Documentation of Chemical Reactions. II. Analysis of the Wiswesser Line Notation

M. OSINGA and A. A. VERRIJN STUART*

Centraal Rekeninstituut der Rijksuniversiteit, Leiden, Holland

Received June 25, 1974

A description is given of the first step toward automatic encoding of chemical reactions, a conversion from a WLN to a kind of connection table. This table is organized in pairs of atoms, as connected by a bond. Instead of the usual element symbols, numbers are given which represent the bond environment of the atom. Some further applications of the system are given.

In a previous article[1] the automatic encoding of chemical reactions was described as the main purpose of our research. As the first part of this research a system was devised by which the reactions were to be classified. A *faceted classification* was considered to be the most suitable. The Wiswesser Line Notation (WLN) was chosen as the coding system for starting materials and end-products.

The next step that is necessary is the analysis of this WLN. One possible way of doing this is described in this paper. Clinging and Lynch[2] have tried to encode reactions without performing such an analysis; however, for more complicated reactions no satisfactory result could be obtained. The analysis of a WLN has been performed by Hyde and coworkers[3-6] and by Granito.[7,8] Such an analysis usually leads to a kind of connection table, which represents the connections of the different atoms in the molecule. In a connection table all connections are explicit, but no difference is made between the different environments of atoms with the same atomic symbol. As some of these differences are already expressed in the WLN, such a connection table was not considered to be completely satisfactory. Therefore another way of representing an atom was chosen.

A NEW REPRESENTATION OF AN ATOMIC ENVIRONMENT

In the first design of representing the environment, use was made of the fact that for each element two words of computer memory were reserved. Thus for a one-letter symbol there was one spare word, which was used for a subdivision.[9]

This soon proved to be too limited for all the aspects we wanted to include. Therefore, a completely numerical representation (of one computer word) was chosen, which will be referred to as the DEAN (Direct Environment Annotating Number). In a DEAN all bonds of an atom are indicated. The list of these is given in Table I.

It proved to be complicated to derive all of the bonds of some atoms in a one-step analysis from the WLN. Especially substituted ring-atoms and branch symbols were difficult to handle, because in WLN a substituent can be many positions away from the ring system or the branch symbol to which it is attached.

Instead of trying to do this in a one-step analysis, a two-step analysis was preferred. For the first step a new type of numerical representation was devised, which did indicate as much of the environment as could be derived in a simple way from the line notation. These representation numbers are called auxiliary numbers (AN's). They are given in Table II.

The computer program performs the "first level analysis" of the WLN of a compound and produces a table which contains these AN's. This is illustrated in Table III, left side, giving the AN's for the compound of Figure 1. From the result of the first level analysis the program continues to derive the DEAN's in the so-called "second level analysis." The result of this second level analysis is presented in a table, which contains the DEAN's (see Table III, right side).

Elements which do not need extensive subdivision, such as the halogens, do not get an AN but are given their DEAN during the first level analysis.

DETAILS OF THE ANALYSIS

A set of computer programs, performing the analysis of a WLN, has been written in Fortran IV, for an IBM 1130, with a memory of 8K 16 bits words and one disk drive.

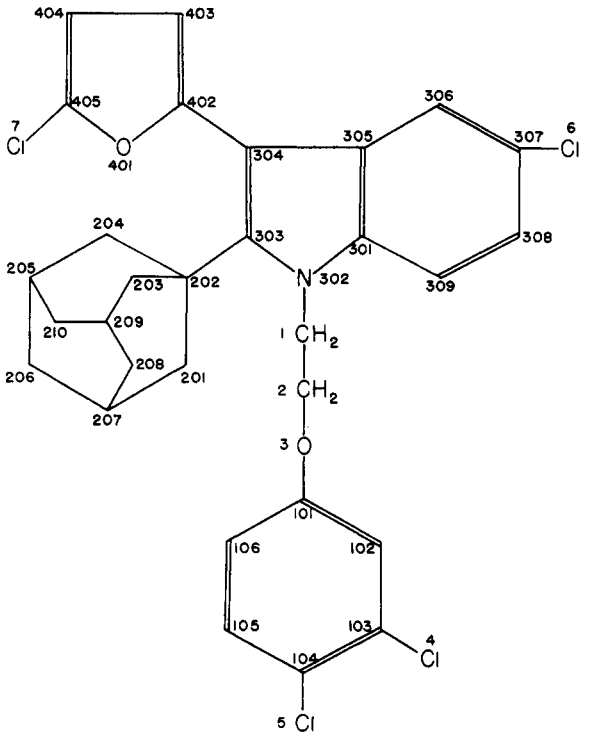
Owing to the core size the possibilities of the system are somewhat limited. These limits are: the total number of bonds should not exceed 100; the total number of rings that are allowed is dependent on their size. If none is bigger than 9 ring atoms, 20 rings are allowed in one system. If only one ring is present, its size could be 180, but this exceeds the total number of bonds allowed. The total number of bridges and multicyclic points may not exceed 10. No more than five hyphenated locants are allowed and no doubly hyphenated locants. Also only five pairs of cited nonconsecutive locants are allowed. These restrictions, although not serious, can easily be relaxed when a bigger computer is available.

The first program reads the punch cards containing the notation and writes them on disk. If a notation requires two consecutive punch cards, they are combined on one record, and the program also "demultiplies" the WLN multiplications. A set of rings attached to all possible positions of another ring leads to a kind of demultiplication which cannot be handled by this program. It has to be fed into the system in an "unmultiplied" form. Demultiplication is also restricted to one level of multiplication; two consecutive slashes cannot be handled either.

Demultiplication does not necessarily lead to the canonic notation, but as long as the correct tables are produced from it, it is not considered urgent to change this.

Since analysis of the addends is usually a waste of time, the addend in general is not included. This is performed by shifting the part of the notation representing the addend one position to the right; the next program stops the analysis after finding two consecutive blanks. Isotope designations can be excluded in the same way. The maximum length allowed for one notation is 100 characters. At the same time the indications of anionic, cationic, and radical character and stereochemistry are dealt with. They are given fixed positions from 101 onwards. The resulting notation is written on disk.

* To whom correspondence should be addressed.



L66 B6 A B- C 1B ITJ B- CT56 BNJ B2OR CG DG& GG D- BT5OJ EG

Converted version
L66 B6 A C D 1B JTJ B- CT56 BNJ B2OR CG DG& GG D- BT5OJ EG

Figure 1. Structure of discussed compound, with Wiswesser Line Notation and converted version thereof.

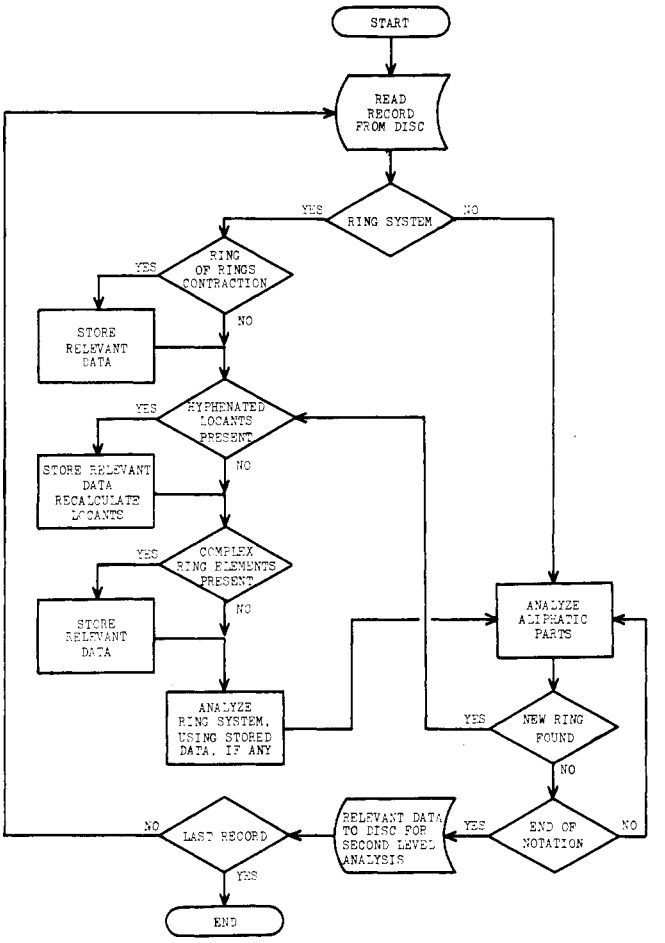


Figure 2. Flowchart of program performing first level analysis.

The main program performs the actual first level analysis. The logic is shown in Figure 2. The first step is to check whether ring-of-rings contraction is present. If so, the notation leads to a table in which the ring character of the "aliphatic" parts is not shown. If not, it is checked to see if hyphenated locants are present. If so, all locants, starting with the hyphenated one, shift one place in the alphabet. This is illustrated in Figure 1. In the meantime this fact is registered.

In the next step the complexities of the ring system (multicyclic points, bridges, cited nonconsecutive locant pairs, spiro atoms) are recorded. This preliminary registration is done in order to keep the subroutine, making the actual ring analysis as small as possible. Some problems were encountered owing to inconsistencies in the WLN rules, *e.g.*, the coding of aliphatic sulfur. It is not appropriate to discuss the actual analysis in more detail.

Apart from the detailed representation, for every atom of a compound a number is derived (H is numbered only in a few special cases). An atom can belong to one of three series: (1) the aliphatic atoms which are numbered from 1 to 99; (2) the benzene atoms which are numbered from 101 to 106 for the first ring, 111 to 116 for the second ring, etc.; (3) the atoms of other ring systems are numbered from 201 to 299 for the first ring, 301 to 399 for the second ring, etc.

An advantage of such a numbering system is the ability to draw conclusions regarding the position of an atom from its number. This will be discussed later on.

With this computer program, the atom representations are not given as an ordinary connection table, but as a list of bonds, which somewhat resemble the augmented pairs, etc., of Lynch and coworkers.[10–16] For the compound of Figure 1, the first and second level analysis is shown in Table III. The list consists of five vertical columns.

The following is indicated in these columns, from left to right: (1) the representation of the first atom of the bond, according to Table I or II; (2) the number of this atom in the compound (see Figure 1); (3) the bond order of the bond; (4) the representation of the second atom of the bond; (5) the number of this atom in the compound (see Figure 1). Of these, only the bond order has not yet been discussed. It is a number which indicates the saturation character of the bond. In aliphatic bonds, a 2 signifies a double bond, and a 3 a triple one. In N-oxides the bond between N and O is always indicated by 2. Fractions are not given.

In rings, double bonds are often delocalized, which makes it impossible to define which bond is single and which is double. Therefore the bond order of a bond in a ring is always represented as 1. Bond order 0 is used to designate a spiro atom. In this case, both ends of the bond represent the same atom.

APPLICATIONS

It can be seen from Table III that this representation has many possibilities in searching, especially for substructures. If one looks for a substructure, in which the ring elements are expressed as defined entities, the result of the first-level analysis is used. If one wants a substructure irrespective of the fact that the atoms form part of a ring, the results of the second level analysis are then more suitable.

Some conclusions can be drawn from the number, especially about the place of substitution. In the example (Figure 1, Table III), it can be seen that the chlorine atoms 4 and 5 are bound to C atoms 103 and 104. These must be adjacent atoms of a benzene ring. A third chlorine (6) is bound to C atom 307, the number indicating that it forms

Table I. List of Direct Environment Annotating Numbers (DEAN's)[a]

Second Level Subdivision of Carbon Atoms

I. C atoms, attached to one cited atom

100	CH_3—C
200	CH_3—X
300	CH_2=C
400	CH_2=X, X not oxygen
500	CH≡C
600	CH≡X
700	C=X, *e.g.*, in isonitriles

II. C atoms, attached to two cited atoms

1100	C—CH_2—C
1200	C—CH_2—X
1300	X—CH_2—Y
1400	C=CH—C
1500	C=CH—X if in ring, then also used for C—CH=X
1600	C—CH=X not in ring, X not O (see 61-6400)
1700	X—CH=Y
1800	C=C=C
1900	C=C=X, X may be O
2000	X=C=Y, X and Y may be O
2100	C—C≡C
2200	C—C≡X
2300	X—C≡C
2400	X—C≡Y

III. C atoms, attached to three cited atoms

3100	C—CH(—C)—C
3200	C—CH(—C)—X
3300	C—CH(—X)—Y
3400	X—CH(—Y)—Z
3500	C—C(=C)—C
3600	C—C(=C)—X, also for C=X isomer of ring
3700	X—C(=C)—Y, also for C=X isomer of ring
3900	C—C(=X)—Y, X not O not used if C of C=X in ring
4000	Z—C(=X)—Y, X not O

IV. C atoms, attached to four cited atoms

5100	C attached to C,C,C,C
5200	C attached to C,C,C,X
5300	C attached to C,C,Y,X
5400	C attached to C,Z,Y,X
5500	C attached to W,Z,Y,X

V. C atoms cited by V

6100	C—C(=O)—C
6200	C—C(=O)—X
6300	X—C(=O)—Y
6400	R—C(=O)—H, R may be H

Second Level of Nitrogen Atoms

N atoms, attached to one cited atom

10100	NH_2—C
10200	NH_2—X
10300	NH=C
10400	NH=X
10500	N≡C
10600	N≡X

N atoms, attached to two cited atoms

11100	C—N—C
11200	C—N—X
11300	X—N—Y
11400	C—N=C
11500	C—N=X
11600	X—N=C
11700	X—N=Y
11700	Other types, including the diazonium N

N atoms, attached to three cited atoms

12100	C—N(—C)—C
12200	X—N(—R)—S
12300	C—N(=C)—C
12400	X—N(=C)—R
12500	C—N(=X)—C
12600	Y—N(=X)—R
12700	C—NO_2
12800	X—NO_2
12900	Other types, including azoxy nitrogen

N atoms, attached to four cited atoms

13100	N attached to C,C,C,C
13200	N attached to C,C,C,X
13300	N attached to C,C,Y,X
13400	N attached to C,Z,Y,X
13500	N attached to W,Z,Y,X

Oxygen

14100	C—OH
14200	X—OH
14300	C—O—C
14400	R—O—X
14500	C=O if 6100–6400 is not applicable
14600	X=O
14700	O=X=O, if at least one other bond is attached to X
14800	Other types of oxygen

Sulfur

15100	C—SH
15200	X—SH
15300	C—S—C
15400	R—S—X
15500	C=S
15600	X=S
15700	C—S(=O)—C
15800	Other tetravalent S
15900	C—SO_2—C
16000	C—SO_2—X
16100	X—SO_2—Y
16200	Other hexavalent sulfur
16300	Other S

Phosphorus

17100	Trivalent P
17200	O=$P(X)_3$, all X's are oxygen functions
17300	O=$P(-C)_3$
17400	O=$P(R)_3$, not all R's may be carbon or oxygen
17500	Other pentavalent phosphorus
17600	Other phosphorus

Table I. *Continued*

Halogens

18100 F monovalent
18200 F polyvalent
18600 Chlorine monovalent
18700 Chlorine polyvalent
19100 Bromine monovalent
19200 Bromine polyvalent
19600 Iodine monovalent
19700 Iodine polyvalent

20000 H

Other elements get the following number 20000 + 100 times their place in the periodic system.

[a] Cited atoms are H atoms cited in the WLN plus all nonhydrogen atoms unless otherwise stated; X, Y, Z, W represent heteroatoms, but not H. R and S may be heteroatoms or carbon, but not H. Different letters do not necessarily imply different atoms.

Table II. List of Auxiliary Numbers, Used to Designate the Elements in the First Level of Coding (AN's)

Carbon atoms

The numbers 1–7 refer to aliphatic atoms

1 CH_3
2 CH_2
3 Wiswesser symbol Y
4 Wiswesser symbol X
5 Wiswesser symbol C
6 Wiswesser symbol V, except aldehydes
7 Wiswesser symbols VH, aldehyde

The numbers 11–20 refer to ring atoms

11 unsaturated, not branched
12 unsaturated, branched
13 saturated, not branched
14 saturated, branched
15 acetylene
16 allene
17 endo spiro
18 exo spiro
19 exocyclic double bond
20 ring V

Nitrogen atoms

10001 $Z(NH_2)$
10002 M(−NH−)
10003 N
10004 K (quat amine)

Oxygen atoms

14001 OH
14002 −O−
14003 =O
14004 W (dioxo)

Sulfur atoms

Sulfur atoms inside a ring system always get 15002

Aliphatic sulfur

15001 SH
15002 −S−
15003 =S
15004 tetravalent S
15005 hexavalent S
15006 other S

Phosphorus

17000 P

Table III. Results of the First and Second Level Analysis of the WLN of the Structure in Figure 1

First level					Second level				
13	201	1	14	202	1100	201	1	5100	202
14	202	1	13	204	5100	202	1	1100	204
13	204	1	14	205	1100	204	1	3100	205
14	205	1	13	206	3100	205	1	1100	206
13	206	1	14	207	1100	206	1	3100	207
14	207	1	13	201	3100	207	1	1100	201
13	203	1	14	202	1100	203	1	5100	202
14	207	1	13	208	3100	207	1	1100	208
13	208	1	14	209	1100	208	1	3100	209
14	209	1	13	203	3100	209	1	1100	203
14	209	1	13	210	3100	209	1	1100	210
13	210	1	14	205	1100	210	1	3100	205
14	202	1	11	303	5100	202	1	3600	303
12	301	1	10003	302	3600	301	1	12100	302
10003	302	1	11	303	12100	302	1	3600	303
11	303	1	11	304	3600	303	1	3500	304
11	304	1	12	305	3500	304	1	3500	305
12	305	1	12	301	3500	305	1	3600	301
12	305	1	11	306	3500	305	1	1400	306
11	306	1	11	307	1400	306	1	3600	307
11	307	1	11	308	3600	307	1	1400	308
11	308	1	11	309	1400	308	1	1400	309
11	309	1	12	301	1400	309	1	3600	301
10003	302	1	2	1	12100	302	1	1200	1
2	1	1	2	2	1200	1	1	1200	2
2	2	1	14002	3	1200	2	1	14300	3
14002	3	1	11	101	14300	3	1	3600	101
11	101	1	11	102	3600	101	1	1400	102
11	102	1	11	103	1400	102	1	3600	103
11	103	1	11	104	3600	103	1	3600	104
11	104	1	11	105	3600	104	1	1400	105
11	105	1	11	106	1400	105	1	1400	106
11	106	1	11	101	1400	106	1	3600	101
11	103	1	18600	4	3600	103	1	18600	4
11	104	1	18600	5	3600	104	1	18600	5
11	307	1	18600	6	3600	307	1	18600	6
11	304	1	11	402	3500	304	1	3600	402
14002	401	1	11	402	14300	401	1	3600	402
11	402	1	11	403	3600	402	1	1400	403
11	403	1	11	404	1400	403	1	1400	404
11	404	1	11	405	1400	404	1	3700	405
11	405	1	14002	401	3700	405	1	14300	401
11	405	1	18600	7	3700	405	1	18600	7

part of a second ring system. The last chlorine (7) is attached to C atom 405, the number indicating that it forms part of a third ring system.

The main problem in calculating a molecular formula from the WLN is associated with the number of hydrogen atoms.

It can be seen from Table I that if one has translated the WLN of a compound in a table using this kind of representation, the problem is easily overcome. In fact, it was more complicated to derive the wanted Hill-order of the elements from the order of the periodic system.

L66 B6 A B- C 1B ITJ B- CT56 BNJ B2OR CG DG& GG D- BT5OJ EG

Mol formula	Mol weight
C30H27CL4N1O2	575.362

Apart from the possibilities in substructure searching and molecular weight calculation based on WLN, the result of our analysis is also useful as the chemical base for the evaluation of structure/property and structure/activity relationships. At the moment we are calculating lipophilic constants based on the work of Nys and Rekker,[17] with the aid of DEAN's. The absence or presence of each individual DEAN of the second level representation could possibly be used as a bit screen.

Our further research will be directed toward comparison of the tables of the analysis of starting material and end products, and using the difference as a mean for automatic encoding or determination of the reactions.

ACKNOWLEDGMENT

The authors wish to express their gratitude to the management of Gist-Brocades Research for the use of their IBM-1130, and to Mr. G. J. Keijser and Mr. A. van der Woude for their valuable discussions concerning this work.

LITERATURE CITED

(1) Osinga, M., and Verrijn Stuart, A. A., "Documentation of Chemical Reactions. I. A Faceted Classification," *J. Chem. Doc.*, **13**, 36–39 (1973).
(2) Clinging, R., and Lynch, M. F., "Production of Printed Indexes of Chemical Reactions. I. Analysis of Functional Group Interconversions," *J. Chem. Doc.*, **13**, 98–102 (1973).
(3) Hyde, E., Matthews, F. W., Thomson, L. H., and Wiswesser, W. J., "Conversion of Wiswesser Notation to a Connectivity Matrix for Organic Compounds," *J. Chem. Doc.*, **7**, 200–204 (1967).
(4) Hyde, E., and Thomson, L. H., "Structure Display," *J. Chem. Doc.*, **8**, 138–146 (1968).
(5) Thomson, L. H., Hyde, E., and Matthews, F. W., "Organic Search and Display Using a Connectivity Matrix Derived from Wiswesser Notation," *J. Chem. Doc.*, **7**, 204–209 (1967).
(6) Campey, L. H., Hyde, E., and Jackson, A. R. H., "Interconversion of Chemical Structure Systems," *Chem. Brit.*, **6**, 427–430 (1970).
(7) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems," *J. Chem. Doc.*, **12**, 190–199 (1972).
(8) Granito, C. E., "CHEMTRAN and the Interconversion of Chemical Substructure Systems," *J. Chem. Doc.*, **13**, 72–74 (1973).
(9) Osinga, M., "Automatic Encoding of Chemical Reactions," Paper presented at the NATO/CNA Advanced Study Institute on Computer Representation and Manipulation of Chemical Information, Noordwijkerhout, June 4–15, 1973.
(10) Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M., "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files," *J. Chem. Doc.*, **13**, 133 (1973).
(11) Adamson, G. W., Creasey, S. E., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in the Common Data Base," *J. Chem. Doc.*, **13**, 158 (1973).
(12) Crowe, J. E., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part I. Noncyclic Fragments," *J. Chem. Soc. C*, 990 (1970).
(13) Adamson, G. W., Lynch, M. F., and Town, W. G., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part II. Atom-Centered Fragments," *J. Chem. Soc. C*, 3702 (1971).
(14) Adamson, G. W., Lambourne, D. L., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part III. Statistical Association of Fragment Incidence," *J. Chem. Soc., Perkin Trans. 1*, 2428–2433 (1972).
(15) Adamson, G. W., Cowell, J., Lynch, M. F., McLure, A. H. W., Town, W. G., and Yapp, A. M., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part IV. Cyclic Fragments," *J. Chem. Soc., Perkin Trans. 1*, 863 (1973).
(16) Adamson, G. W., Creasey, S. E., Eakins, J. P., and Lynch, M. F., "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part V. More Detailed Cyclic Fragments," *J. Chem. Soc., Perkin Trans. 1*, in press.
(17) Nys, G. G., and Rekker, R. F., "Statistical Analysis of a Series of Partition Coefficients with Special Reference to the Predictability of Folding of Drug Molecules. The Introduction of Hydrophobic Fragment Constants (*f* values)," *Chim. Therap.*, **5**, 521–35 (1973).

A Numerical Identifier for the Chemical Elements, Expressing Their Periodic Relationships†

KENNETH W. LOACH

Department of Chemistry, State University College of Arts and Science, Plattsburgh, New York 12901

Received September 10, 1974

Each chemical element can be assigned two numbers related to its periodic table position. The separate or packed numbers are effective identifiers of the element and its periodic relationships. A Fortran subroutine LEMENT has been written that rapidly converts an element symbol into the equivalent numerical identifiers.

Processing of chemical information usually requires the storage and retrieval of chemical element identities. The two common identifiers are the traditional element symbol and the atomic number. The symbol is familiar and the number is easily manipulated, but neither lend themselves readily to the systematic or automatic recognition of periodic relationships. Searches of either identifier for classes of elements (*e.g.*, any transition element, any group Va element) requires considerable additional processing and (usually) multiple tests of each element identifier.

There is a need for an element identifier that is compact, easily manipulated, and which allows the ready recognition of periodic relationships. This can be achieved by assigning

† Presented at the 168th National Meeting of the American Chemical Society, Division of Chemical Literature, Atlantic City, N.J., Sept. 11, 1974.