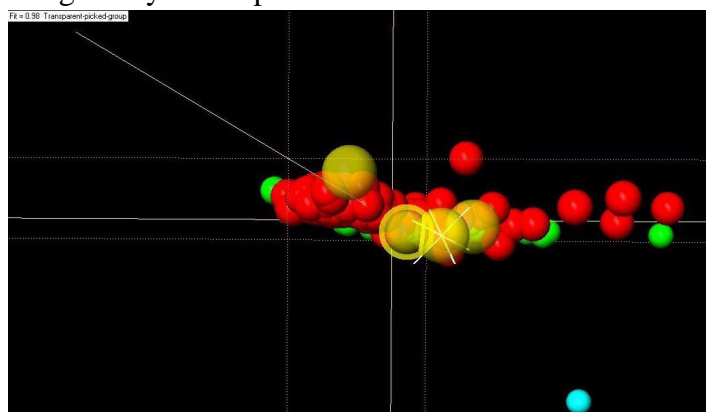
Mask Group, Popup group and Transparent Group

These menu items are demonstrated in the screenshots below:

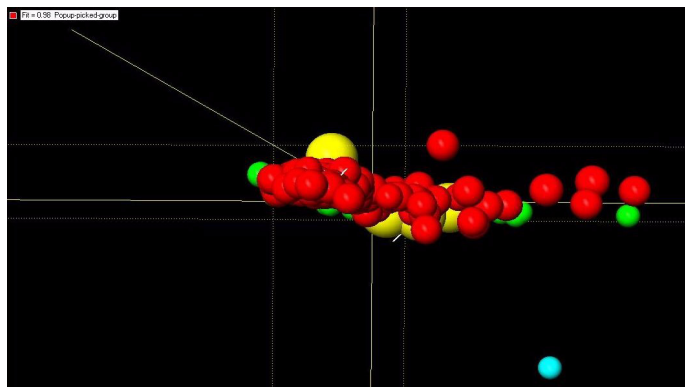
Initially none of these menu items have been selected and the yellow cluster is obscuring some of the red cluster fragments.



In the case below the *Transparent Group* option has been used on the yellow cluster. This allows the user a partial view of the red spheres that lie behind the yellow cluster however it will not allow the user to click through the yellow spheres.

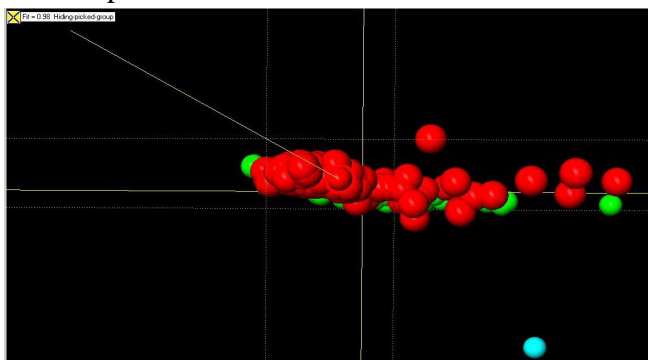


In this case the *Pop-up Group* option has been used on the red cluster. It has now been brought to the foreground and appears in front of the yellow cluster.



In this case the *Mask Group* option has been used on the yellow cluster. It has not been deleted, deselecting this option will cause the

yellow cluster to reappear, however it is temporarily no longer displayed on the plot.

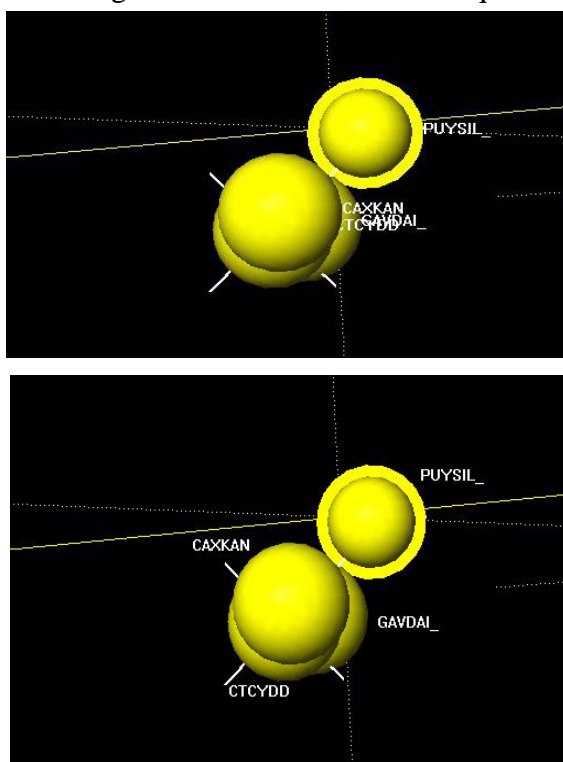


Pressing *space* will toggle if just the selected cluster is masked, or if all other clusters are masked retaining only the selected cluster on the display. The right and left arrow keys can be used to change the group that is currently selected.

Note that the box that displays the fit in the top left corner of the screen will also change to display information on any of the options being used.

Drag label

This menu item is only available if the user has chosen to have the fragment labels displayed. By selecting this menu item all of the fragment labels on the 3D space can be moved individually by clicking on them and dragging to the required new location. This allows for neater diagrams to be created when required for a report.

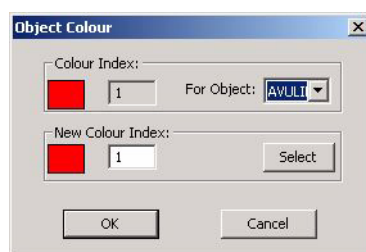


With this option toggled on, the display is fixed and cannot be rotated or zoomed. Therefore it is necessary to use the normal zooming and rotation controls to select a suitable view angle before selecting this option.

Note that when the Drag Labels option is deselected, the labels will move back to their original positions. While in Drag Labels mode, screen shots can be made to use in publications and other printed documents.

Objects colour

While the colours for the 3D plot are normally taken from the dendrogram this option allows the user to change the colour of certain spheres on the 3D plot. Selecting this option opens a new window.

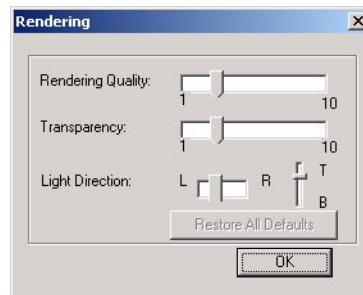


The user can then select the fragment of interest by its label and select its new colour by entering a new colour index. This choice is then confirmed by clicking *Select*. This can then be repeated for other items. The changes do not update until the user selects *OK*. If an error has been made then clicking *Cancel* will cancel the changes. Note that these changes are temporary, they will not update the dendrogram and if the user selects another display then returns to the 3D Plot the changes will have been reset.

This can be used to highlight a fragment for publication purposes or as a more permanent alternative to the *Find Item...* option which stops displaying the location of a fragment as soon as the mouse is moved again. Although *Object colour* changes are reset by leaving the 3D plot it can be used to keep a fragment highlighted while the user rotates, zooms or navigates the plot.

Adjust Rendering

By clicking on the display and pressing *F12* a new window will be opened that allows the user to adjust the drawing quality of the spheres in the plot.

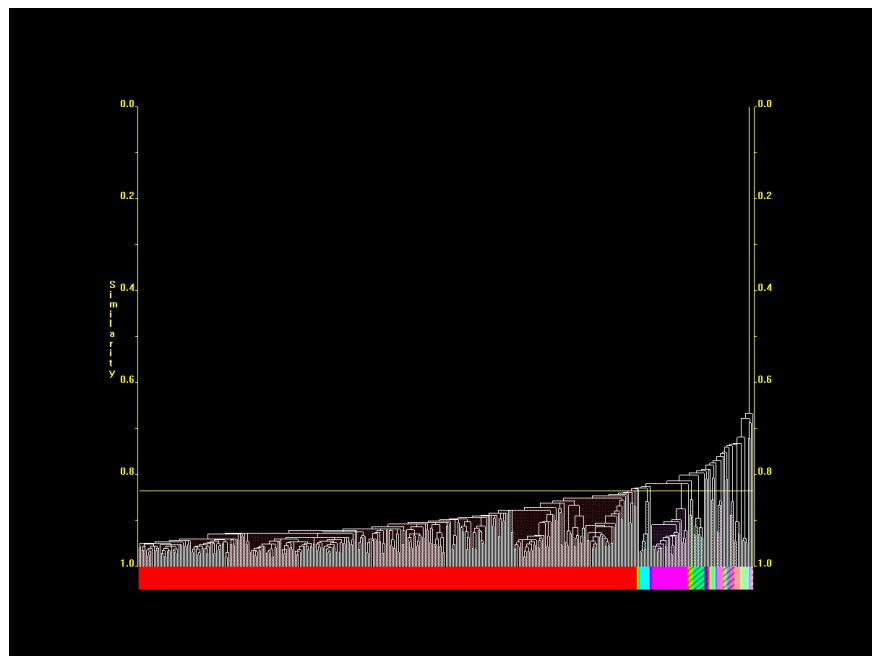


Working with the display can be much faster if the rendering quality is reduced. This is especially the case when dealing with a combination of a low-powered graphics card and a large data set. Equally, increasing the quality may be desirable when producing diagrams for publication purposes.

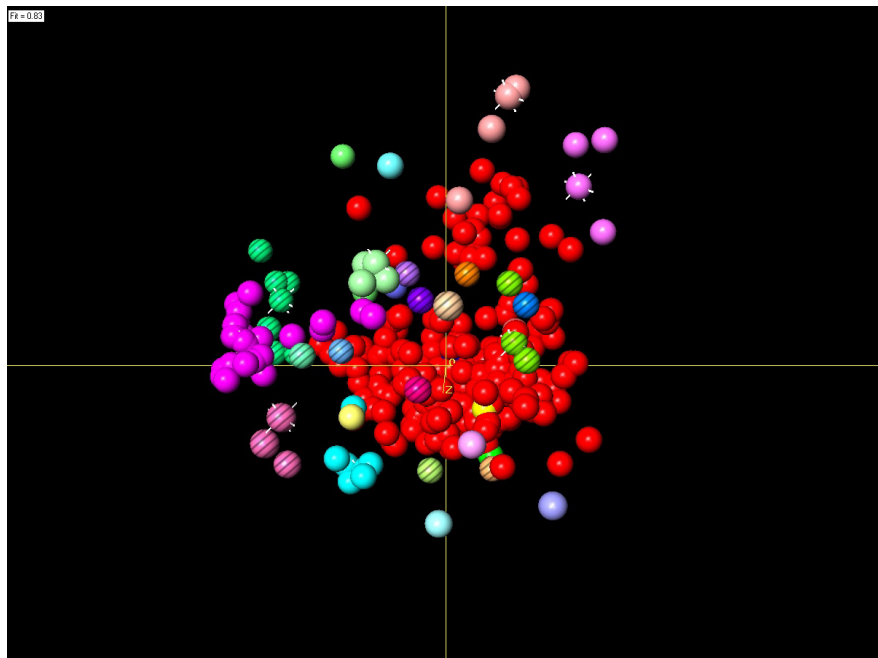
Other drawing properties such as transparency and the set up of the lighting can also be changed to suit the user preference.

6.3.2 Poor clustering

In some cases a dataset may be input into *dSNAP* where there is no real clustering among the structures. In this case the dendrogram may exhibit *chaining*, where every fragment is linked to the next by a slightly higher tie bar:



The 3D plot for this type of dataset is typically characterised by a random, orderless distribution of spheres where no signs of clusters are formed.



In these cases it is possible that the set of structures provided to *d*SNAP cannot be separated into sensible groupings.

6.4 Validation

This tab incorporates four different displays in the one frame.

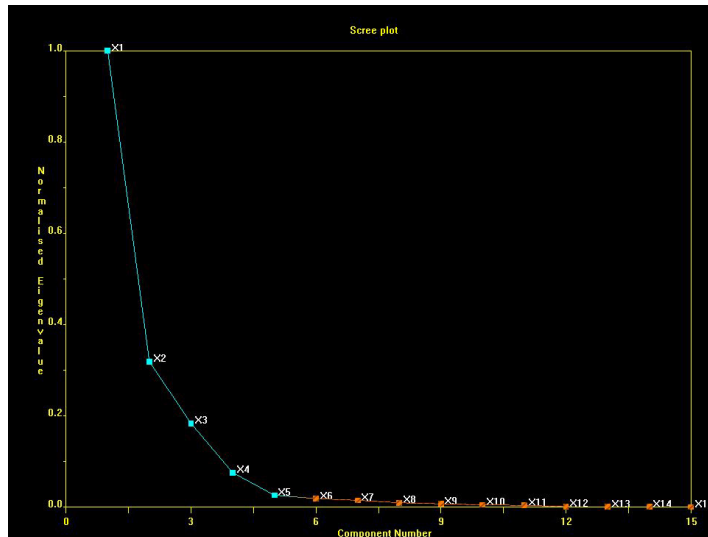


The four displays can be switched between using the pop-up menu in the top left corner. The items here provide further checks on the consistency of the results from the other methods.

6.4.1 Validation > Scree Plot

The Scree plot provides a separate graphical analysis based on principle components analysis, in order to help estimate how many separate groups of fragments are needed to suitably describe the data. Unlike some of the other graphical displays it does not attempt to describe which fragments are likely to belong to each different group.

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



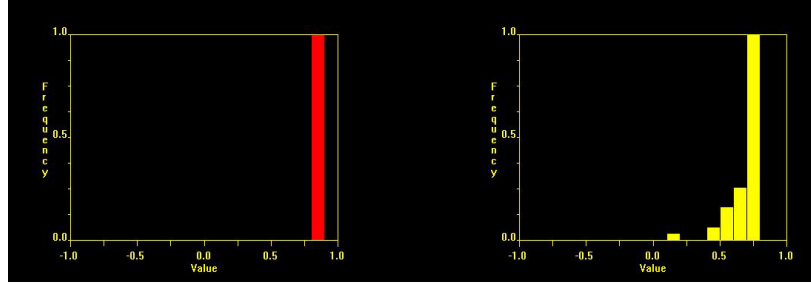
Eigenvalues are derived from the modified correlation matrix (which is first normalised) and are sorted in descending order.

What this represents is the minimum number of clusters that can be used to describe the entire data set being examined. The point where the plotted line changes colour - e.g. between 5 and 6 in the example above, suggests that just 5 clusters are needed to explain over 95% of the variation in the data. A well behaved Scree plot should have a reasonably steep initial descent. A gradual slope or stepped descent indicates difficulty in establishing the number of clusters required, so the program-generated dendrogram cut-level should be examined especially closely.

As with all display screens the Scree plot can be zoomed and moved around in the usual manner. The size of the points on the Scree plot line can be adjusted by holding *Ctrl* (if the alternate command settings are being used then *Alt* is held instead) and moving the mouse up or down with the left mouse button also held. This may be useful for publication purposes.

6.4.2 Validation > Silhouettes

For each of the current clusters as defined by the dendrogram cut-level, this display shows a histogram.



Silhouettes¹ provide an alternative formalism for assessing the compactness and isolation of clusters, and also for identifying those members of a given cluster which are either well established members of the core cluster, or outlying and thus potentially problematic. If the i -th fragment belongs to cluster C_r then we define the silhouette, $s(i)$ as follows:

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

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

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

where there are n_r fragments in cluster r , n_s fragments in cluster s , and d_{ij} is the distance between fragments i and j . The values of $s(i)$ lie between -1 and $+1$.

Therefore each fragment assigned to a particular cluster is given a score from -1.0 to 1.0 , defining its membership of that cluster. Scores nearer 1.0 suggest a strong membership, while scores nearer 0 suggest a weak membership.

1. Rousseeuw, P.J. (1987). *J. Computation & Appl. Math.*, **20**, 53-65.

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



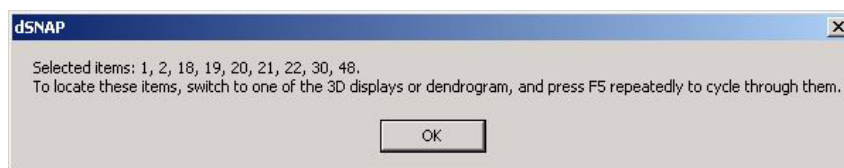
To scroll down to see the next row of histograms, click on the arrow, or use the scroll wheel on the mouse. Similarly, an upwards pointing arrow will appear if there are histograms above the top of the screen.

Allowing the mouse to hover over a particular column shows which fragments correspond to that range of silhouettes. The information provided by the silhouettes should correlate what is seen in other cluster graphics panes. For example a member with a low membership score will often appear more loosely connected to a cluster in the dendrogram, while members with high memberships should be far more representative.

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

Silhouettes can be useful when attempting to determine where to place a cut-level, and if a particular fragment should be included or excluded from a certain cluster.

Clicking the left mouse button on a silhouette will bring up a new window displaying a list of the fragments that are members of that cluster.

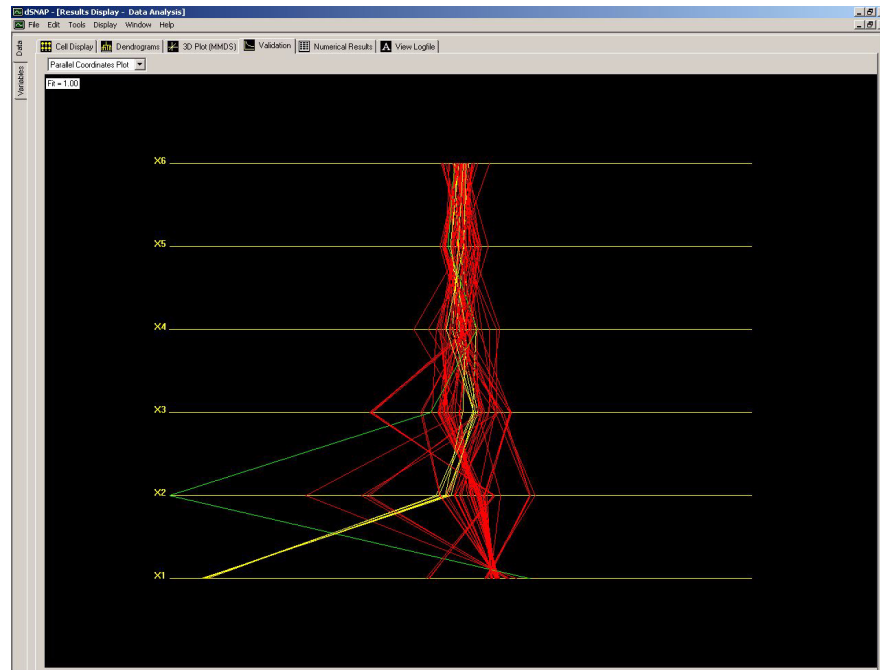


Now that they have been highlighted in this way they can be identified by returning to either the dendrogram or 3D plot and pressing *F4* to scroll through them.

6.4.3 Validation > Parallel Coordinates Plot

Rather than plotting our patterns in a 3D space, here we plot the first 6 dimensions of the clusters in a linear fashion, allowing the user to

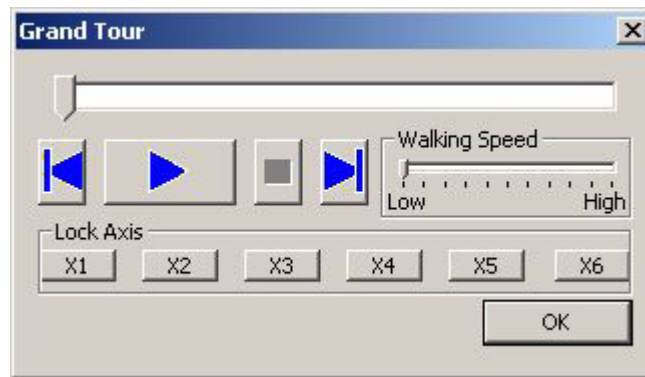
see if the cluster separations evident in the first 3 dimensions still hold together in higher dimensional space. Whereas the first three dimensions are plotted as x, y and z axes on a 3D plot here they are plotted as the first, second and third vertical axes on the plot, with the fourth, fifth and sixth dimensions plotted above.



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

The colours are taken from the data space Dendrogram so it is easy to identify the separate clusters. By right-clicking and selecting *Grand Tour* from the pop-up menu, the display can be animated by rotating

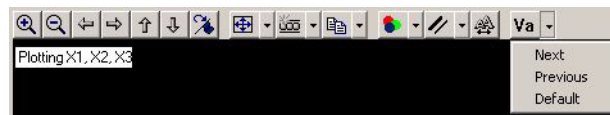
round the different axes simultaneously. This can also be accessed by using the *Ctrl-G* shortcut.



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

6.4.4 Validation > Space Explorer

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



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

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

Plotting X2, X4, X6

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

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

6.5 Numerical Results

This tab displays the correlation matrix generated from the raw numeric data that is used as the basis for the cluster analysis. There are further features in Variables mode which are explained in Section 7.8.

The numerical display screen will normally contain more information than can be fitted on the screen. It can be navigated by use of the scroll bars and will always display the refcodes of the fragments.

A value of 1.0 signifies a perfect 100% correlation, while a 0.0 value signifies a 0% correlation. The closer to 1.0 the better the correlation between the two fragments, and the more similar they are. Negative correlations are also possible, so negative values going to -1.0 are also included. A negative correlation of -1.0 signifies a 100% anti-correlation.

6.6 View Logfile

This tab displays all of the textual output written to the *SNAPlog.txt* file during the importing, processing and analysis undertaken by the program. It is also updated automatically if the user manually adjusts any of the default *dSNAP* results (*e.g.* alters the cut level in the

dendrogram) and therefore provide an audit trail of what has been done.

```

C:\Data\  Debuggers  3D Plot (MNDOS)  Validation  Numerical Results  View Logfile
No additional frequent geometries added to the hit list.
A distance matrix of dimension ( 274, 274) has been generated from an input data matrix of  274 rows and  75 columns
Labels are present on the input file.
When calculating the matrix, lambda =  2.0
Equal weights are applied to all columns.

Metric multidimensional scaling for 274 hits

All eigenvalues less than 0.0 are set to 0.0
There are approximately 5 clusters

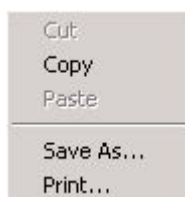
n Eigenvalue Cumulative proportion Change
1 5.23097 0.81261 0.00000
2 0.21129 0.86409 0.73605
3 0.22284 0.89866 0.23805
4 0.17297 0.92553 0.22274
5 0.10924 0.94686 0.24212
6 0.07964 0.95824 0.39165
7 0.05767 0.96720 0.27595

The Pearson correlation coefficient between observed and calculated distance matrices is 0.987
The probability of this occurring by chance is 0.000
The Spearman correlation coefficient between observed and calculated distance matrices is 0.977
The probability of this occurring by chance is 0.000
MNDOS complete.
Current cpu time is 15.62s
Dendrogram generation for 274 hits

Program will use the group average link method.
Top 99 amalgamation steps:
Clusters joined Similarity level Minimax distance Number of clusters Sum of squares Similarity n1 n2
78 71 96.1880 0.0380 99 273.089 96.20000 3 1
44 273 96.1880 0.0382 98 268.320 96.20000 5 1
28 79 96.1122 0.0387 97 263.794 96.18800 16 1
42 43 96.1030 0.0390 96 258.914 96.13125 1 1
123 124 96.1000 0.0390 95 253.716 96.10001 1 1
150 247 96.1000 0.0390 94 247.298 96.10001 1 1
204 225 96.1000 0.0390 93 243.116 96.10001 1 1
1 4 96.0998 0.0390 92 238.428 96.10001 67 7
23 171 96.0960 0.0390 91 233.749 96.09978 10 5
14 25 96.0800 0.0400 90 228.593 96.09600 1 2
60 137 96.0800 0.0400 89 225.313 96.09000 1 1
28 93 95.9796 0.0403 88 221.941 96.09000 17 3
64 154 95.9525 0.0404 87 218.078 95.97559 4 2
7 190 95.9500 0.0405 86 214.508 95.96250 2 1
2 115 95.9214 0.0408 85 211.724 95.95000 7 2
76 154 95.9000 0.0410 84 207.209 95.92143 4 1
123 249 95.9000 0.0410 83 203.080 95.90000 1 1
213 221 95.9000 0.0410 82 198.522 95.90000 3 1
105 241 95.8667 0.0413 81 194.477 95.90000 3 1
105 187 95.8333 0.0417 80 191.078 95.86667 3 1
1 53 95.8167 0.0418 79 188.264 95.83333 74 3
1 76 95.7797 0.0422 78 184.058 95.81622 77 5
217 235 95.5500 0.0435 77 180.063 95.77978 82 2
86 206 95.4500 0.0435 76 175.187 95.65386 2 2
28 14 95.4347 0.0437 75 172.337 95.65000 2 2
8 45 95.6184 0.0439 74 169.110 95.65000 20 4
18 231 95.6800 0.0440 73 164.044 95.62644 11 7
18 231 95.6800 0.0440 72 161.704 95.61033 3 1

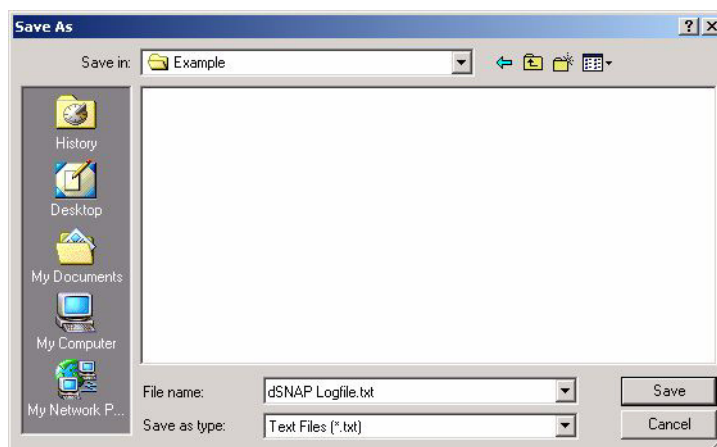
```

A scroll bar appears on the right hand side of the pane to allow the user to view all of the text, as the output is normally quite long. The text can be selected, copied to the clipboard and pasted to other applications by selecting the relevant text to highlight it and then right-clicking in the text pane:



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

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



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

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

There is a separate logfile for both the subject space and variables space.

6.7 Visualising structures

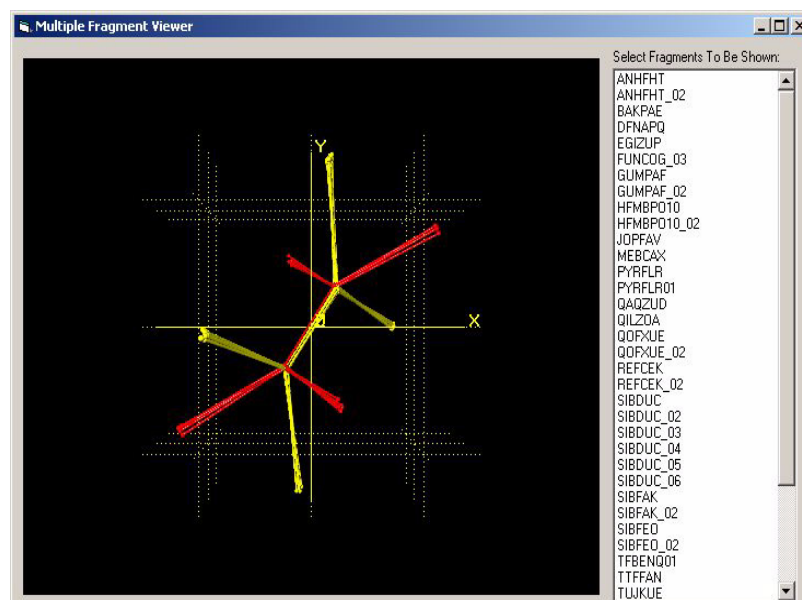
In order to relate the classifications being presented by *dSNAP* back to underlying structural chemistry it is possible for the user to access 3D representations of selected fragments. This can be done as visualisations of the fragment in isolation, the structural variables in the fragment, or as the entire hit structure.

6.7.1 Viewing fragments

This feature can only be accessed from the dendrogram display. It enables the use to select a series of fragment that interest them and compare and assess the structures of those fragments visually.

To do this select the fragments of interest and select *Show Selected Fragments in 3D Viewer* from the *Tools* menu. Alternatively the *F1* short-key can be used. A window showing a diagram of the fragment

opens. To dismiss this, click *OK*. This opens the a new display window:



On opening it will display all of the selected fragments overlaid on each other for easy comparison. In this example two clusters have been selected for viewing in the 3D visualiser and it can be seen there is a clear difference between the two groups. The scroll box on the right of the screen can be used to display each fragment individually.

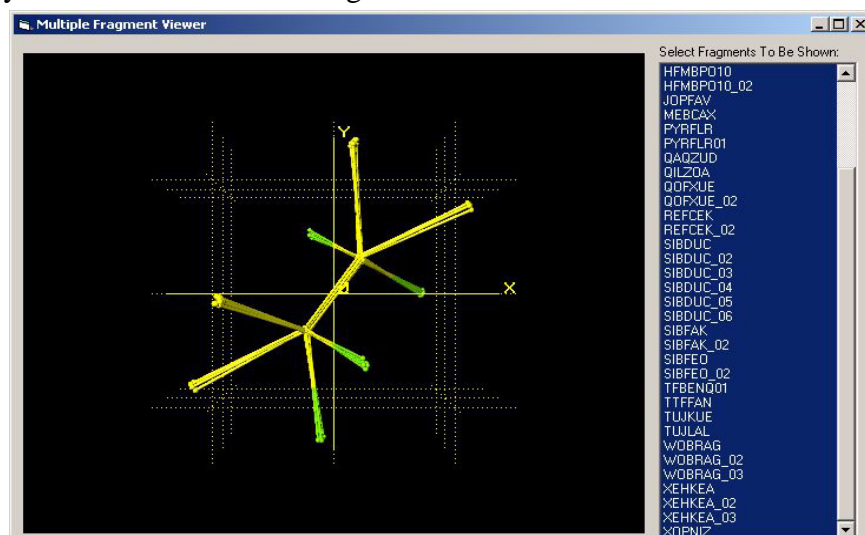
The display can be navigated in the same manner as the others, however the right-click menu now contains extra options relevant to this display.

Colour atoms by:

This option allows the user to decide whether to colour the fragments according to the cluster it has assigned to, the fragment or by the *type* of atoms inside the fragment. When first opened the viewer will display the fragments according to cluster as default. Changing this to *by fragment* gives each fragment a different colour so they can be distinguished between in the overlapped display.

Colouring *by type* means the each atom in the fragment is colour-coded to represent whatever atom it is in the structure. This option can be very useful. In the example above this allows the user to quickly determine that one cluster is comprised of *cis* fragments

while the other cluster is *trans*. Here carbon atoms are coloured yellow and fluorine atoms green.



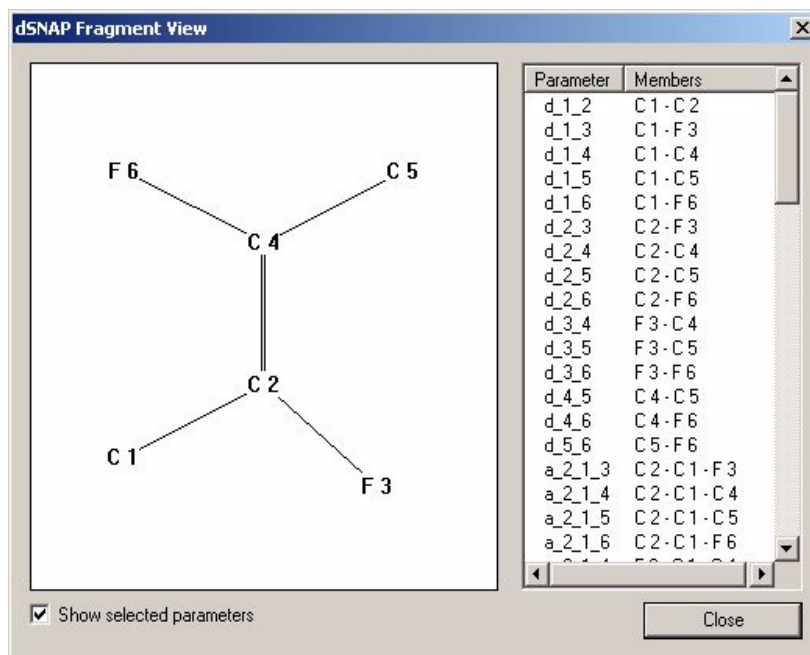
Label atoms by:

This options allows the atoms to be labelled by either *name* (which corresponds to the atom's identity in the complete database structure or included cif) or *number* (which corresponds to the atom number in the search fragment) when either the *Show All Labels* or *Show Selected Labels* option is active.

The 3D viewer feature may not be available if the user attempts to view results saved from a version of *dSNAP* prior to v0.9.5. In such cases it is advised that the user rerun the analysis using the current version of the program.

6.7.2 Viewing variables in structure

This feature can be accessed from any display screen by selecting *Show Fragment Variables in 2D Viewer* from the *Tools* menu or by using the *F2* short-key. It opens a new window that displays a 2D image of the fragment along with a list of all the structural variables.



This is most useful in *Variables* mode and so is described in full detail in that section.

6.7.3 Viewing entire hit structure

It is also possible to view the entire structure that the fragment was obtained from by using either *Mercury* or *ConQuest*. To do this the user must first select the fragments of interest by clicking on them in either the cell display, dendrogram or 3D plot and then select *View Selected Hits...* from *Tools* menu to open the viewing program. Alternatively, once the fragments have been selected the *F3* function key can be used as a short-cut.

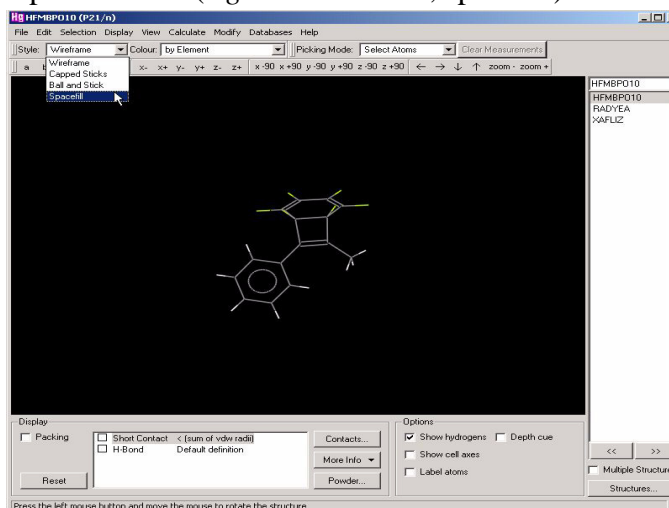
As before it is possible for multiple fragments to be selected and compared to investigate the chemical correlations between fragments in a group. *dSNAP* does not know any chemistry, and requires the user to interpret the clustering results.

Note that if a molecule has two fragments (*e.g.* HAFTEN_01 and HAFTEN_02) then selecting both of these to be viewed in either program will only bring up the parent hit (*e.g.* HAFTEN). Therefore in a case where two different fragments belonging to different clusters are in the same hit, the user must establish which fragment is which. One way in which this can be done is described in Section 6.7.6. Alternatively, the atom names can be displayed in the Multiple Fragments Viewer.

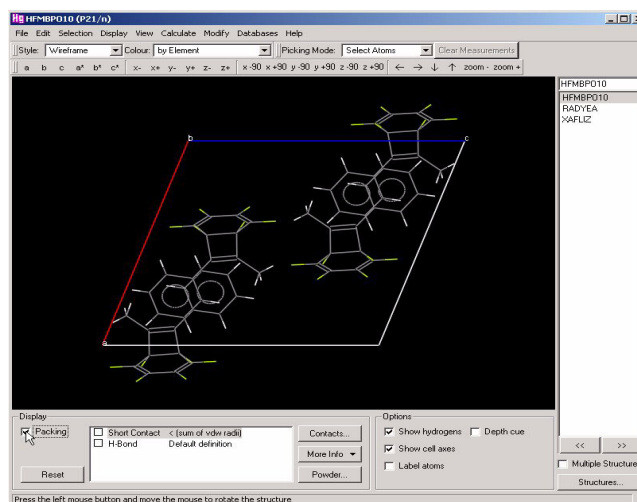
6.7.4 Viewing in Mercury

Mercury is set as the default program for viewing selected hits. There is a list on the right-hand side displaying the list of hits that have been selected by the user.

A useful feature to note is the *Style* menu found in the toolbar running along the top of the screen, where the user can change the style of representation (*e.g.* stick and ball, spacefill):



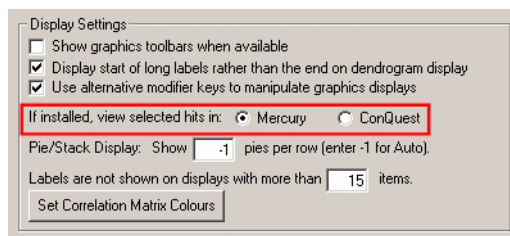
Clicking the *Packing* checkbox in the display box beneath the display allows the user to toggle between single molecule and unit cell views:



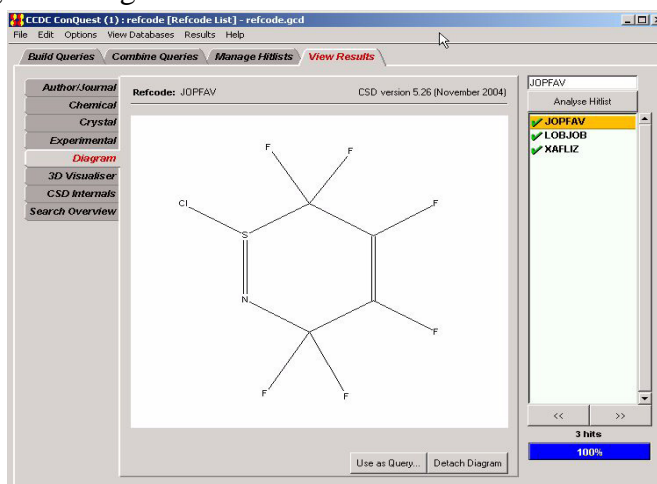
Both of these could be useful for visualising the larger chemical environment.

6.7.5 Viewing in *ConQuest*

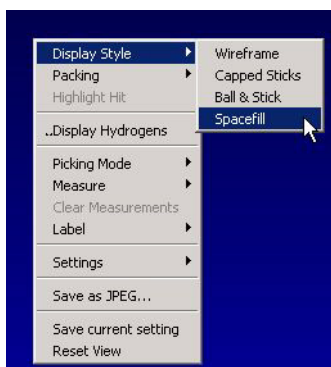
The default program for viewing fragments is *Mercury*, however this can be overridden by selecting *Options* from the *Edit* menu. Any change takes effect for the next run.



When viewing files in *ConQuest* the user is first taken to a screen showing a 2D diagram of the structure:



A 3D model can be accessed by selecting the *3D Visualiser* tab from the left hand tab bar. The structure can now be manipulated as in *Mercury*. To change the style the user must right-click on the display screen and access the *Display Style* menu. Similarly a *Packing* menu can be accessed by right-clicking which allows the user to toggle between a unit cell and single molecule views.



There are a series of further tabs that provide extra information that was entered in the database such as unit cell dimensions, references

and experimental data which may prove useful if a search hit proves to be of interest.

Note that as *ConQuest* has only been given the refcode and has no information about the original query, it is not possible to highlight the fragments in this situation. However it is possible to highlight hits by referencing back to the original saved search file.

6.7.6 Highlighting fragments in *ConQuest*

1 - Find the input folder for the dataset and open the original saved *ConQuest* Search file (.cqs).

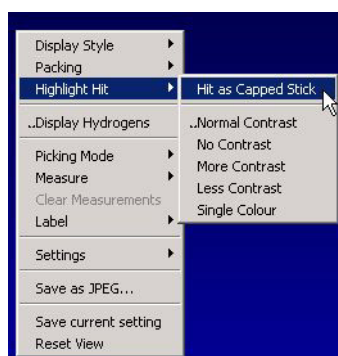
2 - *ConQuest* will open and display the list of hits obtained from the original search query. Select the relevant hit from the right-hand side menu.

3 - The fragment should be highlighted in red just as it would have been after the initial search. The user can scroll through multiple fragments using the *Multiple Hits: Show []* box beneath the display.



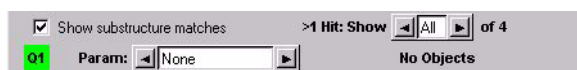
4 - Alternatively to visualise the fragments in 3D select the *3D Visualiser* tab from the left-side tab bar to bring up the 3D representation.

5 - Right click on the display and select *Highlight Hits* and choose the preferred method of highlighting. *Hit as Capped Stick* is particularly useful if the model is still being displayed as wireframe.



The fragments in the molecule are now highlighted. Ensure the parameters box below the screen is set to *none* so the view of the fragments is not obscured. The user can then scroll through the

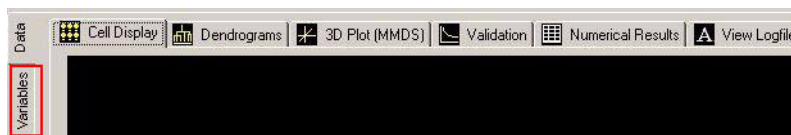
different fragments through the *>1 hit Show:* box that is also located below the display.



These correspond directly to the fragments dealt with in the *dSNAP* clusters, for example 1 of 2 in *ConQuest* refers to HIT_01 in *dSNAP*.

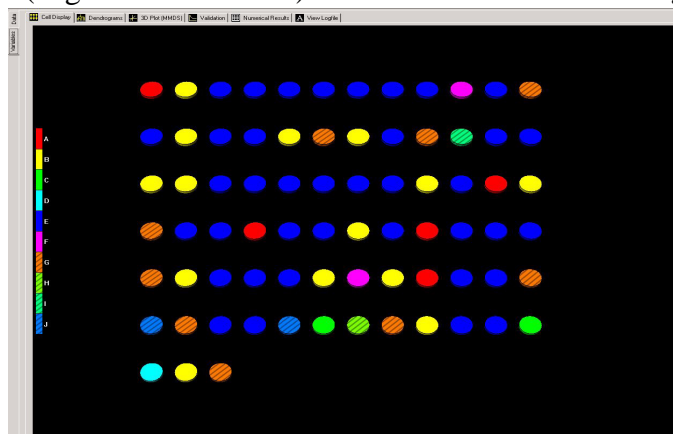
7.1 Results Display

The variables space results can be accessed by clicking on the *Variables* tab on the left-hand tab bar.



7.2 Cell display

The cell display will display a different number of cells than that seen in data space as it is now showing the number of independent variables (angles and distances) instead of the number of fragments:

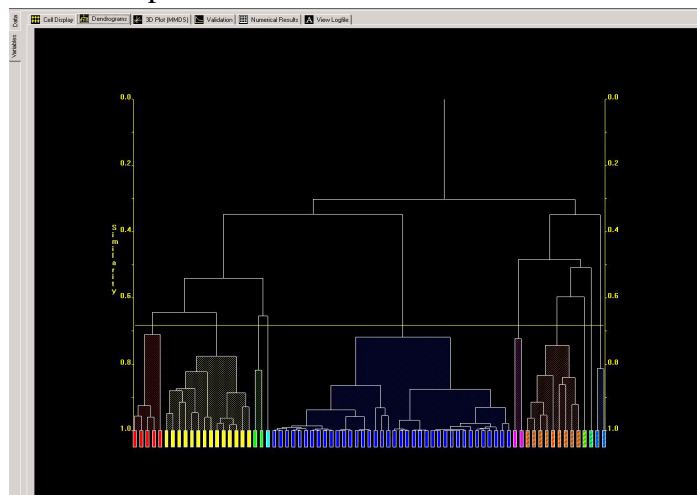


In subject (data) space each cell represents a single fragment, where as in variables space each cell represents a single structural variable (either distance or angle) in the fragment that was defined in the original search query.

The variables are clustered according to how well they correlate to each other. It should be expected that the clusters in variables space will be far more diffuse than in data space. Beyond this, the display pane works exactly like that in data space and an explanation of how to use this display can be found in section Section 6.1.

7.3 Dendrogram

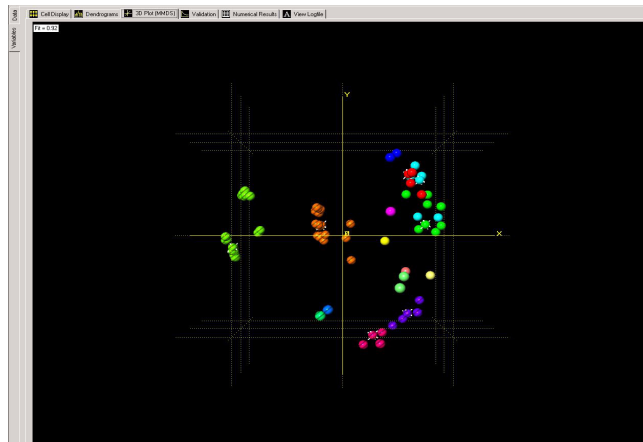
The basic controls in the dendrogram and the basics of how to read and interpret the dendrogram display are the same as that in data space, which is explained in detail in section Section 6.2.



The main difference is that now every box running along the bottom represents an individual defined variable in the fragment. This time the tie bars indicate the level of correlation between the variables it joins. Variables joined by a low tie bar are highly correlated while variables joined by a high tie bar are not correlated. Note that the two dendrograms for the two modes are completely independent of each other. Altering the cut-level in data space will have no effect on the dendrogram in variables space. Variables space can be a useful tool for accessing and interpreting the important factors that give rise to the clustering in data space.

7.4 3D Plot (MMDS)

The controls for manoeuvring the 3D plot and interpretation of the 3D plot are the same as in the data space display, and are explained in detail in section Section 6.3.



While every sphere in data space represents an individual fragment, in variables space each sphere represents a structural variable in the fragment. As in data space they are clustered by a different method to the dendrogram (Metric Multi-Dimensional Scaling) but use the dendrogram colours to allow comparison between the two methods. Spheres that are plotted close to each other are variables that are considered to be highly correlated.

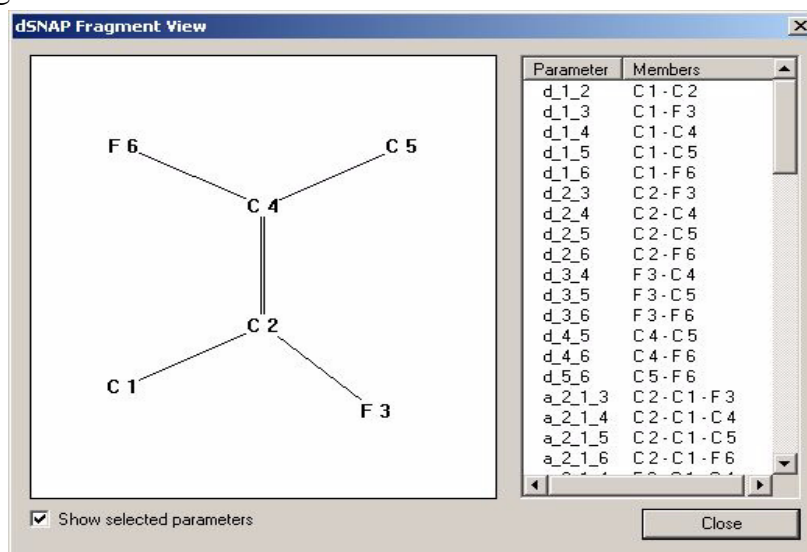
7.5 Validation

All four of the validation methods used in variables space work on the exact same method and principles as those in data space. These are explained fully in Section 6.4.

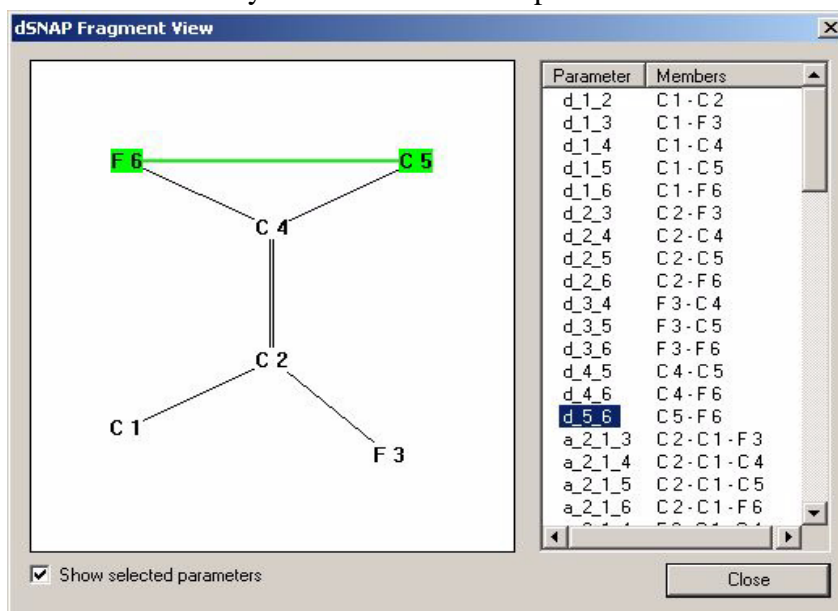
7.6 Viewing Variables in Structures

This feature can be accessed at any time using the *F2* short-key or by selecting *Show Fragment Variables in 2D viewer* from the *Tools* menu. It opens a new window displaying the structure of the

fragment in 2D with a list of all the variables in a scroll box on the right.



This is used to relate the analysis obtained in *Variables* space back to the real structure. Selecting any parameter in the right-hand scroll box will highlight that variable on the 2D display, allowing the user to understand exactly what that variable represents.



Using this window the user can navigate through the *Variables* space results and relate those results back to the real structure.

7.7 Numerical Results

In *Variables* space this screen has some extra features that are not available in *Data* space.

In this display, ranges of values of the correlation can be coloured, which can help to spot trends in the data. By default, these colours are only displayed for fragments containing ten atoms or fewer, but this value can be increased through *Options* in the *Edit* menu. More details can be found in the Program Options chapter.

Additionally, significant correlations can be highlighted in bold:

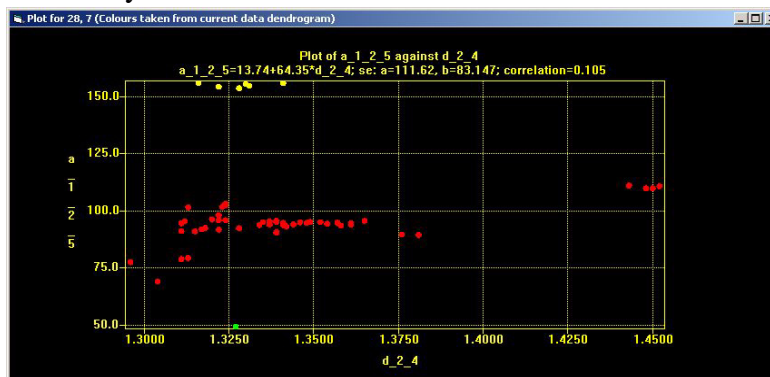
	Cell Display	Dendrogram	3D Plot (MOL5)	Validation	Numerical Results	View Logfile														
Variable	a_1,2	a_1,3	a_1,4	a_1,5	a_1,6	a_1,7	a_1,8	a_1,9	a_1,10	a_1,11	a_1,12	a_1,13	a_1,14	a_1,15	a_1,16	a_1,17	a_1,18	a_1,19	a_1,20	
a_1,1	1.000	0.014	0.013	0.001	0.042	0.103	0.005	0.047	0.225	0.202	0.040	0.009	0.040	0.040	0.040	0.040	0.040	0.040	0.040	0.040
a_1,2	0.014	1.000	0.006	0.003	0.052	0.194	0.203	0.064	0.474	0.495	0.240	0.240	0.014	0.240	0.240	0.240	0.240	0.240	0.240	0.240
a_1,3	0.013	0.006	1.000	0.044	0.221	0.440	0.374	0.041	0.321	0.303	0.203	0.007	0.104	0.104	0.104	0.104	0.104	0.104	0.104	0.104
a_1,4	0.001	0.044	0.044	1.000	0.040	0.440	0.440	0.040	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440	0.440
a_1,5	0.042	0.052	0.221	0.040	1.000	0.241	0.009	0.193	0.307	0.141	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
a_1,6	0.103	0.194	0.440	0.440	0.241	1.000	0.400	0.400	0.107	0.205	0.202	0.023	0.403	0.403	0.403	0.403	0.403	0.403	0.403	0.403
a_1,7	0.047	0.064	0.040	0.040	0.009	0.400	1.000	0.000	0.057	0.054	0.006	0.021	0.044	0.044	0.044	0.044	0.044	0.044	0.044	0.044
a_1,8	0.225	0.474	0.321	0.041	0.307	0.141	0.007	1.000	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029	0.029
a_1,9	0.202	0.495	0.303	0.007	0.141	0.007	0.007	0.029	1.000	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
a_1,10	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	1.000	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
a_1,11	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	1.000	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
a_1,12	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	1.000	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004
a_1,13	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	1.000	0.004	0.004	0.004	0.004	0.004	0.004	0.004
a_1,14	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	1.000	0.004	0.004	0.004	0.004	0.004	0.004
a_1,15	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	0.004	1.000	0.004	0.004	0.004	0.004	0.004
a_1,16	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	0.004	0.004	1.000	0.004	0.004	0.004	0.004
a_1,17	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	1.000	0.004	0.004	0.004
a_1,18	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	1.000	0.004	0.004
a_1,19	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	1.000	0.004
a_1,20	0.040	0.240	0.104	0.440	0.007	0.107	0.057	0.054	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	0.004	1.000

This option may be preferred to the colouring option for larger fragments. By default the cut-off value for a correlation to be considered significant is 0.8 but this can be changed by the user through *Options* in the *Edit* menu.

The column width can also be adjusted to allow the maximum number of values to be displayed, or to read the whole column heading label.

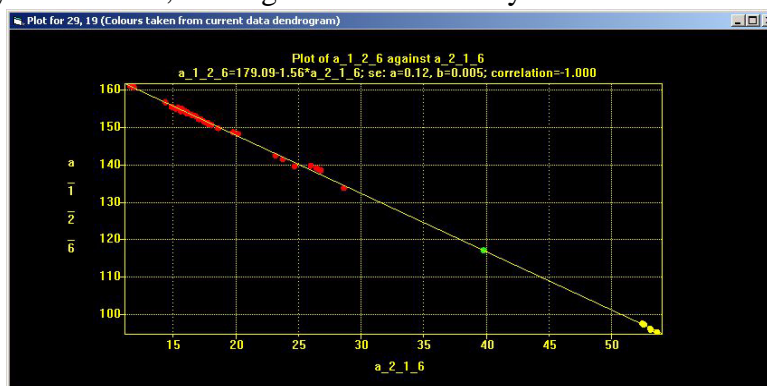
7.8 Variables Space Scatter Plots

By clicking on a cell in the *Numerical Results* pane a new window will open displaying a scatter plot for that the two variables represented by that cell:



As each of these plotted points represents a fragment, the colours used in the scatter plot are taken from the **data** space dendrogram. This allows the user to analyse any possible relationships between trends in particular variable values and the resulting fragment clusters.

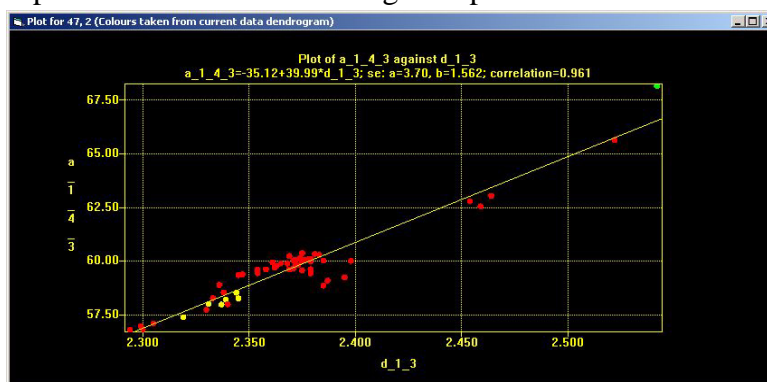
Scatter plots can only be created for two different variables. Although there will be entries in the numerical display where a variable has been correlated to itself and has a correlation of 100% it will not be possible to obtain a scatter plot. If however two different variables have a 100% correlation with each other then a scatter plot may be obtained, although this is reasonably rare.



7.8.1 Features of the scatter plot

As each point on the scatter plot represents a fragment, clicking on any dot on the plot will open the crystal structure that fragment belongs to in either *Mercury* (the default) or *ConQuest*. Holding the mouse over a sphere will display the refcode of the fragment that sphere represents in a tooltip.

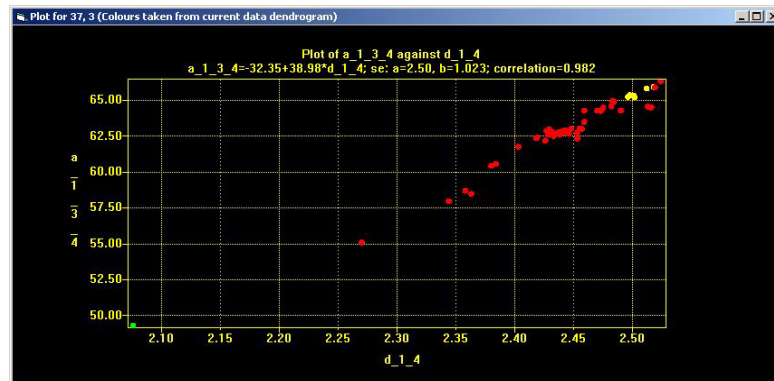
Right-clicking on the scatter plot and selecting *Show line* from the menu plots a line of best-fit through the points.



Finally the size of the spheres can be altered by holding *Ctrl* and the left mouse button while moving the mouse up and down.

7.8.2 Understanding scatter plots for highly correlated values

Two variables which are highly correlated would be expected to have a scatter plot where the points followed a relatively straight line:

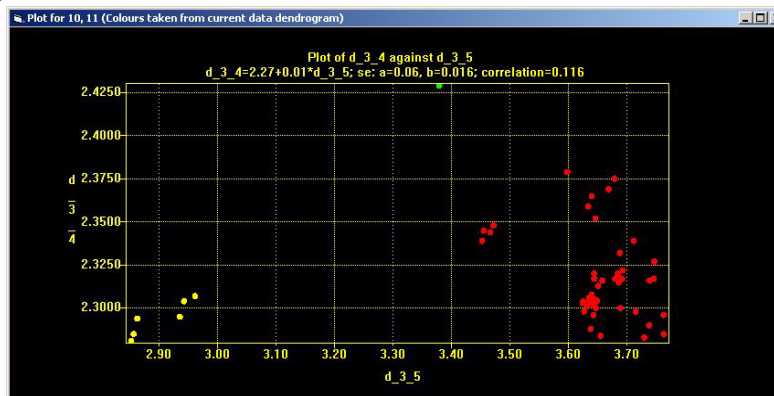


In this case the two variables are very highly related with over a 98% correlation to each other. Using this plot the user can start to interpret the effect that changing one variable has on another. In this case it can be seen that increasing angle $a_{1_3_4}$ is associated with an increase in distance d_{1_4} and *vice versa*.

It should also be noted in this example that along the line itself the points seem to be largely segregated into the different data space clusters. When this occurs it can be understood that the changes in these particular variables may be important to the structural differences between each cluster. Such information can help to make further sense of the results.

7.8.3 Understanding scatter plots for poorly correlated values

Although poorly correlated values do not provide good linear plots they still hold useful information.



In the above case these two variables are very poorly correlated with a correlation of under 12%. Regardless, it is still possible to interpret structural information here. In this example examining the points

relative to the x-axis shows that the yellow cluster typically has a shorter distance d_{3_5} than the red cluster. Of most significance however is the green fragment located at the very top. This fragment has been classified in a cluster of its own and has a far greater distance d_{3_4} than any other fragment on the plot. This suggests distance d_{3_4} may be important to its classification as a lone cluster. Note that the light blue point's position relative to the x-axis is not greatly different from some of those in the red cluster.

By bringing up visualisations of the structures of interest by clicking on the light blue point, and through understanding what distance d_{3_4} represents using the 2D viewer, the user can begin to understand the classification made on this fragment and make decisions on whether this is reasonable. In complex datasets it should be expected that several parameters are required to describe the dataset.

7.9 Logfile

The variables space logfile is a parallel to the data space logfile, so information on uses of the logfile can be found in Section 6.6. The difference is while in data space a hit refers to a fragment, in variables space a hit refers to a variable.

8.1 Other Menu Items

8.1.1 File

This menu allows the user to access the *Run on...* and *View results...* that are used to input data to *dSNAP*. There is also a *Exit* option to shutdown the program.

8.1.2 Display

This menu contains several miscellaneous display options:

Show dendrogram colours on 3D plot

This is on as a default and assigns the colours determined by the dendrogram to the 3D plot. If this option is deselected, all spheres in the 3D plot are brown.

Show fragment identity if available

This option is on as a default and displays the refcode of each fragments in the cell display and dendrogram. When it is deselected the fragments are instead labelled with an index number, according to the order in which they were input.

Show graphics toolbar

When this option is selected the graphics toolbar will be displayed as a default.

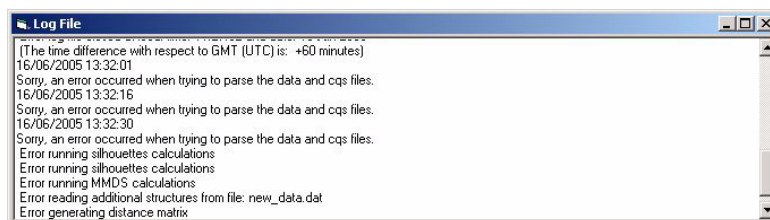
Show current display in new window

This option opens a new moveable window displaying the contents of the current display. The new window can be interacted with (*e.g.* rotated, zoomed) just like the parent display.

8.1.3 Window

This menu contains two menu items. Firstly is the *Windows List* submenu that lists all the currently open windows within the *dSNAP* program. Selecting a window from the list brings it to the front.

Clicking on View Error Log opens a new window with an edited version of the *dSNAP* logfile, which contains a list of errors or problems that have occurred during a run of the program:



It may be a useful aid in the determination of where any problems have occurred.

8.1.4 Help

This menu has several options.

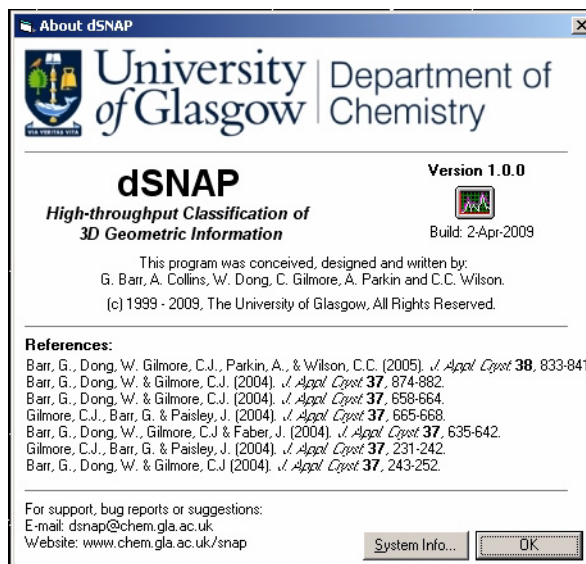
The first seven options bring up Quick Guides, which provide a brief set of instructions, which are of particular use to new users. These are available on the following subjects:

- Nomenclature in dSNAP
- Preparing and Importing Your Data
- Including your own CIF
- Reading a Dendrogram
- Using the 3D Fragment Viewer
- Interpreting and Validating Results
- Preparing Graphics for Publication

View Program Manual can be used at any time to open a PDF version of this manual which can be searched, displayed or printed as required.

View Tutorial can be used at any time to open a PDF version of the tutorial which can be displayed or printed as required.

Finally selecting *About dSNAP* brings up a screen describing the running version of *dSNAP*, and includes useful information such as the version number, build date and contact details for the authors.



Right-clicking on the list of references and selecting *Copy* will add the list to the system clipboard which can be useful when referencing the use of the software in a report.

8.1.5 Table of Shortcut keys

Shortcutkey	Function
Alt-F4	Closes program window
Ctrl-Z	Undo last (one change only)
Alt-F	Find item
Ctrl-F	Find item
Alt-A	Accelerate wheel
F1	Open 3D fragment viewer
F2	Open 2D variable viewer
F3	Open selected hits in Mercury/ConQuest
F5	Reset View
F11	Debugging information
F12	Alter rendering quality
Ctrl-W	Closes current window
Ctrl-F4	Closes current window
Ctrl-C	Copy selection

8.1.6 Table of Graphics Controls in Standard mode

Click on display followed by:

Operation	Control
Zoom in/out	Hold left mouse button and draw area to be zoomed
Move in display	Hold <i>Ctrl</i> and drag with left mouse button
Change object size	Hold <i>Alt</i> then hold left mouse button and move mouse up/down

(Dendrogram only)

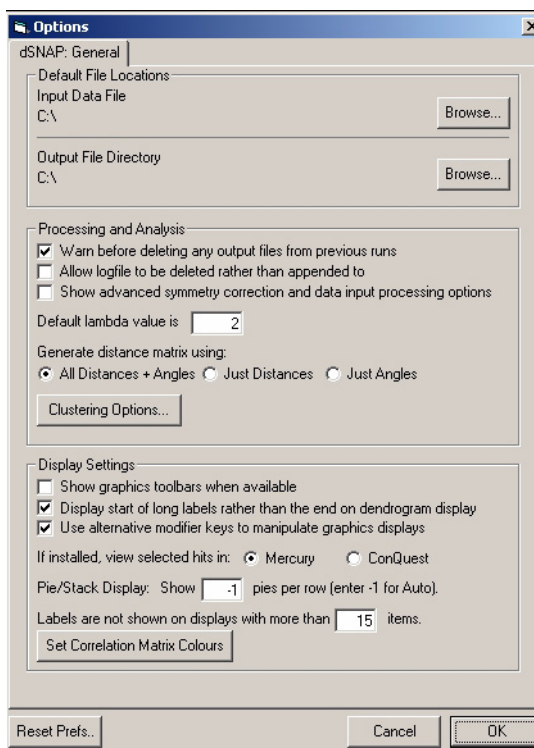
8.1.7 Table of Graphics Controls in PolySNAP-style mode

Click on the display followed by:

Operation	Control
Zoom in/out	Hold <i>Ctrl</i> and draw area to be zoomed
Move in display	Hold <i>Alt</i> then hold left mouse button and drag with left mouse button
Change sphere size	Hold <i>Ctrl</i> then hold left mouse button and move mouse up/down
(Dendrogram only)	
Adjust cut-level	Use mouse scroll wheel <i>or</i> Hold <i>Ctrl</i> and drag with left mouse button
Translate horizontally	Hold <i>shift</i> and use mouse scroll wheel
(3D Plot only)	
Rotate plot	Hold <i>Shift</i> then hold left mouse button and move mouse as required

9.1 dSNAP Options

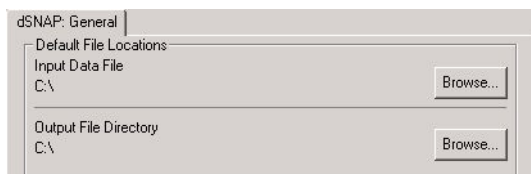
The options are accessed by selecting *Options* from the *Edit* menu. This opens a new window from which several settings and defaults can be changed:



9.1.1 dSNAP General Options

The top section of the window displays the general options concerned with default file locations. This allows the user to select the file

locations for both input files and output folders that will be displayed as defaults whenever *Run On...* is opened.



These can be changed by clicking the *Browse...* boxes and selecting the appropriate file in the window that opens.

9.1.2 Processing and Analysis

Warn before deleting any output files

If the output folder selected has already been used in a previous run then the old output files will be deleted by performing a new run to that location. With this option on, a message will be displayed warning the user when this is about to occur. By default this option is on.

Allow logfile to be deleted rather than appended to

When a fresh run is performed into an old output folder that already contains output files all of these files will be deleted to make way for the new input, apart from the logfile, which is normally only appends the new information onto the old logfile. By default this option is off, however switching it on will mean the old logfile will be deleted. A new logfile will then be created for each following run.

Show advanced data input processing options

This option allows the user to control which atoms are used in the analysis, even if there are less than 20 atoms in the search fragment. By default this option is off. A description of the effects of this option can be found in section 4.2.

Default lambda value is []

The default lambda value refers to the equation:

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

which is used in the cluster analysis. The lambda value affects the way in which the distance between items in the clusters analysis is calculated. Setting lambda to 1 is equivalent to the City Block method, while setting lambda to 2 is equivalent to the Euclidean

method of calculating differences. Other values are also possible but should be used with caution. By default this is set to a value of 2.

Generate distance matrix using:

This option allows the user to select which geometric parameters are included when generating the distance matrix and performing analysis. By default all distances and angles are used, however the user can choose to only use distances or only use angles. This option should be used with caution.

Clustering Options...

Clicking this box will open the Clustering options menu in a new window - see section 9.2 below.

9.1.3 Display Settings

Show graphics toolbars when available

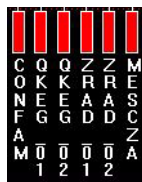
The results display window has several graphics panes that can all display a graphics toolbar. The toolbar consists of many of the common functions that can be performed on the graphic displays and is described in section 5.3.3. Although the user can access the toolbar at any time through the *Display* menu, this option allows the user to choose whether or not this toolbar will appear by default. By default this option is off.

Display start of long labels rather than the end on dendrogram display

When items have larger labels, especially in cases where multiple fragments have suffixes added, the full label does not appear on the dendrogram display. This option allows the user to choose whether the label will be displayed in such a way that the start of the label, and so the majority of the refcode, will be visible;



or the end, and so the fragment number, will be visible. By default the start of the label and the refcode is displayed.



C	Q	Q	Z	M
O	K	R	R	E
N	E	A	D	S
F	G	G	D	C
A				Z
M	0	0	0	A
	1	2	1	2

Use alternative modifier keys to manipulate graphics displays

This options toggles between two different modes, each with a slightly different way of interacting with the graphics displays. This option is described more in section 5.3.1.

If installed, view selected hits in... <Mercury> <ConQuest>

This options allows the user to select which program (either *Mercury* or *ConQuest*) that will be used to view the crystal structures. By default this is set to *Mercury*. If it has not already been installed on the computer, *Mercury* can be downloaded free from the Cambridge Crystallographic Data Centre website (http://www.ccdc.cam.ac.uk/products/csd_system/mercury/). Once the setting for the viewing program has been changed the user must wait until the program has been closed and reopened in order for the change to take come into effect.

Pie/Stack display: Show [] pies per row

For display purposes the number of cells in a row in the *Cell Display* screen can be altered to suit the users needs, *e.g.* for a run with 12 fragments the number entered may be 6, which would create 2 rows of six cells. The minimum value is two and the maximum value is 199. If the value is set to -1, as it is in the default setting, then the program makes its own judgement as to the best number to fit on the screen, based on the current screen size.

Labels are not shown on displays with more than [] items

Although in many cases showing the labels of items on the dendrogram and 3D plot is useful for quick identification, in cases with increasing large numbers of items the display can quickly become congested. This option allows the user to control the number of items above which the associated labels to are displayed. By default this is set to 15.

Set Correlation Matrix Colours

This button brings up another window with options controlling the display of the Numerical Results pane in Variables space.

For performance reasons, these options are only applied for relatively small datasets by default, less than 500 entries. This cut-off value can of course be increased. The number of parameters for a given number of atoms is given in the table:

TABLE 1.

No of atoms	No of parameters
3	6
4	18
5	40
6	75
7	126
8	196
9	288
10	405
11	550
12	726
13	936
14	1183
15	1470
16	1800
17	2176
18	2601
19	3078
20	3610

When the ‘Colour cells to highlight correlations’ option is checked, the Variables space numerical results grid is coloured in according to the settings here.

The default settings have values of between -1.0 and -0.9 as green, between -0.1 and +0.1 as yellow, and between +0.9 and +1.0 as light blue. Values not between one of those ranges are not highlighted.

Both the colours and the highlight ranges can be customised; click on a colour box to bring up a standard Windows colour picker to select an alternative. Colouring the values in this way allows trends in the relationships between parameters to be identified.

When the ‘Highlight absolute correlations’ checkbox is turned on, any values in the table above the specified value (default 0.8) are highlighted in bold. Again, this can be useful to see patterns emerging in the data.

The ‘Minimum Cell Size’ option controls how small the width of each cell in the numerical results grid is allowed to get in order to try to fit more columns into the display without scrolling. The smaller this number, the more columns can be fitted in; the default value is 1000.

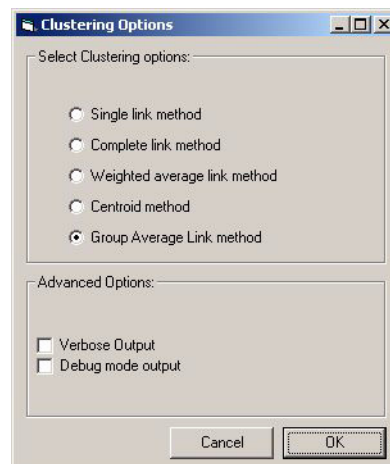
Reset Preferences.

There is a final button at the bottom of the options window named *Reset Prefs*:



Clicking this button will reset all of the options back to their default settings.

9.2 Clustering options



The program can use up to five different methods for computing the dendrogram. By default, the Group Average Link method is selected.

The overall method employed is agglomerative hierarchical clustering, using a distance matrix derived from the geometric data. Initially all the patterns are assigned to individual clusters which are then joined one pattern at a time. When two clusters are so joined, a measure of the distance between them is needed, and each clustering method has its own way of doing this.

In general, when two clusters i, j are combined, a new distance between the cluster and an existing cluster k is calculated as:

$$d_{k(i,j)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}|$$

The parameters α_i , α_j , β , and γ are different for each clustering method used:

Single Link Method:

$$\alpha_i = \frac{1}{2} \quad \beta = 0 \quad \gamma = -\frac{1}{2}$$

Complete Link Method:

$$\alpha_i = -\frac{1}{2} \quad \beta = 0 \quad \gamma = \frac{1}{2}$$

Weighted Average Link Method:

$$\alpha_i = \frac{1}{2} \quad \beta = 0 \quad \gamma = 0$$

Centroid Method:

$$\alpha_i = n_i(n_j + n_j) \quad \beta = \frac{-n_i n_j}{(n_i + n_j)^2} \quad \gamma = 0$$

Group Average Link Method:

$$\alpha_i = \frac{n_i}{(n_i + n_j)} \quad \alpha_j = \frac{n_j}{(n_i + n_j)}$$

$$\beta = 0 \quad \gamma = 0$$

Advanced Options

In addition to the output methods displayed graphically in the main output window, the clustering algorithms also all output textual information to the logfile for the current run of the program. This may be found in the program output folder with the name *SNAPlog.txt*. The detail level of the text output can be modified by the two checkbox options:

Verbose Mode Output

The *Verbose* and *Debug Mode* options control the level of output the cluster analysis section of the program writes to the logfile for each run. In normal program usage, both options are turned off. Occasionally it may be useful to see more detailed information on the clustering results, in which case *Verbose* option may be turned on.

Debug Mode Output

The *Debug Mode* option causes the program to output all of the working matrices generated during analysis to the logfile, to aid in debugging. Note however that use of this option can cause the logfile to become very large - possibly up to tens of megabytes in size. It should be used with caution, as it may also slow the execution of the program, especially when large numbers of hits are being analysed.

Dealing with Structures with Local Symmetry

10.1 The problem of symmetry

Any program that deals with structure matching is likely to have to deal with the problem of symmetry.

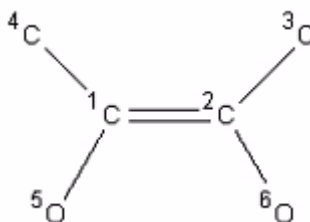
A *ConQuest* search fragment for use in *dSNAP* may frequently possess symmetry. This may affect the whole fragment or the search fragment may contain local symmetry that affects only some of the atoms. This is also known as topological symmetry. Note that the symmetry of the fragment is independent of the symmetry of the whole molecule in which it found, unless the search fragment is a complete molecule.

In *dSNAP*, a topological symmetry correction is applied in when the search fragment has symmetry as the symmetry causes an ambiguity over which atoms in the database structure correspond to which atoms in the search fragment.

The problem is best explained through examples.

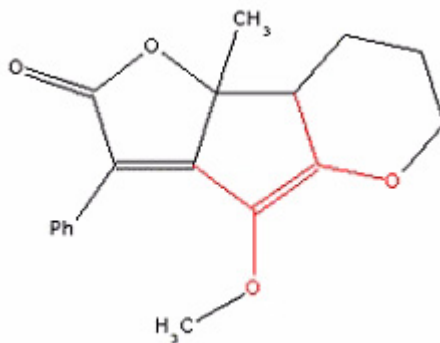
Example 1

This fragment possesses symmetry which affects all the atoms in the fragment, but it is generally found in a molecule which is not.

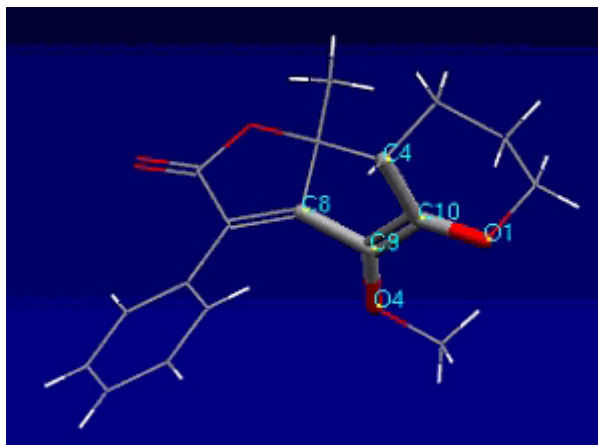


The search fragment was drawn in *ConQuest* as shown in the figure above. This entire fragment is affected by symmetry. Searching for the fragment in the CSD v5.30, with 2 updates and all boxes for search criteria in the search setup checked (including R-factor=<5%, and only Organics) retrieved 170 hits.

One of those hits is refcode KEHHIO, which is drawn below (the hit fragment is highlighted in red):



Although in the search fragment, atom C1 is equivalent to C2, atom C3 to C4 and atom O5 to O6 by the symmetry of the fragment, the atoms occupy distinct chemical environments within the molecule.

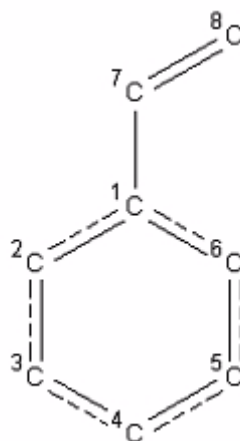


One of the tenets of *dSNAP* is that it does not have any chemical knowledge and so is not prejudicially affected by any preconceptions that the user may have about the fragment would cluster; it therefore provides an unbiased route to establishing structural trends. In this example it may not appear that any renumbering would be expected to have much difference on the clustering; however, not applying a symmetry correction in this case does alter the clustering results.

It is therefore very important to apply a consistent set of criteria to all datasets, regardless of the expectation of the clustering results.

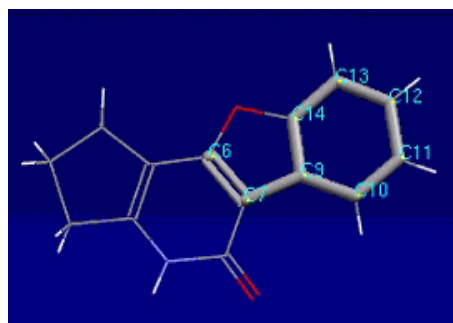
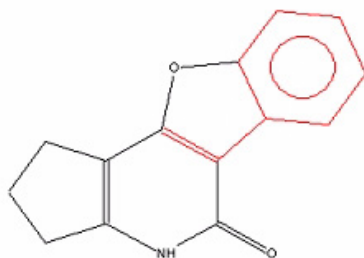
Example 2

This fragment contains only a few atoms that are affected by symmetry.



In this fragment, atom C2 is equivalent to C6, and atom C3 is equivalent to C5. Atoms C1, C4, C7 and C8 are unique. Consider the C2-C8 distance and the C6-C8 distance. Depending on the torsion angle around the C1-C7 bond, these distances are likely to be different.

eg refcode ABACIQ



The numbering scheme that has been applied to the fragment in the published crystal structure is shown in the figure above. Atom C6 in the published structure will correspond to C8 in the search fragment. If atom C14 in the published structure is assigned to be C2 in search fragment, the C2-C8 distance will be small, while if C10 in the published structure is assigned to be C2, then the C2-C8 distance will be larger.

While the problem may appear to be simply one of numbering, as this impacts on the identity of atoms mis-assignment of atoms affects all the parameters which involve those atoms. In order to obtain

meaningful results from the cluster analysis it is essential that steps are taken to correct for any differences that could arise from the problem of topological symmetry.

10.2 Determining the presence of topological symmetry

A modified version of Morgan's Rules (H. L. Morgan, *J. Chem. Doc.*, 1965, **5**, 107-113) is used to determine which atoms, if any, within the search fragment are related by topological symmetry. In this method, each atom in the fragment is assigned an array of values. The first set of values relate to the atom itself, and are based on its element type, the number of atoms explicitly bonded to it (ie not including any hydrogen atoms defined implicitly), the total number of atoms bonded to it including implicit hydrogen atoms, and the sum of the bond types of all the atoms bonded to it. The second set of values relates to atom attributes of the first coordination sphere. The process is repeated for subsequent coordination spheres.

The complete array is then compared for each atom. Atoms that have the same score at every point in the array are assigned a similarity of 1.

Subsequent checks are made to determine whether other attributes of the atoms or bonds defined in the original *ConQuest* search break the topological symmetry. If the symmetry is broken by definitions in the atom or bond cyclicity, the defined atomic charge or the defined number of bonded atoms, the user is alerted to the fact and asked to chose whether or not to apply the symmetry correction. In most cases, it will be most appropriate to apply the symmetry correction, and the decision not to apply the symmetry correction should be justified fully.

10.3 Dealing with symmetry

For a dataset containing n fragments, which each contain j parameters, the mean is calculated for each parameter x .

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Next the variance of the mean for the parameter is calculated:

$$\rho_j = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

This process is repeated for each of the j parameters.

These values of ρ are used to calculate a figure of merit, by taking the mean of ρ over j parameters:

$$\bar{\rho} = \frac{\sum_{i=1, j}^j \rho_j}{j}$$

Starting with the first fragment in the cor file, ρ is calculated for each of the k symmetry possibilities. The symmetry possibility that results in the lowest value of ρ reflects the atom assignment which minimises the variation in the data set, and therefore indicates the best choice of symmetry possibility.

This process is repeated for each of the fragments. Where any fragment has had its symmetry possibility changed, the parameters for that symmetry possibility are subsequently used when calculating x , ρ and $\bar{\rho}$.

If any fragment in the data set has had its symmetry possibility altered, the process is repeated for every fragment in the entire dataset until no fragment has its symmetry possibility altered.

These are then the atom assignments used for the rest of the cluster analysis.

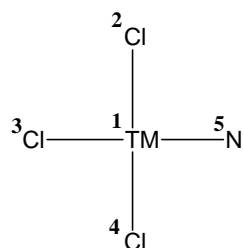
The variable $\bar{\rho}_k$ is termed the MMCV variable in the log files.

Once the presence of topological symmetry has been established, all possible numbering schemes for the fragment which are consistent with the topological symmetry of the fragment are generated; these are referred to as the symmetry possibilities.

10.3.1 Examples of the application of the method

10.3.1.1 Example1

The fragment illustrated in the figure below, showing its numbering scheme from ConQuest, contains three topologically similar chlorine atoms, giving rise to the similarity matrix in the table.



	Atom 1	Atom 2	Atom 3	Atom 4	Atom 5
Atom 1	1	0	0	0	0
Atom 2	0	1	1	1	0
Atom 3	0	1	1	1	0
Atom 4	0	1	1	1	0
Atom 5	0	0	0	0	1

These three topologically similar atoms give a total of six symmetry possibilities:

Symmetry possibility	TM 1	Cl 2	Cl 3	Cl 4	N 5
1	TM 1	Cl 2	Cl 3	Cl 4	N 5
2	TM 1	Cl 2	Cl 4	Cl 3	N 5
3	TM 1	Cl 3	Cl 2	Cl 4	N 5
4	TM 1	Cl 3	Cl 4	Cl 2	N 5
5	TM 1	Cl 4	Cl 2	Cl 3	N 5
6	TM 1	Cl 4	Cl 3	Cl 2	N 5

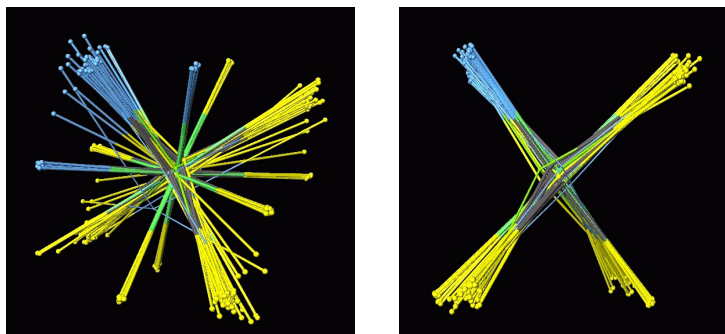
The following tables show first round of calculations used to superimpose the first fragment onto the same region of topological space. All parameters remain the same except for those corresponding to APTCPT, which are highlighted in italics. A single parameter, $\bar{\rho}_k$, is used to decide on the best symmetry possibility, with the best indicated by the lowest value.

CSD refcode	Symmetry Possibility 1			Symmetry Possibility 2			Symmetry Possibility 3		
	d_2_3	d_2_4	d_3_4	d_2_3	d_2_4	d_3_4	d_2_3	d_2_4	d_3_4
<i>APTCPT</i>	4.602	3.275	3.324	3.275	4.602	3.324	4.602	3.324	3.275
ARESAR	3.572	3.524	3.584	3.572	3.524	3.584	3.572	3.524	3.584
ASUJUT_01	3.636	3.750	3.811	3.636	3.750	3.811	3.636	3.750	3.811
ASUJUT_02	3.640	3.782	3.720	3.640	3.782	3.720	3.640	3.782	3.720
BAPXIZ	3.279	4.607	3.273	3.279	4.607	3.273	3.279	4.607	3.273
BENWAS	3.580	3.576	3.671	3.580	3.576	3.671	3.580	3.576	3.671
BIFWUI	3.700	3.658	3.683	3.700	3.658	3.683	3.700	3.658	3.683
BINCOQ	3.280	4.604	3.265	3.280	4.604	3.265	3.280	4.604	3.265
BURGAW01_01	3.516	3.496	3.509	3.516	3.496	3.509	3.516	3.496	3.509
BURGAW01_02	3.514	3.468	3.507	3.514	3.468	3.507	3.514	3.468	3.507
$\bar{\rho}_k$	387.4486			306.4131			384.7895		

CSD refcode	Symmetry Possibility 4			Symmetry Possibility 5			Symmetry Possibility 6		
	d_2_3	d_2_4	d_3_4	d_2_3	d_2_4	d_3_4	d_2_3	d_2_4	d_3_4
<i>APTCPT</i>	3.324	4.602	3.275	3.275	3.324	4.602	3.324	3.275	4.602
ARESAR	3.572	3.524	3.584	3.572	3.524	3.584	3.572	3.524	3.584
ASUJUT_01	3.636	3.750	3.811	3.636	3.750	3.811	3.636	3.750	3.811
ASUJUT_02	3.640	3.782	3.720	3.640	3.782	3.720	3.640	3.782	3.720
BAPXIZ	3.279	4.607	3.273	3.279	4.607	3.273	3.279	4.607	3.273
BENWAS	3.580	3.576	3.671	3.580	3.576	3.671	3.580	3.576	3.671
BIFWUI	3.700	3.658	3.683	3.700	3.658	3.683	3.700	3.658	3.683
BINCOQ	3.280	4.604	3.265	3.280	4.604	3.265	3.280	4.604	3.265
BURGAW01_01	3.516	3.496	3.509	3.516	3.496	3.509	3.516	3.496	3.509
BURGAW01_02	3.514	3.468	3.507	3.514	3.468	3.507	3.514	3.468	3.507
$\bar{\rho}_k$	306.7463			375.7659			378.7582		

The value of $\bar{\rho}_k$ for each symmetry possibility is calculated for fragment 1, and the best symmetry possibility chosen, in this case symmetry possibility 2. This is substituted into the original matrix, and the process is repeated sequentially for fragments 2 to n , minimising the value of $\bar{\rho}_k$ as it goes through. The process is then repeated for fragments 1 to n until there are no more alterations to any of the fragments in the dataset.

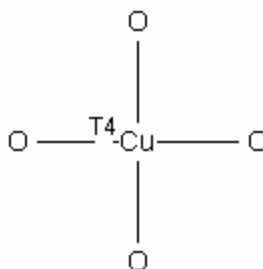
When the process is applied to the full dataset for this fragment, comprising 93 fragments; there are 6 symmetry possibilities. Application of the automatic topological symmetry correction reduces $\bar{\rho}_k$ from 148.165143 to 98.465150 in 4 cycles, and the superimposition of all fragments is improved as shown in the figure below.



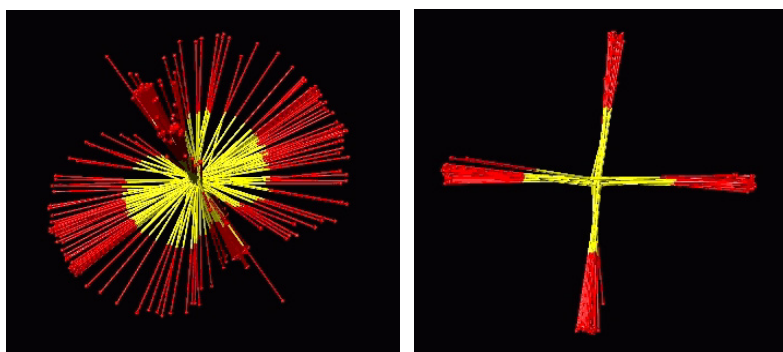
On the left is the superposition of the fragments (coloured by atom type) without applying the correction; on the right is the superposition when the symmetry correction has been applied.

10.3.1.2 Example 2

This example includes 133 fragments containing a 4-coordinate copper atom with four oxygen atoms bound:



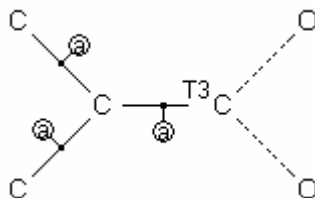
There are 24 symmetry possibilities. Application of the automatic topological symmetry correction reduces from 487.939377 to 11.808393 in 4 cycles, and the superimposition of all fragments is improved as shown in the figure below.



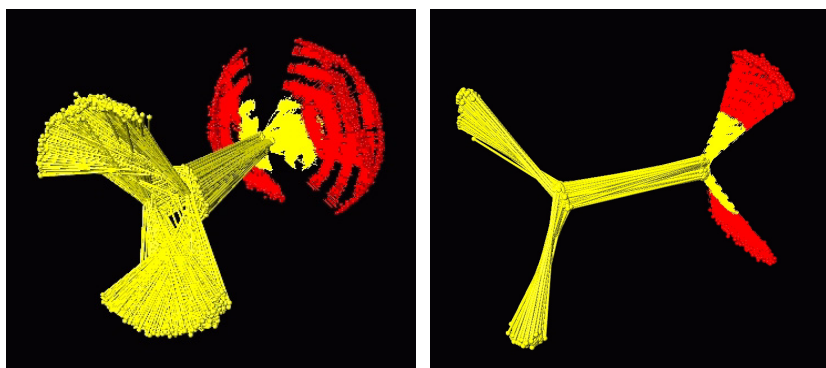
On the left is the superposition of the fragments (coloured by atom type) without applying the correction; on the right is the superposition when the symmetry correction has been applied.

10.3.1.3 Example 3

This example includes 1324 instances of the C_2C-COO fragment shown below:



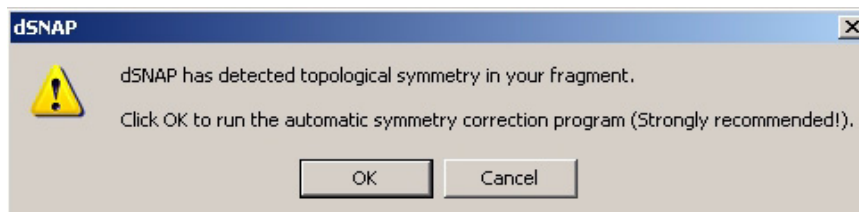
There are 4 symmetry possibilities. Application of the automatic topological symmetry correction reduces from 97.521352 to 27.233092 in 4 cycles, and the superimposition of all fragments is improved as shown in the figure below:



On the left is the superposition of the fragments (coloured by atom type) without applying the correction; on the right is the superposition when the symmetry correction has been applied

10.4 dSNAP output for structures with topological symmetry

If the presence of topological symmetry is detected, a dialog box will appear asking if the user wishes to apply a symmetry correction.



Clicking *OK* means that the symmetry correction is performed. Clicking *Cancel* will run the analysis but without applying a symmetry correction. Applying the correction is very strongly recommended.

The detection of topological symmetry generates a log file called `pre_processing_log.txt`. This contains the calculated similarity matrix, which shows which atoms are related to each other by symmetry, the total number of symmetry possibilities and how they affect the search atom fragment numbering. It also gives the values of the mean of the variance of the mean (MMCV) at the beginning of the minimisation process, and its value at the end of each cycle. Finally, the symmetry possibility assigned to each fragment is listed.

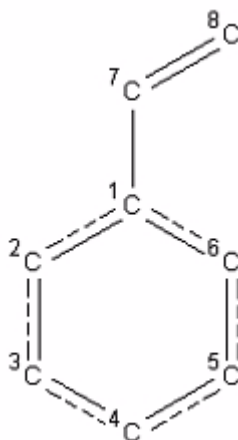
The number of cycles required to minimise the MMCV variable is also given in the main log file.

10.5 The effect on clustering of not dealing with symmetry

The effect of not applying a correction for the symmetry of the fragment is generally apparent from the patterns seen in the dendrogram and MMDS plots. This is particularly apparent when the clustering results are compared for a dataset analysed both applying and not applying the symmetry correction.

The results of clustering the fragment from Example2 in the section above are shown.

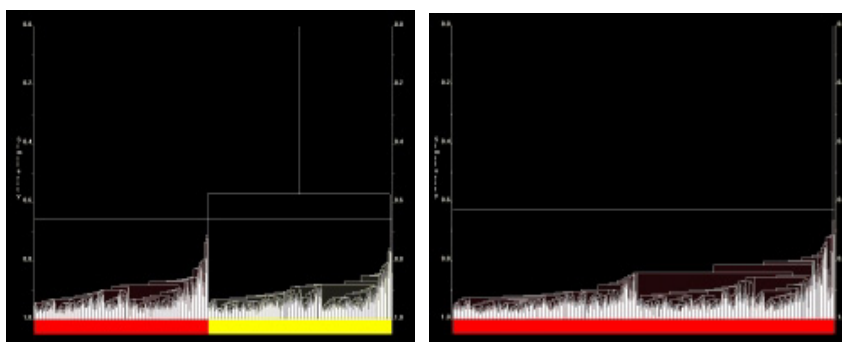
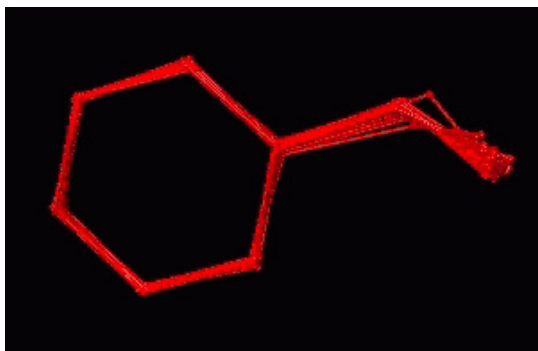
The search fragment was defined in ConQuest as follows:



There were 742 hit fragments.

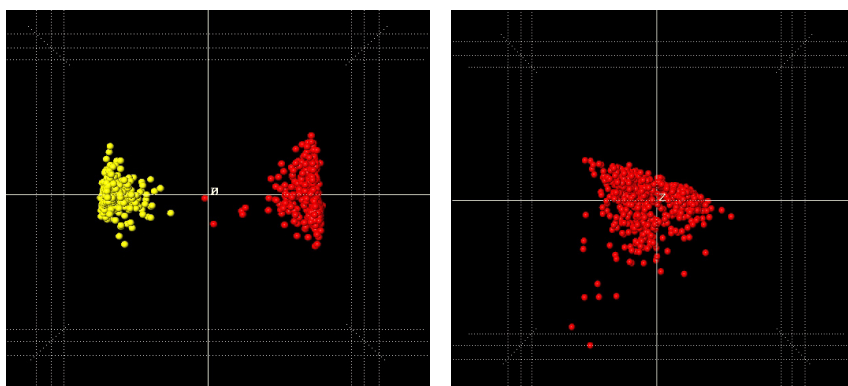
The fragment overlay for the dataset when the symmetry correction has been performed indicates that there is only one main geometry, and so only one cluster (this is also indicated by the form of the

dendrogram, where each fragment is linked to the next by a tie-bar which is at a similarity score which is only slightly lower):



On the left is the dendrogram when symmetry correction has not been applied; the dendrogram for the analysis where the symmetry correction has been applied. Notice how on the left, the dendrogram divides into two main sets which are linked by quite a low similarity score, and that the pattern of the tie-bars in each cluster is very similar to the pattern of tie-bars in the dendrogram on the right.

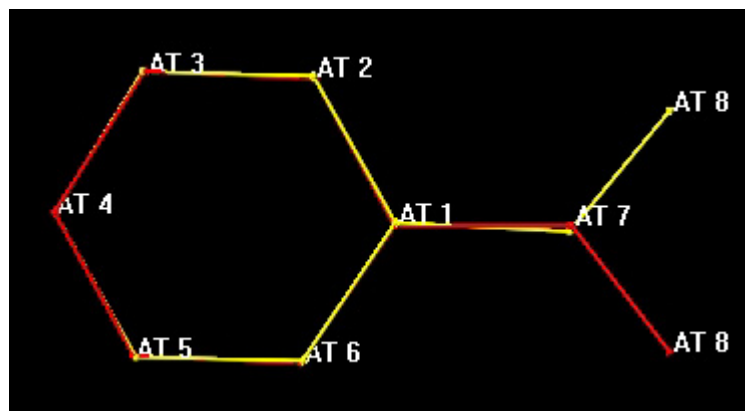
Similarly for the MMDS plot:



Again, the pattern seen in each of the two groups of spheres in the MMDS plot on the left, which is from the analysis where the symmetry correction has not been applied, is highly similar to the

pattern observed in the MMDS plot when the symmetry correction has been applied, shown on the right. The clusters in MMDS plot without the symmetry correction are symmetrical.

Looking at the fragment overlay for the dataset where the symmetry correction has not been applied shows that the two clusters differ only in the orientation of C8:



Here the most representative sample from each cluster has been shown for clarity.

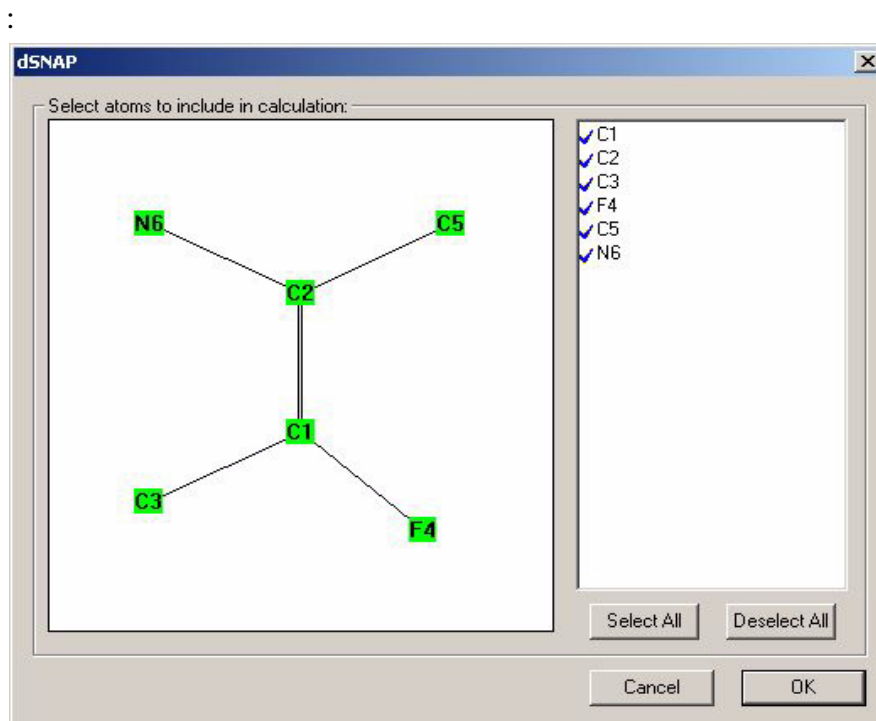
In summary, not correcting for symmetry can result in the presence of spurious clusters that are purely an artefact resulting from the numbering of the data and do not denote real differences between fragments.

This section details further options available. Their use is only recommended by experienced users of the program, and generally only under exceptional circumstances.

11.1 Running *d*SNAP on Selected Atoms

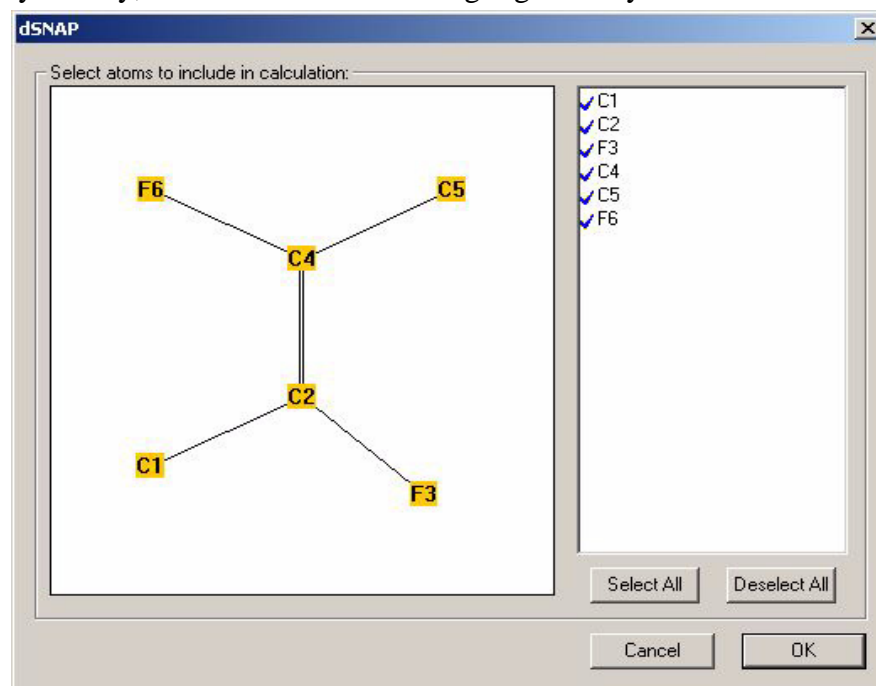
If the number of atoms in the search fragment is greater than 20 the user will be given the opportunity to select which atoms are to be used in the analysis. This is detailed in the earlier chapter on data input.

In cases where less than 20 atoms are present in the fragment this option is not offered by default and analysis will begin immediately. However a similar option is available for fragments with less than 20 atoms, but only if the *Show Advanced Data Input Processing Options* option is turned on in the *Advanced Options* (Section 9.1). In this case, a similar window will open before analysis begins:



All atoms are initially selected. Clicking on them deselects them. Those that are deselected will not be used in the cluster analysis.

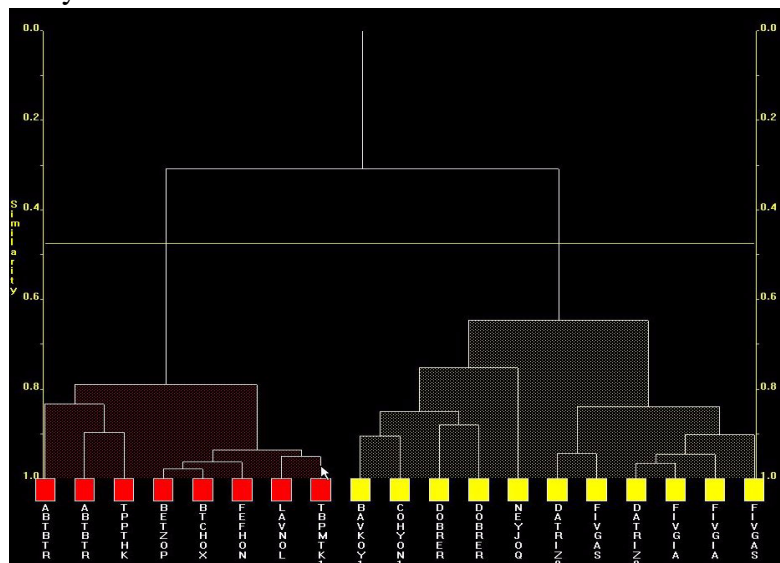
Where the search fragment contains atoms which possess topological symmetry, the affected atoms are highlighted in yellow.



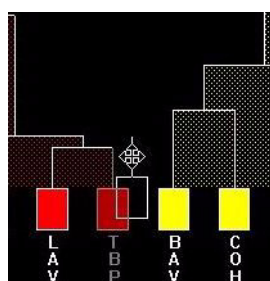
Deselecting one symmetry related atom will also automatically deselect its symmetry equivalent atoms.

11.2 Manually changing the contents of the clusters

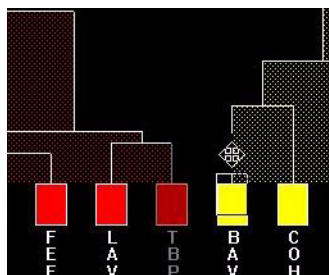
The program makes its assignment of the contents of the clusters using a dendrogram. It should not normally be necessary to override these results, but for the occasions when this is necessary, it is possible to reassign either a fragment or a group of related fragments manually.



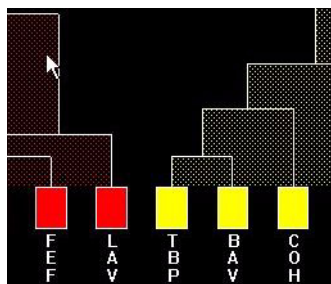
To do this, hold down the *Shift* key and click and hold down on the vertical line attached to the fragment or sub-cluster required to be moved:



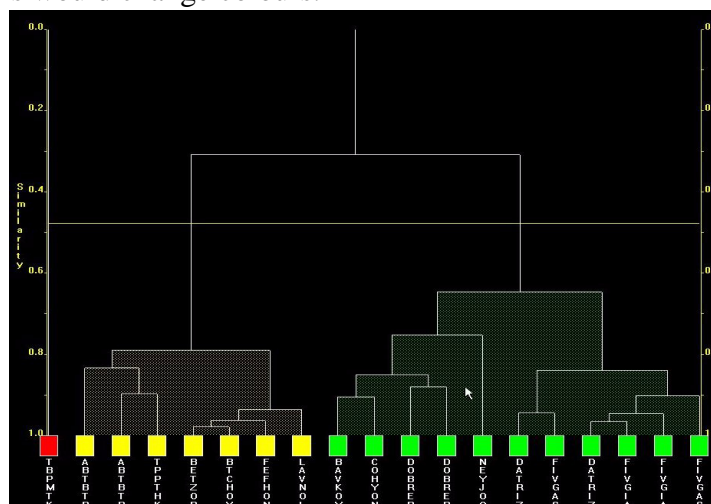
Continuing to hold down both the *Shift* key and the mouse button, drag the unit to the desired location:



To add it to an existing cluster, release the mouse when the cursor is over the dendrogram line the sample is to be added to. The fragment should now be added to the new cluster.



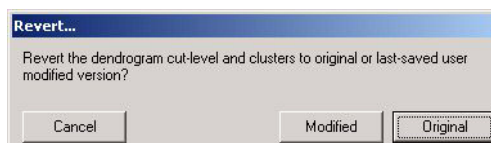
To create an entirely new, separate cluster, drag the fragment to an empty space between existing clusters. For this reason care must be taken when adding fragments to existing clusters to avoid creating new clusters instead. This can be readily seen as the pre-existing clusters would change colours.



To cancel a drag operation part-way through, it is only necessary to release the *Shift* key. As the operation is carried out through the tree lines of the dendrogram only related fragments can be moved together. If the user wishes to move two unrelated fragments, or wishes to move related fragments selectively, perhaps excluding one from a sequence of four, then this can only be achieved in a step-wise manner.

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

To undo multiple operations the user can choose to return to the original dendrogram by selecting *Undo Saved Dendrogram Modifications...* from the display menu.

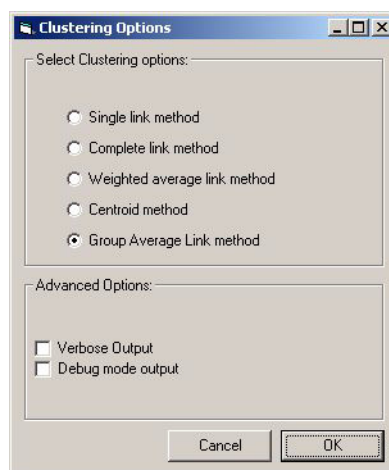


The user must select the *Original* option. Selecting *Modified* will simply refresh the current arrangement, reordering it on the screen.

Note that choosing to save manual changes to the dendrogram clusters will result in the 3D plot being updated. The spatial positioning of the spheres will not change, however they will be coloured accordingly to the new cluster that they have been assigned to and the most representative fragment will have been recalculated accordingly.

11.3 Changing the Clustering Method

The user can choose to re-run just the clustering part of the analysis to experiment with how different individual clustering methods would affect the results. To do this select *Change Cluster Method* from the *Tools* menu. An options dialog box will appear:



A detailed description of how each of these operate is listed in the chapter on Program Options. To select an individual cluster method, choose from the five options presented and click *OK*. The recalculation may take several minutes on larger data sets.

References

The primary reference is:

*d*SNAP: a computer program to cluster and classify Cambridge Structural Database searches

Barr, G., Gilmore, C. J., Parkin, A., and Wilson, C.

J. Appl. Cryst. (2005). **38**, 833-841

Other related references are:

Configurational and conformational classification of pyranose sugars

Collins, A., Parkin, A., Barr, G., Dong, W., Gilmore, C.J., and Wilson, C.C.

Acta Cryst. (2008). **B64**, 57-65.

Using small molecule crystal structure data to obtain information about sulfonamide conformation

Parkin, A., Collins, A., Gilmore, C.J., and Wilson, C.C.

Acta Cryst. (2008). **B64**, 66-71.

Using Cluster Analysis to Study Transition Metal Geometries: 4-Coordinate Transition Metals with Two Salicylaldiminato Ligands

Parkin, A., Barr, G., Collins, A., Dong, W., Gilmore, C.J., Tasker, P.A. and Wilson, C.C.

Acta Cryst. (2007). **B63**, 612-620.

The Application of Cluster Analysis to Enone and Enimine Conformation

Collins, A., Parkin, A., Middlemiss, D.S., Barr, G., Dong, W., Gilmore, C.J., and Wilson, C.C.

Acta Cryst. (2007). **B63**, 469-476

Identifying Structural Motifs in Inter-Molecular Contacts using Cluster Analysis. Part 2. Interactions of Carboxylic Acids with Secondary and Tertiary Amides

Collins, A., Parkin, A., Barr, G., Dong, W., Gilmore, C.J., and Wilson, C.C.

CrystEngComm (2007). **9**, 245-253

Identifying structural motifs in inter-molecular contacts using cluster analysis: 1. Interactions of carboxylic acids with primary amides and with other carboxylic acid group

Parkin, A., Barr, G., Dong, W., Gilmore, C. J., Parkin, A., and Wilson, C. C.

CrystEngComm, (2006). **8**, 257 - 264

High-throughput powder diffraction. IV. Cluster validation using silhouettes and fuzzy clustering

Barr, G., Dong, W. and Gilmore, C. J.

J. Appl. Cryst. (2004). **37**, 874-882

High-throughput powder diffraction. III. The application of full profile pattern matching and multivariate statistical analysis to round-robin-type data sets

Barr, G., Dong, W., Gilmore, C. J., Faber, J.

J. Appl. Cryst. (2004). **37**, 635-642

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

Barr, G., Dong, W., Gilmore, C. J.

J. Appl. Cryst. (2004). **37**, 658–664

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

Barr, G., Gilmore, C. J., Paisley, J.

J. Appl. Cryst. (2004). **37**, 665–668

High-throughput powder diffraction. II. Applications of clustering methods and multivariate data analysis

Barr, G., Dong, W. and Gilmore, C. J.

J. Appl. Cryst. (2004). **37**, 243-252

High-throughput powder diffraction. I. A new approach to qualitative and quantitative powder diffraction pattern analysis using full pattern profiles

Gilmore, C. J., Barr, G. and Paisley, J.

J. Appl. Cryst. (2004). **37**, 231-242

Index

Numerics

2D Fragment Variables viewer 77, 85

3D Fragment Viewer 75

3D Plot (MMDS) 42, 57, 85

A

Accelerate Wheel 47

C

Cell Display 42, 51, 83, 117

Centroid Method 103

CIF files 30

Clip Plane 62

Clustering Options 102

Colour 48, 65

Complete Link Method 103

ConQuest 21, 80, 100

 highlighting fragments 81

Coordinate Files (.cor) 24

Cut-level

 changing 55

 definition 54

D

Data Input 28

Data Space

 definition 6

Debug Mode 104

Dendrogram 42, 52, 84

Drag Label 64

dSNAP

 options 97

E

Eigenvalues 68

F

Find Item 46

Font 49

Fragments

 definition 4

 highlighting 81

 selecting 49

G

GOF Score 58

Grand Tour 71, 73
Graphics Controls
 alternate 95
 default 94
Graphics Toolbar 47, 99
Group Average Link Method 103

H

Help 92
Hits
 definition 4

I

Including your own structures 30
Installation 9

L

Lambda Value 98
Launching dSNAP 13
Licence Agreement 10
Limitations 6
Line of Best Fit 88
Line Size 48
Logfile 42, 73, 98
Long Labels 99

M

Mask Group 56, 62
Mercury 79, 100
Most Representative Member 59

N

Nomenclature 4
Numerical Results 42, 73, 86

O

Opaque Zone 61
Output Files 29

P

Parallel Coordinates Plot 70
Parameters Files (.fgd) 25
Popup Group 62

R

Registration 13
Render
 as dots 60
 as transparent 60
Rendering Quality 65
Requirements 9
Reset Preferences 102

Reset View 45

Run-times 6

S

Scatter Plots 87

Scree Plot 67

Search Files (.cqs) 24

Shortcut-keys 94

Show Axes 56

Show Grid 59

Show Line 88

Silhouettes 69

Single Link Method 103

Space Explorer 72

Stacks 52

Structures

 definition 4

T

Toggle Mode 52, 56

Top View 60

Transparent Group 62

Troubleshooting 17

V

Validation 42

 data 67

 general 85

Variables

 defining 21

 definition 5

Variables Space

 definition 6

Verbose Mode 104

View Manual 92

W

Weighted Average Link Method 103

Welcome Window 28

Z

Zoom 43, 45

