Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Configurational and conformational classification of pyranose sugars

A. Collins,[a,b]* A. Parkin,[b] G. Barr,[b] W. Dong,[b,c] C. J. Gilmore[b] and C. C. Wilson[b]

[a]School of Chemistry, University of Edinburgh, Edinburgh EH9 3JJ, Scotland, [b]Department of Chemistry and WestCHEM Research School, University of Glasgow, Glasgow G12 8QQ, Scotland, and [c]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

Correspondence e-mail: anna.collins@ed.ac.uk

Received 1 August 2007
Accepted 17 December 2007

Automated cluster analysis is used to examine the conformation and configuration of pyranose sugars. Previous findings on this issue are confirmed, importantly from an analysis that requires no prior knowledge of the significant factors determining the conformational classification. The findings on the conformations adopted in the crystalline solid state are found to be different to existing quantum chemical calculations performed for D-glucose in the gas phase, but consistent with empirically determined conformations in the solution state. The use of this clustering analysis in studying chirality in the determined structures is discussed, as is the ability of this type of method to examine higher dimensions within the metric multi-dimensional scaling formalism.

1. Introduction

Hexopyranose sugars found in the Cambridge Structural Database (CSD; Allen, 2002) have been studied previously by cluster analysis (Allen & Fortier, 1993). A total of 17 torsion angles were used to define the 12-atom hexopyranose fragment; the authors note, however, that the method requires some expert knowledge, including the choice of torsion angles to be used in the analysis. The problem is revisited here using a clustering method where all interatomic distances and angles, not just parameters involving bonded atoms, are used in the cluster analysis, and not the torsion angles, which are implicitly defined by the distances and angles. As the process of identifying all the distances and angles is automated, the amount of *a priori* knowledge assumed is minimized. This method has previously been applied to other systems (Parkin *et al.*, 2006; Collins, Parkin *et al.*, 2007; Collins, Barr *et al.*, 2007; Parkin *et al.*, 2007) and here is also applied to related sugar-based chemical fragments.

Perhaps the principal problem with using cluster analysis to probe molecular geometries is in determining a suitable cut-level, which determines the number of clusters (see also Parkin *et al.*, 2006). We demonstrate here that this choice is best made by application of a variety of validation and visualization tools. In this aspect of clustering procedures, even automated clustering procedures, it is currently not possible to eliminate the requirement of chemical knowledge; in order to be of value and use, the clustering should make chemical sense, and this chemical knowledge must be provided by the user. The clustering method employed is implemented in the freely distributed program *dSNAP* and described fully in Barr *et al.* (2005).

© 2008 International Union of Crystallography
Printed in Singapore – all rights reserved

doi:10.1107/S0108768107067341

2. Experimental

A search of the CSD, Version 2.7 (Allen, 2002), using *ConQuest* (Bruno *et al.*, 2002) on the hexopyranose fragment (Fig. 1) yielded 544 hit structures; the additional search parameters required that structures had three-dimensional coordinates determined and were organic. Each hit structure will contain at least one instance of the fragment defined in the search – there were 739 hit fragments within the hit structures. Multiple instances of the search fragment occur in structures with $Z' > 1$ (*e.g.* ACEHIA) or where the fragment motif occurs more than once in a single molecule (*e.g.* BLACTO, the structure of β-lactose, which is formed from glucose and galactose).

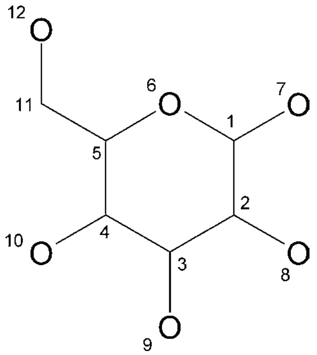

Figure 1
The hexopyranose fragment, with numbering scheme. As each atom in *dSNAP* must have a unique number, the numbering system used differs from the standard IUPAC scheme for hexopyranoses. In the *ConQuest* search, all bonds from the main heterocycle were defined as acylic and the other substituents attached to the ring defined to be hydrogen.

All pyranose rings found were of the chair form, assessed on a visual basis. The ring contains five chiral C atoms, so the fragment has $2^5 = 32$ possible stereoisomers (Allen & Fortier, 1993). In addition, the C11—O12 substituent is expected to take three different conformations, *gauche–gauche* (*gg*; also known as $-g$, where O12 points approximately perpendicular to the plane of the ring), *gauche–trans* (*gt* or $+g$, where O12 points approximately towards O6) or *trans–gauche* (*tg* or *t*, where O12 points approximately towards C4).

Clustering these hits using *dSNAP* (Barr *et al.*, 2005) at a cut-level of 81.6% similarity yields 30 clusters (Table 1, with the corresponding dendrogram shown in Fig. 2). The MMDS (metric multi-dimensional scaling) plot indicates a reasonable fit between the observed and calculated distance matrices of 0.85. Although a fit of over 0.9 is generally considered good, larger data sets are often associated with lower goodness-of-fit. In this three-dimensional plot the fragments are arranged in eight main regions, most of which consist of fragments from several different clusters in the corresponding dendrogram. These are not necessarily clusters linked by a high level of similarity in the dendrogram (*e.g.* clusters *A*, coloured red, and *S*, coloured light brown, which are linked at a similarity level of 0.665). Additionally, in the MMDS plot some of the clusters are quite diffuse. Principal component analysis indicates that more than seven clusters are required to account for over 95% of the data; this is clearly a complex data set. The similarity cut-level has been chosen such that each group of fragments is conformationally and configurationally distinct from all other groups. In the dendrogram, this is indicated as being a suitable choice from the relatively large differences of similarity between tie-bars linking separate groups. In the MMDS plot this is indicated by clear space around the clusters, in the majority of cases. The final check that is made is based on visual examination of the fragments, which is facilitated by overlaying fragments using Procrustes analysis.

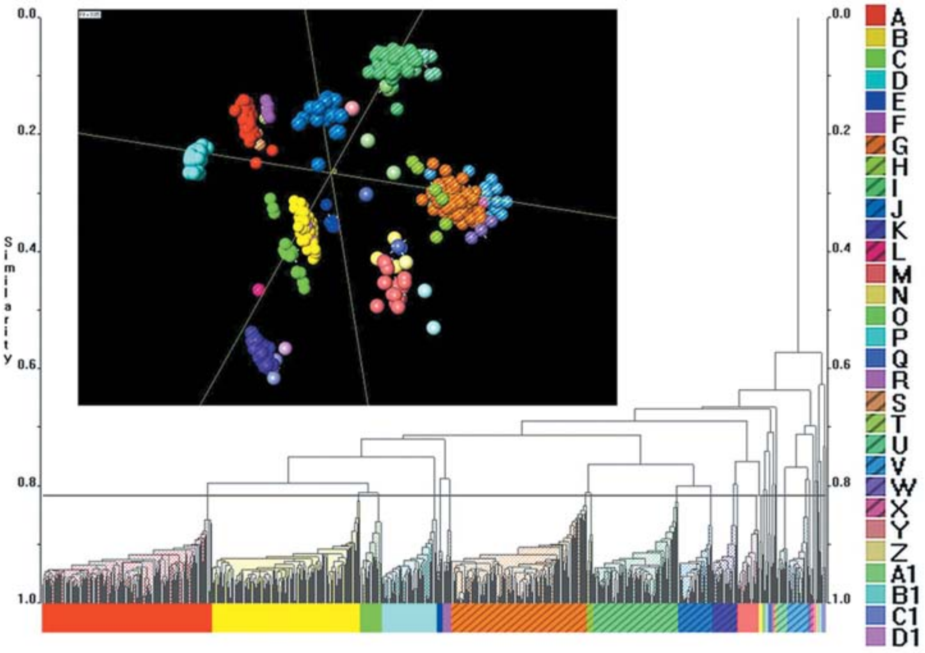

Figure 2
Clustering the hexopyranose dataset with a similarity level of 0.816 in the dendrogram. Each sphere represents a fragment, and the box has orthogonal axes with unit dimensions. The spheres are placed such that there is an optimal fit (in the least-squares sense) between the observed distance matrix generated by the fragment geometries and that calculated from the spheres in the box. The goodness-of-fit measured *via* correlation coefficients between these two matrices is 0.85. The key to the colour-coding of the fragments is also shown.

All but one of the fragments have an equatorial C11 atom. This exception comprises cluster *T*, where C11 is the only axial atom; it is an L sugar. The primary alcohol conformation found is $-g$ ($\simeq -60^\circ$) for the O6—C5—C11—O12 torsion angle (-63.50°) and *t* ($\simeq 180^\circ$) for the C4—C5—C11—O12 torsion angle (171.67°).

Table 1
Types of sugar, the number of axial groups attached to the ring and the relevant clusters.

The geometries of the fragments in each group have been drawn in the supplementary information.

	No. of		Orientation of attached group						No. of
Cluster	fragments		C1	C2	C3	C4	C5	C11	axial groups
A	160	β-Glucose	eq	eq	eq	eq	eq	$+g$	0
B	140	β-Glucose	eq	eq	eq	eq	eq	$-g$	0
C	20	β-Glucose	eq	eq	eq	eq	eq	*t*	0
D	52	β-Galactose	eq	eq	eq	ax	eq	$+g$	1
E	6	β-Mannose	eq	ax	eq	eq	eq	$-g$	1
F	8	β-Mannose	eq	ax	eq	eq	eq	$+g$	1
G	128	α-Glucose	ax	eq	eq	eq	eq	$-g$	1
H	6	α-Glucose	ax	eq	eq	eq	eq	*t*	1
I	80	α-Glucose	ax	eq	eq	eq	eq	$+g$	1
J	32	α-Galactose	ax	eq	eq	ax	eq	$+g$	2
K	23	β-Galactose	eq	eq	eq	ax	eq	*t*	1
L	1	β-Galactose	eq	eq	eq	ax	eq	$-g$	1
M	20	α-Galactose	ax	eq	eq	ax	eq	*t*	2
N	3	α-Galactose	ax	eq	eq	ax	eq	$-g$	2
O	3	α-Allose	ax	eq	ax	eq	eq	$+g$	2
P	3	α-Allose	ax	eq	ax	eq	eq	$-g$	2
Q	3	α-Gulose	ax	eq	ax	ax	eq	$-g$	3
R	2	β-Allose	eq	eq	ax	eq	eq	$-g$	1
S	2	β-Allose	eq	eq	ax	eq	eq	$+g$	1
T	1	α-Idose	eq	eq	eq	eq	ax		1
U	10	α-Mannose	ax	ax	eq	eq	eq	$+g$	2
V	19	α-Mannose	ax	ax	eq	eq	eq	$-g$	2
W	3	α-Mannose	ax	ax	eq	eq	eq	*t*	2
X	3	α-Altrose	ax	ax	ax	eq	eq	$-g$	3
Y	2	α-Talose	ax	ax	eq	ax	eq	$+g$	3
Z	1	α-Talose	ax	ax	eq	ax	eq	*t*	3
*A*1	2	α-Idose	ax	ax	ax	ax	eq	$+g$	4
*B*1	2	α-Idose	ax	ax	ax	ax	eq	*t*	4
*C*1	3	β-Gulose	eq	eq	ax	ax	eq	*t*	2
*D*1	1	β-Idose	eq	ax	ax	ax	eq	*t*	3
Total	739								

There are 39 instances of the uronate fragment (Fig. 3) found in the dataset (5.3%). These have a very small impact on the overall clustering and their impact is greatest when considering the conformation at C11 (see Table 2).

3. Results and discussion – hexopyranose sugars

3.1. Analysis of the clustering

As in Allen & Fortier (1993), it is easy to identify the different types of sugars observed and the most common types are as found in this earlier work (shown in Table 1). In this earlier study, it was suspected that the β-L-glucose structures were in fact the wrong enantiomorph and the coordinate sets of the affected structures had been inverted. Disregarding the enantiomeric type D or L, we find the dataset comprises five principal types, which account for over 90% of the hexopyranose sugars: β-glucose (320; 43.3%); α-glucose (214; 28.9%); β-galactose (76; 10.3%); α-galactose (55; 7.4%) and α-mannose (32; 4.3%). The problem of dealing with enantiomers is explored below.

The different orientations of O12 can be identified by the O6—C5—C11—O12 torsion angle, either by itself or in combination with the C4—C5—C11—O12 torsion angle. It is possible to summarize the different conformations using the O6—O12 and C4—O12 distances. While there is some spread in the data (Fig. 4), there are clearly three groups, corresponding to the orientations $-g$ (short C4,O12 and O6,O12 distances), $+g$ (long C4,O12 and short O6,O12 distances) and *t* (short C4,O12 and long O6,O12 distances). Most of the data-points that deviate from the three principal groups can be attributed to uronate fragments (Fig. 5). It should also be noted that BIKWOH10 (cluster T) has a O6—C5—C11—O12 torsion angle of approximately −60° (*gauche*), and a C4—C5—C11—O12 torsion angle of approximately 180° (*trans*). It has a long C4,O12 distance and a short O6,O12 distance, so it appears as $+g$ in the scatterplot in Fig. 4.

It is not possible to cluster these data on the basis of the sugar configuration only (*i.e.* forming clusters which contain only glucose, galactose *etc.*); the position of O12 relative to C11 has a crucial effect on the overall clustering. Thus, although clusters *A*, *B* and *C* (all forms of β-glucose, but with

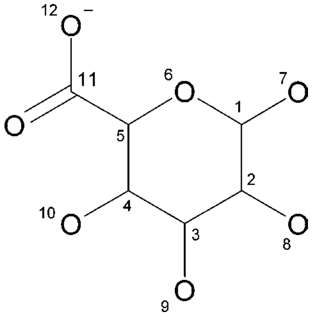

Figure 3
The uronate fragment. Owing to the way that the hexopyranose search fragment was defined in the original search, a uronate fragment defined with delocalized charge over the carboxylate group will not be found.

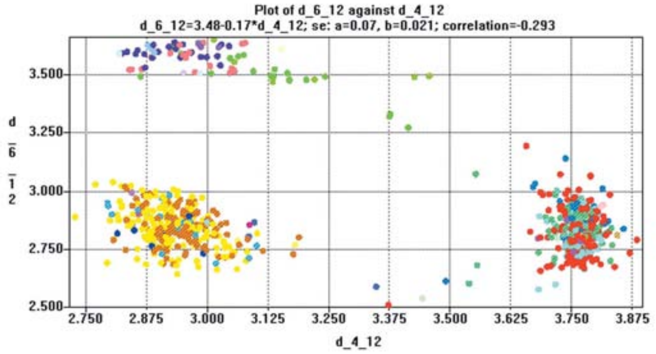

Figure 4
Scatterplot for the C4,O12 and the O6,O12 distances. Although there is some spread in the data, there are three distinct groups, which correspond to the three principal orientations of the C11,O12 group.

different orientations of O12) form a single cluster at a higher cut-level, clusters *D*, *K* and *L* (all of the β-galactose configuration, with different dispositions of O12; see Fig. 6) do not without also incorporating other ring configurations. In this case, raising the cut-level to 0.686 would put all fragments with the configuration of β-galactose into a single cluster, but would also incorporate all the fragments that are found in clusters *A*–*K* at the 0.816 cut-level; such clustering would have little chemical meaning.

3.2. Dealing with chirality

The geometries used by *dSNAP* make no allowance for the absolute stereochemistry of the defined fragment. Within each fragment the relative stereochemistry is retained, but because the fragment is defined as a series of scalar interatomic distances and angles that define the fragment in the distance matrix, enantiomers will be equivalent. This provides additional data points when comparing structures, but there is the loss of potentially valuable data. There are several ways in which this issue can be addressed, dependent on the structural information that is required from the analysis.

(i) The absolute configuration can be disregarded, *i.e.* the *dSNAP* analysis can be used as it stands. This provides the largest dataset, which may be particularly valuable if the search is primarily concerned with identifying outliers on the basis of bond distances and angles. In this situation, the defined stereochemistry of a given structure can still be viewed, for example using *Mercury* (Macrae *et al.*, 2006).

(ii) The initial *ConQuest* search can be restricted to a particular stereochemistry by defining the geometry around any stereocentre. In this case it would only be necessary to define a single stereocentre on the ring as the stereochemistry at any other stereocentre in the structure will follow from this definition.

(iii) *dSNAP* analysis can be performed on the complete dataset, and a subsequent search on the stereochemistry can be performed, as in (ii) above. The clustering data can be combined with this additional stereochemical data to allow enantiomers in a cluster to be identified.

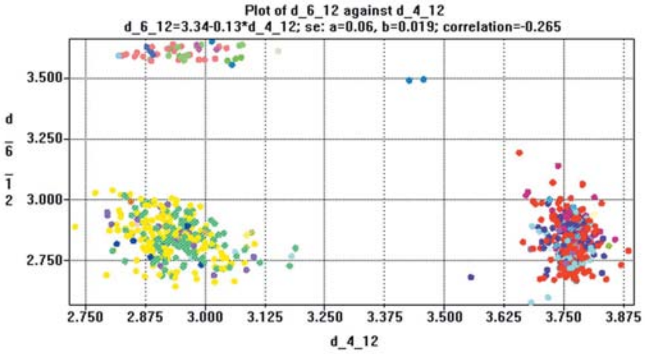

Figure 5
Scatterplot of C4,O12 and O6,O12 distances, excluding uronate fragments. Note that the fragment colouring is different from that in Fig. 4; this is because any change in the data being clustered (in this case the exclusion of uronate fragments) affects the clustering. However, it illustrates that the removal of uronate fragments from the analysis reduces the spread in the data.

However, the absolute stereochemistry is not always determined in a crystallographic experiment and there is not necessarily any way of knowing whether an attempt to identify the correct stereochemistry has been made, even if the original publication is consulted. Therefore there is no guarantee that the stereochemistry found in the CSD is the correct absolute stereochemistry.

The majority of hit structures in the present study are found in Sohncke space groups, mainly $P2_1$ (36.9%), $P2_12_12_1$ (46.0%), $C2$ (6.6%) and $P1$ (4.4%). Only two structures are found in centrosymmetric space groups: FIXYOA ($P\bar{1}$) has Z' = 1 and CUXFAC ($P2_1/c$) has Z' = 2. Thus, there is one hit fragment in FIXYOA, while there are two in CUXFAC, *i.e.* the hit fragment found is dependent only on the contents of the asymmetric unit. Even though both enantiomers are present in the crystallographic unit cell, only one hit fragment is found per independent molecule as the enantiomers are crystallographically identical.

The pseudo-torsion angle O6,C5,C11,C4 takes approximate values of −120° for D sugars and +120° for L sugars. There are 18 structures which are found to have a torsion angle of approximately +120°, suggesting that they are L sugars; a list of CSD refcodes of these structures and the fragment number assigned in *dSNAP* where appropriate, is given in the supplementary information. Of these, BIKWOH10 has an axial C11 atom. By defining the geometry at C5, it is possible to establish that, out of the 739 fragments in the full dataset, only 17 (2.3%) appear to be L sugars; because there are so few of these fragments it is simple to isolate them by hand for further examination.

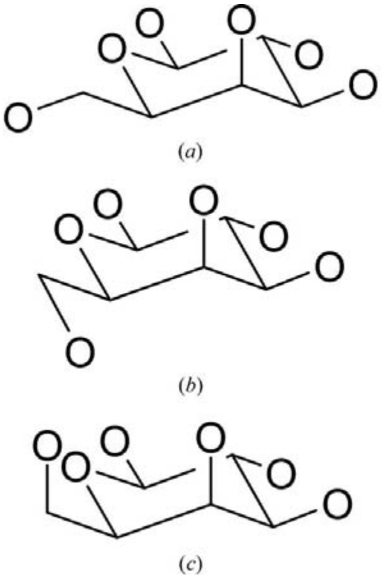

Figure 6
The geometries of clusters (*a*) *D*, (*b*) *K* and (*c*) *L*.

Table 2
The number of uronate fragments found in each cluster.

Cluster	No. of uronate fragments
C	14
H	2
I	3
J	1
K	2
M	10
Q	1
*A*1	3
*B*1	2
*D*1	1

3.3. Cluster identification at higher dimensions

The MMDS plot shows that some of the clusters identified in the dendrogram appear in the same locations in the three-dimensional space representation, for example clusters *G* and *P*, where fragments in cluster *P* appear embedded in the fragments belonging to cluster *G* (see Fig. 7). MMDS plots are not restricted to them; it is quite possible to place the spheres in a box of arbitrary dimensionality up to the size of the distance matrix. Usually the first three dimensions are sufficient and the correlation coefficient, quoted as a figure of merit, is a useful indicator of this. In very complex cases, such as this, three dimensions may not suffice and the *dSNAP* software permits the exploration of higher dimensionality. Viewing the data using the higher dimensions available in the analysis (Fig. 7, right) shows that the embedding of clusters is not symptomatic of incorrect clustering (and the different geometries exhibited by these two clusters confirm that it is inappropriate for them to be clustered together), but that the distinction between them cannot be adequately described in the three-dimensional metric space of the analysis.

3.4. Anomeric effect

The anomeric effect leads to a systematic shortening of the C1—O7 bond relative to other C—O bond lengths. Exporting these bond lengths directly from the CSD is probably the best route to studying this effect, but it can also be observed using the validation tools that are included with the *dSNAP* clustering software (Barr *et al.*, 2007). This may facilitate the correlation of effects such as this bond shortening on the overall clustering of the entire dataset.

Table 3
The relative energies of different dispositions of O12, calculated assuming Boltzmann statistics for the relative populations, for the commonly occurring sugar configurations, glucose and galactose in both α and β forms.

Configuration	Proportion of dataset	Relative energy (kJ mol^{-1})	Proportion of dataset	Relative energy (kJ mol^{-1})
	β-Glucose		α-Glucose	
$-g$	0.4375	0.33	0.5981	0
$+g$	0.5000	0	0.3738	1.17
t	0.0625	5.16	0.0280	7.59
	β-Galactose		α-Galactose	
$-g$	0.3026	2.02	0.594	0
$+g$	0.6842	0	0.313	1.59
t	0.0132	9.80	0.094	4.58

In this case the anomeric effect can be observed (see Fig. 8), although it is not possible to discern from these scatterplots that the β anomer is more strongly affected.

3.5. Energy calculations

Estimates of the relative energies of the different conformations for a given type of sugar can be made on the basis of their relative populations in the CSD (Table 3). A Boltzmann distribution with a temperature of 298 K is assumed. Similar estimates for enone fragments in the CSD were found to show good agreement with gas-phase calculations (Collins, Barr *et al.*, 2007). The method assumes that the database is not biased, and randomly samples the relevant sample space; this is unlikely to be the case, but by considering each type of sugar separately this bias can be minimized. The O12 orientation is less likely to be under the control of the synthesis and more likely to be governed by crystal packing, whereas the configuration of the sugar under investigation will be much more heavily determined by the synthesis.

Theoretical studies of glucose indicate that in the gas phase the α anomer is more stable than the β anomer (see, for example, Corchado *et al.*, 2004), while in solution this situation is reversed. Empirical results show the α:β anomers ratio to be 36:64 (Angyal, 1968, 1969; which corresponds to an energy difference of 1.42 kJ mol^{-1}, assuming a Boltzmann distribution), while theoretical values of the ratio in solution estimate an energy difference of 0.84 kJ mol^{-1} (Barrows *et al.*, 1998). In our study,

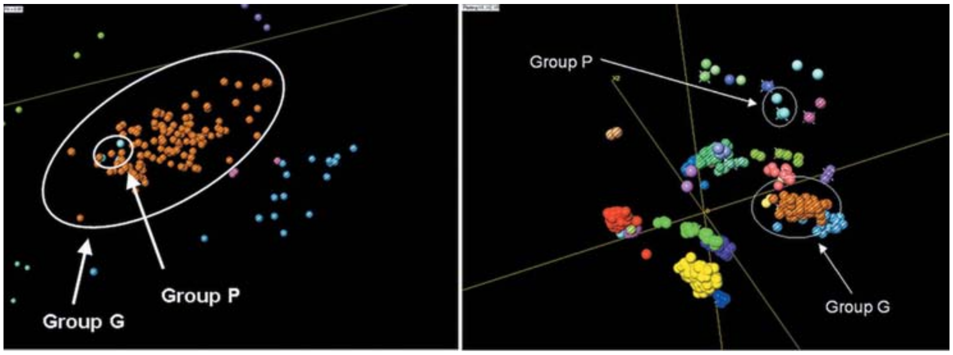

Figure 7
(*a*) The MMDS plot for the sugar clustering using dimensions 1, 2 and 3, showing group *P* (light blue) embedded in group *G* (orange–brown). Each sphere represents a single fragment. The size of the spheres has been reduced in order that those representing fragments in group *P* can be seen. (*b*) The three-dimensional MMDS plot using dimensions 1, 2 and 5. Group *P* is now distinct from group *G*, confirming the distinct nature of these clusters.

D-α-glucose-based sugars (214 fragments) are less common than D-β-glucose-based sugars (320 fragments) in a 40.1:59.9 ratio. This corresponds to an energy difference of 1.00 kJ mol^{-1}, assuming a Boltzmann distribution and random sampling of the sample space.

Also noteworthy is the effect of the orientation of the hydroxymethyl group. In the gas phase the *t* conformer of α-glucose is the lowest energy form, while in the solid state, on the basis of the incidence of the conformations in D-α-glucose-based fragments in the CSD, this form is the highest-energy conformer, with the lowest incidence. Similarly, gas-phase calculations indicate that the highest energy form of glucose is the β +*g* conformer, which appears to have the lowest energy in the solid state, with the highest incidence within D-β-glucose-based fragments. Interestingly, in solution, NMR experiments indicate that the *t* conformer is the least common and crystal structures derived from initial association in solution could be expected to reflect this.

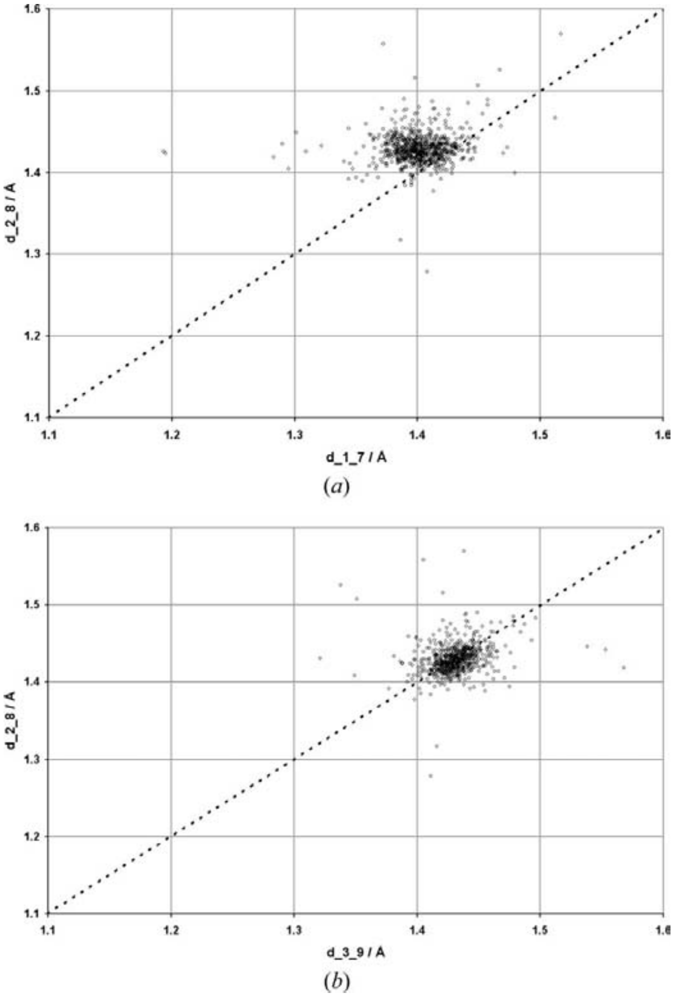

Figure 8
(*a*) Plot of the C2—O8 distance against the C1—O7 distance. The anomeric effect would be expected to affect the C1—O7 distance and this bond length is shown to be shorter than the unaffected C2—O8 distance. (*b*) The C2—O8 distance against the C3—O9 distance, showing these bonds are of approximately the same length, and emphasizing the influence of the anomeric effect on the C1—O7 distance seen in Fig. 8(*a*).

Table 4
Types of deoxyhexopyranose sugar, the number of axial groups attached to the ring and the relevant clusters.

Cluster	No. of fragments	O7	O8	O9	O10	C11	α or β anomer
A	37	ax	eq	eq	eq	eq	α
B	7	ax	eq	eq	ax	eq	α
C	2	ax	eq	ax	eq	eq	α
D	33	ax	ax	eq	eq	eq	α
E	1	ax	ax	eq	ax	eq	α
F	9	eq	eq	eq	eq	eq	β
G	2	eq	ax	eq	eq	eq	β
H	1	eq	ax	ax	eq	eq	β
I	4	eq	eq	ax	eq	eq	β
J	6	eq	eq	eq	ax	eq	β
K	1	eq	eq	eq	ax	eq	β
L	1	ax	eq	eq	eq	eq	α
M	1	ax	ax	ax	ax	eq	α

However, there are several important caveats. The content of the CSD is not a reflection of stability. A crystal structure may reflect a metastable state, for example. There is also the 'social bias' of the crystal structures in the database; materials contained in this resource are those that are of interest to chemists, biochemists and solid-state scientists, and so the sugars in the database are likely to be biased towards those that have pharmaceutical applications or are involved in important biochemical pathways, for example; this may be particularly important when comparing the occurrence of α and β anomers.

Also, importantly, the theoretical energy calculations described were carried out on glucose, $C_6H_{12}O_6$, while the fragment used for the database search is less restrictive, as the ring O atoms, O6–O10, could carry an *R* group other than hydrogen. This greatly increases the number of fragments for cluster analysis, but also introduces a large number of variables, such as an increase in steric factors that can affect the conformation within the fragment, and the presence of additional groups that may affect both inter- and intramolecular hydrogen bonding.

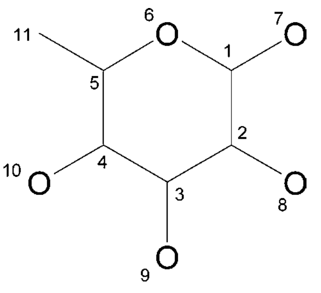

Figure 9
6-Deoxyhexopyranose fragment, with numbering scheme. In the *ConQuest* search, all bonds from the main heterocycle were defined as acylic and the other substituents attached to the ring defined to be hydrogen. Note that the numbering scheme is that used in the clustering in *dSNAP* and does not correspond to the conventional numbering of 6-deoxyhexopyranose

Table 5
Types of pyranose sugar, the number of axial groups attached to the ring and the relevant clusters.

Group *G* does not have a ring in the chair form; it is a skew boat.

Cluster	No. of fragments	O7	O8	O9	O10	
A	2	ax	eq	ax	eq	α-Ribopyranose
B	18	ax	eq	eq	ax	β-Arabinose
C	13	ax	eq	eq	eq	α-Xylopyranose
D	5	eq	eq	eq	ax	α-Arabinose
E	36	eq	eq	eq	eq	β-Xylopyranose
F	3	eq	ax	eq	eq	β-Lyxopyranose
G	2	ax	eq	eq	eq	β-Xylopyranose
H	6	ax	ax	eq	ax	β-Ribopyranose
I	2	ax	ax	eq	eq	α-Lyxopyranose
J	2	ax	ax	ax	ax	β-Xylopyranose

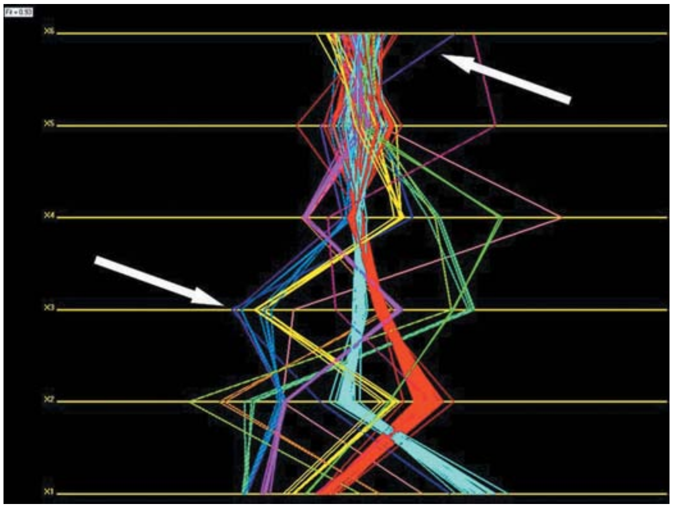

Figure 11
Parallel-coordinates plot in six dimensions. Note how the purple line, corresponding to group *K*, only diverges from group *J* (mid-blue) in the sixth dimension (at the top of this plot). The white arrow on the left at the third dimension indicates where the groups are coincident; the arrow at the top right indicates where the groups diverge.

4. Other pyranose sugars

The analysis performed on the hexopyranose sugars has also been performed on other pyranose sugar derivatives.

4.1. 6-Deoxyhexopyranose derivatives

The fragment was defined as in Fig. 9. As drawn, the fragment corresponds to atoms 1–11 of the hexopyranose fragment. In order to derive the appropriate dataset, the refcode list of this new search was compared with the refcode from the previous search using *ConQuest* to give a hitlist consisting only of deoxyhexopyranose structures. There were 84 resulting hit structures yielding a total of 105 fragments for cluster analysis. Clustering at a cut-level of 81.2% similarity yields 13 clusters (Fig. 10 and Table 4). The MMDS plot indicates good fit (0.93). The similarity level was determined using a combination of examination of the dendrogram, MMDS plot and validation scatterplots, as in the hexopyranose case discussed above. The cut-level is different from the hexopyranose case because the similarity can only be assessed between the fragments of a given dataset, and so similarity values are not directly comparable between datasets.

A *ConQuest* search of this fragment based on the pseudo-torsion angle O6—C5—C11—C4 shows that approximately equal numbers of hits had this angle as +120 and −120° (52 and 53 fragments, respectively). Interestingly, this class of deoxyhexose sugars includes rhamnose, which is found in group *G*, and fucose which is found in group *B*, both of which occur in the L form in nature.

As in the hexopyranose case, this problem is not adequately described in the three-dimensional MMDS plot. As previously described, MMDS is not, however, constrained to three dimensions and the space explorer tools in *dSNAP* allow the user to explore up to six dimensions.

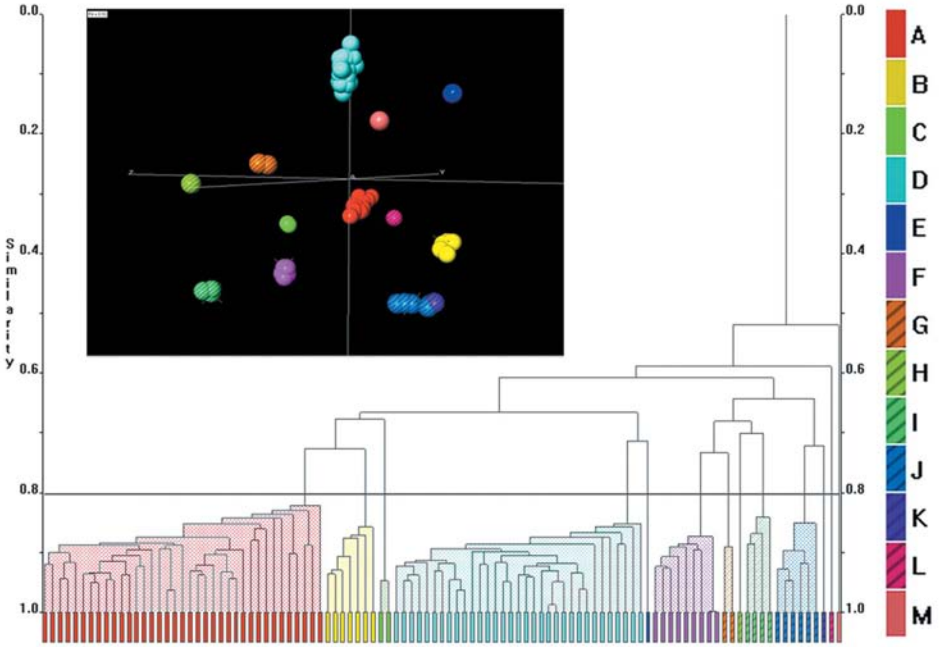

Figure 10
Clustering the dataset with a dendrogram cut-level of 0.812 in the dendrogram. The inset shows the corresponding MMDS plot, where fragments are coloured as in the dendrogram. The goodness-of-fit is 0.93.

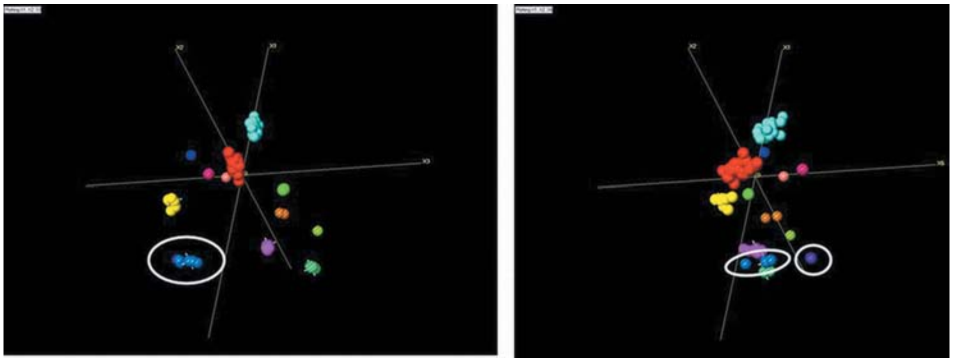

Figure 12
dSNAP ‘Space explorer’ plots using different dimensions in the MMDS analysis. In the plot on the left, which are the standard *x*, *y*, *z* coordinates, groups *J* and *K* appear in the same cluster. They are separated in the plot on the right when looking at dimensions 1, 2 and 6. The relevant areas are circled in white.

In addition, parallel-coordinate plots can be used to investigate higher dimensions (Inselberg, 1985). The principle is simple and elegant: to show a set of points in an *n*-dimensional space, a set of *n* parallel, equally spaced, horizontal lines is drawn. A point in *n*-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the *i*th axis corresponds to the *i*th coordinate of the point. These tools show that groups *J* (blue) and *K* (purple) only appear to lose their equivalence in the sixth dimension (see Figs. 11 and 12). Assigning them to separate clusters is indicated by their large difference in similarity in the dendrogram.

However, unlike the case in the hexopyranoses where different types of sugar were clustered together in the MMDS plot, groups *J* and *K* are of the same type (see Fig. 13), and the differences arise as a result of a very short C1—O7 bond distance, and a longer C1—O6 distance, which suggests that there is a structural distortion that distinguishes these two clusters. The scatterplot (see Fig. 14) indicates that cluster *K* is a wide outlier, and so that structure should be carefully examined.

4.2. Pyranose sugars

The dataset based on the search fragment for the pyranose ring, defined in Fig. 15, contained 89 fragments from 79 structures in the CSD. At a similarity cut-level of 76.3% there are ten clusters (Fig. 16 and Table 5). At this cut-level, there is a large jump in similarity to the next tie-bar, indicating that above this point the fragments are quite different. In the MMDS plot, the clusters are clearly separated in space, which is another indication that this is a good choice of cut-level. The goodness-of-fit in the MMDS plot is 0.94.

Groups *B*, *D*, *H* and *J* occur as L sugars. Group *G* contains structures where the pyranose ring is of the skew boat form (from refocde MTBZXP10).

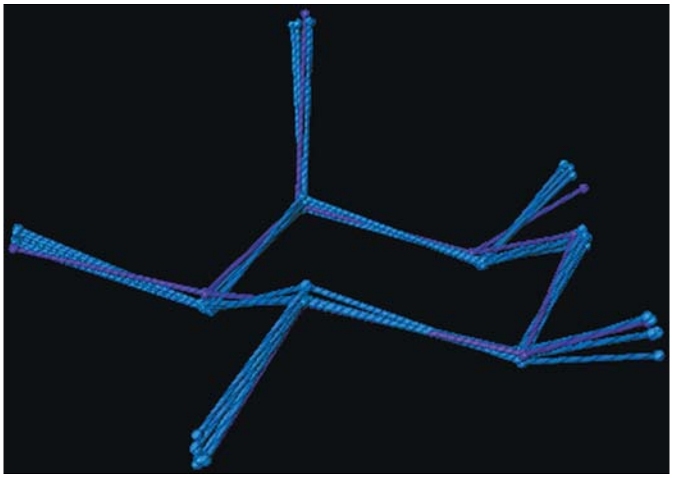

Figure 13
Overlay of group *J* (blue) and *K* (purple), showing the high degree of similarity in the conformation – the difference between these two clusters lies in a pair of bond lengths and cluster *K* is indicated as an outlier from the additional validation tools in *dSNAP*.

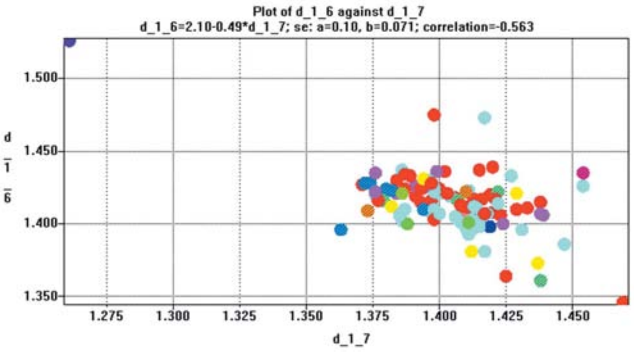

Figure 14
Plot of C1—O6 distance against C1—C7 distance, indicating that group *K* (purple point, top left) may represent a structural error.

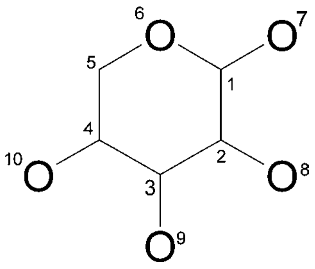

Figure 15
Pyranose fragment, with numbering scheme. In the *ConQuest* search, all bonds from the main heterocycle were defined as acylic and the other substituents attached to the ring defined to be hydrogen, including both substituents on C5.

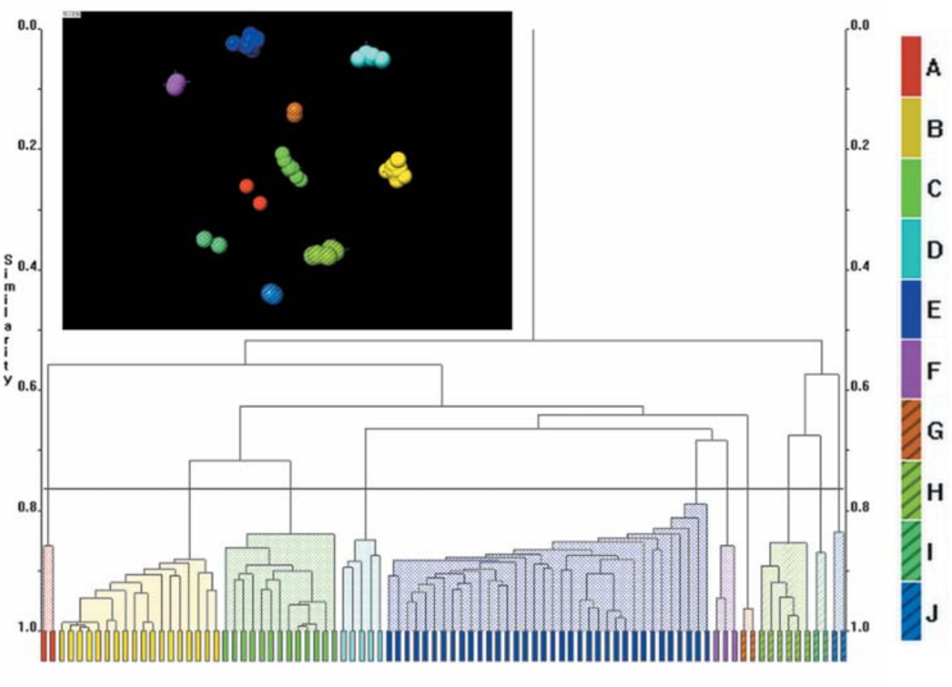

Figure 16
Dendrogram and MMDS plot (inset) for clustering the pyranose derivatives. The dendrogram cut-level is 0.763. Colours in the MMDS are taken from the dendrogram. The goodness-of-fit is 0.94.

Six fragments occur in centrosymmetric space groups $P2_1/c$ (ABINOR, ABINOR01-04, all in group *B*) and $C2/c$ (BIBWUE, in group *H*).

5. Conclusions

This work has shown the power of automated cluster analysis, as implemented in *dSNAP*, in revealing structural trends within six-membered sugar ring fragments. In augmenting previous studies of these fragments, the various tools available within the *dSNAP* program have been found to be invaluable, including higher-dimensional MMDS plots, which this work shows can distinguish clusters which appear coincident in the standard three-dimensional MMDS plot, and the use of scatterplots for visualization of individual parameters, which can help in the rapid identification of outliers – and is a major strand in the validation aspects of *dSNAP* (Barr *et al.*, 2007). Any structure that is indicated as an outlier by these visualization tools should be carefully examined.

The issue of chirality and absolute configuration has also been investigated – treating these aspects of a database structural analysis can suffer from incomplete or inaccurate information being deposited in the CSD. The presence of such problematic information can be identified within the methodology used here by defining a critical parameter or parameters to be included in the clustering and which will clearly distinguish the chirality of the fragment of interest. The combination of fully automated clustering based on all distances and angles – the default in *dSNAP* – together with intelligent interpretation based on chemical and structural knowledge has revealed the anticipated distribution of sugar conformations in the various systems studied.

With the populations of each of these geometries readily available from such an analysis, simple energy calculations have been carried out and compared with gas- and liquid-state estimates of the energy difference between α and β anomers of D-glucose-based structures, and between conformations of the C11—O12 group. Interestingly, the solid-state energy rankings are in close agreement with that indicated from solution studies, but reversed from that indicated in the gas state. This perhaps reflects the fact that the crystal structure generally emerges from molecular association in solution.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Allen, F. H. & Fortier, S. (1993). *Acta Cryst.* B**49**, 1021–1031.

Angyal, S. J. (1968). *Aust. J. Chem.* **21**, 2737–2746.

Angyal, S. J. (1969). *Angew. Chem.* **81**, 172–182.

Barr, G., Dong, W., Gilmore, C. J., Parkin, A. & Wilson, C. C. (2005). *J. Appl. Cryst.* **38**, 833–841.

Barr, G., Dong, W., Gilmore, C. J., Kern, A., Parkin, A. & Wilson, C. C. (2007). *Z. Kristallogr. Suppl.* **26**, 209–214.

Barrows, S. E., Storer, J. W., Cramer, C. J., French, A. D. & Truhlar, D. G. (1998). *J. Comput. Chem.* **19**, 1111–1129.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Collins, A., Barr, G., Dong, W., Gilmore, C. J., Middlemiss, D. S., Parkin, A. & Wilson, C. C. (2007). *Acta Cryst.* B**63**, 469–476.

Collins, A., Parkin, A., Barr, G., Dong, W., Gilmore, C. J. & Wilson, C. C. (2007). *CrystEngComm*, **9**, 245–253.

Corchado, J. C., Sanchez, M. L. & Aguilar, M. A. (2004). *J. Am. Chem. Soc.* **126**, 7311–7319.

Inselberg, A. (1985). *Vis. Comput.* **1**, 69–91.

Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & van de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.

Parkin, A., Barr, G., Dong, W., Gilmore, C. J. & Wilson, C. C. (2006). *CrystEngComm*, **8**, 257–264.

Parkin, A., Barr, G., Collins, A., Dong, W., Gilmore, C. J., Tasker, P. A. & Wilson, C. C. (2007). *Acta Cryst.* B**63**, 612–620.