Computer analysis of molecular geometry, Part VII: The identification of chemical fragments in the Cambridge Structural Data File

P Murray-Rust and J Raftery*

Glaxo Group Research Ltd, Greenford Road, Greenford, Middlesex, UB6 OHE, UK
* Napier College, Colinton Road, Edinburgh, EH10 5DT, UK

The geometrical information in the Cambridge Structural Data File can be interpreted automatically to provide information about chemical bond types and atom types. Methods for using the Data File as a criterion for selecting or rejecting molecular fragments and also to identify substituents are outlined. Facilities in the program Geostat are described. Problems that can arise because of the different nature of the information in the data files and the different operation of the Connser and Geom substructure searches are outlined, and some procedures to overcome them are reported.

Keywords: data analysis, molecular geometry, Geostat

received 31 March 1985, accepted 15 April 1985

The Cambridge Structural Data File (DAT) contains the atomic coordinates from over 45 000 published crystal structures, and is now an established research tool for many workers in analysing and using molecular geometry. It is commonly used in conjunction with the bibliographic (Bib) and connectivity (Con) files, which are used to decide which datasets on DAT should be retrieved for the particular problem. The use of the program Connser, an atom-by-atom search for molecules or molecular fragments in the Con file, is now very common. Datasets can be retrieved (program Retrieve) for the 'hits' produced by Connser and processed by the program Geom. Details of these files and programs are described in several papers[1–5], the most recent of which gives a bibliography of the techniques used, and in the publications of the Cambridge Crystallographic Data Centre (CCDC).

One of the commonest procedures is the following:

(i) Select a molecule or molecular fragment, and code its structure as a set of atom and bond properties. Run Connser on Con, and save the refcodes (hits) corresponding to structures that contain this fragment.

This work was carried out at: Department of Chemistry, University of Stirling, Stirling FK9 4LA, UK

(ii) Retrieve the corresponding datasets from DAT (if they have coordinates) for input to Geom.

(iii) Set up a Frag input for Geom that defines the molecular fragment of interest. This will usually be the substructure defined in (i) or part of it. Geom will then tabulate requested parameters for every occurrence of the fragment that it finds on the DAT file.

For many molecules or fragments, particularly those with complex structures or those that occur rarely, this procedure works perfectly. Several examples are given in the CCDC guide[6]. In some cases, however, particularly where fragments are small and occur several times in a DAT entry (because of their chemical frequency), serious problems can occur if the procedure described above is used without considerable modification. These problems arise from the fundamentally different nature of the information in the Con and DAT files, and from the different operation of the Connser and Geom substructure searches. In the present paper these problems are outlined, and some additional procedures that should overcome many of them are reported.

CON FILE AND CONNSER

The Con file has been described in detail (CCDC User Manual[6]), and only a brief review of salient points will be given. For each Con entry, three main types of information are given:

- the chemical type of each atom, given by its elemental symbol, the number of coordinated non hydrogen atoms (NCA), the number of coordinated hydrogen atoms (NH), the formal (integral) atomic charge and the number of the residue to which it belongs,
- the properties of each bond, given as the two terminal atoms and the bond type (1 = single, 2 = double, 3 = triple, 5 = aromatic, 7 = delocalized and 9 = π-bonds to metals), which is negative if the bond is cyclic,
- the number of residues of each type in the overall molecular formula.

From this information, it is possible to reconstruct a

 0263-7855/85/020060-09 $03.00

complete chemical formula of the compound studied in the publication but without any stereochemical information. (Sometimes it is possible to deduce the stereochemistry from the corresponding Bib entry.) In a very few cases, it will not be possible to generate a complete chemical formula, and this only occurs when the authors have not provided enough information in their paper for the CCDC to assign an unambiguous formula. The exact coding on Con chosen by the CCDC is often only one of many alternatives; this occurs frequently with coordination compounds, π-bonded ligands in organometallic compounds, electron deficient compounds and systems where two or more resonance structures can easily be written.

DAT FILE AND GEOM

The information on the DAT file is fundamentally different. It includes checked cell dimensions, symmetry operators and atomic coordinates (for each of which an elemental symbol is given). If necessary, additional symmetry-related atoms are generated to represent the whole of the crystal chemical unit. There is also a connectivity record (generated by the Unimol program), based on the distances between atoms. This record does not incidate bond type nor (explicitly) whether a bond is cyclic. It is important to appreciate that the Frag search routine of Geom does not use this record, and it will not be referred to again.

The data on DAT will not, in many cases, represent the complete chemical formula of the crystal studied. There are many reasons for this, most common of which is that not all the atoms in the structure have been located by crystallographic analysis. This is particularly common for hydrogen atoms, but can also be true of other atoms that are partially disordered or that have high thermal motion. Where there is partial disorder, then, even where the alternative atomic sites have been identified, the positional information is often excluded. Where hydrogen atoms have been misplaced (or misreported) to the extent that they do not form chemically reasonable geometrical connections, they are excluded. In some cases, as in the metal cyanides M–C≡N, it may not be possible to distinguish the identities of atoms close in atomic number. In many structures of low accuracy, the X-ray data is quite insufficient to establish the chemical formula. In these cases, presumably, the formula published by the authors and coded in Con has been deduced from additional chemical information.

CHEMICAL STRUCTURE AND DAT ENTRIES

In principle, therefore, the Con file cannot be totally mapped onto the DAT, and, for this reason, Connser and Geom retrieve fragments in different ways. Geom must search for fragments solely on geometrical, not chemical, criteria, while Connser does the opposite. In using Geom, atoms are assigned default bonding radii, which can be altered with the Intra or Inter commands or overwritten with values in the particular DAT entry, and these are used to decide which atoms are bonded. For each atom, a table of connected atoms is then set up. (In the present version of the program, a maximum of ten connected atoms is allowed for each atom. If this is exceeded, additional atoms are not considered as being connected, although their distance from the central atom may be lower than the cut off. If radii in Intra or Inter are set high, then this can happen frequently and many fragments will not be retrieved.) The program then searches for fragments embedded in this network of connected atoms that fulfil the criteria in the frag definition. Unlike Connser, this definition does not allow the following:

- explicit bond type (eg 1, 2, 3, 5, 7, 9 in Con),
- whether a bond is cyclic or acyclic,
- the chemical nature of each atom in terms of its known chemical connectivity.

Bond types can, in many cases, be determined by a consideration of lengths, especially for types 1, 2, 3 and 5 (aromatic), but the distinction between 5 and 7 (delocalized) is not possible, nor is 9 (metal-ligand π-bond) easy to treat. As a result of the uncertainty as to which atoms, especially H, have been included in DAT, it is difficult to specify the hybridization state and number of H ligands of a given atom. Furthermore, although the option E (for exact coordination of an atom) can be used (see Appendix 1), its operation is different from that in Connser in that H atoms must be explicitly considered. For example, ATn C 2 E will retrieve CH_2 groups only if hydrogen atoms have not been located for that particular atom or if ExclH has been used to remove all H atoms from the Frag search. If hydrogen atoms are included, the string ATn C 4 E would be necessary. If Inter or Intra are used to calculate distances greater than normal bonded ones, extra connections will be generated that will result in the test failing.

An example may clarify the problem. In a study of the conformation of simple fragments[7], the frequency with which the acyclic fragment, $-CH_2-CH_2-CH_2-CH_2-$, adopts a gauche conformation (ie τ (CCCC) was in the range $-90°$ to $+90°$) was analysed. Possible approaches are shown in Appendix 1.

Connser can easily be used to retrieve (Appendix 1(a) (i)) all compounds that contain acyclic $-(CH_2)_4-$ groups. The problem then is that many of these compounds will also contain other groups that may satisfy the Frag criteria in Geom. The two main classes are:

- $-(CH_2)_4-$ as part of a ring (e.g. in cyclohexyl derivatives),
- polysubstituted butyl groups (as in the cholesterol side chain).

There are several ways of proceeding.

- Additional Connser searches can be run (Appendix 1(a), (ii)) to retrieve compounds that contain any of the unwanted groups. These compounds are then excluded from the original list of hits (Appendix 1(a), (i)). This has the advantage of decontaminating the data at an early stage but the drawback that many compounds with the desired features are excluded. The standard Frag search (Appendix 1(b), (i)) is, however, completely satisfactory when this has been done.
- Hydrogen atoms can be explicitly considered (Appendix 1(b), (ii)), so that substituted butyl groups will be excluded. However, useful structures for which some or all of the eight hydrogen atoms were not published will be excluded.
- All possible fragments are retrieved without any screening (Appendix 1(a), (i) and 1(b)(i)), and the

data are tabulated. Contaminating fragments can then be rejected by human examination of the chemical nature of each. Even with interactive graphics systems, this is labour-intensive and error-prone unless the dataset is small (when it is probably the best method). For $-(CH_2)_4-$ groups, a great deal of work is involved.

- The Frag search routine can be modified to include extra tests and to interpret (as far as is possible) geometrical data in terms of chemical bonding. Some of the modifications that have been made are detailed below.

CHIRALITY

The chirality of a fragment can easily be determined by Geom. In some cases, this is straightforwardly done using Test Tors on a proper torsion angle. In most cases, the test is best carried out on improper torsion angles, particularly for determining whether substitution at an atom is R or S. Thus, if a carbon atom is substituted by four different ligands all of which (n_1, . .,n_4) are unambiguously located in the fragment, the record

TEST TORS $n_1\ n_2\ n_3\ n_4\ t_1,t_2$

(t_1,t_2 being −180/0 and 0/180 as appropriate) will test the chirality.

This can be applied at any number of centres. Occasionally, a molecule will be reported in the file with a chirality opposite to that required, and it would be useful to have enantiomeric coordinates. Thus, a routine has been written to test a reference torsion angle and, if this test fails, to invert the structure. Subsequent Tests and Defs then operate on the inverted structure. In principle, it should be possible, using Geom alone, to describe the complete stereochemistry of most of the molecules on the file.

COVALENT BONDING FROM GEOMETRY

For most of the structures on the file, the bonding can be regarded as covalent, and its properties calculated from the geometry of the crystal structure. The major exceptions are found in the following: ionic compounds, irregular coordination geometry, organometallic π-bonding, electron deficient compounds and cluster compounds. Apart from these, it is possible to define a geometrical covalent connectivity for the structure, which is the set of interatomic distances that are less than the sum of the appropriate covalent radii (with a tolerance, in this case, of 0.4 Å). This connectivity is essentially that in the residues identified by the Unimol program (which is used by CCDC to process every structure before entry on the DAT file[3]). This connectivity (which is part of each DAT entry) will be referred to as the Unimol connectivity. It will be calculated afresh in Geom using the built-in covalent radii and tolerance; it will not be alterable by changing radii in the Inter or Intra record. This Unimol connectivity will then be used to determine bond orders, the hybridization state of atoms and whether bonds in residues are cyclic. For many purposes, the connection table searched by the Frag routine will be the Unimol connectivity (i.e. when the default radii are used).

GEOSTAT — MODIFICATIONS TO GEOM78

Substantial modifications have been made to Geom78, which overcome many (but not all) of the problems discussed in earlier sections. They can be used in any combination with themselves or with the published Geom options. They are now included in a program Geostat, which is being distributed by the CCDC to subscribers to the file. At present, Geostat is limited to FORTRAN 77 (Vax version).

Identification of cyclic bonds

An algorithm[8] has been implemented for determining all the cyclic bonds in the Unimol connectivity before entering other routines in Geostat (especially the Inter or Intra options). In most cases, these will be the cyclic bonds in the standard chemical formula of the molecule. When a bond is identified as cyclic, its length is made negative (cf. the Connser convention for cyclic bonds). If at a later stage, the connection table is increased (by Inter, Intra or Coord), the signs of the lengths of these bonds are not changed, even if they now form part of cycles. The algorithm (which is only called if the cyclic nature of bonds is queried by the Frag coding, i.e. keywords A, C, Allbond, or Nocs, or if the Label option is used) is based on the spanning tree of the graph of the connection table, and the time taken is roughly proportional to the number of vertices (atoms). The input coding is exactly as in Connser (see Appendix 1.)

Rejection of embedded fragments

It is often difficult to retrieve small fragments that correspond to a single chemical class of compounds. Thus, if one wishes to retrieve ethers, C—O—C, from DAT, one would also get esters, C—O—C=O, ketals, C—O—C—O, and other groups, the properties of which might be sufficiently different from ethers that one might wish to exclude them. A facility has been introduced that allows a fragment to be retrieved if it meets all the normal criteria in Frag and only if, in addition, it fails extra criteria defining an unwanted substitution pattern (see Appendix 2). Atoms that would occur in an unwanted superfragment are flagged by the keyword N.

A search is first made, as normal, for the fragment defined by the unflagged atoms and by the Bonds and Tests relating only to unflagged atoms. If this search fails, the program moves to the next DAT entry. If it is successful, the atoms flagged N are then used to generate a superfragment, in which the desired fragment might be embedded. If all the N-flagged atoms are found and Bonds and Tests relating to them satisfied, then the fragment is rejected and the search backtracks to find a new potential fragment. When these are all exhausted, the search moves to the next DAT entry. Notice that the Bonds and Test records are thus involved in rejection criteria if they relate to flagged atoms. The example in Appendix 2 should clarify the logic.

Hybridization and bond number

In the Con file, each atom record contains the atomic symbol, the number of nonhydrogen ligands, the num-

Table 1. BNEs for some octet structures involving first-row atoms

BNE	Approximate hybridization	Symbol				
0	sp^3	Q	—C—	—N:	—N—	i.e. QC and QN
1	sp^2	T	=C—	= N:	=N+	i.e. TC and TN
2	sp	B	=C=	≡C–	=N⁺	≡N: i.e. BC and BN

The character symbols Q (quaternary), T (ternary) and B (binary) are used for BNE = 0, 1 and 2, respectively. Lines from atoms represent bonded ligands.

ber of hydrogen ligands and the formal charge. For most compounds of p-block elements (nonmetals), it is possible to deduce from this information the number of multiple bonds (localizable in a resonance formulation) in which the atom is involved. In most cases, this will dictate the hybridization state of the atom, but this is not always the case: in particular, 3-coordinate uncharged N can be found in both pyramidal (e.g. in Me_3N, sp^3) and planar (e.g. R_2NAr, sp^2) arrangements. Confining the argument to octet structures of the first row, a bond number excess (BNE) can be defined as:

BNE(atom)=B + formal charge − number of ligands − group of Periodic Table (1)

Hence, the BNE is the number of p-orbitals that the atom contributes towards multiple bonding (see examples in Table 1). All the terms on the right-hand side of equation (1) are given in Con (the number of ligands is NCA + NH), but in DAT only the element symbol is given. Nevertheless, BNEs can normally be derived for many atoms on the DAT file from their coordination geometry in the Unimol connectivity, as follows.

For each bond AB, the length of a (sometimes hypothetical) single bond can be calculated from covalent radii as:

$$r_0(AB) = r(A) + r(B) \quad (2)$$

Pauling[9] defined the bond number (n_{AB}) of a bond AB, length r (AB) as:

$$r(AB) = r_0(AB) - c \log n_{AB} \quad (3)$$

where c is a constant (which is taken to be 0.7 Å). This has been used in several systems[10–12] to show that the sum of the bond numbers at a central atom is fairly constant despite considerable variations in individual lengths. As an extension of this, it is proposed here that the BNE can be deduced as:

BNE(A) = nearest integer to (sum of n_{Ax_i} − number of ligands) (4)

where the X_i are the ligands of atom A.

Equations (2) and (4) have been used to derive BNEs for atoms (at present confined to C and N) in structures on the DAT file. Bond numbers (n_{AB}s) are calculated for C–C, C–N, C–O, C–S, N–N, N–O, N–S connections in the Unimol table (using the covalent radii, r(A): C, 0.77; N, 0.70; O, 0.65; S, 1.02 Å). Bond numbers for other connections involving C or N are set to 1. The BNEs for all C and N atoms in the structure are then calculated. It is found that, for precise structures in which all the atoms are accurately determined, the sum in brackets in equation (4) (for C or N) is often close to integral (see Figure 1). In principle, therefore, for these structures, it is possible to determine the number of p-orbitals that each C and N contributes to multiple bonding. For isolated multiple bonds and for many delocalized systems this works well, but, particularly where planar N is involved, BNEs are often higher than the expected integer. This can give poor results for many heterocyclic compounds.

Use of geometry to calculate BNEs

As an additional check, therefore, the valence angles around the central atom are investigated. The bond angle for two-coordinate C or N is calculated, and a value of 150° is used to distinguish between sp and sp^2 hybridization, if required. For 3-coordinate C and N, the improper torsion angle (L_1)–(C,N)–(L_2)–(L_3) is calculated, and a value of 155° used to decide whether the coordination is planar or pyramidal. The BNE is then converted to an integer if it falls in certain ranges with certain coordination geometries (Table 2). Otherwise it is flagged as unidentifiable (BNE = −99).

For example, this combination of bond and angle tests when applied to the geometry of pyrrole[14] gives a BNE of 0.8 for the N, which can clearly be classified as sp^2 hybridized. Moreover, the sum of the BNEs for the five ring atoms (5.3) is a good indication of the aromatic character.

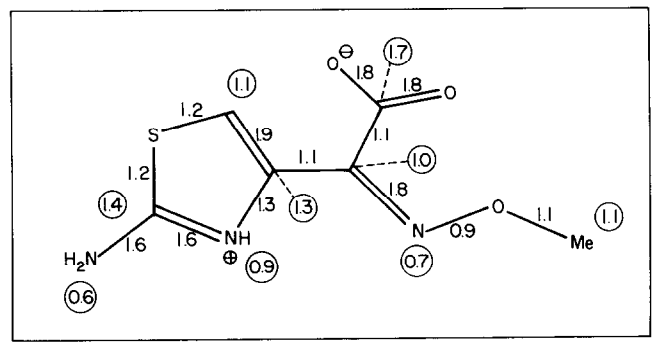

Figure 1. 2-(2-amino-4-thiazolyl)2(Z)-(methoxy)-imino acetic acid, a recent light-atom structure, where R = 0.038 and σ(C–C) = ca. 0.004 Å (i.e. AS on DAT would be 1)[13]. The BNEs for all the C and N atoms and the n_{AB}s for bonds to these atoms have been calculated. All C and N atoms including the NH_2 group, would be labelled as T by the algorithm

Table 2. BNE coordination geometries

	BNE range		
Coordination	**0–0.5**	**0.5–1.8**	>1.8
>4	−99	−99	−99
4	0	−99	−99
3 (nonplanar)	0	−99	−99
3 (planar)	−99	1	−99
2 (bent)	0	1	−99
2 (planar)	−99	−99	2
1	0	1	2

Table 3. Symbols for bonds with bond numbers (n_{AB}) in certain ranges

n_{AB}	*Symbol*
<0.75*	:
1.0 ± 0.25	–
1.5 ± 0.25	*
2.0 ± 0.25	=
2.5 ± 0.25	&
>2.75	#
indeterminate†	.
cyclic	!

* If n_{AB} is very small, r(AB) may be longer than the limit r(A) + r(B) + TOL set up in Geom and the connection will not be made.
† Default for all bonds whose number is not calculated.
‡ This symbol precedes the symbol for the bond number if the bond is cyclic. No symbol is printed for acyclic bonds. Thus a bond in an aromatic ring might be: TC!*TC.

For most structures on DAT, chemical labels can therefore be given to bonds and atoms from geometrical considerations alone. A code is suggested (Tables 1 and 3) for labelling C and N atoms and for bonds involving them. If the connection table is output with these labels, it is usually possible to reconstruct the chemical formulas of the residues fairly easily. The main exceptions are when hydrogen atoms are missing, when groups are disordered, where they show very high thermal motion (e.g. in steroid side chains it can be impossible to locate double bonds from the geometry), or when there are large experimental errors.
These hybridization labels (Q, T and B) can then be used to qualify atoms in the Frag search (see Appendix 3). Bond numbers can already be tested with the Test Dis facility.

In many structures, hydrogen atoms are totally or partially omitted from the DAT file. This does not affect the BNEs of the atoms to which they should have been attached. This is clearly seen by comparing, for example R_3N and R_3NH+, which both have the same BNE (i.e. 0) and which might be indistinguishable in structures of low accuracy.

Substituents on fragments

When a fragment is retrieved by Geom, it often has one or more substituent groups. It is often desirable to know what these are, since they may have large steric or electronic effects, or even totally change the chemical nature of the fragment. Thus, if ether groups, defined as a C–O–C fragment, are being retrieved, one will also get esters, anhydrides, ketals, aminals and other groups that may be considered chemically distinct from ethers. With large, infrequently occurring fragments, it is possible to deduce the substituents by inspecting the Pluto plots and/or the diagrams of chemical connectivity that will soon be on file. But there are still many cases where it is valuable to have an automatic tabulation of the substituents on a fragment, and a method has been developed for this which, though not catering for every situation, is fairly powerful.

All bonds and atoms are labelled (by the methods and symbols used in the previous section. If a fragment is then retrieved, its substituent atoms are all identified together with the type of the connecting bond; the substituent atoms and bonds (if any) of these substituents are also identified. They are sorted, compared with a dictionary of common substituent groups and listed (see Table 4). As a result of experimental errors and different amounts of conjugation, there is not always a unique label for a given functional group. The most common labels are given in Table 4. The dictionary often maps several labels at one synonym.

The substituent tree could, in principle, be taken to a depth greater than two, but FORTRAN 77 (in which Geostat is written) does not easily allow recursion, and the output becomes very complicated. For some groups, the number of substituents can be quite high (e.g. at a depth of two, there are 12 atoms in the tree of substituents of P in PPh_4^+).

Most common organic functional groups can be identified unequivocally at this level. Some of the most common examples are given in Appendix 4, where, with a little practice, it is easy to interpret the notation. For some of the commonest groups, a dictionary of synonyms has been prepared.

For a few specialized types of substituent, such as side chains in amino acids, the level of two branches is not adequate (the 20 amino acids fall into 11 classes). Even this can be valuable, since, with the graphic for-

Table 4. Examples of some labels produced by Geostat and their shortened dictionary representation

Label output by Geostat	Dictionary synonym	Functional group
—TC(=O*O)	—COO(X)	acid/ester/anion
–TC(*O*O)	–COO–	carboxylate anion
–TC(–QC=O)	–COR	aliphatic acyl (ketone)
–TC(*TN=O)	–CON(X)	amide
–TC(!*TC!*TC)	–AR	aryl
–QC(!–QC!–QC)	–CY–R	cycloalkyl
–QC()	–CH3	methyl
–QC(–O)	–CH20(X)	hydroxymethyl
–BC(#BC)		acetylene
–TC(=TN)		imine
–BC(#BN)	–CN	nitrile/cyanide
–BC(#O)		(metal)carbonyl
–TN(*O*O)	–NO2	nitro
–TN(*O)		nitroso
=TN(*O)		oxime
=TN(–C)		imine
–QN(–O)		hydroxylamine
–TN(=TC)	–N=C	imine
–QN(–C–C)		amine(aliphatic)
–TN(*TC)		amide or aromatic amine
–TN(=TN)	–N=N(X)	azo
–BN(#BN)		diazo
=BN(=BN)		diazo
=O	=O	carbonyl
–O	–O(H)	hydroxyl(or o–)
–O(–QC)	–OR	alkoxy
–S(.S)		disulfide
–S(.O.O–C)		sulfone
–S(.O–C)		sulfoxide
–S(.O.O.N)	–SO2N(R)	sulfamide
–S(.O.O.O)	–SO3(R)	sulfonate

Hydrogen atoms have not been considered in this table

mula of the whole molecule, it is then usually possible to tell which fragment has been retrieved.

Treatment of hydrogen atoms

The standard code in Geom78 includes the E routine (although this is not explicitly shown in the Geom78 description, its format is identical to that in Connser). It operates on the full connection table (not the Unimol connectivity), so that if this table has been expanded beyond Unimol there may be unpredictable results for many atoms. Unlike Connser, no distinction is made between H and other atoms, and since H atoms are often omitted from structures on the DAT file, the E option will not always be easy to use. One solution to this (Appendix 1(b) (i)) is to include a card with Excl H, which will effectively erase all H atoms from the calculation (i.e. as if they had never been on file). This has the disadvantage that H atoms cannot be used in any other operation of Geom. An operation has therefore been included (see Appendices 1 and 4) that excludes H atoms from consideration only when the connectivity of a particular atom in Frag is being analysed (additionally only the Unimol connectivity is considered, connections introduced by Intra or Inter being ignored).

APPLICATIONS

The facilities described here are all compatible with each other and with standard Geom78 routines. The keywords A, C, E and XH bring the Frag routine closer to the power and accuracy of Connser, and the rejection facility (some or all of which will soon be available in Connser) will also increase the accuracy of searches. The label can also be used for several purposes, its great advantage being that its output is machine-readable.

An obvious example is the determination of geometrical substituent constants[15]. The accurate geometry of benzene rings can be represented by a linear combination of the effects of all the substituents on the ring. With several thousand benzene rings on the DAT file, it is prohibitively time-consuming to examine each ring geometry output by Geom. Instead, the substituents can be deduced from the first two levels of the Label tree, and this is normally enough to interpret the distortions of ring geometry. These Labels, although sometimes rather cumbersome, are unique, and can be automatically read into a subsequent program for statistical analysis.

The Label facility can be used for the automatic identification of a base attached to a sugar ring. In Figure 2, Frag has been used to search for the substructure (bold lines) that is part of all ribonucleosides. seven key atoms are Labeled (numbers in circles), and the substituents will almost alwauys identify the base and type of ring substitution. The light lines indicate a fragment consistent with the labels given below; uncircled numbers represent the normal numbering.
bering.

Adenyl derivatives will give:

(1) !*TC(!*TN)
(2) –N()
(3) !*TN(!*TC)
(4) !*TN(!*TC)

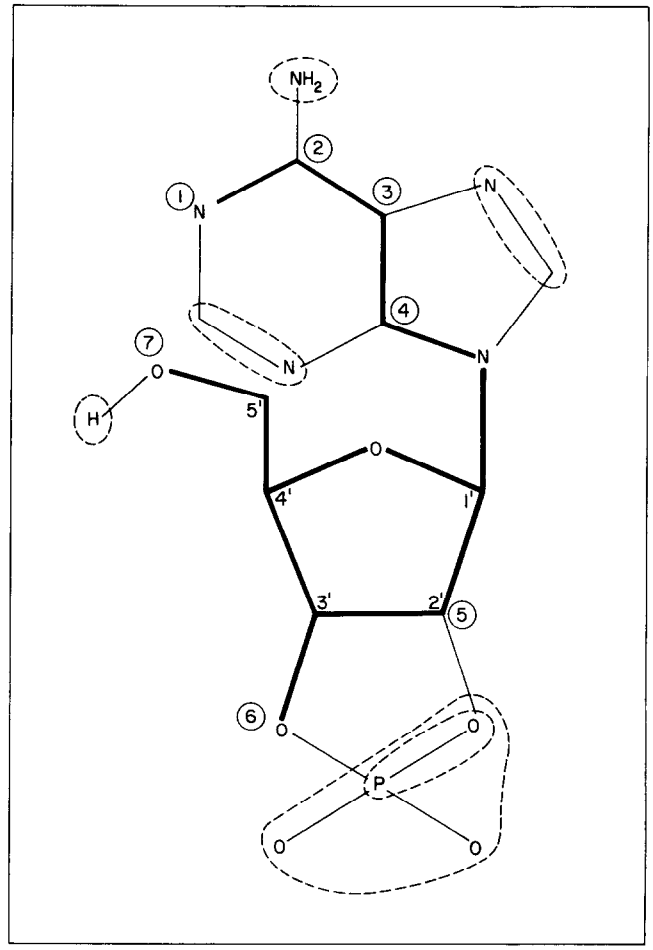

Figure 2. Substructure of ribonucleosides within a fragment

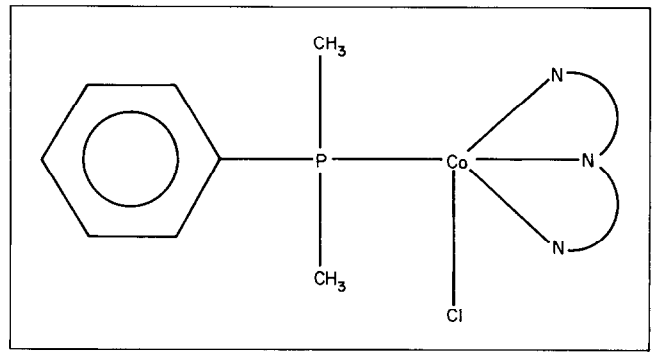

Figure 3. Structure represented by the substituents of a coordinated P atom

Common bases (e.g. A,G,U,T,C) can all be differentiated by the labels 1–4. Metal coordination or protonation (if H atoms are explicitly considered) will also be revealed in labels 1,3 and 4.

A ribose 2′,3′-cyclic phosphate will give:

(5) !.O(!.P)
(6) !.P(.O.O!.O)

An unsubstitued 5′ OH will give:

(7) *?*?(if H not located) or –H

The substituents of a coordinated P atom:

–AR;–CH_3;CH_3;–CO(!–TN!–TN!–TN–CL)

represent the structure in figure 3. Note that the ligand must be tridentate (or the Co–N bonds could not all be cyclic) and that the Ns are sp^2 hybridized (e.g. as in terpyridyl).

Other examples of Label might be: the nature of a metal atom and its attached ligands coordinated to a ligand defined in Frag, or the substitution pattern of a central metal atom. The labels can be grouped automatically into classes by a statistical package, and used as a powerful set of selection and rejection criteria. Label can also be used with Inter, and this is potentially extremely powerful.

CONCLUSION

The number and percentage of structures on DAT for which all atoms have been given accurate parameters is rapidly increasing. With light-atom structures, the accuracy is now high enough for essentially all the structures to give reliable BNEs and n_{AB}s. Thus, 29 structures have been reported[16] with no atom heavier than Cl. All of them had H coordinates reported, 12 had σ (C–C) <0.005 Å (AS = 1 on DAT), 12 had 0.005 Å <σ (C–C) <0.01 Å (AS = 2 on DAT) and only five had 0.01 Å <σ (C–C) <0.02 Å (AS = 3 on DAT). All would give easily interpretable BNEs.

The number of structures on DAT will almost certainly exceed 100 000 by the end of the decade, and automatic methods of analysing geometry will clearly be necessary. The potential techniques of analysis are so varied that, as has already happened, Geom will be repeatedly modified to meet the requirements of particular problems. It is already a very large program, and it is hoped that different versions of it do not proliferate, causing confusion in the community of users. Users that have successfully tested enhancements to Geom/Geostat are strongly urged to collaborate with the CCDC in updating the program. The modifications described in this paper are therefore available through CCDC and not from the authors.

ACKNOWLEDGEMENT

James Raftery wishes to thank the SERC for a post doctoral fellowship.

APPENDIX 1. FRAGMENT RETRIEVAL PROBLEMS

Several problems exist in the retrieval of acyclic–$(CH_2)_4$– fragments. The following examples describe various types of Connser and Frag searches. The inputs will work on the standard Geom except where stated on the comment (C) records.

(a) Connser inputs

(i) Acyclic tetramethylenes

```
Q ACYCLIC TETRAMETHYLENES
AT1 C 1 2,3
AT2 C 2 2 E
AT3 C 2 2 E
AT4 C 1 2 E
BO 1 2 1
BO 2 3 1
BO 3 4 1
ALLBOND ACYCLIC
C THIS WILL RETRIEVE ALL NORMAL BUTYL GROUPS
AND
C ALL 1, 4 DISUBSTITUTED N-BUTANES WHICH ARE NOT
IN A RING
END
```

(ii) Cyclic tetramethylene groups

```
Q CYCLIC TETRAMETHYLENE GROUPS
C THESE ARE RETRIEVED SO THAT ANY REFCODE IN
C THIS CATEGORY CAN BE EXCLUDED FROM ANALYSIS
AT1 C 2 2 E
AT2 C 2 2 E
AT3 C 2 2 E
AT4 C 2 2 E
BO 1 2 1
BO 2 3 1
BO 3 4 1
ALLBOND CYCLIC
END
```

(b) Alternative Geom inputs

In the following examples only the Frag section is shown.

(i) Tetramethylenes search excluding hydrogens

```
FRAG TETRAMETHYLENES
C HYDROGENS ARE EXCLUDED FROM THE SEARCH
C ANY COMPOUND WITH SUBSTITUTION OTHER THAN 4-
OR 1,4
C WILL BE EXCLUDED
C THIS FACILITY, THOUGH NOT IN THE PRESENT WRITE-
UP
C WORKS AS IN CONNSER
C THE EXCLUSION OF HYDROGENS MEANS THAT THEY
WILL NOT BE
C COUNTED AS SUBSTITUENTS
EXCL H
AT1 C
AT2 C 2 E
AT3 C 2 E
AT4 C 2 E
BO 1 2
BO 2 3
BO 3 4
TEST DIS 1 2 1.46,1.7
TEST DIS 2 3 1.46,1.7
TEST DIS 3 4 1.46,1.7
C THESE TESTS MAY EXCLUDE SOME VERY INACCU-
RATE STRUCTURES
TEST TORS 1 2 3 4 −90, 90
C THIS EXCLUDES TRANS CONFORMATIONS
C CYCLIC STRUCTURES WILL STILL CONTAMINATE
C SO WILL STRUCTURES WITH TRISUBSTITUTED AT1
C THE DATA
END
```

(ii) 1,4-disubstituted butane retrieval

```
FRAG TETRAMETHYLENES
C THIS WILL ONLY RETRIEVE 1,4-DISUBSTITUTED
BUTANES
C BUT CANNOT DISTINGUISH CYCLIC STRUCTURES
C ONLY COMPOUNDS WHERE ALL HYDROGENS HAVE
BEEN LOCATED WILL
C BE RETRIEVED
AT1 C
AT2 C
AT3 C
AT4 C
AT5 H
AT6 H
AT7 H
AT8 H
AT9 H
AT10 H
AT11 H
AT12 H
BO 1 2
BO 2 3
```

```
BO 3 4
BO 1 5
BO 1 6
BO 2 7
BO 2 8
BO 3 9
BO 3 10
BO 4 11
BO 4 12
C TRISUBSTITUTED AT1 OR AT4 ARE NOT POSSIBLE
TEST TORS 1 2 3 4 -90,90
END
```

(iii) Special keyword search

```
FRAG TETRAMETHYLENES
C ***NOT*** AVAILABLE IN STANDARD GEOM
C THE KEYWORD XH STRIPS H ATOMS OFF SPECIFIED
ATOM
C ONLY. THE KEYWORD EXCL H WOULD EXCLUDE ALL
C H ATOMS. THE KEYWORD E WORKS NOW ONLY FOR
C NON-HYDROGEN SUBSTITUENTS.
AT1 C 2 XH E
AT2 C 2 XH E
AT3 C 2 XH E
AT4 C
C AT4 MAY HAVE 1 OR 2 SUBSTITUENTS
C NOW BONDS ARE TESTED FOR CYCLIC PROPERTIES
C ON GEOMETRIC GROUNDS, I.E. ON THE UNIMOL
C CONNECTIVITY
BO 1 2 A
BO 2 3 A
BO 3 4 A
TEST DIS 1 2 1.46,1.7
TEST DIS 2 3 1.46,1.7
TEST DIS 3 4 1.46,1.7
TEST TORS 1 2 4 -90,90
C LABEL SUBSTITUENTS ON AT1 AND AT4
C THIS CAN BE USED FOR AUTOMATIC REJECTION OF
C TRISUBSTITUTED AT1, AND THE LABELS MAY
C BE USED FOR DIVISION OF THE DATA SET INTO
GROUPS
LABEL 1 4
END
```

APPENDIX 2. PATTERN REJECTION

It is possible to reject unwanted substitution patterns in a fragment. The following example retrieves ethers and esters, but not aminals or ketals.

```
FRAG ETHERS/ESTERS BUT NOT KETALS OR AMINALS
C THE BASIC DESIRED FRAGMENT IS DEFINED
AT1 C
AT2 O
AT3 C
C NOW POSSIBLE ADDITIONAL SUBSTITUENTS ARE
FLAGGED
AT4 N, O N
BO 1 2
BO 2 3
C IF AT1-AT3 HAVE BEEN FOUND THE BASIC FRAGMENT
C HAS BEEN LOCATED. NOW IT IS TESTED TO
C SEE IF IT IS PART OF AN UNWANTED SUPERFRAGMENT
BO 3 4
TEST DIST 3 4 1.3,1.6
C IS THERE AN ADDITIONAL C-N OR C,-O BOND WITHIN
C RANGE 1.3-1.6
C IF THERE IS, THE FRAGMENT IS REJECTED IF DIST IS IN
C THIS RANGE THEN
C FRAG IS PART OF
C A KETAL/ACETAL/AMINAL SUPERFRAGMENT AND IS
C REJECTED
C ESTERS WILL NOT SATISFY THIS TEST AND SO WILL
C BE INCLUDED
END
```

APPENDIX 3. USE OF HYBRIDIZATION CODE IN FRAG SEARCHES

The example retrieves ketones where both ligands of the carbonyl are sp^3 hybridized. The coding for AT2 bypasses the TEST DIS and speeds up the search.

```
FRAG SATURATED KETONES
C THIS IS ***NOT*** STANDARD GEOM
AT1 C Q
C THIS IS AN SP3 CARBON
AT2 C T 3 E
AT3 C Q
BO 1 2
BO 2 3
END
```

APPENDIX 4. TREATMENT OF HYDROGEN ATOMS IN FRAG (GEOSTAT)

```
FRAG-CH2NH3+
AT1 N 4 E
C THIS SEARCHES FOR AN N ATOM WITH EXACTLY 4
C LIGANDS, HYDROGEN INCLUDED
AT2 C 2 XH E
C THIS LOOKS FOR A C WITH EXACTLY 2 NON-HYDRO-
GEN LIGANDS
C ***NOT*** STANDARD GEOM78
AT3 H
AT4 H
AT5 H
C NOTE THAT IF 'EXCLUSH' HAD BEEN USED
C CH2NH2 GROUPS WOULD HAVE BEEN RETRIEVED AS
WELL
BO 1 2
BO 1 3
BO 1 4
BO 1 5
END
```

APPENDIX 5.

Geom 78 has been modified to include the above features (and many others, some of which are described in Part VI[17]). THe resulting program Geostat, is written in FORTRAN 77 and is available through the CCDC.

REFERENCES

1 **Kennard, O, Watson, D G and Town, W G** *J. Chem. Doc.* Vol 13 (1973) pp 14–19
2 **Allen, F H et al** *J Chem. Doc.* Vol 13 (1973), pp 119–123
3 **Allen, F H, et al.** *J Appl. Crystallogr.* Vol 7 (1974) pp 73–78
4 **Allen, F H, et al.** *J Chem. Doc.* Vol 13 (1973) pp 211–218
5 **Allen, F H** *Acta Crystallogr. B* Vol B37 (1981) pp 890–900
6 *Cambridge Crystallographic Data Centre User Manual* (1978) 2nd Ed
7 **Murray-Rust, P** *in Molecular Structure and Biological Activity* **(Eds: Duax, W L and Griffen, J)** Elsevier, Netherlands
8 **Tarjan, R** *SIAM J Comput.* Vol 1 (1973) pp 146–160
9 **Pauling, L** *J Am. Chem. Soc.* Vol 69 (1947) pp 542–553
10 **Burgi, H B** *Inorg. Chem.* Vol 12, (1973) 2321–2325

11 **Murray-Rust, P, Burgi, H B, and Dunitz, J D** *J. Am. Chem. Soc.* Vol 97 (1975) 921–922
12 **Dunitz, J D** (1978) *X-Ray Analysis and the Structure of Organic Molecules-Ithaca* Cornell University Press, USA (1978)
13 **Laurent, G, and Durant, F** *Cryst. Struct. Comm.* Vol 10 (1981) pp 1 015–1 023
14 **Chiang, J F** and **Kratus, M T** *Acta Crystallogr. A* Vol A28, S 206
15 **Domenicano, A and Murray-Rust, P** *Tetra. Lett.* (1979) pp 2283–2286
16 *Acta Crystallogr. B* Vol B37 No 9 (1981)
17 **Murray-Rust, P and Raftery, J** *J.* No 2 (June 1985) *Mol. Graph.* Vol 3