Acta Cryst. (1996). B52, 535–549

Conformational Analysis from Crystallographic Data using Conceptual Clustering

DARRELL CONKLIN,[a]† SUZANNE FORTIER,[a,b] JANICE I. GLASGOW[a] AND FRANK H. ALLEN[c]*

[a]*Department of Computing and Information Science, and* [b]*Department of Chemistry, Queen's University, Kingston, Ontario, Canada K7L 3N6, and* [c]*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England*

(*Received 23 June 1995; accepted 15 December 1995*)

Abstract

The rapid growth of crystallographic databases has created a demand for novel and efficient techniques for the analysis of molecular conformations, in order to derive new concepts and rules and to generate useful classifications of the available data. This paper presents a conceptual clustering approach, termed IMEM (image memory), which discovers the conformational diversity present in a dataset of crystal structures. In contrast to numerical clustering methods, IMEM views a molecular structure as comprising qualitative relationships among its parts, *i.e.* the structure is viewed as a molecular scene. In addition, IMEM does not require the user to have any *a priori* knowledge of an expected number of conformational classes within a given dataset. The IMEM approach is applied to several datasets derived from the Cambridge Structural Database and, in all cases, chemically correct and sensible conformational classifications were discovered. This is confirmed by a rigorous comparison of IMEM results with published conformational data obtained by energy-minimization and numerical clustering methods. Conformational analysis tools have an important part to play in the conversion of raw molecular databases to knowledge bases.

1. Introduction

The rapid growth of crystallographic databases has created the need for computational techniques to structure, manage and compress the accumulated data and transform them into knowledge bases. These techniques may be used to obtain meaningful generalizations, deriving useful classifications and perhaps discovering new concepts or rules about molecular structures. For example, the Cambridge Structural Database (CSD; Allen *et al.*, 1991) contains several instances of each of the 20 commonly encountered amino acids. From these, it is possible to derive generic concepts for each residue which summarize common features and point to differences. In this way we may realize the acquisition of knowledge from data (see *e.g.* Engh & Huber, 1991; Benedetti, Morelli, Nemethy & Scheraga, 1983).

At present, the crystallographic databases can generally be described as flat files. The information is stored in the form of separate and unlinked entries (apart from implicit links contained in the bit-encoded screens of entries in the CSD). Clearly, the most appropriate way to organize and link information depends on the questions being asked. Thus, it is important to have in hand flexible classification algorithms that can help not only to analyze the data so as to find similarities and differences, but also to retain the results through the creation of structured knowledge bases. Finally, although several general concepts, rules and constraints about molecular structures have been explicitly formulated, many still remain buried within the databases. Given that the crystallographic databases now contain vast amounts of data (the CSD now contains more than 150 000 entries), there exists a clear and crucial need for computer-aided techniques that will contribute to the discovery of new knowledge.

The demand for more convenient access to molecular data has led to the restructuring of flat files into relational (Huysmans, Richelle & Wodak, 1991) and object-oriented database systems. This latter approach has been employed by Gray, Paton, Kemp & Fothergill (1990) for the purpose of protein structure analysis. Here, Prolog queries are used to retrieve information by navigating through a network of objects that represent the primary, secondary and tertiary structure of proteins. A similar system, which uses an extensible object-oriented class library for representing, verifying and rendering macromolecular structures, has recently been developed in C++ (Chang, Shindyalov, Pu & Bourne, 1994). As the quantity and range of use of molecular knowledge increases, there will be a growing need for data storage and retrieval by systems such as these, which rely on the organization of molecules according to their structural hierarchies.

Most of the techniques used so far to acquire knowledge from the geometrical information stored in the databases have been numerical or statistical in origin (Taylor & Allen, 1992). Techniques more closely allied to artificial intelligence approaches have also been used,

† Present address: ZymoGenetics Inc., 1201 Eastlake Avenue East, Seattle, WA 98102, USA.

particularly for protein structure classification and prediction (Blundell, Sibanda, Sternberg & Thornton, 1987; Rooman & Wodak, 1988; Hunter & States, 1992). Over the last few years we have been investigating the use of machine-learning techniques for application to crystallographic data, as part of a project in molecular scene analysis (Fortier *et al.*, 1993). In this paper we report on the development and application of IMEM, a concept formation approach designed specifically for objects or scenes described in terms of their parts and the interrelationships among these parts. The next section provides background information on machine-learning and concept-formation approaches and describes the theoretical foundation of IMEM. We then present applications of IMEM to the conformational classification of six- and seven-membered rings, steroid C17-side chains and hexopyranose sugars. The results obtained in these classification and learning exercises are compared with those generated by numerical clustering techniques. The paper concludes with a discussion of the potential role of conceptual clustering methodologies in the construction of crystallographic knowledge bases.

2. Background information and methodology

Machine learning encompasses several paradigms of automated learning, such as the inductive, analytic, connectionist and genetic approaches (Carbonell, 1989). Although these approaches differ significantly in both their knowledge representation and in their learning engine, they all share a fundamental objective: to automate the acquisition of knowledge, thereby improving the performance of computing systems. In the molecular scene analysis project (Fortier *et al.*, 1993) we have focused on conceptual clustering and concept formation approaches, which belong to the inductive paradigm of machine learning.

2.1. *Conceptual clustering approaches*

Clustering techniques are usually divided into two categories: *numerical* and *conceptual*. Although the machine-learning branch of artificial intelligence is concerned mainly with the latter, numerical techniques will also be described briefly to highlight differences and similarities between the two approaches. In numerical clustering the samples are viewed from a geometrical perspective as a set of data points in an n-dimensional space, where n is the number of attributes used to characterize each data point. The goal of the clustering exercise is to partition the data points, grouping together points that are similar. Distance metrics are used to measure dissimilarity, while criterion functions help measure the quality of the data partition. Thus, numerical clustering techniques normally rely exclusively on *quantitative* attributes. In structural chemistry these are typically the geometrical parameters that are commonly used to describe the 3D (three-dimensional) image of a molecule (see, for example, Table 1*a*).

Table 1. *Relational attributes for molecular conformation (a) quantitative and (b) qualitative*

The orientation relation is due to Wirth (1986); the clinal and periplanar relations are defined by Klyne & Prelog (1960).

Relational attribute	Arity
(*a*) Quantitative	
Interatomic distance	2
Bond angle	3
Torsion angle	4
(*b*) Qualitative	
Bonded	2
Proximal	2
Linear	3
Angular	3
syn-, antiperiplanar (*sp*,*ap*)	4
+, − clinal (+*c*, −*c*)	4
Orientation	4
R/*S*	5
cis/*trans*	≥ 5
Axial/equatorial	≥ 5

Conceptual clustering techniques share with their numerical counterparts the goal of partitioning the data into 'natural' groupings. They have, however, an additional goal, which is to characterize the clusters in terms of simple and meaningful concepts, rather than in terms of a set of statistics. These methods predominantly use *qualitative* attributes. Some qualitative concepts that are commonly used to describe chemical structure, and particularly stereochemical relationships, appear in Table 1(*b*). In clustering approaches the tasks of classification and learning do not usually rely on externally predefined categories or labeled examples. For that reason, they are termed *unsupervised* approaches.

Both *agglomerative* and *divisive* clustering techniques exist. The agglomerative techniques use a bottom-up approach with a starting point consisting of as many clusters as instances. In divisive techniques a top-down approach is used. The starting point consists of a single cluster containing all instances. Clustering techniques may be differentiated on the basis of whether they allow for *overlapping* of clusters or whether they produce only *disjoint* partitions. A clustering technique may be *nonincremental* depending on whether all the observations are available at the outset of a clustering exercise or are presented as a stream. In the latter case the classification evolves as each new observation is presented to it. Conceptual clustering techniques come with various combinations of the attributes mentioned above (incremental *versus* nonincremental, agglomerative *versus* divisive and overlapping *versus* disjoint). The term *concept formation* is normally used to refer to incremental conceptual algorithms.

In conceptual clustering learning proceeds through the generalization, characterization and organization of a set of observations. Concept formation approaches translate a stream of observations into a concept hierarchy that organizes and summarizes the observations. Thus, following the definition of Gennari, Langley & Fisher (1989), concept formation can be described in terms of the following set of tasks:

Given: a sequential presentation of objects and their associated description;

find: (1) clusters that group these objects into classes;
(2) a summary description (*i.e.* a concept) for each class;
(3) a hierarchical organization for these concepts.

Several useful concept formation algorithms already exist, for example *UNIMEM* (Lebowitz, 1987) and *COBWEB* (Fisher, 1987). These systems, however, rely on object representations expressed as a list of attribute-value pairs. This representation is not the most suitable one for structured domains where the salient features of an object are not only its attributes, but also the relationship among its parts. An emerging area of interest in machine learning is the design of structured concept formation algorithms, in which structured objects are formed and then organized in a knowledge base (Thompson & Langley, 1991).

3. The IMEM approach

IMEM is a concept formation method specificially designed for objects or scenes described in terms of their parts and the interrelationships among those parts (Conklin, 1995; Conklin & Glasgow, 1992). These relationships may be topological (*e.g.* connectivity, proximity, nestedness) or spatial (*e.g.* direction, relative location, symmetry). The IMEM approach has been implemented as a system to perform conceptual clustering specifically with molecular structure data.

3.1. *Knowledge representation in IMEM*

A molecular structure is represented in IMEM as an image, which comprises a set of parts with their 3D coordinates, and as a set of relations that are preserved for the image. These relations may be expressed as functions that operate on the image. Although these functions may return quantitative values, as in the case of bond angles or interatomic distances, they are represented qualitatively in terms of attributes that are true or false. Table 1 illustrates a variety of molecular n-ary relations, for $n = 2, 3$ *etc*. Parts of an image may be considered at various levels of complexity. In applications to molecular structure classification parts could be selected, for example, at the atomic, the functional group or the secondary structure motif levels.

IMEM also requires some background knowledge to be defined. This includes a set of predefined primitive concepts which specify the nondecomposable units used to describe images and a definition of all the relevant relations which will be used for clustering.

3.2. *Classification and learning in IMEM*

The notions of *equivalence*, *subsumption* and *similarity* are at the core of the IMEM approach to classification and learning. Two images are considered equivalent with respect to a set of predefined relations if their parts are identically related. More formally, two images C and D are *equivalent* with respect to an n-ary relation r if there exists a bijective function f such that for all n-tuples $c_1, \ldots, c_n$ of parts in C

$$r(c_1, \ldots, c_n) \text{ if and only if } r(f[c_1], \ldots f[c_n]).$$

That is, for every part c_i in image C there exists a corresponding (unique) part $f(c_i)$ in image D. If, for example, parts c_i and c_j are bonded in C, then their corresponding parts $f(c_i)$ and $f(c_j)$ must also be bonded in D. Thus, equivalence with regards to the connectivity relation reduces to full-graph isomorphism. Two images are equivalent with respect to a *set* of relations if there exists a single function f that preserves equivalence for each of the relations in the set. Equivalence of images under rigid geometric transformations (*e.g.* rotation, translation) is accomplished by choosing relations which do not depend on the particular coordinate frame of an image.

An image C *subsumes* an image D with respect to a set of relations if and only if C is equivalent to a subimage of D, *i.e.* there is an injective function f from C to D that preserves the specified relations. If connectivity is the only relation of interest, then subsumption reduces to subgraph isomorphism, *i.e.* C is a substructure of D. Subsumption is illustrated in Fig. 1 and is further discussed below in the context of a subsumption hierarchy.

Finally, *similarity* between two images can be measured by determining the least common subsumers of the images. Formally, an image M is a least common subsumer of images C and D if and only if M subsumes both C and D and there does not exist an image M' such that M subsumes M' and M' subsumes C and D. Note that two images may have a number of different least common subsumers.

We now define similarity using a variant of the Dice association coefficient (Salton & McGill, 1983)

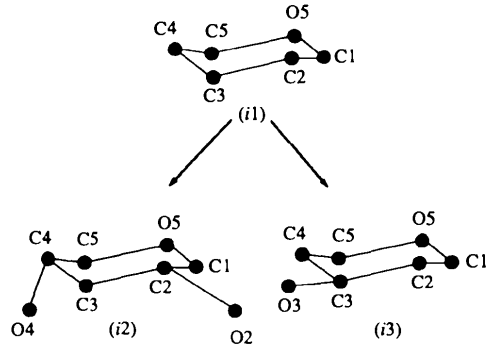
$$S(C, D) = 2|M|/(|C| + |D|), \quad (1)$$

where $S(C, D)$ denotes the similarity between the images C and D, M is the least common subsumer of C and D with the largest number of parts, and $|X|$

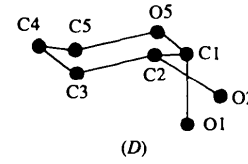
denotes the number of parts in an image X. Once again, if connectivity is the only relation considered then the similarity function defined in (1) corresponds to the definition of similarity based on maximal common substructure (Willett, 1990).

The IMEM algorithm uses an incremental, top-down, divisive approach to build a subsumption hierarchy that summarizes and classifies a dataset. This hierarchy is represented as a directed graph, where each node is labeled by an image: a leaf node corresponds to an image of an individual instance of a molecular structure and an intermediate node corresponds to an image that denotes the class of structures which it subsumes (*i.e.* all leaf node images under it in the hierarchy). Edges in the graph denote subsumption. Edges implied by transitivity are implicitly represented in the hierarchy. Thus, if A subsumes B and B subsumes C, there is no need for an edge between A and C.

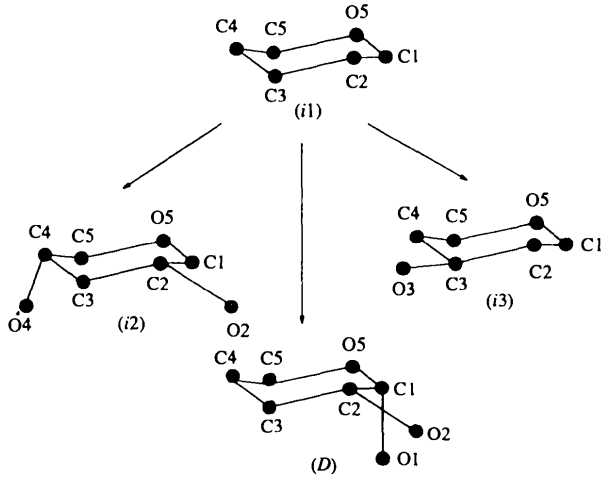
The initial state of a subsumption hierarchy consists of a single most general image. A new input image D is incorporated into the hierarchy by being recursively passed down to the children of each image that are found to subsume it. This process is followed until an equivalent image is found or a level is reached at which none of the child images are found to subsume the input image. In this manner the most specific subsumers of the input image are identified. Links are then made from the most specific subsumers to the input image. At this



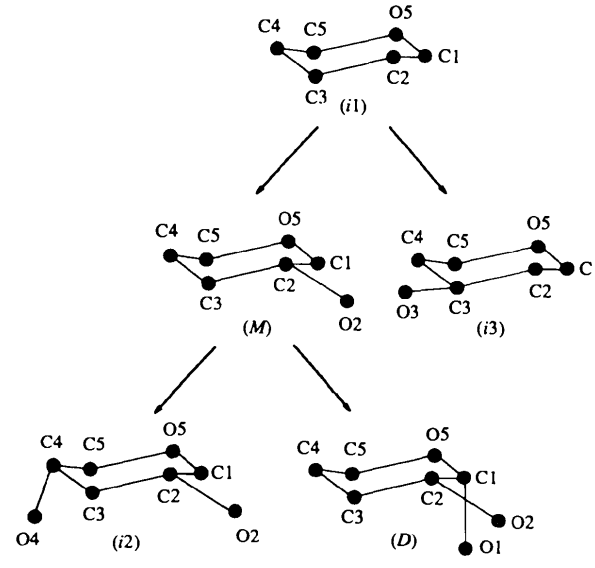
(*a*) Initial subsumption hierarchy



(*b*) Input image



(*c*) Updated subsumption hierarchy



(*d*) Generalization of D and $i2$

Fig. 1. Development of a subsumption hierarchy for molecular images (see text): (*a*) initial hierarchy, (*b*) a new input image D, (*c*) hierarchy after D is added and (*d*) final hierarchy after inclusion of generalization of D and $i2$.

point, IMEM attempts to generalize the input image with each of the other children of the most specific subsumers. If a child C is similar enough to the input image, *i.e.* $S(C, D)$ is greater than some predefined threshold value, then a least common subsumer M is constructed for C and D. M is then classified into the current hierarchy: inserted just below all most specific subsumers and above all most general subsumers.

Fig. 1(*a*) provides a graphical illustration of an initial subsumption hierarchy in which image $i1$ subsumes images $i2$ and $i3$, *i.e.* $i1$ is a subgraph of both $i2$ and $i3$. Assume that IMEM is now given an input image D, as depicted in Fig. 1(*b*). Image D is compared with $i1$, which is found to subsume it. It is then passed down to the child images $i2$ and $i3$, neither of which subsume D. Thus, $i1$ is the most specific subsumer of D and D is entered into the hierarchy as a child of $i1$, as illustrated in Fig. 1(*c*). Similarity measures are then calculated for D and each of the images $i2$ and $i3$. Assuming that $i2$ is determined to be similar to D [*i.e.* $S(i2, D) >$ threshold], then a least common subsumer M is constructed and classified into the hierarchy, resulting in the graph displayed in Fig. 1(*d*). If the threshold had been set so that D was not considered similar to either $i2$ or $i3$, then the final hierarchy would have remained as illustrated in Fig. 1(*c*).

Thus, the choice of a similarity threshold value for this process is important as it determines, to a large extent, the shape of the classification tree. The value can range between 0 and 1. At the extremes, a value of 1 produces a maximally broad tree while a value of 0 yields a maximally deep tree.

3.3. *IMEM applied to the CSD*

In the set of examples presented only a few of the qualitative relations of Table 1 will be used by IMEM for the CSD datasets. A relation used in all datasets is that of atomic connectivity. Letting *erad*(*a*) be the covalent radius (Cambridge Structural Database, 1994) of an atom of type a, the *bonded* relation is defined as

$$\text{bonded}(x, y) \stackrel{\text{def}}{=} [erad(x) + erad(y) - \alpha \leq d(x, y)] \wedge [d(x, y) \leq erad(x) + erad(y) + \alpha],$$

where d is the interatomic distance (in Å). Thus, two atoms are bonded if their interatomic distance is less than the sum of their atomic radii, plus a constant factor α. Taking the radius of carbon to be 0.68 and setting $\alpha = 0.4$ determines that two carbons are considered to be bonded if their interatomic distance is between $0.68 + 0.68 - 0.4$ and $0.68 + 0.68 + 0.4$ (0.96–1.76) Å. Since oxygen is also assigned a covalent radius of 0.68 Å, the same range is used for carbon–oxygen and oxygen–oxygen bonding. This same definition of atomic connectivity in crystal structures is used by the CSD software.

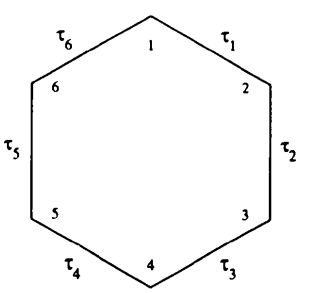
All of the molecular graphs exhibited by the CSD datasets studied in this paper have previously been classified manually and/or automatically using numerical clustering methods. Thus, for each CSD dataset it is possible to ask two questions: 'are the archetypal classes representable by a small set of disjoint concepts?' (the question of *representational* adequacy) and 'are there sufficient numbers of instances of these concepts in the data?' (the question of *empirical* adequacy).

4. Results

4.1. *Six-membered carbocycles*

A general six-membered carbocycle adopts a small number of typical conformations, and conformation classes discovered by an unsupervised learning procedure can be validated against known classes. A dataset of 222 six-membered rings, chosen to exhibit a broad spectrum of chemical environments and conformations, has been retrieved from the CSD and previously analyzed using various numerical clustering procedures (Allen, Doyle & Taylor, 1991*a*,*b*,*c*). This dataset is reanalyzed here using IMEM and compared with the Jarvis–Patrick numerical clustering scheme of Allen *et al.* (1991).

Fig. 2 displays the 2D topology of the six-membered ring. For the purposes of numerical clustering, the six intraannular torsion angles $\tau_1, \ldots, \tau_6$ (Fig. 2, Table 2) were used. The six attributes τ_1–τ_6 take on different values with each application of a 2D symmetry permutation and this introduces complications for numerical clustering routines which require a metric for the computation of dissimilarities between examples. Since the six-membered ring has extensive permutational symmetry (Fig. 2), the similarity of two structures under one permutation may be different from their similarity with respect to another permutation. The solution presented by Allen, Doyle & Taylor (1991*a*,*b*) is to fix one conformation, say C in the similarity expression (1) above, and use the maximum similarity, $S(C, D)$, obtained from all possible permutations of D.



1	2	3	4	5	6	
2	3	4	5	6	1	
3	4	5	6	1	2	
4	5	6	1	2	3	
5	6	1	2	3	4	
6	1	2	3	4	5	
6	5	4	3	2	1	
5	4	3	2	1	6	
4	3	2	1	6	5	
3	2	1	6	5	4	
2	1	6	5	4	3	
1	6	5	4	3	2	

Fig. 2. Six-membered carbocycle: torsional descriptors and atomic permutational symmetry group.

Table 2. *Conformations of six-membered carbocycles: torsional and conceptual descriptions*

Name	Torsional description						IMEM conceptual description					
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
Phenyl	0	0	0	0	0	0	*sp*	*sp*	*sp*	*sp*	*sp*	*sp*
Chair	60	−60	60	−60	60	−60	+*c*	−*c*	+*c*	−*c*	+*c*	−*c*
Boat	0	−60	60	0	−60	60	*sp*	−*c*	+*c*	*sp*	−*c*	+*c*
Twist-boat	33	33	−70	33	33	−70	+*c*	+*c*	−*c*	+*c*	+*c*	−*c*
Sofa	30	0	0	−30	60	−60	+*c*	*sp*	*sp*	−*c*	+*c*	−*c*
Half-chair	45	−15	0	−15	45	−62	+*c*	*sp* or −*c*	*sp*	*sp* or −*c*	+*c*	−*c*
Screw-boat	40	0	−22	0	40	−60	+*c*	*sp*	*sp* or −*c*	*sp*	+*c*	−*c*

This is termed 'symmetry-modified' clustering. Symmetry is automatically detected and handled by IMEM.

To represent six-membered rings, cycloheptanes and steroid side chains in IMEM, we use the Klyne–Prelog relations, which partition the circular torsion angle space into four regions (Klyne & Prelog, 1960; see Fig. 3). The partitioning gives rise to four qualitative spatial relations or, equivalently, one functional relation. A connected chain of four atoms is planar if their torsion angle is in the *sp* or *ap* ranges. Klyne & Prelog (1960) used a planarity value of 30° in their paper, so that *sp* is −30 to +30°, and *ap* is 150–210°. Tests with IMEM have shown, however, that the value of 30° is not equally appropriate for all problems, because it often overgeneralizes planarity. Thus, the Klyne–Prelog framework has been used to define a *family* of functional relations, each given by a particular choice of a planarity parameter γ (see Fig. 4).

Table 2 lists the seven standard canonical conformations of a six-membered ring. These were derived by energy minimization calculations and are taken from Table 1 of Allen & Taylor (1991). The rightmost column lists the *structural sequence* for the canonical series of torsion angles: a listing of the corresponding Klyne–Prelog relations. When there is some ambiguity about the structural sequence, *i.e.* when a canonical angle is close to a boundary, this is noted with a disjunction of possibilities. It can be seen that IMEM unequivocally represents the four major conformation classes of phenyls, chairs, boats and twist-boats. That is, no concept for any of these archetypes could possibly denote another archetype. However, the half-chair concept subsumes the sofa concept. These two conformations are actually quite close in conformational space (Allen & Taylor, 1991). Also, some screw-boats could be interpreted as half-chairs, if τ_3 is *sp*. Again, these two conformations are close in conformational space, so this occasional equivocation is permissable.

Having considered the question of representational adequacy for the six-membered rings, we now turn to the question of empirical adequacy. The IMEM method was used to cluster 222 six-membered rings, using a similarity threshold of $t = 1.0$ and a planarity value of $\gamma = 15°$. Ten concepts, covering 219 instances, were created. Table 3 presents the discovered concepts. The second column shows the structural sequence under a permutation from the symmetry group chosen to match the standard conformation classes listed in Table 2. The third column gives a standard name to the discovered concept. These were derived from a comparison with the archetypal concepts given in Table 2. The three singletons in the IMEM clustering are instances 58 (BABPIP), 116 (BEWNOG) and 21 (ACCITR10).

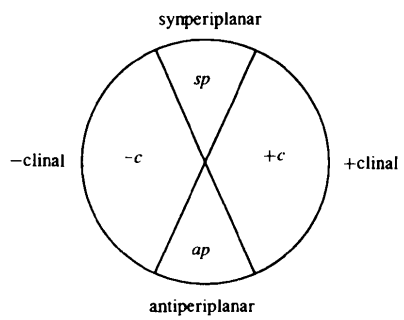


Fig. 3. The Klyne–Prelog (1960) relations which partition torsion angle space to generate four qualitative relational attributes.

$sp(p, q, r, s) \stackrel{\text{def}}{=}$

if bonded (p, q) and bonded (q, r) and bonded (r, s) then

$t(p, q, r, s) \geq 360 - \gamma$ or $t(p, q, r, s) < \gamma$

$+c(p, q, r, s) \stackrel{\text{def}}{=}$

if bonded (p, q) and bonded (q, r) and bonded (r, s) then

$\gamma \leq t(p, q, r, s) < 180 - \gamma$

$ap(p, q, r, s) \stackrel{\text{def}}{=}$

if bonded (p, q) and bonded (q, r) and bonded (r, s) then

$t(p, q, r, s) \geq 180 - \gamma$ and $t(p, q, r, s) < 180 + \gamma$

$-c(p, q, r, s) \stackrel{\text{def}}{=}$

if bonded (p, q) and bonded (q, r) and bonded (r, s) then

$180 + \gamma \leq t(p, q, r, s) < 360 - \gamma$

Fig. 4. Concept definitions for the Klyne–Prelog (1960) relations: $t(p, q, r, s)$ is the torsion angle.

Table 3. *Conformations of six-membered carbocycles discovered by IMEM (Id is the IMEM concept number)*

Id	Discovered concept						Standard name(s)	Frequency
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6		
1	*sp*	*sp*	*sp*	*sp*	*sp*	*sp*	Phenyl	35
2	*sp*	$-c$	$+c$	*sp*	$-c$	$+c$	Boat	52
3	$+c$	$-c$	$+c$	$-c$	$+c$	$-c$	Chair	64
4	$+c$	$+c$	$-c$	$+c$	$+c$	$-c$	Twist-boat	8
5	$-c$	*sp*	*sp*	*sp*	$-c$	$+c$	Half-chair	5
6	$-c$	$+c$	*sp*	$+c$	$-c$	$+c$	Half-chair	5
7	*sp*	$-c$	$+c$	$+c$	$-c$	$+c$	Boat-like	12
8	$-c$	*sp*	$+c$	*sp*	$-c$	$+c$	Screw boat	6
9	$-c$	*sp*	*sp*	$+c$	$-c$	$+c$	Half-chair, sofa	29
10	$+c$	$-c$	$+c$	$-c$	$-c$	$+c$	Boat-like	3
11	*sp*	*sp*	$-c$	$+c$	*sp*	*sp*	Phenyl-like	1

Table 4. *Comparison of concept discovery results for six-membered carbocycles from IMEM and from numerical clustering (Allen, Doyle & Taylor, 1991b) denoted as ADT (Id is the IMEM or ADT concept number, Np is the combined membership of a concept class)*

	ADT		IMEM	
Class	Id	Np	Id	Np
Phenyl	1	35	1	35
Boat	2	63	2	67
	3		7	
	4		10	
	5			
	6			
Chair	7	59	3	64
	8			
	9			
Twist-boat	12	9	4	8
Half-chair/sofa/	10	38	5	45
screw-boat	11		6	
			8	

Instances 58 and 116 are heavily constrained by complex bridging of their basic six-membered rings. The third singleton, instance 21, has torsion angles (5,7,−18,17,−6,−6) and is very close to the phenyl classification.

Using the same dataset of 222 six-membered carbocycles, the Jarvis–Patrick numerical clustering method used by Allen, Doyle & Taylor (1991*b*) created 14 clusters with membership greater than one, including two doublet clusters, and 14 singletons. To compare the numerical results with those of IMEM, the 12 numerical clusters with membership greater than two (see Table 4) were compared with the ten IMEM concepts. A comparative analysis of the numerical and IMEM clusterings reveals similarities. These can be considered for each archetypal concept depicted in Fig. 4(*a*).

Phenyl. The numerical and IMEM clustering results are identical: the same 35 instances are assigned to the phenyl class by both methods.

Boat. The numerical method uses five concepts to describe 63 boats. IMEM uses three concepts to describe 67 boats. All of the entries of the five numerical clusters appear in the three IMEM concepts, except for instance 58 (a singleton in IMEM). The five extra instances in the three IMEM boat concepts are (IMEM cluster identifier: numerical cluster identifier in parentheses)

instance 83 (BEHFAV): (−41, 78, −15, −60, 103, −37; 7:12)

instance 85 (BEHFAV): (−46, 80, −15, −60, 99, −31; 7:12)

instance 120 (BEWNOG): (22, 30, −32, −17, 60, −62; 10:singleton)

instance 48 (ACONTN10): (−22, 27, 20, −67, 73, −30; 10:singleton)

instance 175 (AMHPEN10): (−16, 16, 3, −21, 21, −3; 2:doublet).

Chair. The numerical method finds 59 chairs, IMEM 64. The three numerical chair clusters are completely contained within a single IMEM concept. The five extra IMEM chairs, all singletons in the numerical analysis, are

instance 20 (ACAMYA): (56, −57, 55, −50, 49, −53)

instance 34 (ACLYCA10): (55, −63, 63, −54, 43, −44)

instance 76 (BCYLON10): (65, −70, 70, −53, 45, −56)

instance 117 (BEWNOG): (66, −92, 89, −92, 62, −44)

instance 181 (BEWNOG): (28, −22, 37, −54, 59, −47).

Twist-boat. As mentioned above, the numerical method places two IMEM boats (83 and 85) in this class. IMEM places instance 78 (BCYLON10) (−31, −30, 56, −18, −44, 71), a singleton in the numerical method, in the twist-boat concept. The numerical and IMEM clusters are otherwise identical.

Half-chair/sofa/screw-boat. The two methods differ most for these classes. The numerical method discovers 29 half-chairs and assigns nine instances to a sofa/screw-boat class. IMEM assigns 39 instances to a half-chair/sofa class, but also discovers a distinct screw-boat concept covering six instances. There are four possible IMEM concepts for the half-chair (see Table 2) and only two of these occur in the dataset.

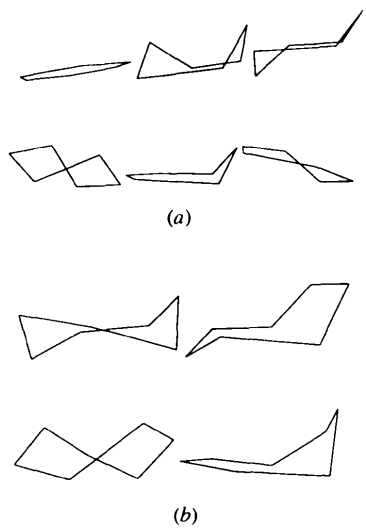
Table 5. *Conformations of seven-membered rings: torsional and conceptual descriptions*

Name	Torsional description							IMEM conceptual description						
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7
Chair	64	−84	66	0	−66	84	−64	+c	−c	+c	*sp*	−c	+c	−c
Twist-chair	54	−72	88	−39	−39	88	−72	+c	−c	+c	−c	−c	+c	−c
Boat	−58	−31	70	0	−70	31	58	−c	−c	+c	*sp*	−c	+c	+c
Twist-boat	−64	−18	75	−18	−64	45	45	−c	*sp* or − c	+c	*sp* or − c	−c	+c	+c

The proportions of the archetypal conformations in the 222 six-membered carbocycles are not representative of the whole CSD. To clarify this point, the CSD search program *QUEST*3D (Cambridge Structural Database, 1994) was used to extract all rings of six carbons connected by single bonds, regardless of their chemical environment. Only structures with an R factor ≤ 0.10 were considered. This search produced 16466 examples. Processing this dataset using IMEM gave the following results

chair	11892	(72.2%)
boat	2030	(12.3%)
twist-boat	503	(3.1%)
planar	103	(0.6%)
others	289	(1.9%).

We note that this simple search definition for a six-membered carbocycle does not exclude all occurrences of rings containing Csp^2 atoms, *i.e.* those with exocyclic double bonds: the primary consideration was to test IMEM with a very large dataset of unknown composition. Nevertheless, the chair is the energetically preferred conformation of a six-membered carbocycle and this is reflected in its dominant relative population.



(*a*)

(*b*)

Fig. 5. Conformational concepts discovered by IMEM: (*a*) for six-membered rings (left to right): phenyl, boat, chair, twist-boat, half-chair/sofa and screw-boat; (*b*) for seven-membered rings: twist-chair, chair, twist-boat and half-chair.

4.2. *Cycloheptane*

The cycloheptane dataset (dataset 7C1 of Allen, Howard & Pitchford, 1993) was processed by IMEM in an analogous manner to the six-membered carbocycles: the only difference being the ring size and, hence, the order of the permutation symmetry group (see Fig. 6). A planarity value of $\gamma = 15°$ was also used to represent cycloheptane conformations. Table 5 lists the four archetypal forms of cycloheptane and shows that the twist-boat concept subsumes the boat concept, that is, IMEM does not uniquely represent the twist-boat class using the Klyne–Prelog relations. Aside from this minor equivocation, IMEM can adequately represent the major cycloheptane conformations.

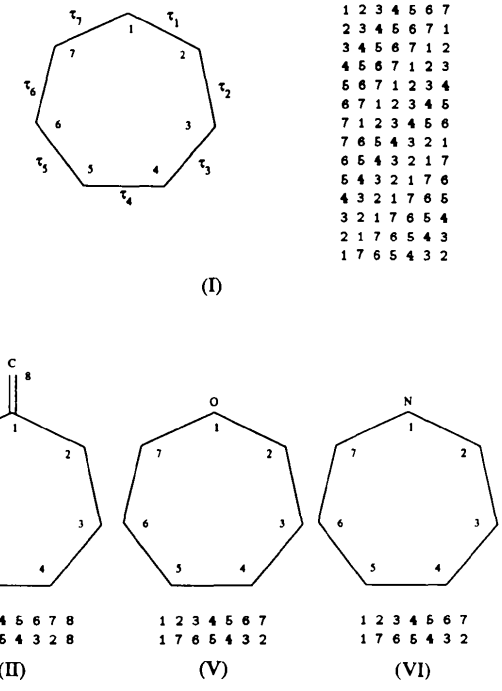


Fig. 6. Seven-membered rings: torsional descriptors and atomic permutational symmetry groups for cycloheptane (I), methylene-cycloheptane (II), oxacycloheptane (V) and azacycloheptane (VI), from Allen, Howard & Pitchford (1993) for (I) and Allen *et al.* (1994) for (II), (V) and (VI).

Table 6. *Conformations of seven-membered rings discovered by IMEM (Id is the IMEM concept number)*

Id	Discovered concept							Standard name(s)	Frequency
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7		
1	$+c$	$-c$	$+c$	$-c$	$-c$	$+c$	$-c$	Twist-chair	56
2	$+c$	$+c$	$-c$	*sp*	$-c$	$+c$	$-c$	Chair	29
3	$-c$	$-c$	$+c$	*sp*	$-c$	$+c$	$+c$	Twist-boat, boat	10
4	$-c$	*sp*	$+c$	*sp*	$-c$	$+c$	*sp*	Boat-like	2
5	$-c$	$-c$	$+c$	$-c$	$-c$	$+c$	$+c$	Boat-like	1
6	$-c$	*sp*	*sp*	$+c$	$-c$	$+c$	$+c$	Novel	1
7	*sp*	$-c$	$+c$	$-c$	$-c$	$+c$	$-c$	Twist-chair-like	1

Allen, Howard & Pitchford (1993) extracted a dataset of 101 cycloheptane structures from the CSD. They note that there are three main conformational subgroups in this particular dataset: chairs, twist-chairs and boat/twist-boats. Jarvis–Patrick numerical clustering generated four clusters with membership greater than two, two doublet clusters and nine singletons. IMEM created five concepts and four singleton concepts (Table 6). One singleton was subsequently found to have crystallographic coordinate errors. The other three singletons are instances 100 (HPAPTX) (−31, 50, −5, 0, −54, 102, −48), 17 (BUHXAD) (−32, 83,−52, −36, 76, −35, 1) and 57 (HYMINA) (−21, 75, −22, −64, 55, 29, −53). A visual inspection revealed that singleton instance 100, concept 6 in Table 6, looks very much like a half-chair due to its two contiguous *sp* relations (see Fig. 5*b*). As with the six-membered carbocycles, the results of IMEM can be compared with the numerical results for each archetypal class (see Table 7).

Twist-chairs. IMEM discovers 56 twist-chairs, the numerical method generates 48. Of the extra IMEM twist-chairs, six appear in doublet clusters or as singletons in the numerical study and instances 19 (CEBBUG) (63, −88, 59, 18, −78, 85, −62) and 32 (DIPHEP10) (67, −85, 55, 16, −74, 76, −59) are numerically classified as chairs. Both these have a τ_4 torsion angle just slightly into the $+c$ range, causing them to fall under the IMEM twist-chair concept.

Chairs. As mentioned above, the numerical method classifies two IMEM twist-chairs as chairs. IMEM also adds instance 15 (BOLWUU), a member of a numerical doublet cluster, to the chair class.

Boats. The boat clusters from numerical analysis and from IMEM are identical.

Fig. 5(*b*) depicts the IMEM images for the three archetypal concepts discussed above and for the unique half-chair instance.

4.3. *Exo-unsaturated and heterocyclic seven-membered rings*

In a second study of seven-membered rings, Allen, Howard, Pitchford & Vinter (1994) applied Jarvis–Patrick numerical clustering to *exo*-unsaturated and heterocyclic systems. Fig. 6 illustrates three of these structures which are analyzed here by IMEM. Note that their topological symmetry is constrained by the presence of an exocyclic atom at C1 (II), or a heteroatom at position 1 (V, VI). The archetypal classes for these structures are the same as for cycloheptane (Table 5), except that there are four possible forms of each archetype. These are produced by a symmetry element passing through one of the pairs of atoms 4–5, 3–4/5–6, 2–3/6–7 or 1–2/7–1.

Table 7. *Comparison of concept discovery results for seven-membered rings from IMEM and from numerical clustering (Allen, Howard & Pitchford, 1993), denoted as AHP (Id is the IMEM or AHP concept number, Np is the combined membership of a concept class)*

	AHP		IMEM	
Class	Id	Np	Id	Np
Twist-chair	1	48	1	56
	3			
Chair	2	30	2	29
Boat/twist-boat	4	10	3	10

Table 8. *Comparison of concept discovery results for seven-membered rings (II), (V) and (VI) (Fig. 5) from IMEM and from numerical clustering (Allen, Howard, Pitchford & Vinter, 1994) denoted as AHPV*

The comparison is in terms of numbers of rings assigned to each class. Class descriptors are those defined by Allen, Howard, Pitchford & Vinter (1994).

Class	AHPV	IMEM
	(II)	
TC2	14	15
TC3	15	15
C1	12	12
TC4	8	8
C2	2	2
	(V)	
TC3	17	17
TC2	12	16
C2	8	6
C3	5	4
	(VI)	
TC3	11	11
C1/TC4	6	6

Table 9. *Conformations of steroid C17 side chains: torsional and conceptual descriptions (conformation names are from Duax, Griffin, Rohrer & Weeks, 1980)*

Name	Torsional description						IMEM conceptual description					
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6
A	180	180	180	180	180	3–120	*ap*	*ap*	*ap*	*ap*	*ap*	*sp* or $+c$
B	180	180	180	40–70	180	43–85	*ap*	*ap*	*ap*	$+c$	*ap*	$+c$
C	180	180	56–85	180	180	−39 to −64	*ap*	*ap*	$+c$	*ap*	*ap*	$-c$
D1	180	60–73	180	−67 to −99	180	−39 to −64	*ap*	$+c$	*ap*	$-c$	*ap*	$-c$
D2	180	60, 86	180	180	180	−3, −59	*ap*	$+c$	*ap*	*ap*	*ap*	*sp* or $-c$
D3	180	57, 65	180	61, 64	180	56, 70	*ap*	$+c$	*ap*	$+c$	*ap*	$+c$

Table 10. *Conformations of steroid C17 side chains discovered by IMEM (Id is the IMEM concept number)*

Id	Discovered concept						Standard name	Frequency
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6		
1	*ap*	*ap*	*ap*	*ap*	$+c$	*ap*	A	93
2	*ap*	*ap*	*ap*	*ap*	*ap*	*sp*	A	12
3	*ap*	*ap*	*ap*	$-c$	$-c$	*ap*	B	7
4	*ap*	*ap*	$-c$	*ap*	$+c$	*ap*	C	4
5	*ap*	$+c$	*ap*	$-c$	$-c$	*ap*	D1	3
6	*ap*	$+c$	*ap*	*ap*	$-c$	*ap*	D2	8
7	*ap*	$-c$	*ap*	*ap*	*ap*	*sp*	D2	2
8	*ap*	$+c$	*ap*	$+c$	$+c$	*ap*	D3	4
9	*ap*	$+c$	*ap*	*ap*	$+c$	*ap*	D2-like	5
10	*ap*	*ap*	$+c$	*ap*	$+c$	$-c$	C-like	4
11	*ap*	*ap*	$+c$	*ap*	*ap*	$+c$	C-like	2
12	*ap*	*ap*	*ap*	*ap*	$+c$	$-c$	A-like	3

The IMEM method was applied to three datasets: II (64 examples), V (50 examples) and VI (26 examples), using a planarity value of $\gamma = 15°$. The cluster population results from IMEM and from the numerical analysis are given in Table 8; the two methods are seen to be in very close agreement.

4.4. *Steroid C17 side chain*

The six-membered carbocycle and the cycloheptane datasets were chosen to illustrate high topological symmetry (orders 12 and 14, respectively) combined with the conformational constraints imposed by ring cyclicity. The steroid C17 side chain is a structure unconstrained by ring cyclicity and it exhibits low-order topological symmetry, as shown in Fig. 7.

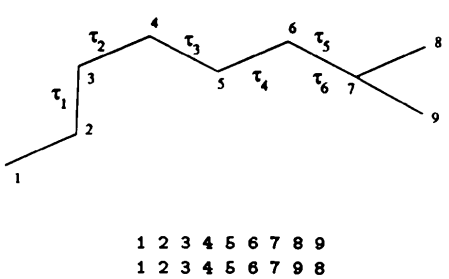


Fig. 7. Steroid C17 side chain and atomic permutational symmetry group. The C2 atom here corresponds to C17 in standard steroid numbering.

The Klyne–Prelog relations are also used to represent steroid C17 side chain conformations (Table 9). However, a planarity value of $\gamma = 15°$, as used for the previous two datasets, was found to be inappropriate. In the steroid side chains many of the 180° values in the second column of Table 9 fluctuate well outside the *ap* range (165–195°), which is defined by $\gamma = 15°$. Torsion angles in ring systems are more constrained than those in fully flexible acyclic chains, thus it is appropriate to allow a wider definition of planarity in this case.

The biological importance of the steroid side chain conformation has led to considerable crystallographic activity in this area. The conformations of the C17 side chains of Fig. 7 were analyzed manually by Duax, Griffin, Rohrer & Weeks (1980), who generated the set of archetypal classes reproduced in Table 9. The corresponding IMEM concepts, using a planarity value of $\gamma = 30°$, are in the rightmost column of the table. In contrast to the previous two cyclic datasets, there is no equivocation of canonical forms for the C17 side chains: each class has a unique description. Processing a dataset of 151 C17 side chain examples with IMEM yielded 12 concepts with two or more instances and four singletons (see Table 10). All of the singletons were subsequently identified as having slightly abnormal geometry due to high thermal motion and/or unresolved disorder. Small adjustments to the standard covalent radius of carbon had been made in the CSD to generate correct connections in these few special cases. Inadvertently, these adjusted radii were not passed to IMEM, hence this program was unable to

Table 11. *Comparison of concept discovery results for steroid C17 side chains from IMEM and from numerical clustering (Allen, Bath & Willett, 1995), denoted as ABW (Id is the IMEM or ABW concept number, Np is the combined class membership)*

	ABW		IMEM	
Class	Id	Np	Id	Np
A	1	112	1	105
			2	
B	7	6	3	7
C	9	5	4	4
D1	2	3	5	3
D2	6	7	6	10
			7	
D3	3	4	8	4

assess bondedness correctly. These four individuals are omitted from Table 10. Given that τ_5 and τ_6 can be interchanged due to the topological symmetry of the fragment, then IMEM concepts 1–8 correspond to the archetypal classes (Table 9) identified by Duax, Griffin, Rohrer & Weeks (1980). Concepts 9–12, covering 14 instances, are minor variations on these major classes, as indicated in Table 10.

Allen, Doyle & Taylor (1991*c*) processed a set of 101 examples of steroid side chains using single-linkage and Jarvis–Patrick clustering algorithms. They considered both *chiral* and *achiral* clustering, where the latter regards enantiomorphs as equivalent. Allen, Bath & Willett (1995) repeated the achiral experiment (which is formally equivalent to the IMEM procedure) using the augmented set of 151 examples that were used in the IMEM experiment above. The numerical single-linkage clustering algorithm was used and, since this is linked directly to the CSD master files, connectivities were correct for all 151 examples. The numerical clustering generated ten clusters with two or more instances and covering 148 of the 151 examples; there were three singleton clusters. The numerical and IMEM results are compared in Table 11 and minor differences are detailed below for each archetypal concept:

Class A. The numerical method assigns 112 instances to the *A* class, IMEM 105. Of the seven extra numerical instances, one is an IMEM singleton, three others (instances 23, 28 and 92) are assigned by IMEM to the D2 class, since τ_2 is well out of the *ap* range at −110.9, −120.2 and −99.9°, and one (116) is assigned to the *B* concept, since τ_4 is out of the *ap* range at 122.0°. The remaining two instances appear in IMEM concept 12.

Class B. The *B* clusters are identical, except that IMEM adds instance 116, as noted above.

Class C. The *C* clusters are indetical, except that the numerical method incorporates an IMEM singleton.

Class D1. The clusters are identical.

Class D2. As noted above, IMEM assigns instances 23, 28 and 92 to the *D2* class. Otherwise, the clusters are identical.

Class D3. The clusters are identical.

4.5. *Hexopyranose sugars*

The previous examples have illustrated applications of IMEM to the conformational clustering of 3D molecular graphs and, in each case, the effects of topological permutational symmetry was an additional complexity. By contrast, the hexopyranose sugars, which have the hydrogen-depleted 2D graph of Fig. 8, are topologically asymmetric by virtue of the exocyclic C6 atom. Further, a classification of the set of 3D graphs that correspond to the 2D representation of Fig. 8 involves dividing this 3D graph set into configurational rather than conformational classes (Allen & Fortier, 1993). Thus, each ring C atom is stereogenic, yielding $2^5 = 32$ possible stereoisomers. However, although we may assume *a priori* that the hexopyranose ring adopts a chair conformation, the ring can exist in a + or − form ($^1C_4$ or $^4C_1$, depicted in Allen & Fortier, 1993). The inversion of the ring does not alter the local (*R*,*S*) chirality at any of the C atoms so that the set of 3D graphs can contain up to 64 conformational/configurational subclasses.

In this experiment, the Klyne–Prelog relations (*sp*, +*c etc.*) are not used. Instead, a binary relation of nonbonded contact or *proximity* is used; this relation computes whether or not two atoms are within a van der Waals contact distance of each other. Letting *vrad*(*a*) be the van der Waals radius of an atom of type *a*, the *proximal* relation is defined as

$$proximal(x, y) \stackrel{\text{def}}{=} [\text{not bonded}(x, y)] \wedge [d(x, y) \leq vrad(x) + vrad(y) + \beta],$$

that is, two atoms are *proximal*: (*a*) if they are not bonded and (*b*) if their interatomic distance is less than the sum of their van der Waals radii plus a constant factor β. A family of proximity relations arises by varying β. Taking the van der Waals radii of carbon and oxygen to be 1.70 and 1.52 Å, the upper bounds of some proximity relations are shown in Table 12(*a*). A value of $\beta = 0.07$ is used for the experiments reported here

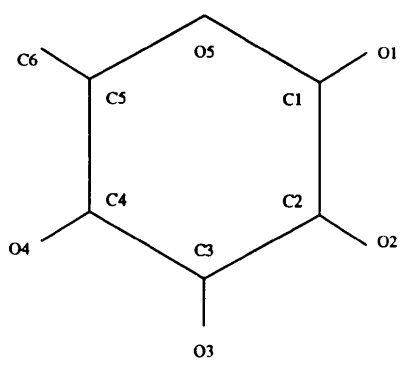


Fig. 8. Atomic nomenclature for hexopyranose sugars.

Table 12. *Hexopyranose sugars: (a) upper-bound distances for proximity between different pairs of atoms for different values of β and (b) the five unique proximity values at β = 0.07*

(*a*)

Atom	Atom	$\beta = 0.00$	$\beta = 0.05$	$\beta = 0.07$	$\beta = 0.10$
C	O	3.22	3.27	3.29	3.32
C	C	3.40	3.45	3.47	3.50
O	O	3.04	3.09	3.11	3.14

(*b*)

	O1	O2	O3	O4	C6
O5		δ_1		δ_2	
C1			δ_3		δ_4
C3	δ_5				

so, for example, a carbon and an oxygen are proximal if their interatomic distance is between 1.76 and 3.29 Å.

A dataset of 249 hexopyranose structures, described by Allen & Fortier (1993), was processed using IMEM. A similarity threshold of $t = 0.75$ was used to encourage the formation of a multilayer concept taxonomy, that is, concepts with fewer parts than the training examples. IMEM only searched for *connected* images, where every atom has a transitive path to every other atom via the *bonded* relation. The concept taxonomy of Fig. 9 was created from 249 examples. When all immediate successors of a concept are individuals, these are not displayed in the taxonomy. The taxonomy clearly shows 14 concepts (at the leaves) present in the dataset. Comparing these concepts with manual labelings of the 249 examples revealed that IMEM exactly reproduced the standard chemical classification of the pyranose sugars. The classification was achieved using the simple and intuitive relation of proximity. Note that the subsumption hierarchy (Fig. 9) is not strictly a tree: instances (*e.g.* COKBIN_88) may be subsumed by more than one parent class.

The numbers in parentheses for each concept of Fig. 9 are the numbers of parts in the image. The two concepts uniq-31 and uniq-8 both have eight parts: the central pyran ring and substituents O3 and C6. In uniq-31 C1 and O3 are proximal and in uniq-8 C1 and O3 are not proximal. The atoms C1 and C6 are proximal in the individual BIKWOH10 and not in uniq-31 or uniq-8 (therefore, all other 248 examples), so it is not subsumed by either of these concepts.

Table 13. *Hexopyranose configurational classes discovered by IMEM (Id is the IMEM concept number)*

Id	Discovered concept δ_1	δ_2	δ_3	δ_4	δ_5	Standard name	Frequency
uniq-1	0	0	0	0	0	β-D-Glucose	104
uniq-19	0	0	0	0	1	α-D-Glucose	73
uniq-16	0	1	0	0	0	β-D-Galactose	25
uniq-7	0	1	0	0	1	α-D-Galactose	19
uniq-10	1	0	0	0	1	α-D-Mannose	8
uniq-14	1	1	0	0	1	α-D-Talose	3
uniq-17	0	0	1	0	1	α-D-Allose	2
uniq-22	0	1	1	0	0	β-D-Gulose	3
BIKWOH10*	0	0	0	1	0	α-L-Idose	1
uniq-26	1	0	0	0	0	β-D-Mannose	3
COKBIN*	0	0	1	0	0	β-D-Allose	1
uniq-34	0	1	1	0	1	α-D-Gulose	2
uniq-33	1	0	1	0	1	α-D-Altrose	3
PAIDOP*	1	1	1	0	1	α-D-Idose	1

After clustering, the 14 discovered concepts were inspected to see which proximity values were invariant. It was found that only seven interatomic proximity values varied across the discovered concepts. These are the proximity relations between O5—O2, O5—O4, C1—O3, C1—C6, C—O4, C3—O1 and C4—O2. Among the instances of any particular discovered concept, all other proximity values are equivalent. Furthermore, the O5—O2 and C4—O2, and O5—O4 and C2—O4 proximities were found to be functionally equivalent. Thus, a total of five interatomic proximity values, listed in Table 12(*b*), might be used to cluster the 249 pyranose sugars. For example, referring to Table 12(*b*), δ_3 is the attribute for the proximity value between C1 and O3. Table 13 uses these attributes to

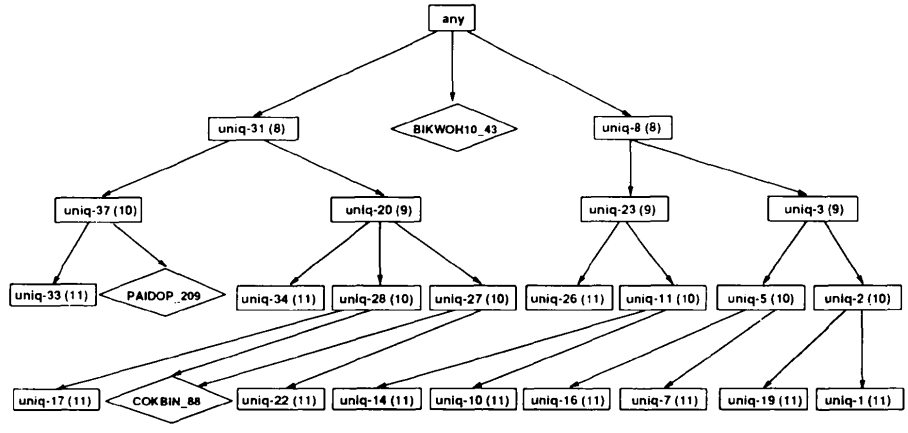


Fig. 9. The hexopyranose concept hierarchy for 249 instances. Individual (singleton) concepts are in diamonds. The numbers in parentheses for each concept are the numbers of parts in the image.

give a compact structural sequence description to each concept. A '1' is an abbreviation for true, '0' for false.

The attributes of Table 12(*b*) can be correlated with the standard D, L and α, β attributes for the pyranose sugars. It can be seen from Table 13 that BIKWOH10, the only L-form sugar in the dataset, is also the only one where C1 and C6 are proximal (attribute δ_4). This is perfectly reasonable, considering that the configuration at the stereocenter C5 determines the D or L designation for a sugar and that C6 is the substituent of C5. Thus, the attribute δ_4 determines the D or L designation. The configuration at the stereocenter C1 determines the α or β designation and it appears that one attribute, δ_5, can distinguish between α and β forms (see, for example, uniq-1 and uniq-19). The three remaining attributes, δ_1, δ_2 and δ_3, can be used to designate the eight major sugar classes (glucose, galactose, mannose *etc.*). The five qualitative attributes will give rise to a total of 32 (2^5) concepts.

An analysis was performed to determine the sensitivity of the clustering to the van der Waals proximity parameter β. It was found that values of β in the range 0.06–0.09 produced perfect clusterings. Below this range, at $\beta = 0.05$, instance 121 (refcode FITKAU) becomes a singleton concept. An inspection of this structure revealed that it has an O2—O3 distance of 3.09 Å, sufficient to indicate proximity only at higher β values (see Table 12*a*). In all other β-D-glucose instances, O2 and O3 are proximal. At $\beta = 0.0$, the same taxonomy as for $\beta = 0.05$ is produced, with the exception that two extra concepts are fomed. One is an α-D-galactose split (two instances), the other is a β-D-glucose split (three instances). Above the 0.06–0.09 range, at $\beta = 0.1$, instance 70 (refcode CELGIJ) becomes a singleton concept, since O1 and C4 are then proximal, with a distance of 3.32 Å (see Table 12*a*). The O1 and C4 atoms are not proximal in any other α-D-glucose instances. Thus, all values of β in the range 0.00–0.10 produce reasonable clusterings of the hexopyranose dataset. As β increases past this range, clustering quality slowly but progressively deteriorates.

Allen & Fortier (1993) cluster the data into the same categories as those of Table 13. As numerical descriptors, they used six intraannular (*e.g.* O5—C1—C2—C3) and five improper (*e.g.* O5—C1—O1—C2) torsion angles. Chemical expertise was necessary to determine these quantitative parameters from the many possible available descriptors. In fact, only the five improper torsion angles (projected valence angles) were necessary to determine configuration and these are related to the standard *R* and *S* descriptors of stereochemistry. To test this hypothesis using IMEM, two relations were defined as

$$R(p,q,r,s) \stackrel{\text{def}}{=} \text{if bonded}(p,q) \text{ and bonded}(p,r) \text{ and bonded}(p,s) \text{ then } t(p,q,r,s) \geq 0$$

$$S(p,q,r,s) \stackrel{\text{def}}{=} \text{if bonded}(p,q) \text{ and bonded}(p,r) \text{ and bonded}(p,s) \text{ then } t(p,q,r,s) < 0.$$

These 4-ary relations are similar to the 5-ary *R* and *S* relations of stereochemistry, with the only difference being that a substituent hydrogen is assumed present. They are also very similar to the Klyne–Prelog relations, except that they are not defined among chains of atoms. Rather, they are defined between three atoms and a substituent, as in the improper torsion angles used by Allen & Fortier (1993), *e.g.* C5—C4—O4—C3 in Fig. 8. These *R* and *S* relations also led to perfect partitionings of the hexopyranose dataset. Two experiments were performed: a chiral and an achiral clustering. In the chiral approach IMEM created 16 rather than 14 concepts. The two extra concepts represented enantiomorphs of β-D-glucose (nine instances) and α-D-glucose (one instance). These ten instances were previously identified by Allen & Fortier (1993) as examples where the coordinates in the published papers had been accidentally inverted. These enantiomorphs were not discovered using the *proximal* relation, as proximity is invariant under mirror inversion. In the achiral approach 13 rather than 14 concepts were discovered, with the D and L forms of the α-idose sugars merged into a single concept.

5. Discussion

This paper has described the application of the IMEM structured concept discovery approach to several datasets drawn from the CSD. These were chosen to illustrate different points: the conformational clustering of cyclic structures (datasets 1 and 2), extended chain structures (dataset 3) and structures whose 3D shapes are determined by stereochemistry (dataset 4). In each case the discovered concepts were compared with published results from numerical clustering techniques and also with accepted chemical classifications. For all datasets, the concepts discovered by IMEM are chemically sensible and this has been illustrated by a rigorous comparison with archetypal conformation classes. For the hexopyranose dataset, a technique for stereochemical partitioning not based directly on standard stereochemical descriptors proved to be effective. All discoveries were made using simple and chemically intuitive relations, such as Klyne–Prelog torsion angle partitioning, van der Waals proximity and relations that mimic *R*,*S* stereochemical descriptors.

The IMEM approach has discovered concepts which correspond to actual low-energy conformers. Similar to the numerical clustering approach of Allen, Doyle &

Taylor (1991*b*), this was carried out with no *a priori* knowledge of the potential energy surface. The success of both approaches reflects and depends upon an ability to adequately represent and relate conformation classes. While the conceptual approach uses qualitative relations and logical subsumption, the numerical approach uses quantitative numerical features and distance metrics. Both techniques will have difficulties if conformations are not adequately represented in a training set or if there are no apparent populated peaks in conformation space for the molecule in question. However, in the latter case, IMEM may still be able to identify similar 3D substructures in examples. In addition, both techniques will often be able to express 3D arrangements which do not correspond to true conformation classes, *e.g.* rare conformers or instances that occur along an interconversion pathway between highly populated low-energy conformations. Further, discovered concepts could be used as starting points for computational energy minimizations, perhaps removing the need for searches over the complete conformational space (Shah & Dolata, 1993). In any case, discovered concepts could be used as templates or conformational units for structure generation and for small-molecule model building.

The IMEM approach has addressed a central problem that exists in the conformational clustering scheme of Allen, Doyle & Taylor (1991*b*), in that it removes the need for substantial user intervention during the discovery process. IMEM was not guided by *a priori* chemical knowledge of the number and form of the classes to expect in a dataset. Also, it appears that less *a posteriori* processing of the clustering results is necessary to extract meaningful classes. The numerical clustering method creates many singletons and doublet clusters and these are excluded from subsequent analyses using expert intuition and not by an autonomous process.

The IMEM approach does not, however, completely remove the need for user intervention in the clustering process. Indeed, as with any clustering program, the results of IMEM should be closely scrutinized. If a poor clustering is being presented, *e.g.* on the basis of chemical sensibility, then this is an indication that the relations used to express IMEM concepts should be modified. IMEM is as much a statement about knowledge representation as it is about learning. When the right representation (*i.e.* the correct relations) is used, the learning task becomes simplified and an appropriate clustering tends to appear immediately from IMEM. For this reason, we have started to explore automated methods for detecting the best relations for expressing IMEM concepts in a given problem domain (Conklin, 1995; Guo & Fortier, 1995). However, we feel that this paper does provide some empirical evidence that a small library of parameterized relations may be sufficient to describe the conformations of many cyclic and extended chain systems. Further experiments in knowledge discovery, particularly for systems that have not already been addressed by other methods, will be needed to fully substantiate this claim.

The clustering method of Allen, Doyle & Taylor (1991*b*) is best suited to the task of conformational clustering where all examples have the same constitution and are described by the same number of quantitative features. Although not illustrated in this paper, IMEM is in theory capable of more general discovery tasks, where recurrent patterns among disparate molecular graphs can be sought. This is part of the *pharmacophore discovery* task: identifying the common structural features in drugs of known and related activity (Golender & Rozenblit, 1983). Provided that pharmacophores of interest could be represented adequately by qualitative concepts, the IMEM structured concept discovery method could be applied.

IMEM provides a hierarchical approach to image classification. The subsumption hierarchy constructed by IMEM generally contains intermediate concept nodes as well as the final clusters of images. These intermediate levels provide for more efficient classification and retrieval techniques. Also they provide added insight into how conceptual clusters are related. For example, in the subsumption hierarchy depicted in Fig. 1(*d*) we can see that images *D* and *i*2 are related in terms of their most common subsumer *M*, which is a subimage for which one atom and bond have been deleted in each of the subsumed images.

Finally, although it cannot yet be claimed that IMEM applies equally well to all conformational clustering tasks, this paper has successfully considered a variety of molecular graphs illustrating different aspects of molecular shape. There are certainly many other graphs for which knowledge of conformational preferences remain buried within the entries of the Cambridge Structural Database, for example, higher-order and fused ring systems, hydrogen-bonded motifs, amino acids rotamers *etc.* Molecular database mining tools such as IMEM should have a role to play in uncovering these concepts.

References

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

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

Allen, F. H., Bath, P. & Willett, P. (1995). *J. Chem. Inf. Comput. Sci.* **35**, 261–271.

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., Howard, J. A. K., Pitchford, N. A. & Vinter, J. G. (1994). *Acta Cryst.* B**50**, 328.
Benedetti, E., Morelli, G., Nemethy, G. & Scheraga, H. A. (1983). *Int. J. Peptide Protein Res.* **22**, 1–15.
Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E. & Thornton, J. M. (1987). *Nature*, **326**, 347–352.
Cambridge Structural Database (1994). *CSD User's Manual.* Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.
Carbonell, J. (1989). Editor. *Machine Learning: Paradigms and Methods.* Amsterdam: Elsevier.
Chang, W., Shindyalov, I. N., Pu, C. & Bourne, P. E. (1994). *Comput. Appl. Biosci.* **10**, 575–586.
Conklin, D. (1995). *Knowledge Discovery in Molecular Structure Databases.* Ph.D. Dissertation, Queen's University, Canada.
Conklin, D. & Glasgow, J. (1992). *Machine Learning: Proceedings of the Ninth International Conference (ML92)*, pp. 111–116. New York: Morgan Kaufmann.
Duax, W. L., Griffin, J. F., Rohrer, D. C. & Weeks, C. M. (1980). *Lipids*, **15**(9), 782–792.
Engh, R. A. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.
Fisher, D. H. (1987). *Mach. Learn.* **2**, 139–172.
Fortier, S., Castleden, I., Glasgow, J., Conklin, D., Walmsley, C., Leherte, L. & Allen, F. H. (1993). *Acta Cryst.* D**49**, 168–178.
Gennari, J. H., Langley, P. & Fisher, D. (1989). *Artific. Intell.* **40**, 11–61.
Golender, V. E. & Rozenblit, A. B. (1983). *Logical and Combinational Algorithms for Drug Design.* Letchworth, England: Research Studies Press.
Gray, P. M. D., Paton, N. W., Kemp, J. L. K. & Fothergill, J. E. (1990). *Protein Eng.* **3**, 235–243.
Guo, S. & Fortier, S. (1995). *Abstracts of the American Crystallographic Association*, Vol. 23, p. 130 (Abstract M143, Montreal Meeting).
Hunter, L. & States, D. J. (1992). *Proc. Seventh IEEE Conf. on AI Applications: The Biotechnology Computing Mini-track.*
Huysmans, M., Richelle, J. & Wodak, S. J. (1991). *Proteins*, **11**, 59–76.
Klyne, W. & Prelog, V. (1960). *Experientia*, **16**, 521–523.
Lebowitz, M. (1987). *Mach. Learn.* **2**, 103–138.
Rooman, M. J. & Wodak, S. J. (1988). *Nature*, **335**, 45–49.
Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval.* New York: McGraw-Hill.
Shah, A. V. & Dolata, D. P. (1993). *J. Comput.-Aided Mol. Des.* **7**, 103–124.
Taylor, R. & Allen, F. H. (1992). *Structure Correlation*, edited by H.-B. Bürgi & J. D. Dunitz, pp. 111–161. Weinheim: VCH Publishers.
Thompson, K. & Langley, P. (1991). *Concept Formation: Knowledge and Experience in Unsupervised Learning*, pp. 127–161. New York: Morgan Kaufmann.
Willett, P. (1990). *Concepts and Applications of Molecular Similarity*, edited by G. M. Maggiora & M. A. Johnson, pp. 43–64. New York: Wiley.
Wirth, K. (1986). *J. Chem. Inf. Comput. Sci.* **26**, 242–249.