Documentation of Chemical Reactions. IV. Further Applications of WLN Analysis Programs: A System for Automatic Generation and Retrieval of Information on Chemical Compounds (AGRICC)

M. OSINGA *

Gist-Brocades N.V. Research & Development, P.O. Box 523, Haarlem, The Netherlands

A. A. VERRIJN STUART

Instituut voor Toegepaste Wiskunde en Informatica, Rijksuniversiteit, Leiden, The Netherlands

Received April 25, 1977

A description is given of the AGRICC (Automatic Generation and Retrieval of Information on Chemical Compounds) system, a set of computer programs with the following functions: registration, derivation of a bond table, consistency checks on the WLN used as input, calculation of a fragmentation code, derivation of a graphical representation, and calculation of physico-chemical properties. The fragmentation code contains 2146 fragments. The result is presented in the form of a KWOC index. The fragments are also used for the calculation of physico-chemical properties, e.g., NMR spectra according to Zürcher, lipophilicity, molecular refraction, molecular volume, and parachor.

INTRODUCTION

In a previous article of this series[1] a computer program was described which derives a bond table from Wiswesser Line Notation (WLN). The atoms were represented as "Direct Environment Annotating Numbers" (DEANs). This was done to compare the structures of starting materials and end products and provide a basis for encoding the reactions in a faceted classification. The bond tables are also useful for the documentation of further information, especially on starting materials and end products.

The WLN of a compound implicitly carries all structural information. The programs make this information explicit in the form of a bond table. However, the structure of this table does not provide an easily perceptibile visual image. The programs therefore were combined with additional routines that make the explicit information more accessible to users, forming a system that was called AGRICC, Automatic Generation and Retrieval of Information on Chemical Compounds. AGRICC deals with the following six aspects:

I. Registration
II. Derivation of a connection table (bond table)
III. Consistency checks
IV. A fragmentation code
V. A graphical representation
VI. Calculations of physico-chemical properties

In Figure 1 the successive steps involved in generating explicit information and its human accessible forms by the programs from WLN are shown.

The following parts of this article start off with an overview of the system as a whole. Therefore two aspects are dealt with in more detail, viz., the fragmentation code and the calculation of physico-chemical properties.

OVERVIEW OF THE SYSTEM

I. Registration. Most chemical laboratories have different sources of information on chemical compounds. These include records of compounds synthesized in the laboratory, present in store and in manufacturer's catalogues, and of collections of spectra and other physico-chemical properties. If information from all these sources is added to a central file, this would require multiple referencing of the same compound. If this occurs often, it would not be economic to process the corresponding WLN more than once. Therefore we developed a registration system, in which the WLN is registered a single time, together with all of its source indications. This makes it possible to produce a cross-index for the source indications.

A registry number is necessary for linking the different files. As usual, our registry number contains a check digit, and the compound information is allowed to be entered into the system only if no error was found by the consistency checks (see section III). Because our computer, an IBM 1130 with two disk drives, has only limited data storage capacity, much attention has been paid to a storage system which is efficient in terms of space rather than in terms of processing time. Consequently, three WLN symbols are stored in one 16-bit computer word. The file in which the notation is stored is determined by the number of words occupied by the notation. The record length of these files is equal to the length of the notation, so no space is lost. The registry number is based on file number and record number within the file. Thus no separate storage of the number is necessary.

Although in a bigger computer the necessity of using as little space as possible is less stringent, even then an increase of the storage efficiency by a factor of 10 would be important.

II. Derivation of the Bond Tables. A description of the programs performing this has been published.[1]

III. Consistency Checks. Since the WLN code contains redundancy, it is possible to perform consistency checks. Mistakes in coding a compound into a WLN can be of two types.

A. Mistakes that produce correct WLNs, but of another compound. Examples are the coding of a methyl instead of an ethyl, the coding of a benzene ring rather than a pyridine ring, wrong position of substitution, etc. Some of these mistakes can be checked by molecular formula comparison. In our system this is not done by the program, because it requires much effort to produce the man-calculated molecular formula. However, as the molecular formula is calculated by our program, comparison with the man-calculated molecular formula is possible, and automatic comparison could be implemented in the system easily. For other mistakes of this type, visual inspection of the computer-generated graphical representation is necessary.

B. Mistakes that produce incorrect WLNs. In principle these can all be detected by a suitable computer program.

0095-2338/78/1618-0026$01.00/0

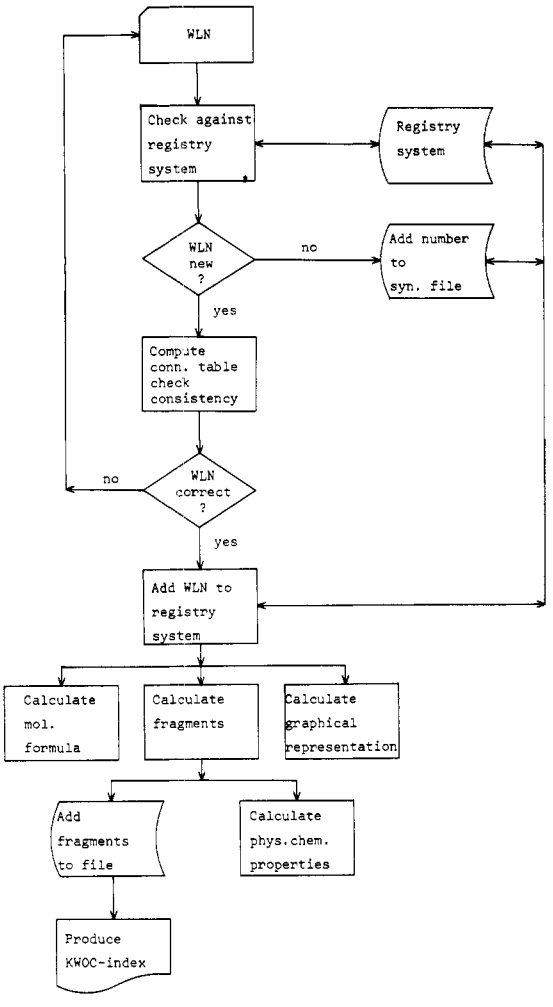


Figure 1. Flow of data in the AGRICC system.

However, in practice completeness is almost impossible owing to the enormous number of possibilities. Therefore, a selection has to be made. In this respect two approaches are possible.

1. The program is intended to help coders to code complicated compounds. This is actually not a way of checking a notation, but rather a way of preventing mistakes. Of course, it could be used routinely as well, but usually such a program takes a lot of computer time. Therefore the decision whether or not to use the program is made by the coder. A program of this type is PATHFINDER.[2-4] It helps in finding the locant path in complicated ring systems.

2. The program is intended to do routinely less complicated checks on WLNs. This last principle is applied in our system. It bears some resemblance to WLKEN.[3]

Examples of mistakes that can be discovered by the program are given in Figure 2. A further explanation of Figure 2 is given in the part "molecular formula and molecular weight calculations". Checks included are simple ones in aliphatic groups, such as the place of "H", the use of the U symbol, the elements between hyphens, etc. A check on the set of WLN rules that describe the coding of branched aliphatic compounds (rules 6, 7, 8, and 2) is also included. The rule numbers mentioned here refer to Smith.[5] The intermediate results of these checks are useful in solving crowding problems when deriving the graphical representation. Other checks deal with rings: the locant path is verified in some systems and the number of rings is checked against the number of saturation symbols, if present. Checks regarding the occurrence of ring elements as well as the order of citation are also included.

The last series of example deals with rule 39. In this case the intermediate results are also useful for solving crowding problems for the graphical representation. Just for reference a few correct WLNs are included in Figure 2. As can be seen, some of the errors were not only detected, but also corrected. In general, the correction of mistakes needs a much more complicated program than detection. In our program correction is possible only in simple cases.

Table I

Designation	Representation	Designation	Representation
Single bond	·	CH_3	C3
Double bond	=	CH_2	C2
Aromatic bond	+	NH_2	N2
Triple bond	*	Aromatic C atom	C4

IV. The Fragmentation Code. One purpose of developing our fragmentation code was to enable the resulting codes to be used by the scientists at the bench. This can be accomplished in one of two ways: by a terminal or by using a computer printout. When no terminals are available computer listing have to be used. This influences the selection of the fragments. A fragmentation code that has to fulfill this purpose must include a great number of rather selective fragments, because visual searching does not allow simultaneous references to many search terms.

Another purpose served by the fragmentation code was the calculation of physical-chemical properties. All the fragments necessary for these calculations were included in the fragmentation code. They are generally quite small and are present in many compounds (not very selective). This will be discussed in more detail in part V below. In the literature no indication was found of any previous fragmentation code based on this principle.

The requirements following from these two purposes are somewhat contradictory. As a consequence, the code system contains simple as well as complicated fragments. Some of the latter may contain one or more of the former. The simple fragments (such as methyl group) will hardly be used as primary search terms in the list searching, but they can be useful for specifying a question within the primary search term. The fragments found are expressed as numbers.

V. The Graphical Representations. The graphical representation of a molecule is certainly the most familiar one to a chemist. Therefore, it is very useful to include in an information system the possibility of generating a graphical representation as output from the system. However, a computer with standard output devices cannot approach the clarity of the chemist's drawings. For this reason, in our program the concessions in Table I were made. A bond is considered aromatic if double bonds can be drawn in two directions in a ring, starting from the same atom. This occurs, e.g., in benzene rings. The representation contains only one symbol for each type of bond. This does not mean that the program does not distinguish between different directions of single bonds. In fact, it recognizes eight of them.

Crowding of the graphical representation is diminished by using the consistency tables built during the checking phase. Examples of the structures output by our system are shown in Figures 3 and 4.

VI. Calculation of Physico-Chemical Properties. These days, much time and effort is devoted to the calculation of physico-chemical properties on the basis of the chemical structure. As mentioned, one of the purposes for which the fragmentation code was developed is the possibility of calculating such properties. With respect to the influence of the various atoms, two possibilities can be distinguished.

A. The property is carried by one, or a few closely connected atoms. If there are many such atoms, multiple values are generated creating a kind of spectrum. Examples of these properties are: UV, IR, and NMR. spectra, etc.; degree of

	MOL.FORMULA	MOL.WEIGHT	
NO AMPERSAND OR SPACE FOLLOWING UM IN ZYUMQ		EX 01	
NON-ACCEPTABLE POLYVALENT ATOM IN 2-S-2G2		EX 02	
MISTAKE IN UNUSUAL ELEMENT IN Q-PH-		EX 03	
BENZENE RING FOLLOWED BY UNACCEPTABLE SYMBOL QRBO1		EX 04	
INCORRECTLY PLACED H FOUND IN HPHR		EX 05	
INCORRECTLY PLACED H FOUND IN EX 06 CORRECTED TO VH1VQ AND FURTHER PROCESSED			
VH1VQ	C 3H 4 O3 0 0 0 0 0 0 0	88.061	88.016037 EX 06
INCORRECTLY PLACED H FOUND IN EX 07 CORRECTED TO SH2Q AND FURTHER PROCESSED			
SH2Q	C 2H 6 O1 S1 0 0 0 0 0 0	78.132	78.013931 EX 07
RULE 2 NOT OBSERVED IN QR DZ		EX 08	
LONGEST CHAIN WAS NOT SELECTED IN G1X1R DF&R DF&R DG		EX 09	
LONGEST CHAIN WAS NOT SELECTED IN FR BF CF D1G E1Q FE		EX 10	
LONGEST CHAIN WAS NOT SELECTED IN QR BYR DO1&R DZ		EX 11	
LONGEST CHAIN WAS NOT SELECTED IN 1OVY-LI-&2OR		EX 12	
INVALID SYMBOL FOUND IN T6NY B1 CQ		EX 13	
MISTAKE IN RING SIZE ORDER IN L E6 B665TJ		EX 14	
SUM OF LOCANTS NOT CORRECT IN L E6 D665TJ		EX 15	
SUM OF LOCANTS NOT CORRECT IN L E5 D666TJ		EX 16	
SUM OF LOCANTS NOT CORRECT IN L C656 BHJ		EX 17	
MISTAKE IN RING SIZE ORDER IN T65 BNVVN GHJ		EX 18	
MISTAKE IN RING SIZE ORDER IN L-16--14-J		EX 19	
T I5 F6 D6 C665 JO VOJ	C22H12 O2 0 0 0 0 0 0 0 0	308.335	308.083741 EX 20
ALPHABETIC ORDER OF LOCANTS IS NOT LOWEST T I5 G6 D6 B665 JO VOJ		EX 21	
NUMBER OF SATURATION SIGNS NOT CONSISTENT WITH THE NUMBER OF RING NUMERALS IN L E5 B666T&TJ		EX 22	
NUMBER OF SATURATION SIGNS NOT CONSISTENT WITH THE NUMBER OF RING NUMERALS IN T C6666T&TJ		EX 23	
CONSECUTIVE LOCANTS CITED INCORRECTLY T66 AN BVTJ		EX 24	
LOCANTS IN WRONG ORDER IN T66 CO AMTJ		EX 25	
INCORRECTLY PLACED O-SYMBOL IN T56 AOJ		EX 26	
INCORRECTLY PLACED Y-SYMBOL IN L56 AYJ		EX 27	
INCORRECTLY PLACED V-SYMBOL IN L E5 B666 IVTJ		EX 28	
INCORRECTLY PLACED O-SYMBOL IN L E5 B666 MOTJ		EX 29	
INVALID RING ELEMENT IN T66 BGTJ		EX 30	
RING ATOM WITH ONE ATTACHMENT IN L E5 B666TJ FR H1		EX 31	
INVALID QUATERNARY CARBON ATOM IN T E5 B666TJ A1 CR DF DF		EX 32	
RING SYSTEM OF MOST RING NUMERALS NOT FIRST IN T7N CN ES AUTJ C1- CL66J		EX 33	
START WITH WRONG RING SYSTEM BY ALPHABETIC ORDER IN T6NJ B1- BT6N CNJ		EX 34	
T5N CNJ B1- AT5NJ	C 8H 8 N3 0 0 0 0 0 0 0	146.173	146.071838 EX 35
T5SJ C1- BT5MJ	C 9H 9 N1 S1 0 0 0 0 0 0	163.241	163.045563 EX 36
START WITH WRONG RING SYSTEM BY ALPHABETIC ORDER IN T5MJ C1- BT5SJ		EX 37	

Figure 2.

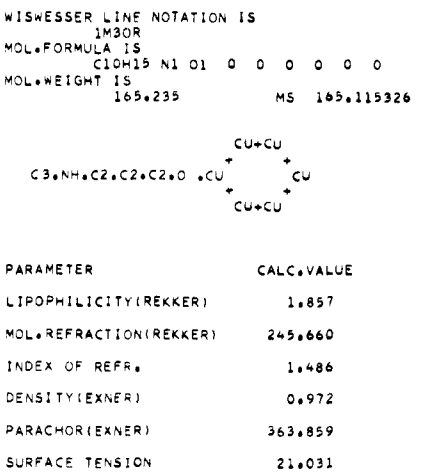


Figure 3.

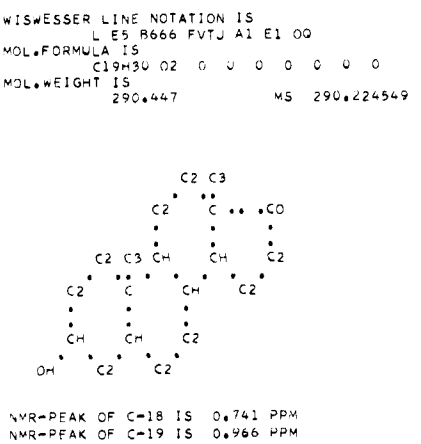


Figure 4.

dissociation. The influence of other atoms is strongly dependent on their distance to the atom under investigation. We will call these "local properties".

B. The property is carried by the molecule as a whole. Examples of these are molecular weight, partition coefficient, molar refraction, molar volume, parachor, etc. Here the influence is not dependent on the distance between atoms. This does not imply that groups cannot exert influence on each other. We will call these the "systemic properties". These can be calculated by assigning a value to each fragment, and then adding the values together. More details on these calculations can be found in the last part of this paper.

Some properties do not belong to either group, (e.g., melting point, boiling point, etc.). In general, these are far more difficult to calculate with a certain degree of accuracy than the local and the systematic properties.

DETAILS OF THE FRAGMENTATION CODE

The numbers used in the fragmentation code as mentioned in the system overview are derived in two steps. The first step is the calculation of the "computer numbers", a nearly closed set also used for calculation of the physico-chemical properties. In the second step these numbers are translated by the computer into the "user numbers", which from a hierarchically arranged set. This is done to make it easier for the user to find the appropriate number, and to permit the searching of a group of numbers. In the following, "number" will refer to "user number".

At this moment our fragmentation code contains 2146 fragments, but it can easily be expanded, and probably will be. The fragmentation code consists of two parts: the general code and special codes.

Table II

Central atom(s)	Code range	No. of codes assigned
1 carbon atom	10 001–13 000	160
>1 carbon atom	13 001–14 000	87
Nitrogen	14 001–16 000	82
Oxygen	16 001–17 000	22
Sulfur	17 001–18 000	43
Phosphorus	18 001–19 000	26
Halogens	19 001–19 800	16
Other atoms	19 801–20 000	16

I. The General Code. The general code is subdivided into three parts: fragments designating basic rings, fragments designating basic atoms, and fragments designating types of condensation.

a. Basic Ring Fragments. Basic rings are the rings coded by a number in the WLN code. Isolated benzene rings are also included as basic rings. The basic rings are separated in condensed and isolated rings. A first subdivision is made in carboxylic and heterocyclic rings. A further subdivision is made according to the ring size and the type of heteroatoms. Some of the isolated rings, such as benzene rings, are subdivided even further. The numbers from 1 to 10 000 are reserved for these fragments. Of these 192 numbers have been assigned.

b. Basic Atomic Fragments. These are fragments describing one atom within its environment, in effect, a somewhat enlarged DEAN.[1] The atom to which the DEAN belongs is called its "central atom". In general, each DEAN produces one fragment, but some do not, such as the hydroxy part of a carboxylic acid group.

The basic atomic fragments are divided into two groups, depending on whether their central atom is aliphatic or forms part of an alicyclic or heterocyclic ring. All fragments are calculated, but in the latter group not all are added to the system, because some do not have much information value and would only overload the system.

The fragments are divided according to the central atom:

1. Carbon atoms. These are further subdivided in two classes:

1A. Fragments consisting of one carbon atom. Sometimes the environment is to some degree included in the fragment, e.g., whether a fragment is attached to a saturated or an unsaturated carbon atom.

1B. Fragments comprising two or more carbon atoms

2. Nitrogen
3. Oxygen
4. Sulfur
5. Phosphorus
6. Halogens
7. Other atoms. This includes hydrogen atoms attached to atoms other than C, N, O, or S.

The code range and the number of codes assigned to these fragment groups are represented in Table II.

A problem arises sometimes when double bonds can be drawn in two directions in rings comprising heteroatoms. In this case it is sometimes possible to derive different fragments, depending on the way the double bonds are drawn. The program determines whether a double bond can be drawn in different directions starting from the same atom. If so, a fragment is chosen with a preference for single bonds. This applies especially to the two-atom fragments of type 1B. An example is given in Figure 5.

c. Fragments Describing Condensed Ring Systems. Two subgroups are distinguished.

1. Fragments designating the number of rings and their sizes. No indication is given of the heteroatomic character of the rings. The total number of rings apart from the isolated

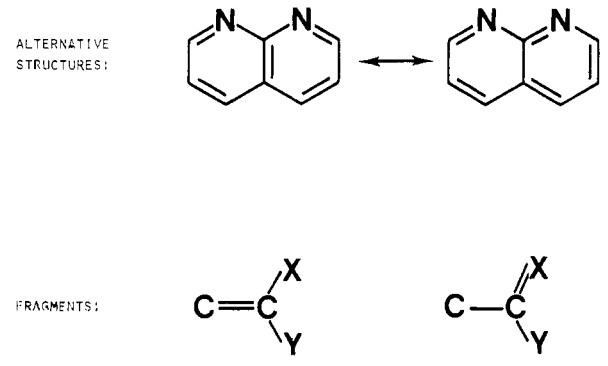
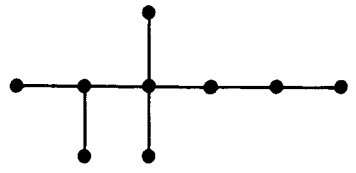
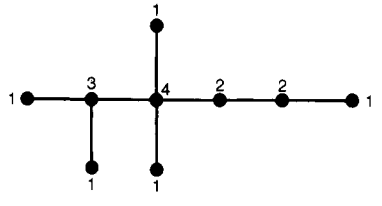


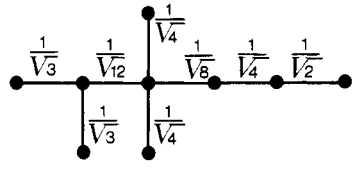
Figure 5.



ASSIGN TO ALL ATOMS, IRRESPECTIVE OF TYPE, A NUMBER THAT REPRESENTS THE NUMBER OF BONDS NOT LEADING TO HYDROGEN.



ASSIGN TO ALL BONDS A NUMBER BY MULTIPLYING THE NUMBERS FOUND IN THE PREVIOUS STEP FOR THE ADJACENT ATOMS, AND THEN TAKING THE REVERSE OF THE SQUARE ROOT OF THIS PRODUCT.



CALCULATE THE SUM OF THESE NUMBERS TO OBTAIN THE CONNECTIVITY INDEX. IN THIS EXAMPLE THE CONNECTIVITY INDEX $X = 4.004$

Figure 6.

benzene rings and the number of isolated benzene rings are also coded. The numbers 20 001–24 000 are reserved for these fragments. Of these 71 numbers have been assigned.

2. Fragments indicating the way the rings are connected, e.g., spiro, bridge, etc. These fragments are derived from the WLN during the analysis. The numbers 24 001–25 000 are reserved for these fragments. Of these, 6 numbers have been assigned.

II. Special Codes. These are codes, used only for coding compounds with special characteristics. Their absence or presence is determined by topological comparison. The most important of these codes is the steroid code. Only the usual 5,6,6,6 ring system is coded as a steroid; thus no homo- or azasteroids are included. However, extra rings may be condensed to the ring system.

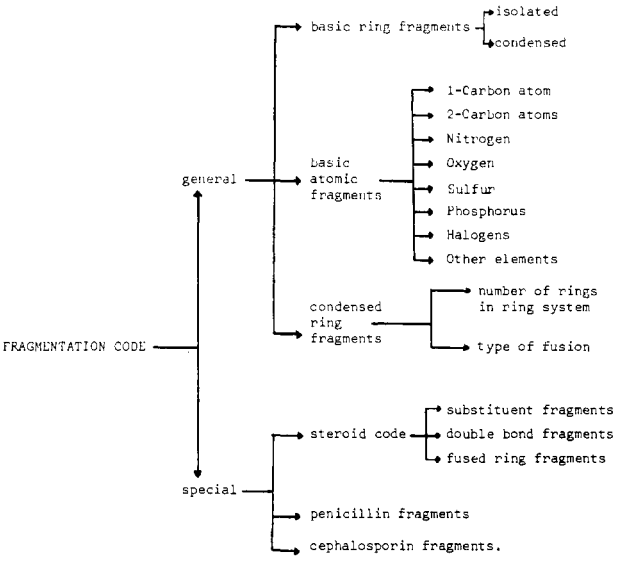
In this code the type of substituent, the place of substitution on the standard steroid skeleton, and the stereochemistry are reflected in one number. For this purpose 81 numbers per position are available. Of these four imply a double bond between the substituent and the skeleton. These cannot be given to the six points of fusion, thus producing 1353 fragments for all the substituents. Apart from this, there are 20 fragments indicating double bonds in the standard steroid skeleton, and 57 fragments for rings fused to it. The standard features of the steroid skeleton are not coded again in the general code. Thus the fragment of 5,6,6,6 ring system of the general code is used only for nonsteroids. The numbers 25 001–30 000 have been reserved for the steroid fragments. Of these 1429 numbers have been assigned.

Other special codes are the penicillin and the cephalosporin codes. For the penicillin code the numbers 30 001–31 000 are reserved. Of these 17 numbers have been assigned. For the cephalosporin code the numbers 31 001–32 000 are reserved. Of these, 26 numbers have been assigned.

Further Processing of the Fragments. The fragment numbers from one compound form a record, which is added to a file. From this file a KWOC index is produced. In order to save space on the disk, a second file containing the rotated versions of the records is not produced. Such a file would be at least ten times as big as the original file. Therefore the rotated versions are produced at the moment the KWOC index is printed.

Apart from using a computer listing for retrieval purposes it is also possible to query the file of fragments via a computer program. This program accepts normal Boolean logic, but it is also possible to specify a range of numbers. For example, 14 001/16 000 indicates all numbers from 14 001 to 16 000 that satisfy the condition, i.e., all nitrogen functions. The fragmentation code may be summarized as shown in Scheme I.

Scheme I



DETAILS OF THE PHYSICO-CHEMICAL PROPERTY CALCULATION

In the system overview a distinction was made between local and systematic properties. An example of a calculation of a local property is the work of Zürcher,[6,7] who developed a system for the calculation of the C-18 and C-19 NMR peaks in the spectra of steroids. Many other publications in this field are known.[8,9] Other examples include work leading to the Hammett and related constants.

The difference between a systemic and a local property may be illustrated by Table III, in which the C-19 parameter values of 5α-steroid ketones are given, as used in the Zürcher calculations.[9] For a systemic property there would be no difference between the values for the different positions.

Table III. C-19 Zürcher Values of Different 5α-Steroid Ketones[a]

Position	Shift, ppm	Position	Shift, ppm
1	0.375	7	0.275
2	−0.025	11	0.217
3	0.242	12	0.100
4	−0.033	15	0.800
6	−0.050	17	0.017

[a] For the 16-ketones no shift parameter was given.

We will now discuss in more detail the calculation of the properties included in our program.

a. NMR Spectrometry Calculations. As mentioned, it is possible to calculate the NMR shifts of the C-18 and the C-19 methyl groups of steroids more or less accurately. This work was initiated by Shoolery and Rogers.[10] The calculation is possible owing to the rigid skeleton of the steroid, which fixes the distance between a substituent and the methyl group, as well as their angle. Different types of steroids, however, have their own set of parameter values. NMR spectrometry calculations are included in our programs.

b. Molecular Formula and Molecular Weight Calculations. The molecular weight is a perfectly additive property. Apart from the molecular weight in the normal sense of the word, the molecular weight from mass spectrometry is also calculated. This is not based on the average weight of the isotopes of an atom, but on the weight of the most frequently occurring isotope of that atom. Additionally, a molecular weight is calculated based on the molecule with its addends. These addends are not included in the molecular formula. An example of type *a* calculations is given in Figure 3; calculations of the *b* type are found in Figures 2, 3, and 4. In Figure 2 the second number in the row "MOL.WEIGHT" represents the above-mentioned molecular weight based on the most abundant isotope. In Figure 3 this number is preceded by the letters "MS".

c. Lipophilicity. One of the most intensively studied properties is the partition coefficient of a compound. This is defined as the quotient of the solubilities of a compound in two theoretically immiscible solvents. One of the solvents is usually water. The property is considered to be related to the membrane-penetrating potential of that compound. The logarithm of this value, the lipophilicity, can be calculated more or less accurately from group contributions of the fragments of which the compound is composed.

In the literature two systems of assigning group contributions can be found. The system proposed by Fujita[11] is based on the equation:

$$\Pi(x) = \log P(\mathrm{SX}) - \log P(\mathrm{SH})$$

in which P(SH) is the partition coefficient of the structure SH, and P(SX) is that of the structure in which H is replaced by X. Π(X) is the so-called hydrophobic substitution constant of X. This infers that Π(H) is equal to zero. In principle, this is only true in relative calculations, but if one wants to calculate the lipophilicity of a whole structure, and not only the influence of a substituent, an absolute value of Π(H) has to be calculated. This was not done by Fujita, and therefore many correction terms are necessary.

Rekker et al.,[12] who developed the view on the value of Π(H) given above, calculated other group contributions, which were based on the principle that hydrogen has its own contribution to the lipophilicity. They found that the correction terms of Fujita were no longer necessary. To improve the predicting quality of their system new correction terms were introduced. These were largely based on the proximity of electron-attracting groups.

An aspect that did not receive the attention it deserves in the literature is the selection of the fragments. (It is also applicable for the properties mentioned in the rest of this article.) In all systems, for example, $-CH_2-$ always gets the same group contribution. However, it is not very likely that $C-CH_2-C$ contributes exactly the same as $C-CH_2-X$ and $X-CH_2-X$ (in which X represents a polar group). For $X-CH_2-X$ Rekker introduced a proximity correction. Therefore, it is unlikely that $C-CH_2-X$ should have a contribution equal to that of $C-CH_2-C$. This aspect should become apparent especially when two polar groups are neighbors. However, if different types of $-CH_2-$ fragments, etc., are introduced in the calculations, the equations necessary to calculate the group contributions become interdependent. This means that extra information is necessary to solve them. It might well be derived from charge distribution information. The molecular volume or its surface should be taken into account as well. This aspect, however, lies outside the scope of this paper.

d. Molecular Refraction. The molecular refraction is the product of the index of refraction and the molecular weight. This property is also additive. Two ways are possible for using this additivity: (1) by determining atomic refractions, and (2) by determining bond refractions. Since the first approach requires a considerable number of correction terms, preference was given to the latter. The fragment parameters were determined based on the tables of Vogel,[13] together with the corrections of Rekker.[14]

e. Molecular Volume. By assuming that a fragment of a molecule, e.g., a methyl or a hydroxy group, has the same volume wherever it is placed in the molecule a "partial volume" can be determined for each of the fragments. The sum of these partial volumes of a molecule gives the molecular volume which can be calculated for a compound. In our system the values of Exner[15] were used. The molecular volume is rather dependent on the temperature. Exner calculated the partial volumes at 20 °C. From the molecular volume the density can easily be calculated, which is included in our programs.

f. The Parachor. The parachor, introduced by Sugden,[16] is defined as:

$$P = [M/(D_1 - D_v)]\sqrt[4]{\gamma}$$

in which P = parachor, M = molecular weight, D_1 = density in liquid phase, D_v = density in vapor phase, and γ = surface tension. The density of the vapor is usually neglected, giving the formula:

$$P = (M/D)\sqrt[4]{\gamma}$$

in which D represents the density of the liquid phase. The parachor in an additive property which was supposed to be temperature independent

In our program we use the fragment values as calculated by Exner.[17] Exner states that the parachor is not completely temperature independent, especially for polar compounds.

Apart from the physico-chemical properties mentioned, it would be simple to calculate the connectivity index of Kier et al.[18] This connectivity index is calculated as follows. Each nonhydrogen atom is given a number, equal to the number of nonhydrogen atoms attached to this atom. Subsequently a value, which could be called the bond connectivity, is assigned to each bond, equal to the square root of the product of the numbers assigned to the atoms adjacent to the bond. The connectivity index is the sum of all the bond connectivities; an illustration is given in Figure 6.

It is clear, that the DEANs easily provide all the information necessary for the calculation of the connectivity index. However, the present authors are of the opinion that its value is not as great as Kier et al. suggest. The reason for this opinion can be illustrated by the third article by Kier's group,[19] in which a description is given of the correlation between the connectivity index and the lipophilicity. This article contains

a table giving the connectivity index, the experimental lipophilicity, and the calculated lipophilicity. However, the table was not accompanied by a derivation of how the calculation was made. A careful study of the footnotes reveals that for each functional group a separate function is necessary, because the functional group influences the intercept of the straight line representing the equation. Thus, instead of a straightforward calculation for a new functional group, a multiple regression analysis of derivatives with this functional group has to be made to find the intercept. The fact that this is not clearly stated in the article is reason for us to doubt the usefulness of the connectivity index.

Many other kinds of properties might be calculated. One of these would be the distances of the atoms in a molecule. It would be possible to derive parameters needed for quantum mechanical calculations. We are working at the moment on an extension to combine the fragments using pattern recognition techniques.

LITERATURE CITED

(1) M. Osinga and A. A. Verrijn Stuart, "Documentation of Chemical Reactions. II. Analysis of the Wiswesser Line Notation", *J. Chem. Doc.*, **14**, 194–98 (1974).
(2) C. M. Bowman, F. A. Landee, N. W. Lee, and M. H. Reslock, "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Line Notation of Complex Polycyclic Structures", *J. Chem. Doc.*, **8**, 133–8 (1968).
(3) T. Ebe and A. Zamora, "Wiswesser Line Notation Processing at Chemical Abstracts Service", *J. Chem. Inf. Comput. Sci.*, **16**, 33–5 (1976).
(4) A. Zamora and T. Ebe, "Pathfinder II. A Computer Program That Generates Wiswesser Line Notation for Complex Polycyclic Structures", *J. Chem. Inf. Comput. Sci.*, **16**, 36–39 (1976).
(5) E. G. Smith, "The Wiswesser Line-Formula Chemical Notation", McGraw-Hill, New York, N.Y., 1968.
(6) R. F. Zürcher, "171. Protonenresonanzspektroskopie und Steroidstruktur. I. Das C-19-methylsignal in Funktion der Substituenten", *Helv. Chim. Acta*, **44**, 1380–95 (1961).
(7) R. F. Zürcher, "232. Protonenresonanzspektroskopie and Steroidstruktur. II. Die lage der C-18 and C-19-methylsignale in Abhängigkeit von den Substituenten am Steroidgerust", *Helv. Chim. Acta*, **46**, 2054–88 (1963).
(8) J. E. Page, "Nuclear Magnetic Resonance Spectra of Steroids", *Annu. Rep. NMR Spectrosc.*, **3**, 149–210 (1970).
(9) N. S. Bhacca and D. H. Williams, "Application of NMR Spectroscopy in Organic Chemistry", Holden-Day San Francisco, Calif., 1964.
(10) J. N. Shoolery and M. T. Rogers, "Nuclear Magnetic Resonance Spectra of Steroids", *J. Am. Chem. Soc.*, **80**, 5121–351 (1958).
(11) T. Fujita, J. Twasa, and C. Hansch, "A New Constant Derived from Partition Coefficients", *J. Am. Chem. Soc.*, **86**, 5175–80 (1964).
(12) G. G. Nys and R. F. Rekker, "Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The Introduction of Hydrophobic Fragmental Constants (*f* Values)", *Chem. Therap.*, **5**, 521–35 (1973).
(13) A. I. Vogel, W. Greswell, G. Jeffery, and I. Leicester, "Bond Refractions and Bond Parachors", *Chem. Ind. (London)* 358 (1950).
(14) R. F. Rekker and D. J. Reiding, "Over de praktische bruikbaarheid van Vogel's moleculaire bindings-refracties", *Chem. Weekblad*, **58**, 513–19 (1962).
(15) O. Exner, "Additive Physcial Properties. II. Molar Volume as an Additive Property", *Collect. Czech. Chem. Commun.*, **32**, 1–23 (1967).
(16) S. Sugden, "A Relation between Surface Tension, Density and Chemical Composition", *J. Chem. Soc.*, 1177–89 (1924).
(17) O. Exner, "Additive Physical Properties. III. Re-examination of the Parachor", *Collect. Czech. Chem. Commun.*, **32**, 1–23 (1967).
(18) L. B. Kier, L. H. Hall, W. J. Murray, and M. Randić, "Molecular Connectivity. I. Relationship to Nonspecific Local Anesthesia", *J. Pharm. Sci.*, **64**, 1971–74 (1975).
(19) W. J. Murray, L. H. Hall, and L. B. Kier, "Molecular Connectivity. III. Relationship to Partition Coefficients", *J. Pharm. Sci.*, **64**, 1978–81 (1975).

Hash Functions for Rapid Storage and Retrieval of Chemical Structures

W. T. WIPKE,* S. KRISHNAN, and G. I. OUCHI

Board of Studies in Chemistry, University of California, Santa Cruz, California 95064

Received August 8, 1977

A method is described for determining if a given chemical structure or its enantiomer is contained within a file in time essentially independent of file size. The stereochemically extended Morgan algorithm (SEMA) name is used as a key for directly computing the address of the compound. Three separate files of compounds are used to study the effectiveness of four different hash functions. Various subsets of the SEMA name were also used as keys to study effect of information loss on hashing efficiency. A work function is used to compare the amount of work required to access a compound in the file.

INTRODUCTION

"Is this compound commercially available?" and "Is this a new compound?" are frequent questions in chemistry. To answer such questions the chemist generates a name for the compound and "searches" for an identical name in an ordered list of compound names. The names might be IUPAC or WLN[1] and they are normally ordered alphabetically. Even with this ordering the time required to determine if a compound is contained within a file of compounds increases as the size of the file increases. The time required for such a search becomes very important when the number of searches being performed is large.

The same questions arise within our Simulation and Evaluation of Chemical Synthesis (SECS) program.[2,3] When a synthetic precursor is generated, SECS must determine if the compound already exists in the synthesis tree or if it exists in a library of available starting materials or if it is a common precursor existing in other synthesis trees.[4] Since this determination must be made for every precursor and the files being searched may be large, we were interested in an efficient search method.

Hash table methods are well known for searching files with a search time that is independent of the number of records in the file.[5] When the number of records is large, a good hash table method is faster than a linear search method for which the search time is proportional to n or a binary search method whose search time is proportional to $\log_2 n$, where n is the number of records in the file.

The reason for the greater speed of hash table methods is that the hash function, h, operates directly on the "search key", K, to produce an address, $h(K)$, or position in the hash table where the requested entry must be if it exists in the file. Of course, the same hash function must be used for retrieval as was used for creating the file.

Occasionally two different keys will hash to the same address, $h(K) = h(K')$. This situation is called a *collision*. The colliding keys are attached to that entry in the table as a linear list. A linear search of that short list is required when

0095-2338/78/1618-0032$01.00/0