Acta Cryst. (1993). B49, 1021–1031

Stereochemical and Conformational Classification of the Hexopyranose Sugars Using Numerical Clustering Methods

BY FRANK H. ALLEN

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

AND SUZANNE FORTIER

Departments of Chemistry and Computing and Information Science, Queen's University, Kingston, Ontario, Canada K7L 3N6

(*Received* 22 *April* 1993; *accepted* 5 *August* 1993)

Abstract

The methods of single-linkage and complete-linkage cluster analysis have been used to generate, in a rapid and semi-automatic manner, a variety of configurational and conformational classifications of 249 hexopyranose fragments retrieved from the Cambridge Structural Database. Conformational aspects of the fragment were described using standard torsion angles, while configurations adopted at the ring carbons were described by five projected valence angles. The classification experiments show that (*a*) the pyran ring adopts a $^4C_1$-chair conformation in all but ten of the fragments, where it is found that the conformation adopted is that of a $^1C_4$ chair, (*b*) only 14 (of a possible 32) hexopyranose stereoisomers are represented in the crystallographic data, and (*c*) the C6—O6 side chain adopts only *+gauche*, *−gauche* and *trans* conformations. Clustering methods were also used to obtain representative orthogonal coordinates that permit the building of approximate models of hexopyranose units for which crystal structures are not yet available. The analyses indicate that the ten crystal structures having a $^1C_4$ pyran-ring conformation may be reporting coordinates which describe the wrong enantiomorph. This automated data analysis indicates desirable extensions of the cluster-analysis methodology and also suggests improvements to the evaluation, search and display facilities of the Cambridge Structural Database System.

Introduction

The ability to depict, classify and compare two-dimensional (2D) and three-dimensional (3D) patterns is an integral part of human learning and understanding. Once we have accumulated a personal pattern library, we may recognize and classify new patterns by noting their points of similarity and dissimilarity to existing entries in the library. The representations we use for these patterns are crucial (Marr & Nishihara, 1978; Marr, 1982), since they permit us to segment complex patterns into their meaningful parts and to recognize those parts, and the relationships between them, at various levels of detail. These fundamental processes of pattern representation, recognition and classification are clearly exemplified in the field of structural chemistry. The representations that we use there may be spatial, *e.g.* a formal 2D chemical structural diagram, visual, *e.g.* a 3D plot of a molecular structure, or they may be numerical, *e.g.* a set of internal (geometrical) coordinates. The classifications that we deduce may describe general concepts, such as steroid, metal complex, protein *etc.*, or they may be more specific and describe localized concepts, such as functional groups, *cis–trans* relationships, boat or chair conformations, helices and β-sheets, *etc.*, concepts that describe subpatterns and their relationships.

Nowadays, pattern recognition and classification are major activities related to the sub-area of artificial intelligence research that is concerned with machine learning (see *e.g.* Carbonell, 1990; Schalkoff, 1992). Pattern recognition (PR) is important in many diverse areas such as image processing, computer vision, speech recognition and fingerprint analysis. PR methodologies can involve statistical and numerical techniques, syntax-analysis or neural-network approaches. 'Statistical' PR is used when patterns can be represented numerically. In some cases, possible categories for the classification may have been predefined and the task consists of assigning instances to those categories. This is a form of supervised learning. In the absence of externally predefined categories, we must resort to the techniques of unsupervised learning. The methods of cluster analysis (Everitt, 1980) fall into this latter category.

Some earlier papers (Allen, Doyle & Taylor, 1991*a*,*b*,*c*; Allen & Taylor, 1991) have described the development of clustering algorithms for application

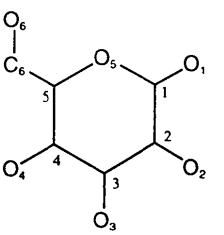
to crystallographic data contained in the Cambridge Structural Database (CSD; Allen, Davies *et al.*, 1991). These algorithms were applied to the partitioning of the conformations of a set of six-membered rings, where each conformation was described by the six intra-annular torsion angles. More recently, these methods have been applied to an analysis of seven- and eight-membered rings (Allen, Howard & Pitchford, 1993) to learn about the conformational preferences of these systems as observed in crystal structures. By this means, we 'discover' the archetypal conformational concepts (such as chair, twist-chair, twist-boat–chair, *etc.*) that apply in these substructures, and can visualize conformational interconversions in relation to low-energy features of the relevant potential energy hypersurface.

In the present paper, we apply cluster-analysis techniques to examples of the hexopyranose structure (I) that were retrieved from the CSD. Here the classification problem involves both conformational and configurational variations arising (*a*) from the different conformations that may be adopted by the pyran ring and by the side-chain oxygen O6, and (*b*) from the alternative stereochemical arrangements that may be adopted by the ring substituents O2—O5 and C6. Obviously, a considerable body of knowledge concerning hexopyranose structure and stereochemistry already exists: it is copiously treated in standard texts (a brief summary is given below) and available crystallographic data have recently been reviewed by Jeffrey (1990). Here, however, we are more concerned with the general problem of automated knowledge acquisition from crystallographic data (see *e.g.* Allen, Rowland, Fortier & Glasgow, 1990). In particular, we wish to answer three questions: (i) can we classify the hexopyranose structures using only the information available in the CSD [this information does not, at present, include codified stereochemical descriptors, for example the *R,S* descriptions of Cahn, Ingold & Prelog (1956)], (ii) can we gain insight from the classification, and (iii) can we predict any missing classes. In assessing the success of any automated methodology, we must begin by using well known examples. In the event, this work also has important implications for the CSD itself, by identifying potentially incorrect stereochemical assignments in the crystallographic literature and by indicating desirable upgrades to CSD search, display and data-analysis procedures.

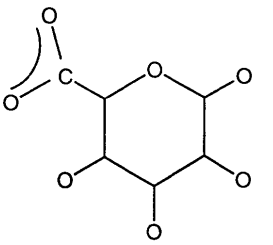
The hexopyranose sugars

Configurational aspects

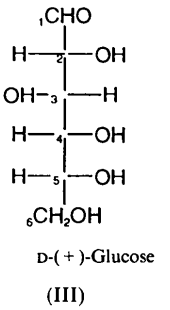
The hexopyranose sugars of general formula (I) represent cyclic hemiacetal forms of the straight-chain aldohexoses (III). The latter have four asymmetric centres and, hence, $2^4 = 16$ stereoisomers. These belong to two enantiomeric families, the D-series and the L-series, corresponding to *R* and *S* chirality at C5 (III), respectively. The 16 stereoisomers are D- or L-allose, altrose, glucose, mannose, gulose, idose, galactose and talose. Cyclization of (III) generates an additional asymmetric centre at C1 in (I). The number of stereoisomers rises to $2^5 = 32$ and each of the original 16 stereoisomers now has an α- or β-anomer due to the alternative *S* or *R* chiralities at C1. The α-, β-, D-, L-conventions are those in common use in carbohydrate chemistry. However, for this paper, it is important to describe each asymmetric centre uniquely in all 32 stereoisomers and the *R,S* designators of Cahn, Ingold & Prelog (1956) are summarized in Table 1.



(I)

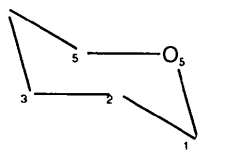


(II)



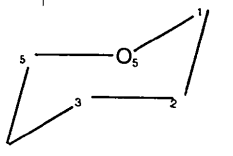
D-(+)-Glucose

(III)



$^4C_1$

(IV)



$^1C_4$

(V)

Conformational aspects

The pyran ring can exist in either of two enantiomeric chair forms [(IV), (V)]. If we view the plane formed by C2, C3, C5, O5 along the C3—C5 or C2—O5 vectors, then (IV) has C4 above this plane and C1 below it and is denoted as the $^4C_1$ conformation; this situation is reversed in (V), which is denoted as $^1C_4$. Thus, the ring and its immediate substitutents in (I) can exist in any one of 64 config-

Table 1. *R,S-configurational descriptors for the 32 stereoisomers of parent hexopyranose* (I)

The axial (*a*) or equatorial (*e*) dispositions of the ring substituents with respect to a $^4C_1$ pyran ring conformer are also indicated in parentheses.

(1) α-D-Hexopyranoses

	C1	C2	C3	C4	C5
α-D-Allose	*S*(*a*)	*R*(*e*)	*R*(*a*)	*S*(*e*)	*R*(*e*)
α-D-Altrose	*S*(*a*)	*S*(*a*)	*R*(*a*)	*S*(*e*)	*R*(*e*)
α-D-Glucose	*S*(*a*)	*R*(*e*)	*S*(*e*)	*S*(*e*)	*R*(*e*)
α-D-Mannose	*S*(*a*)	*S*(*a*)	*S*(*e*)	*S*(*e*)	*R*(*e*)
α-D-Gulose	*S*(*a*)	*R*(*e*)	*R*(*a*)	*R*(*a*)	*R*(*e*)
α-D-Idose	*S*(*a*)	*S*(*a*)	*R*(*a*)	*R*(*a*)	*R*(*e*)
α-D-Galactose	*S*(*a*)	*R*(*e*)	*S*(*e*)	*R*(*a*)	*R*(*e*)
α-D-Talose	*S*(*a*)	*S*(*a*)	*S*(*e*)	*R*(*a*)	*R*(*e*)

(2) β-D-Hexopyranoses

As for the α-D-series (1), but with an *R* configuration (equatorial to a $^4C_1$ ring conformation) at C1 in all cases.

(3) α-L-Hexopyranoses

Change $R \rightarrow S$ and $S \rightarrow R$ in the α-D-series descriptors (1); exchange $a \rightarrow e$ and $e \rightarrow a$ in (1) to obtain conformational dispositions on a $^4C_1$ frame.

(4) β-L-Hexopyranoses

As for the α-L-series (3), but with an *S* configuration (axial to a $^4C_1$ conformation) at C1 in all cases.

(5) $^1C_4$ pyran conformers

All *R,S* descriptors identical to those for $^4C_1$ ring, but $a \rightarrow e$ and $e \rightarrow a$ in changing from any $^4C_1$ conformer to the corresponding $^1C_4$ conformer.

urational/conformational subgroups although, of course, the *R,S* configurational descriptors are unaltered by a change in ring conformation. Simple energy considerations, confirmed by X-ray and other experiments, show that the $^4C_1$ conformation (IV) is preferred and the axial, equatorial (*a,e*) disposition of ring substituents with respect to a fixed $^4C_1$ frame are also included in Table 1. Finally, the conformation adopted by O6 must be considered. Here, our chemical knowledge would indicate that the ± synclinal (±*gauche*) and antiperiplanar (*trans* or *anti*) relationships between C6—O6 and C5—O5 would be preferred arrangements, facts confirmed by crystal structure analyses (see Jeffrey, 1990). Thus, simple *a priori* chemical knowledge would indicate that at least $3 \times 64 = 192$ shape classifications of (I) may possibly exist.

Previous systematic analyses

Arnott & Scott (1972) used available crystallographic results to generate mean geometry and model coordinates for α- and β-D-glucose. Jeffrey & Taylor (1980) derived a more extensive list of mean geometries and compared these data with results from a modified molecular-mechanics procedure. Sheldrick & Akrigg (1980) used the CSD to generate model coordinates for hexopyranose sugars *via* least-squares superposition methods. Variations in geometry involving the anomeric centre C1 are frequently noted in these references, and are specifically addressed by Kirby (1983) and by Allen *et al.* (1987).

Methodology

The July 1990 release of the CSD was used throughout this work. The graphical search program *QUEST* (now extended as *QUEST3D*) was used for search and data retrieval. Data analyses were performed *via* the program *GSTAT*. All software is fully described in the *CSD User Manual* (1992).

CSD chemical search and data retrieval

A 2D substructure search was carried out for the hexopyranose fragment (I) defined and constrained as follows: (i) only C and O atoms were explicitly defined to permit substitution (by H or by non-H atoms) at any site, (ii) all bonds were specified as single, (iii) exocyclic C—O, C—C bonds were required to be acyclic, primarily to avoid dioxa-ring formation through vicinal hydroxy oxygens, (iv) direct cyclic bonds between atoms of the fragment and (unspecified) substituents were not permitted, and (v) hits were only selected from CSD entries which (*a*) were organic compounds, (*b*) had no residual numerical errors following CSD check procedures, (*c*) did not exhibit crystallographic disorder, and (*d*) had *R* factors ≤ 0.120. This search strategy located 249 hexopyranose fragments in the 184 structures whose CSD reference codes are given in Table 2.* These search criteria also located a small number (ten) of uronic acids and their derivatives. These entries have a carboxylic acid or carboxylate function (II) at C5 and were retained in order to study differences between the C6—O orientations in (I) and (II). The uronates (II) are included as normal examples of (I) in all other experiments that do not involve the C6—O6 orientation.

Limitations of the CSD chemical search

The current substructure search mechanisms of the CSD software system operate on simple 2D representations of chemical structure. Specifically, these representations do not contain atomic configurational descriptors, for example the *R,S* descriptors of Cahn, Ingold & Prelog (1956) or their more algorithmic counterparts proposed by Petrarca, Lynch & Rush (1967) or by Wippke & Dyott (1974). This means that a search for (I) will locate all hexopyranose derivatives permitted by the chemical constraints (i)–(iv) above, but will not permit the direct location of individual hexopyranose families, *e.g.* glucoses, mannoses, galactoses *etc.*, in their various stereoisomeric forms. Within the CSD system, this direct

* Full literature citations for the 184 CSD entries used in this study and an extended version of Table 8 have been deposited with the British Library Document Supply Centre as Supplementary Publication No. SUP 71322 (16 pp.). Copies may be obtained through The Technical Editor, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

Table 2. *CSD reference codes for the* 184 *instances of fragment* (I) *used in this study*

ABGPON	BOSREG	DIKTEW	GLUCSA01	MOGLPR
ACCELL10	BOSWOV	DIKTIA	GLUCSE	MTAGLV
ACELLO	BOSXIQ	DIZPEH	GLUCUR20	NABDGC
ACGLPR	BOTRAD	DMGALP	GURXPX10	NAPAGQ
ACLACT	BUFTIF	DOCVOG	JAPHUD	NPGLPN
ACSARM	BUHSEC	DOTMUU	KACJED	OACGAP
ACTDGN	BUWXEW	DOYFOM	KGLUCD	OLGOSE
ADGALA01	CAGALA10	DOYFOM10	KGLUCP02	PACALP
ADGALA10	CAGLUC10	DOZMIO	KGLUCP13	PACDGP
ADMANN	CANAGL10	DUDXOP	KSCOSF	PAFLEB
ADMHEP	CAPZAU	DUKSIL	LACBCB	PAIDOP
ADTALO01	CAPZAU10	FAHHEB	LACTOS03	PATPYS
ADTALO10	CARZAW	FENSUM	LOGANN	PAUCIN
AGPAGP10	CEKLUZ	FIRZEL	MAGALP10	PHMALT
ALTRCA	CELGIJ	FITKAU	MALTOS10	PLANTE10
APTERN	CELLOB01	FIWLIG	MALTOS11	PULCHA10
BAGDIW	CELLOB02	FIXYOA	MALTPY	PULCHB11
BAGZEO	CEWWAC01	FOJLOF	MALTPY01	RAFINO
BAKLOO	CEWWAC20	FONYUC	MBDGAL	SAPMEB
BALBOF	CHONDM	FUJWAI	MBDGAL01	SEBXUS
BAVCAC	CILBII	FUNWOA	MBDGAL02	SOPROS
BAXNAP	CIMDUX	FUSICO10	MBDGPH10	SRGALU
BAXNET	CIVCUF10	FUSXIA	MCELOB	SRHXGU
BAXSEY	CIVDAM10	FUWTEW	MCRZMA	SUCROS03
BAXSEY01	CIZWOX	GAFTUC	MEACGU10	SUCROS04
BDGHEP	COFMEP10	GAFVIS	MEAGPY	THCLCS
BDGLOS01	COKBIN	GAFVOY	MELEZT01	TRECAB
BDGLOS10	COSHEX	GAPRHM10	MELEZT02	TREHAL01
BDGPGL	CUXFAC	GEMDOR	MEMANP	TREHAL02
BEJRAJ	DACHEU	GENTBS	MEMANP11	TREHAL03
BESTIC	DACHEU01	GENTBS01	MGALPY	TREHAL10
BIKWOH10	DAKKEF	GGHPCA	MGALPY01	TURANS
BIZHIB	DAGPLU	GIFRIW	MGLUCP	TURANS01
BLACTO	DECGPY10	GLPHYC10	MGLUCP11	VAFPIB
BOLNUL	DEKYEX	GLPMAC10	MHALAM	VAFPOH
BONPAV	DEZNEB	GLUCMH11	NMALTS	VASKOP
BOPXEJ	DIGTAO	GLUCSA	MOAGLP10	

location of individual stereoisomers can, at present, only be effected by a suitable choice of geometric constraints. Obviously, this choice will depend entirely on the 2D substructure that is being studied.

Choice of geometrical descriptors

The conformational and configurational aspects of (I) were described using the 17 torsional descriptors listed in Table 3. The angles τ_1–τ_6 and τ_{12} described the conformation of the pyran ring and of the —C6—O6 side chain. To define the configurations at C1—C5 we have used the projected valence angles τ_7–τ_{11} [improper torsion angles in the nomenclature of Allen & Rogers (1969)]. These parameters, which represent a projection of the intra-annular valence angle at C*n* onto a plane perpendicular to the C*n*—O or C*n*—C exocyclic substituent bond are clearly related to the Cahn, Ingold & Prelog (CIP; 1956) description of configuration. For a tetrahedral centre C*n*, these projections will have numerical values close to +120 or −120°, using the Klyne & Prelog (1960) sign convention. The sense of rotation is independent of the ring conformation ($^1C_4$, $^4C_1$), and there are a possible 2^5 unique sign sequences for C1—C5. In the CIP convention (Table 1), *R*,*S* descriptors for (I) are assigned by viewing each asymmetric centre along the C*n*—H vector, the descriptors then indicate the sense of rotation that relates the substituents of highest priority. In our scheme for (I), we view each asymmetric centre along the vector connecting ring C*n* with its exocyclic substituent and describe the configuration in terms of the sense of rotation relating the two ring atoms attached to C*n*. Thus for (I), *and only for* (I), the 1:1 relationships listed in Table 4 link the sign of the projected valence angle *at a given centre* C*n* and its CIP *R*,*S* descriptor. Other relationships will occur for other variants of the substructure, *e.g.* those with further substitution at the C or O atoms.

The torsion angles τ_1–τ_{12} form the basis for the classification experiments that are the primary focus of this work. However, for the predictive work based on these classifications it was also necessary to employ τ_{13}–τ_{17} to describe the axial/equatorial orientations of the direct ring substituent bonds.

Table 3. *Torsional descriptors for the hexopyranose fragment* (I) *used in the various classification experiments*

Descriptors marked with an asterisk are projected valence angles.

Torsion angle	Atom sequence
τ_1	O5—C1—C2—C3
τ_2	C1—C2—C3—C4
τ_3	C2—C3—C4—C5
τ_4	C3—C4—C5—O5
τ_5	C4—C5—O5—C1
τ_6	C5—O5—C1—C2
τ_7*	O5—C1—O1—C2
τ_8*	C1—C2—O2—C3
τ_9*	C2—C3—O3—C4
τ_{10}*	C3—C4—O4—C5
τ_{11}*	C4—C5—C6—O6
τ_{12}	O5—C5—C6—O6
τ_{13}	C5—O5—C1—O1
τ_{14}	O5—C1—C2—O2
τ_{15}	C1—C2—C3—O3
τ_{16}	C2—C3—C4—O4
τ_{17}	C3—C4—C5—C6

Table 4. *Relationship between the signs of the projected valence angles* τ_7–τ_{11} (*numerical values ca* 120°) *and the R,S descriptors at* C1–C5 *in* (I), *respectively*

		Sign of angle	
Angle	Ring substituent	+	−
τ_7	C1—O1	*R*	*S*
τ_8	C2—O2	*S*	*R*
τ_9	C3—O3	*S*	*R*
τ_{10}	C4—O4	*R*	*S*
τ_{11}	C5—C6	*R*	*S*

Data analysis

Values of τ_1–τ_{17} were calculated for all 249 instances of (I), using the program *GSTAT*, to yield a torsional data matrix $T(249,17)$. A variety of different clustering experiments were carried out, each experiment using a specific subset of torsion angles, $T(249,N_t)$, where the N_t angles defined some specific conformational and/or configurational feature(s) of (I).

In each case, partitioning of $T(249,N_t)$ was carried out using single-linkage, complete-linkage and Jarvis–Patrick (1973) clustering algorithms, with torsional dissimilarities calculated according to Allen, Doyle & Taylor (1991*a*,*b*,*c*). All algorithms generated similar clustering structures and the data reported here are derived from single-linkage analyses using dissimilarities generated using the city-block metric, except for the results in Table 8 for which the complete-linkage method proved preferable. For each experiment, the tabulated results consist of mean torsion angles relevant to each cluster, together with values of the circular concentration parameter ($\bar{R}$) and torsional e.s.d.'s calculated using the methods of circular statistics (Upton & Fingleton, 1989; Allen & Johnson, 1991).

The single- and complete-linkage methods are agglomerative clustering algorithms. Initially, all objects (here, chemical fragments) are regarded as forming $N_f(=249)$ individual clusters of unit occupancy. Larger clusters are then formed stepwise by merging existing clusters, the merging being dictated by the next lowest dissimilarity value available at that step (see Everitt, 1980). Ultimately, such a process will place all objects in a single cluster of occupany N_f. Statistical and numerical indicators generated at each step, together with the chemical acceptability of the current classification, are used to assess an optimum clustering point between step 1 and the final step N_f-1. In the present implementation of these algorithms the principal indication of optimum clustering is provided by a plot of fusion dissimilarity (D_f: the dissimilarity value that triggers the merging of two clusters at a given step) *versus* the step number. A sharp rise in such a plot is indicative that clusters representing different classifications are being merged into a single, and probably chemically unacceptable, superset.

In assessing the success of any clustering experiment, it is important to know (*a*) that each cluster (partition) contains a group of objects that are closely similar to one another, and (*b*) that different clusters are well separated in the appropriate parameter space. Here we address these criteria by tabulating (where appropriate) the following quantitative measures: for (*a*), D_{max} and D_{ave} are, respectively, the maximum and average torsional dissimilarities of cluster members from the centroid of each cluster and, for (*b*), C_{min} is the minimum torsional dissimilarity between the centroid of a given cluster and the centroid of any other cluster identified in that experiment. Using the city-block metric gives D_{max}, D_{ave} and C_{min} as the sum of the N_t absolute torsional differences in degrees. Division by N_t then yields the semi-normalized 'mean dissimilarity per torsion angle' values that are cited in Tables 5, 6 and 8. In the case of torsion angles, full normalization into the range 0.0 (identical objects) to 1.0 (maximally dissimilar objects) may be achieved through further division by 180°.

Table 5. *Conformation of the pyran ring in hexopyranose sugars: mean torsion angles* (°) *and their e.s.d.'s* (*in parentheses*)

Additional tabulated items are fully described in the text. (*a*) refers to before and (*b*) after inversion of erroneous configurations (see text).

	Conformation		
	$^4C_1$ (*a*)	$^1C_4$ (*a*)	$^4C_1$ (*b*)
τ_1	56.6 (3)	−57.4 (12)	56.7 (3)
τ_2	−52.8 (2)	51.2 (12)	−52.7 (2)
τ_3	52.4 (2)	−50.9 (9)	52.3 (2)
τ_4	−55.7 (3)	56.5 (7)	−55.7 (2)
τ_5	62.0 (3)	−65.5 (12)	62.2 (3)
τ_6	−62.5 (3)	66.1 (13)	−62.6 (3)
	Cluster		
	1	2	1
N_{obs}	239	10	249
$\bar{R}_{max}$	0.998	0.999	0.998
$\bar{R}_{min}$	0.997	0.997	0.997
D_{max}	9.5	4.3	9.5
D_{ave}	3.1	2.7	3.1
C_{min}	114.9	114.9	-

Within the present implementation of the clustering algorithms (Allen, Doyle & Taylor, 1991*c*), each cluster is characterized by its most representative fragment (MRF) and atomic coordinates for each MRF can be output in a variety of forms for use in molecular modelling applications. In Table 9 we report orthogonal coordinates with respect to internal fragment axes having their origin at O5 in all cases: X is along the O5–C1 vector, Y is an axis orthogonal to X and closest to the O5–C4 vector, Z is orthogonal to X,Y such that it forms a right-handed set.

Results and discussion

Preliminary analysis: errors in reported configurations

Given the stated objectives of this work, the first clustering experiment was designed to classify instances of (I) on the basis of ring conformation and on the basis of the configuration observed at each ring carbon. The data matrix $T(249,11)$ was constructed, using τ_1–τ_{11} of Table 3, and optimum clustering was obtained at step 233 in the agglomerative process to yield 16 well defined clusters ranging in population (N_p) from 1 to 95. The top six clusters were readily identified as: (*a*) β-D-glucose ($N_p=95$), (*b*) α-D-glucose (73), (*c*) β-D-galactose (25), (*d*) α-D-galactose (19), (*e*) β-L-glucose (9) and (*f*) α-D-mannose (8). The pyran ring conformation was $^4C_1$ for all fragments in clusters (*a*)–(*d*) and (*f*), but $^1C_4$ in cluster (*e*).

The occurrence of cluster (*e*) with a $^1C_4$ ring conformation and an *SSRRS* configuration was a surprise, not least because the corresponding compound names indicated the β-D-glucose enantiomer.

Table 6. *Stereochemical partitioning of the hexopyranose unit based on* τ_1–τ_{11}

(*a*) Mean torsion anlges (°) and their e.s.d.'s (in parentheses) for the five major clusters. Other tabulated items are described in the text.

	Cluster				
	1	2	3	4	5
N_p	104	74	25	19	8
τ_1	57.8 (4)	56.2 (4)	56.3 (6)	56.0 (11)	55.5 (9)
τ_2	−52.2 (4)	−53.8 (4)	−52.2 (6)	−53.4 (11)	−54.8 (9)
τ_3	51.7 (4)	53.2 (4)	52.5 (5)	51.5 (11)	54.6 (8)
τ_4	−56.4 (4)	−54.7 (5)	−57.0 (5)	−54.4 (9)	−54.8 (10)
τ_5	64.7 (3)	59.1 (4)	63.8 (5)	60.5 (6)	58.0 (9)
τ_6	−65.7 (3)	−59.8 (2)	−63.4 (5)	−60.3 (7)	−58.4 (5)
τ_7	117.4 (2)	−121.2 (1)	118.8 (2)	−120.5 (4)	−121.8 (5)
τ_8	−119.7 (2)	−123.0 (2)	−119.7 (4)	−121.9 (5)	119.5 (7)
τ_9	121.1 (2)	120.3 (2)	123.4 (4)	123.5 (10)	121.0 (5)
τ_{10}	−119.9 (2)	−121.2 (2)	120.0 (4)	120.1 (6)	−118.9 (8)
τ_{11}	119.5 (2)	121.0 (2)	120.6 (3)	122.4 (5)	120.9 (8)
$\overline{R}_{max}$	1.000	1.000	1.000	1.000	1.000
$\overline{R}_{min}$	0.998	0.997	0.999	0.996	0.999
D_{max}	7.5	4.8	3.5	5.5	2.9
D_{ave}	2.2	2.1	1.8	2.6	1.6
C_{min}	11.9	11.6	11.9	11.8	11.6

(*b*) Mean torsion angles (°) for clusters with $N_p = 2, 3$, individual values for singleton fragments. The mean e.s.d. over all angles (where computed) is *ca* 1.5°.

	Cluster				
	6	7	8	9	10
N_p	3	3	3	3	2
τ_1	54.1	47.9	56.9	57.6	51.1
τ_2	−55.3	−45.1	−50.2	−53.2	−47.5
τ_3	56.1	50.3	49.0	52.9	49.6
τ_4	−56.5	−57.7	−54.6	−55.3	−56.6
τ_5	59.2	62.2	64.4	61.0	62.4
τ_6	−57.1	−57.3	−65.2	−63.9	−58.1
τ_7	−121.4	−123.4	117.6	119.7	−121.8
τ_8	121.7	122.0	−122.5	119.8	−124.7
τ_9	124.1	−121.2	−121.7	122.2	−120.5
τ_{10}	118.3	−120.5	121.7	−121.0	120.8
τ_{11}	121.8	120.4	120.4	120.4	125.3

	Cluster			
	11	12	13	14
N_p	2	1	1	1
τ_1	48.1	61.6	45.2	63.3
τ_2	−50.0	−58.0	−41.0	−61.3
τ_3	55.2	57.0	45.5	56.8
τ_4	−57.8	−57.3	−52.9	−52.5
τ_5	58.9	62.0	58.6	58.0
τ_6	−53.7	−65.1	−54.8	−64.9
τ_7	−123.0	119.8	−123.9	116.4
τ_8	−123.2	−120.0	120.4	−114.4
τ_9	−120.4	−118.2	−120.8	118.2
τ_{10}	−122.1	−123.9	122.3	−120.4
τ_{11}	119.8	119.9	120.4	−124.8

Further investigation revealed a further instance of a $^1C_4$ pyran ring in a singleton cluster identified as α-L-glucose (*RSRRS*). Again, the compound name indicated the α-D-enantiomer. To check these observations, cluster analysis was applied to the reduced data matrix $T(249,6)$ using τ_1–τ_6: the six intra-annular torsion angles of the pyran ring. Relevant results (Table 5) from clustering step 247 show two highly compact $^4C_1$ ($N_p = 239$) and $^1C_4$ ($N_p = 10$) conformational clusters well separated in the six-dimensional parameter space.

We suspect that the ten conflicting instances of (I) are taken from structures in which an erroneous choice of enantiomorph has been made. This point is now being investigated further but, for the purpose of this study, these ten coordinate sets were inverted in all subsequent operations. All 249 pyran rings are now $^4C_1$ chairs (Table 5) with a mean absolute torsion angle of 57.0°. The mean conformation shows almost perfect symmetry about O5–C3 but with the expected enhanced puckering (Riddell, 1980) closer to the ring oxygen: $|\tau_5,\tau_6| = 62.4 > |\tau_1,\tau_4| = 56.2 > |\tau_2,\tau_3| = 52.5°$.

Table 7. *Chemical name assignments for the* 14 *stereochemical partitions of* $T(249,11)$ *presented in Table* 6

Cluster	N_p	Signs (τ_7–τ_{11})	*R,S* descriptors*	Chemical name
1	104	+ − + − +	*RRSSR*	β-D-Glucose
2	74	− − + − +	*SRSSR*	α-D-Glucose
3	25	+ − + + +	*RRSRR*	β-D-Galactose
4	19	− − + + +	*SRSRR*	α-D-Galactose
5	8	− + + − +	*SSSSR*	α-D-Mannose
6	3	− + + + +	*SSSRR*	α-D-Talose
7	3	− + − − +	*SSRSR*	α-D-Altrose
8	3	+ − − + +	*RRRRR*	β-D-Gulose
9	3	+ + + − +	*RSSSR*	β-D-Mannose
10	2	− − − + +	*SRRRR*	α-D-Gulose
11	2	− − − − +	*SRRSR*	α-D-Allose
12	1	+ − − − +	*RRRSR*	β-D-Allose
13	1	− + − + +	*SSRRR*	α-D-Idose
14	1	+ − + − −	*RRSSS*	α-L-Idose

* See Table 4.

Stereochemical partitioning

The revised data matrix $T(249,N_t)$, with $N_t = \tau_1$–τ_{11} of Table 3, was partitioned into 14 clusters at step 235 of the single-linkage process, as indicated very clearly on the plot of Fig. 1(*a*). Five of these clusters (Table 6*a*) had $N_p \geq 4$ and encompassed a total of 230 instances of (I). The remaining 19 fragments spanned nine smaller partitions having $N_p \leq 3$ (Table 6*b*: here, it is not possible to generate meaningful summary statistics). Formal chemical names (Table 7) can readily be assigned to the 14 partitions and it is seen that only 13 of the possible 16 D-hexopyranose stereoisomers were present in the CSD of July 1990: β-D-allose, β-D-idose and β-D-talose are not represented. The 14th partition contains a single L-stereoisomer (BIKWOH10: Neuman, Becquart, Avenal, Gillier-Pandraud & Sinay, 1985) in which an α-L-idose fragment adopts a $^4C_1$-pyran ring conformation to avoid unfavourable 1,3-diaxial interactions.

Obviously, only the configuration-sensitive descriptors τ_7–τ_{11} in (I) are necessary for the automated partitioning of stereoisomers. Use of $T(249,5)$ does indeed generate a partitioning of the revised data set that is identical to that described above. Such an approach would have been effective even if some of the pyran rings *had* adopted a conformation that differed from $^4C_1$. However, in view of the inversions applied above, we also included τ_1–τ_6 in the cluster analysis in order (additionally) to detect

Table 8. *Cluster analysis of* —C6—O6 *side-chain conformations in* (I), (II) *based on* τ_7–τ_{12} *and using the complete-linkage algorithm and city-block metric for calculation of dissimilarities*

Results are reported as τ_{12} subpartitions of the 14 stereochemical clusters of Tables 6 and 7, which are identified by chemical name in the subheading (an asterisk indicates that O4 is equatorial to a $^4C_1$ ring). The individual subpartitions are identified by a subcluster number (N_{sc}), a conformational descriptor (Conf.) and a subpartition population (N_{sp}). [The mean value of τ_{12}, the circular concentration ($\bar{R}$) of τ_{12}, together with values of D_{ave}, D_{max} and C_{min} (see text) for the 12 subpartitions with $N_{sp} \geq 4$ are cited in an extended version of this table which has been deposited].† (*a*), (*b*), (*c*) and (*d*) include 4, 2, 1 and 4 uronate fragments (II), respectively (see text).

N_{sc}	Conf.	N_{sp}	N_{sc}	Conf.	N_{sp}	N_{sc}	Conf.	N_{sp}
β-D-Glucose*			α-D-Glucose*			β-D-Galactose		
1.1	*−gauche*	48	2.1	*−gauche*	41	3.1	*+gauche*	14
1.2	*+gauche*	47	2.2	*+gauche*	29	3.2	*trans*	9[c]
1.3	*trans*	5[a]	2.3	*trans*	2[b]	3.3	*−gauche*	1
1.4	*cis*	4[a]	2.4	*cis*	2[b]	3.4	*cis*	1[c]
α-D-Galactose			α-D-Mannose*			α-D-Talose		
4.1	*trans*	8[d]	5.1	*+gauche*	5	6.1	*+gauche*	2
4.2	*+gauche*	6	5.2	*−gauche*	3	6.2	*trans*	1
4.3	*cis*	5[d]						
α-D-Altrose*			β-D-Gulose			β-D-Mannose*		
7.1	*−gauche*	3	8.1	*trans*	3	9.1	*+gauche*	2
						9.2	*−gauche*	1
α-D-Gulose			α-D-Allose*			β-D-Allose*		
10.1	*−gauche*	2	11.1	*+gauche*	1	12.1	*−gauche*	1
			11.2	*−gauche*	1			
β-D-Idose			α-L-Idose*					
13.1	*trans*	1	14.1	*−gauche*	1			

† See deposition footnote.

small conformational variations that may occur between the various stereoisomers.

In the event, Table 6 shows that the main conformational variations occur at τ_5, τ_6: the torsion angles about intra-annular bonds that involve O5 in (I). Values of τ_5, τ_6 are enhanced in β-pyranoses (clusters 1, 3, 8, 9, 12) and decreased in the α-epimers, a reflection of the anomeric effect. Additional ring flattening is induced by 1,3-syndiaxial interactions (Hassell & Ottar, 1947; Jeffrey, 1990) as in clusters 7, 10, 11 and 13, where an axial O1 is involved in the 1,3-interaction in all cases. In cluster 6, where O3, O4 are syndiaxial, flattening is apparent at τ_5, τ_6, but for a very small sample.

Conformational preferences of the >C5—C6—O6 *side chain*

The stereochemical partitions were further subdivided according to the conformational preferences of their >C5—C6—O6 side chains by using $T(249,6)$ as the basis for clustering. Here, the torsional descriptors comprised τ_7–τ_{11} to provide configurational partitioning as before, together with τ_{12} which describes the orientation of the C6—O6 bond. Single-linkage analysis yielded optimum clustering at step 214 (see Fig. 1*b*). Beyond this point the algorithm chooses to merge two larger clusters whilst

Table 9. *Orthogonal coordinates for use in combinatorial model building of hexopyranose units*

All coordinates are referred to the common molecular axis set described in the text.

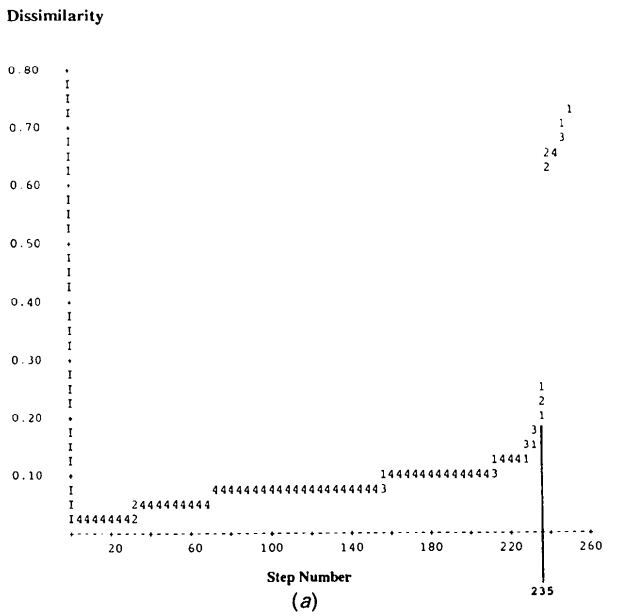
	x	*y*	*z*
(*a*) $^4C_1$-pyran ring			
O5	0.0	0.0	0.0
C1	1.4094	0.0	0.0
C2	1.9379	1.2169	−0.7543
C3	1.3795	2.4822	−0.1296
C4	−0.1284	2.4165	0.0
C5	−0.5465	1.1274	0.7003
(*b*) Direct ring substituents (*e* = equatorial, *a* = axial)			
O1 *e*	1.8053	−1.1626	−0.6789
O1 *a*	1.9487	−0.0190	1.2947
O2 *e*	3.3512	1.2940	−0.6836
O2 *a*	1.5615	1.1359	−2.0467
O3 *e*	1.7040	3.6339	−0.9021
O3 *a*	1.8778	2.7565	1.2097
O4 *e*	−0.6157	3.4990	0.7863
O4 *a*	−0.7644	2.5001	−1.2836
C6 *e*	−2.0407	0.9429	0.7725
C6 *a*	−0.4247	1.0673	2.1861
(*c*) Terminal O6 attached to C6			
C6(*e*)—O6 + *g*	−2.4100	−0.1539	1.4390
−*g*	−2.5111	0.8704	−0.6696
t	−2.7181	1.8932	1.3430
C6(*a*)—O6 − *g*	−1.0988	−0.0832	2.6699

leaving eight individual fragments in singleton clusters in a chemically unreasonable way. This problem did not arise with the complete-linkage algorithm. This method uses the dissimilarity values in a slightly different way (see Everitt, 1980) which is designed to delay the formation of larger clusters until later in the analysis. Fig. 1(*c*) (compare with Fig. 1*b*) and chemical criteria now indicated an optimum clustering point at step 220 for which detailed results are presented in Table 8.

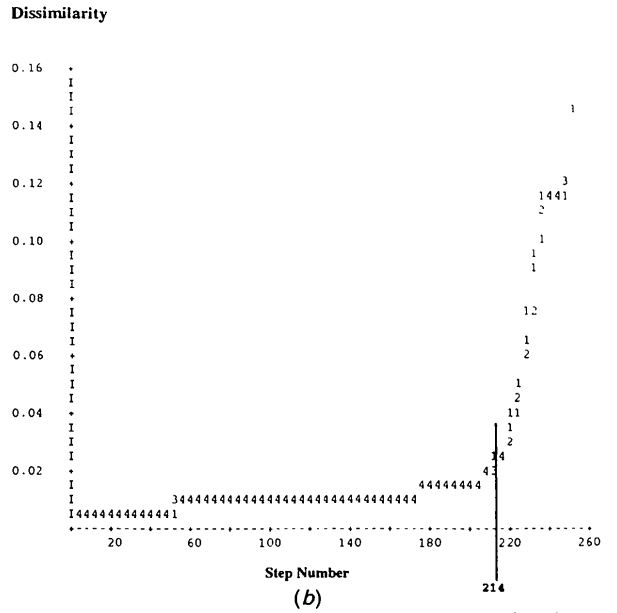
The problems experienced with the single-linkage algorithm in this case have their origins in the relatively broad spread of τ_{12} in the *−gauche*, *+gauche* classifications for β-D-glucose. This is reflected in lower values for the concentration parameter ($\bar{R}$). This gave rise to a chaining effect along the τ_{12} axis, a common failing of the single-linkage method (Allen, Doyle & Taylor, 1991*b*). Indeed, it has been found (Allen, Howard & Pitchford, 1993) that the Jarvis–Patrick (1975) algorithm is a satisfactory and computationally efficient alternative for the classification of flexible ring systems. In the present context, intracluster torsional variance is at a maximum for τ_{12} (defining the —C6—O6 side-chain orientation), is less apparent in those angles (τ_1–τ_6) that define the ring conformation, and is at a minimum for τ_7–τ_{11}, which define configurations at ring C atoms. In cases where classification involves torsional descriptions of freely rotatable bonds, it is important to consider carefully both the algorithmic basis of the experiment and the chemically sensibility of the results generated by such unsupervised methods.

The results of Table 8 include 11 uronate fragments (II). Each of these fragments has one C6—O

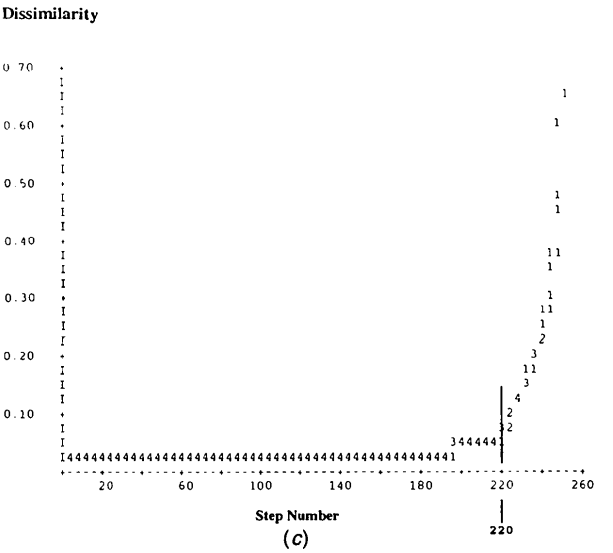
bond in a pseudo-*cis* orientation with respect to the C5—O5 ring bond (torsion angles in the range −32.7 to 13.7°) while the other C6—O bond adopts a pseudo-*trans* arrangement. Both of these arrangements occur in the clusters of Table 8. If we eliminate the uronate fragments, then a breakdown of C6—O orientations in true hexopyranoses (I) can be obtained. When C4—O4 is equatorial to the $^4C_1$ ring, the $+g$ (84 instances) and $-g$ (99 instances) orientations of C6—O are dominant and only a single instance (FIWLIG) of the 'forbidden' t conformer is observed (see Jeffrey, 1990; Hassell & Ottar, 1947). For C4—O4 axial, the $+g$ (22 instances) and t (16 instances) dominate. Here, four fragments adopt the 'forbidden' $-g$ conformation: BAGDIW, cluster 3.3; GGHPCA and PACDGP, cluster 10.1; PLANTE10 from cluster 4.3 (τ_{12} = −48.6°) is also best described as a $-g$ conformer. These results confirm, and update, an earlier study of Perez, St Pierre & Marchessault (1978), in which the limited occurrence of 'forbidden' conformations (Hassell & Ottar, 1947) is ascribed to strong intramolecular hydrogen bonding.



(*a*)



(*b*)



(*c*)

Fig. 1. Plots of dissimilarity *versus* step number (*a*) for the stereochemical partitioning of hexopyranose sugars by single-linkage clustering, (*b*), (*c*) for the study of —C6—O6 orientations. Part (*b*) is from the single-linkage algorithm, part (*c*) is from the complete-linkage algorithm. The dissimilarities are those used by each algorithm in fusing existing clusters at a given step number. The step number selected as representing optimum clustering is indicated on each plot.

The discovered features of hexopyranose structure

The clustering experiments described so far have enabled us to survey available crystallographic data for (I) and (II) and to discover, very rapidly, three basic features of hexopyranose structure.

(1) The pyran ring adopts a $^4C_1$ chair conformation in all but ten of the available examples. Thus the study identifies ten outliers which we believe result from incorrect enantiomorph selection: this is supported by name/structure conflicts in the retrieved entries.

(2) The dataset partitions perfectly into 14 subgroups, each identified by a specific sequence of the five projected valence angles τ_7–τ_{11}. Obviously, 18 sign sequences are missing in the current dataset and it would be reasonable to infer, without *a priori* knowledge, that these sign sequences define a further 18 subgroups not yet exemplified in the crystallographic literature. Because the ring framework has a

constant conformation (after inversion of the ten outliers) it is also possible to relate the sign sequence of τ_7–τ_{11} to sequences of *a* (axial) or *e* (equatorial) descriptors of substituent orientations.

(3) The orientation of the C6—O bond with respect to the C5—O5 bond in (I) falls into three clear conformational groupings: $+g$ ($\tau_{12} \simeq +60°$), $-g$ ($\tau_{12} \simeq -60°$) and t ($\tau_{12} \simeq 180°$).

The only *a priori* chemical knowledge that has been used in these analyses is the distinction between conformational and configurational effects and the selection of suitable numerical parameters to describe them.

Model coordinates for hexopyranose fragments

The three basic features of hexopyranose structure identified above provide the basis for the derivation of orthogonal coordinate approximations that are suitable for building models of any hexopyranose unit, irrespective of its occurrence (or not) in the current crystallographic literature. For models based on a $^4C_1$ pyran ring, we accomplish this by taking (*a*) the most representative coordinates for the ring atoms, (*b*) adding substituents O1 to C6 in the (axial or equatorial) orientations appropriate to the required stereoisomer (see Table 1), and (*c*) adding O6 in a $+g$, $-g$ or t conformation to complete the required model. By this means, we can (theoretically) construct 96 models of the $^4C_1$-D,L stereoisomers of (I).

The necessary orthogonal coordinates are presented in Table 9, referred to the molecular axes described in the *Data analysis* section. Table 9(*a*) reports the coordinates of the most representative $^4C_1$ pyran ring fragment obtained from run (*b*) of Table 5. Table 9(*b*) reports coordinates of the substituents O1—C6 in both equatorial and axial orientations to the $^4C_1$ ring. These data were the most representative values obtained by partitioning the complete dataset (single-linkage clustering) on the basis of τ_1–τ_6 and (successively) τ_{13}, τ_{14}, τ_{15}, τ_{16} and τ_{17}. Summary results of this 'nodal partitioning' are given in Table 10, which also includes a breakdown of the observed C6—O6 orientations for C6(*e*) and C6(*a*) in (I). The most representative coordinates for the four observed orientations of O6 are reported in Table 9(*c*).

The data of Table 9 may be used combinatorially to build $^4C_1$-hexopyranose models. Combination of the ring and direct substituent coordinates according to the *a/e* relationships of Table 1 generates the 32 stereisomeric frameworks. For those 16 stereoisomers having C6(*e*) we may then model three possible O6 orientations, but only the single ($-g$) orientation of O6 is available for C6(*a*) stereoisomers. Thus, 64 of the possible 96 models of $^4C_1$-D,L-hexopyranoses are derivable from Table 9. A further 64 models of $^1C_4$-D,L-hexopyranose units may also be derived from Table 9 by coordinate inversion.

Table 10. *Nodal partitioning of* $T(249,7)$ *for* τ_1–τ_6 *and (successively)* τ_{13}, τ_{14}, τ_{15}, τ_{16}, τ_{17}, *to obtain most representative coordinates for equatorial (e) and axial (a) substituents at each carbon node of the* $^4C_1$ *ring*

Conformational subdivisions at C6 were obtained using $T(238,8)$ for τ_1–τ_6, τ_{17} and τ_{12} (uronate fragments omitted). Single-linkage clustering (city-block metric) was used in all cases. Not all C6 conformers were classified in this process.

Node	Substituent	Orientation		Mean τ	N_{obs}
C1	O1	*e*	τ_{13} =	177.7 (3)	137
		a	τ_{13} =	60.4 (3)	112
C2	O2	*e*	τ_{14} =	177.4 (3)	231
		a	τ_{14} =	−66.4 (8)	18
C3	O3	*e*	τ_{15} =	−173.7 (3)	237
		a	τ_{15} =	71.3 (14)	12
C4	O4	*e*	τ_{16} =	171.7 (3)	196
		a	τ_{16} =	−68.6 (6)	53
C5	C6	*e*	τ_{17} =	−174.6 (3)	248
		a	τ_{17} =	74.5 (–)	1
C6(*e*)	O6	$+g$	τ_{12} =	66.1 (7)	105
		$-g$	τ_{12} =	−64.7 (6)	101
		t	τ_{12} =	174.8 (22)	17
C6(*a*)	O6	$-g$	τ_{12} =	−63.5 (–)	1

It should be stressed that models built in this way are approximate only: they will not, for example, reflect the minor differences in ring conformation that exist (Table 6) between the various stereoisomers. The models are, however, perfectly adequate for use in molecular graphics operations, or as starting points for computational procedures or crystal structure determination.

Nodal partitioning at C1: *the anomeric effect*

The nodal partitioning of the complete dataset on the basis of axial (112 instances) or equatorial (137 instances) orientations of the C1—O1 bond (see Tables 9, 10) is, of course, a separation of the α (axial) and β (equatorial) anomers. It is these different orientations of C1—O1 with respect to the C5—O5 bond (α = synclinal, $\tau_{13} \sim 60°$, β = antiperiplanar, $\tau_{13} \sim 180°$) that give rise to the anomeric effect (Kirby, 1983) and to the significant geometrical differences that exist between the α- and β-anomers. For completeness, we present mean geometry for the α- and β-anomers in Table 11. The results of this analysis compare well with other published data (see *e.g.* Jeffrey, 1990; Jeffrey & Taylor, 1980; Arnott & Scott, 1972). We observe: (*a*) C1—O1 is significantly shortened in the β-anomers, (*b*) C1—O5 is marginally lengthened in the β-anomers, (*c*) valence angles involving O5, particularly O5—C1—O1, are all smaller in the β-anomers, and (*d*) there is enhanced puckering of the pyran ring about bonds involving O5 (τ_5, τ_6) in the β-anomers. This enhancement is also apparent, but to a lesser degree, in τ_1, τ_4 which also involve O5. The mean puckering angle ($|\bar{\tau}|$) in

Table 11. *The anomeric effect: mean bond angles* (Å), *valence and torsion angles* (°) (*e.s.d.'s in parentheses*) *for the α- and β-anomers of hexopyranosides*

	Cluster 1(α,axial)	Cluster 2(β,equatorial)	All fragments
N_{obs}	112	137	249
O5—C1	1.418 (1)	1.423 (1)	1.421 (1)
C1—O1	1.413 (2)	1.395 (2)	1.403 (1)
C1—C2	1.523 (1)	1.518 (1)	1.520 (1)
C2—C3	1.521 (1)	1.524 (1)	1.523 (1)
C3—C4	1.519 (1)	1.520 (2)	1.519 (1)
C4—C5	1.530 (2)	1.528 (1)	1.528 (1)
C5—O5	1.439 (2)	1.435 (1)	1.437 (1)
C5—O5—C1	113.9 (1)	111.9 (1)	112.8 (1)
O5—C1—O1	111.5 (1)	107.3 (1)	109.2 (2)
C2—C1—O1	108.0 (2)	108.2 (1)	108.1 (1)
O5—C1—C2	110.3 (1)	109.2 (1)	109.7 (1)
C1—C2—C3	110.7 (1)	109.7 (1)	110.2 (1)
C1—C3—C4	110.3 (2)	110.9 (1)	110.6 (1)
C3—C4—C5	110.2 (2)	110.1 (1)	110.1 (1)
C4—C5—O5	110.4 (2)	108.8 (1)	109.5 (1)
τ_1	55.5 (4)	57.6 (3)	56.7 (3)
τ_2	−53.3 (4)	−52.3 (3)	−52.7 (2)
τ_3	52.9 (3)	51.9 (3)	52.3 (2)
τ_4	−54.8 (4)	−56.4 (3)	−55.7 (2)
τ_5	59.4 (3)	64.4 (3)	62.2 (3)
τ_6	−59.5 (2)	−65.2 (3)	−62.6 (3)

the pyran ring is 55.9 (3)° in the α-anomers and 58.0 (3)° in the β-anomers.

Concluding remarks

The methods of single- and complete-linkage cluster analysis have been used to generate a variety of classifications of the hexopyranose sugars. In particular, these methods have generated a perfect classification of the hexopyranose stereoisomers, have confirmed and enhanced earlier manual analyses of the C6—O6 orientation, and formed the basis for the derivation of model coordinates for 128 $^4C_1$, $^1C_4$-D,L-hexopyranose fragments, many of which are not yet represented in the crystallographic literature. The analytical methods used have, to a large degree, answered the questions posed in the *Introduction*: classifications have been achieved, knowledge has been acquired, and predictions have been made on the basis of that knowledge.

However, these results have not been obtained without the use of some expert *a priori* chemical knowledge, firstly in the choice of torsion angles and projected valence angles to define conformational and configurational concepts, respectively, and secondly in determining a suitable end-point for each clustering experiment. Further, the knowledge acquired in each experiment is not retained in a structured and machine-readable form to be accessed by later investigations. If clustering techniques are to be used routinely then some automation of end-point detection is obviously desirable. A variety of clustering criteria have been described in the literature (see *e.g.* Everitt, 1980), and the work described here, and in related crystallographic applications, will provide valuable tests of the efficacy of these criteria. We are also addressing the problem of knowledge organization and storage and have described a semantic network model for a molecular knowledge base for use in molecular scene analysis (Fortier *et al.*, 1993; Allen, Rowland, Fortier & Glasgow, 1990). In these contexts, we are now examining a conceptual clustering approach to classification (Conklin, Fortier, Glasgow & Allen, 1992) which describes structured objects in terms of their parts, and the relationships among those parts, and which generates structured results that are suitable for direct inclusion in a knowledge base. Initial results from these experiments will be reported shortly (Fortier, Conklin, Glasgow & Allen, 1993).

Finally, the hexopyranose analysis has generated two important pointers for the CSD itself: (*a*) it appears that a number of configurational assignments in the crystallographic literature are incorrect and methods must be devised to locate these errors, and (*b*) the systematic assignment of sense-of-chirality descriptors to stereogenic atoms would improve significantly the search and display facilities of the CSD software system. Both of these developments are now being actively pursued.

We thank Dr Olga Kennard OBE FRS for her interest in this work. We gratefully acknowledge financial assistance in the form of travel grants from NATO, The Royal Society of London and NSERC, and for an NSERC operating grant for SF. All computations were carried out on an IBM3084Q at the University of Cambridge Computing Service; we thank UCCS Staff for their assistance.

References

ALLEN, F. H., DAVIES, J. E., GALLOY, J. J., JOHNSON, O., KENNARD, O., MACRAE, C. F., MITCHELL, E. M., MITCHELL, G. F., SMITH, J. M. & WATSON, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*a*). *Acta Cryst.* B**47**, 29–40.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*b*). *Acta Cryst.* B**47**, 41–49.

ALLEN, F. H., DOYLE, M. J. & TAYLOR, R. (1991*c*). *Acta Cryst.* B**47**, 50–61.

ALLEN, F. H., HOWARD, J. A. K. & PITCHFORD, N. A. (1993). *Acta Cryst.* B**49**, 910–928.

ALLEN, F. H. & JOHNSON, O. (1991). *Acta Cryst.* B**47**, 62–67.

ALLEN, F. H., KENNARD, O., WATSON, D. G., BRAMMER, L., ORPEN, A. G. & TAYLOR, R. (1987). *J. Chem. Soc. Perkin Trans.* 2, pp. S1–S19.

ALLEN, F. H. & ROGERS, D. (1969). *Acta Cryst.* B**25**, 1326–1330.

ALLEN, F. H., ROWLAND, R. S., FORTIER, S. & GLASGOW, J. I. (1990). *Tetrahedron Comput. Methodol.* **3**, 757–774.

ALLEN, F. H. & TAYLOR, R. (1991). *Acta Cryst.* B**47**, 404–412.

ARNOTT, S. & SCOTT, W. E. (1972). *J. Chem. Soc. Perkin Trans.* 2, pp. 324–335.

CAHN, R. S., INGOLD, C. K. & PRELOG, V. (1956). *Experientia*, **12**, 81–86.

CARBONELL, J. (1980). *Machine Learning: Paradigms and Methods.* Cambridge, MA: MIT Press.
CONKLIN, D., FORTIER, S., GLASGOW, J. I. & ALLEN, F. H. (1992). *Proceedings of the Machine Learning* 1992 *Workshop on Machine Discovery*, Aberdeen, Scotland.
CSD User Manual (1992). Cambridge Crystallographic Data Centre, Cambridge, England.
EVERITT, B. (1980). *Cluster Analysis*, 2nd ed. New York: Wiley.
FORTIER, S., CASTLEDEN, I., GLASGOW, J. I., CONKLIN, D., WALMSLEY, C., LEHERTE, L. & ALLEN, F. H. (1993). *Acta Cryst.* D**49**, 168–178.
FORTIER, S., CONKLIN, D., GLASGOW, J. I. & ALLEN, F. H. (1993). *Acta Cryst.* In preparation.
HASSELL, O. & OTTAR, B. (1947). *Acta Chem. Scand.* **1**, 929–942.
JARVIS, R. A. & PATRICK, E. A. (1973). *IEEE Trans. Comput.* **22**, 1025–1034.
JEFFREY, G. A. (1990). *Acta Cryst.* B**46**, 89–103.
JEFFREY, G. A. & TAYLOR, R. (1980). *J. Comput. Chem.* **1**, 99–108.
KIRBY, A. J. (1983). *The Anomeric Effect and Related Stereoelectronic Effects at Oxygen.* Heidelberg: Springer.
KLYNE, W. & PRELOG, V. (1960). *Experientia*, **16**, 521–523.
MARR, D. (1982). *Vision.* San Francisco: Freeman.
MARR, D. & NISHIHARA, H. K. (1978). *Proc. R. Soc. London Ser. B*, **200**, 269–294.
NEUMAN, A., BECQUART, J., AVENEL, D., GILLER-PANDRAUD, H. & SINAY, P. (1985). *Carbohydr. Res.* **139**, 23–34.
PEREZ, S., ST PIERRE, R. & MARCHESSAULT, R. (1978). *Can. J. Chem.* **56**, 2866–2871.
PETRARCA, A. E., LYNCH, M. S. & RUSH, J. E. (1967). *J. Chem. Doc.* **7**, 154–165.
RIDDELL, F. G. (1980). *The Conformational Analysis of Heterocyclic Compounds.* London: Academic Press.
SCHALKOFF, R. (1992). *Pattern Recognition: Statistical, Structural and Neural Approaches.* New York: Wiley.
SHELDRICK, B. & AKRIGG, D. (1980). *Acta Cryst.* B**36**, 1615–1621.
UPTON, G. J. G. & FINGLETON, B. (1990). *Spatial Data Analysis by Example*, Vol. 2, *Categorical and Directional Data.* New York: Wiley.
WIPPKE, W. T. & DYOTT, T. M. (1974). *J. Am. Chem Soc.* **96**, 4834–4842.

Acta Cryst. (1993). B**49**, 1031–1039

Electron-Density Distribution in Crystals of *p*-Nitrobenzene Derivatives

BY MASAHIKO TONOGAKI, TAKASHI KAWATA AND SHIGERU OHBA*

Department of Chemistry, Faculty of Science and Technology, Keio University, Hiyoshi 3, *Kohoku-ku, Yokohama* 223, *Japan*

AND YUTAKA IWATA AND IWAO SHIBUYA

Research Reactor Institute, Kyoto University, Kumatori, Sennan-gun, Osaka 590-04, *Japan*

(*Received* 15 *February* 1993; *accepted* 19 *May* 1993)

Abstract

The electron-density distributions in five nitrobenzene derivatives have been examined by the multipole expansion method based on X-ray diffraction data measured with Mo $K\alpha$ radiation (λ = 0.70926 Å) at 120 K. *p*-Dinitrobenzene, $C_6H_4N_2O_4$, (I), M_r = 168.1, monoclinic, $P2_1/n$, a = 10.941 (2), b = 5.3813 (5), c = 5.6684 (4) Å, β = 92.116 (8)°, V = 333.51 (7) Å^3, Z = 2, D_x = 1.67 Mg m^{-3}, μ = 0.135 mm^{-1}, R = 0.030 for 2045 unique reflections. 4-Nitrobenzoic acid, $C_7H_5NO_4$, (II), M_r = 167.1, monoclinic, $A2/a$, a = 12.857 (1), b = 5.0272 (2), c = 20.997 (2) Å, β = 97.072 (8)°, V = 1346.8 (2) Å^3, Z = 8, D_x = 1.65 Mg m^{-3}, μ = 0.130 mm^{-1}, R = 0.030 for 3930 reflections. 4-Nitrobenzamide, $C_7H_6N_2O_3$, (III), M_r = 166.1, monoclinic, $P2_1/c$, a = 7.393 (2), b = 6.8005 (9), c = 13.814 (2) Å, β = 90.88 (1)°, V = 694.4 (2) Å^3, Z = 4, D_x = 1.59 Mg m^{-3}, μ = 0.119 mm^{-1}, R = 0.033 for 3811 reflections. 4-Nitrobenzaldehyde oxime, $C_7H_6N_2O_3$, (IV), M_r = 166.1, monoclinic, $P2_1/c$, a = 6.2336 (6), b = 4.8377 (5), c = 24.352 (2) Å, β = 94.87 (8)°, V = 731.7 (1) Å^3, Z = 4, D_x = 1.51 Mg m^{-3}, μ = 0.113 mm^{-1}, R = 0.043 for 2493 reflections. 4-Nitroaniline, $C_6H_6N_2O_2$, (V), M_r = 138.1, monoclinic, $P2_1/n$, a = 12.122 (2), b = 6.0276 (9), c = 8.487 (1) Å, β = 92.72 (1)°, V = 619.4 (2) Å^3, Z = 4, D_x = 1.48 Mg m^{-3}, μ = 0.107 mm^{-1}, R = 0.040 for 2573 reflections. Single-crystal neutron diffraction studies were also made for (I) at 120 K, and for (II) at both 120 and 302 K. For the nitro group, the N—O bonding electrons and the lone pairs of the O atoms are clearly observed. The π-donating nature of the amino group in (V) can be seen by the polarization of the electron density of the benzene ring. The carboxyl H atom in (II) and the oxime H atom in (IV) have pronounced positive effective charges, which are reflected in the relatively large thermal parameters in the X-ray conventional refinement using neutral-atom scattering factors. Neutron diffraction study of (II) indicated that the disorder of the COOH group due to the double proton transfer is 11 (3)% at 302 K and is not observed at 120 K within experimental error.

* To whom correspondence should be addressed.