

- (2) (a) L. B. Kier, "Molecular Orbital Theory in Drug Research", Academic Press, New York, N.Y., 1971; (b) L. Farnell, W. G. Richards, and C. R. Ganellin, "Conformation of Histamine Derivatives. 5. Molecular Orbital Calculation of the H₁-Receptor "Essential" Conformation of Histamine", *J. Med. Chem.*, **18**, 662-666 (1975).
- (3) (a) R. C. Bingham, M. J. S. Dewar, and D. H. Lo, "Ground States of Molecules. XXV. MINDO/3. An Improved Version of the MINDO Semiempirical SCF-MO Method", *J. Am. Chem. Soc.*, **97**, 1285-1293 (1975); (b) M. J. Dewar, "Prediction of Properties and Behaviour of Materials", NTIS Report, AD-A003 698, 1974.
- (4) E. M. Engler, J. D. Andase, and P. v. R. Schleyer, "Critical Evaluation of Molecular Mechanics", *J. Am. Chem. Soc.*, **95**, 8005-8025 (1973).
- (5) O. Exner, in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Ed., Plenum Press, London, 1972, Chapter 1, p 1.
- (6) J. Hine, "Structural Effects on Equilibria in Organic Chemistry", Wiley, New York, N.Y., 1975, Chapter 3, p 55.
- (7) O. Exner, ref 5, p 46.
- (8) C. Hansch, in "Drug Design", Vol. 1, E. J. Ariens, Ed., Academic Press, New York, N.Y., 1971, Chapter 2, p 271.
- (9) G. J. Janz, "Thermodynamic Properties of Organic Compounds", Academic Press, London, 1967; (b) J. Hine, ref 6, Chapter 1, p 1.
- (10) (a) S. M. Free and J. W. Wilson, "A Mathematical Contribution to Structure-Activity Studies", *J. Med. Chem.*, **7**, 395-399 (1964); (b) P. J. Harrison, "A Method of Cluster Analysis and Some Applications", *J. Appl. Statistics*, **17**, 226-236 (1968).
- (11) J. R. Koskinen and B. R. Kowalski, "Structure-Reactivity Correlations for Organic Molecules by Pattern Recognition", NTIS Report AD-785 913, 1974.
- (12) (a) S. A. Hiller, V. E. Golender, A. B. Rosenblit, L. A. Rastrigin, and A. B. Glaz, "Cybernetic Methods of Drug Design. 1. Statement of the Problem - The Perceptron Approach", *Comp. Biomed. Res.*, **6**, 411-421 (1973); (b) B. R. Kowalski and C. F. Bender, "The Application of Pattern Recognition to Screening Prospective Anticancer Drugs. Adenocarcinoma 755 Biological Activity Test", *J. Am. Chem. Soc.*, **96**, 916-918 (1974); (c) R. D. Cramer, G. Redl, and C. E. Berkoff, "Substructural Analysis. A Novel Approach to the Problem of Drug Design", *J. Med. Chem.*, **17**, 533-535 (1974); (d) A. J. Stuper and P. C. Jurs, "Classification of Psychotropic Drugs as Sedatives or Tranquilizers Using Pattern Recognition Techniques", *J. Am. Chem. Soc.*, **97**, 182-187 (1975); (e) K. C. Chu, R. J. Feldman, M. B. Shapiro, G. F. Hazard, and R. I. Greran, "Pattern Recognition and Structure-Activity Relationship Studies", *J. Med. Chem.*, **18**, 539-545 (1975).
- (13) (a) G. W. Adamson and J. A. Bush, "Method for Relating the Structure and Properties of Chemical Compounds", *Nature, (London)*, **248**, 406-408 (1974); (b) "A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures", *J. Chem. Inf. Comput. Sci.*, **15**, 55-58 (1975); (c) "The Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics", *J. Chem. Soc. Perkin Trans. 1*, 168-172 (1976); (d) G. W. Adamson and D. Bawden, "A Method of Structure-Activity Correlation Using Wiswesser Line Notation", *J. Chem. Inf. Comput. Sci.*, **15**, 215-220 (1975).
- (14) (a) V. B. Bond, C. M. Bowman, N. L. Lee, D. R. Peterson, and M. H. Reslock, "Interactive Searching of a Structure and Biological Activity File", *J. Chem. Doc.*, **11**, 168-170 (1971); (b) E. Hyde, D. R. Lambourne, and L. A. McArdle, Abstracts of Papers, 163rd National Meeting of the American Chemical Society, Boston, Mass., April 1972; (c) C. Hansch, A. Leo, and D. Elkins, "Computerized Management of Structure-Activity Data. 1. Multivariate Analysis of Biological Data", *J. Chem. Doc.*, **14**, 57-61 (1974).
- (15) M. F. Lynch, J. M. Harrison, W. G. Town, and J. E. Ash, "Computer Handling of Chemical Structure Information", Macdonald-Elsevier, London-New York, 1971.
- (16) G. W. Adamson, S. E. Creasey, J. P. Eakins, and M. F. Lynch, "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*, 2071-2076 (1973).
- (17) G. M. Kellie and F. G. Riddell, "The von Auwers Boiling Point Rule. A New Approach", *J. Chem. Soc. Perkin Trans. 1*, 740-744 (1975).
- (18) (a) O. Exner, ref 5, p 41; (b) J. Hine, ref 6, p 16.
- (19) J. E. Dubois, J. J. Aaron, O. Alcais, J. P. Doucet, F. Rothenberg, and R. Ucan, "A Quantitative Study of Substituent Interactions in Aromatic Electrophilic Substitution. I. Bromination of Polysubstituted Benzenes", *J. Am. Chem. Soc.*, **94**, 6823-6828 (1972).
- (20) Statistical Analysis Mark II Applications Package, International Computers Limited Technical Publication 4301, London, 1971.
- (21) P. B. D. de la Mare and J. H. Ridd, "Aromatic Substitution", Butterworths, London, 1959.
- (22) B. R. Kowalski and C. F. Bender, "Pattern Recognition. A Powerful Approach to Interpreting Chemical Data", *J. Am. Chem. Soc.*, **94**, 5633-5639 (1972).
- (23) M. Sjostrom and S. Wold, "Statistical Analysis of the Hammett Equation. II. A Unified Inductive Sigma Scale", *Chem. Scripta*, **6**, 114-121 (1974).
- (24) J. Shorter in "Advances in Linear Free Energy Relationships", N. B. Chapman and J. Shorter, Ed., Plenum Press, London, 1972, Chapter 2, p 71.
- (25) J. E. Ash and E. Hyde, "Chemical Information Systems", Ellis Horwood, Chichester, 1975.

Documentation of Chemical Reactions. III. Encoding of the Facets

M. OSINGA* and A. A. VERRIJN STUART**

Gist-Brocades N. V., Research & Development, Haarlem, Holland
and Centraal Rekeninstituut der Rijksuniversiteit Leiden, Leiden, Holland

Received June 27, 1975

A computer program is described for the automatic encoding of chemical reactions into the following faceted classification: (1) the chain facet, (2) the ring facet, (3) the rearrangement facet, and (4) the unusual element facet. The main problem for the chain facet is to determine the actual change in starting material and end product. The first step in the solution is to compare the pairs of DEAN's of two bonded atoms. Sometimes this leads to inconsistent equivalences. It was necessary to develop tests to find them and guidelines to find the correct equivalences. When encoding the ring facet, the main problem encountered was the delocalization of double bonds. When the correct equivalences are established, the rearrangement facet and the unusual element facet do not present serious problems.

INTRODUCTION

The automatic encoding of chemical reactions is the ultimate purpose of our research. The reactions to be encoded make use of the faceted classification described before.¹ In this classification five facets are distinguished: (1) the chain facet (present in all reactions); (2) the ring facet; (3) the rearrangement facet; (4) unusual elements (elements other than C, H, O, N, S, P, F, Cl, Br, I); and (5) other facets (these are

not dealt with in the automatic analysis).

The Wiswesser Line Notation (WLN) was selected as the chemical coding system for the starting material and end-product. A computer program for analyzing a WLN has been described.² This analysis leads to a bond table, in which a bond and two atoms on each end of the bond is described, as illustrated in Figure 1. In the bond table the numbers on a horizontal line represent a "bond pair".

The type of atom, after a first level analysis, is expressed as an auxiliary number (see left part of table of Figure 1). From this auxiliary number the definitive DEAN or Direct

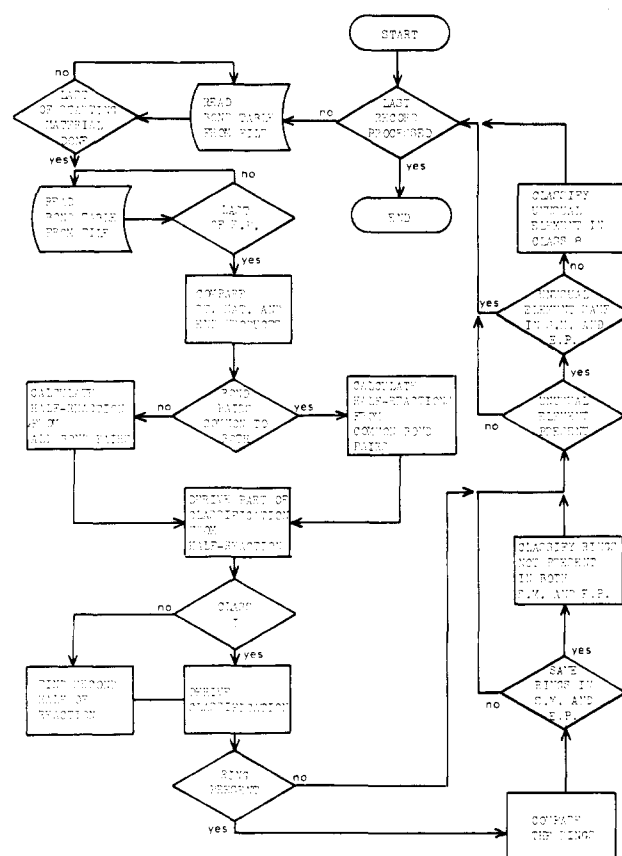
* Author to whom correspondence should be addressed at Gist-Brocades N.V.

** Centraal Rekeninstituut der Rijksuniversiteit Leiden.

BOND TABLE AFTER FIRST LEVEL					TRANSLATION	BOND TABLE AFTER SECOND LEVEL						
ANALYSIS					TABLE	ANALYSIS						
AUX.	TOP.	BOND	AUX.	TOP.	TOP.	BEAN	BEAN	TOP.	BOND	BEAN	TOP.	
NO.	NO.	VALUE	NO.	NO.	NO.			NO.	VALUE		NO.	
10002	201	1	13	202	1	14100		11100	201	1	1300	202
11	202	1	10003	203	2	18700		13	202	1	12100	203
10003	203	1	13	204		120	203	1200	203	1	1000	204
13	204	1	13	205	101	3600		1200	204	1	2000	205
13	205	1	10002	201	102	1400		1200	205	1	1100	201
10003	203	1	11	101	103	3600		12100	203	1	3600	101
11	101	1	11	102	104	3600		3600	101	1	2400	102
11	102	1	11	103	105	3600		2400	102	1	3600	103
11	103	1	11	104	106	3600		3600	103	1	3600	104
11	104	1	11	105	201	11100		3600	104	1	3300	105
11	105	1	11	106	202	1300		3500	105	1	1400	106
11	106	1	11	101	203	12100		1400	106	1	3600	101
11	101	1	10001	1	204	1000		3600	103	1	2400	0
11	101	1	18600	2	205	1200		3600	104	2	1800	2
11	105	1	1	3				3500	105	1	100	3

A chemical structure of pyridine, a six-membered aromatic heterocycle. The nitrogen atom is at the bottom position and is labeled with the number 1. Moving clockwise from the nitrogen, the carbon atoms are labeled 2, 3, 4, 5, and 6. The double bonds are located between atoms 2 and 3, 4 and 5, and 6 and 1.

In this article a computer program is described that performs the encoding of four of the facets, based on the bond tables



2. Only one of the atoms of the changing bond is described in both the starting material or end product. This will be referred to as "*partial description*". Hydrolysis may serve as an example of a partial description, if water is not encoded

as a starting material (as in reaction II).

If the chain facet of a reaction has to be determined by the computer program, the first thing that has to be established is which atoms are involved in the reaction. Of course it would be ideal to do an atom-by-atom comparison, but in practice this is not possible because of the enormous number of substructures that have to be compared. Therefore, the bond tables of starting material and end product are compared. Using DEAN's instead of atomic symbols reduces the probability of finding incorrect equivalences.

Every reaction causes a change in the way the molecule is numbered. Now two possibilities can be distinguished: (a) the change affects only the reacting part of the molecule; (b) the change affects other parts, or even the whole molecule. The latter could happen, if a ring closure changes the locant path.

In the first case there is no problem; in the second case problems might arise owing to the fact that two atoms are compared which have the same DEAN, but which are not identical from the point of view of the molecule as a whole. Therefore a check had to be developed, which determines whether such incorrect equivalences are present. This check is based on the principle that, if atom A is attached to atom B, atom B has to be attached to atom A. If the second attachment is not found inconsistency is concluded. If so, the ideal solution would again be an atom-by-atom comparison of all substructures present in starting material and end product, but this would be extremely time consuming. Therefore it is useful to try to find guidelines for these comparisons. These depend on the cause of the inconsistency. An illustration of the procedure will be given for a case of ring closure, which is an important source of inconsistencies.

The technique developed for finding the equivalent DEAN's is based on the equivalences found during the ring comparison. As will be described later, the first step in the ring comparison is the change of DEAN's to numbers not indicating anything outside the ring. Subsequently, these numbers are rotated and contrarotated until the equivalent atoms are recognized.

Using the results of this comparison it is necessary to keep in mind that the numbers can be equal, while the corresponding DEAN's are not. Therefore, the DEAN's of the ring found to be equivalent during the ring analysis have to be rotated anew. We have to take into account that there may be two or more rings present in both starting material and end product which are equal with respect to the ring facet, but not with respect to the DEAN's. In this case "cross-comparison" is necessary. Of course, this happens most often in systems with more than one benzene ring.

After having found a new possible set of equivalences, again a consistency check is needed to determine whether the new one is correct. Apart from the check already mentioned, it is also possible to use the newly formed ring as a check. The atoms it comprises should form one or more chains in the starting material. The problem is illustrated in Figure 4. In this case the sixth set of equivalences is considered to be the correct one.

If there are two or more starting materials or end products, problems might arise because the same topological number can be present in starting material and end product twice or even more often. In the computer program this is solved by adding 1000 for each extra compound of starting material or of the end product to the topological number of the atoms in that compound. With this, no atom in different compounds of the starting materials can have the same topological number, nor in the end product. Complications can arise due to the delocalization of double bonds (see, e.g., Figure 5). Both end products are identical, owing to the delocalization of the double bonds.

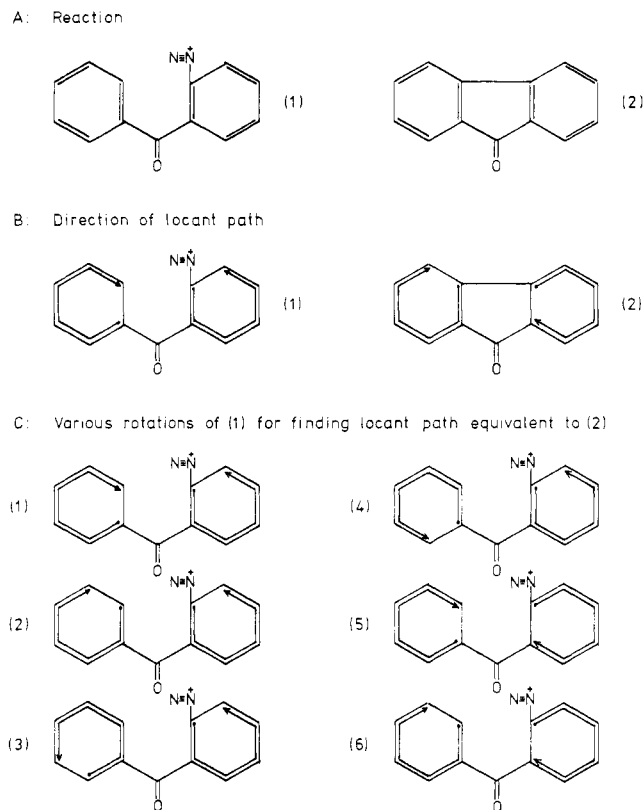


Figure 4.

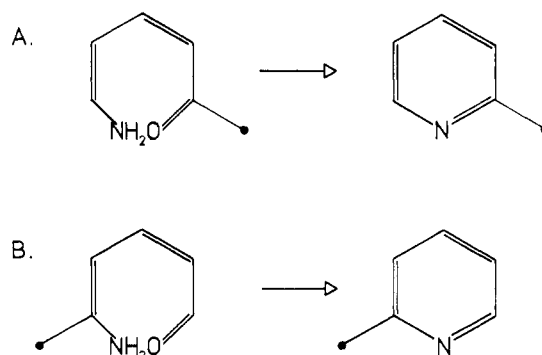





Figure 5. Due to delocalization, the end products of the two reactions are identical, contrary to the suggestion of the diagram.

As was stated before, the problem of the bond value was solved by defining the DEAN's equal to 1. However, this leads to another complication, because sheer comparison of the DEAN's would miss the equivalence between the aldehyde C=O group and the corresponding C=N of the ring. Therefore not only cyclic double bonds between a carbon and a heteroatom have to undergo a kind of enolization, but this has to be applied to aliphatic ones as well. Before the actual determination of the reaction, this has to be reversed.

When the computer program has established which bond pairs are common to starting material and end product (which represent the set of common DEAN pairs) and which are not common, those bond pairs are selected for which one DEAN belongs to the set of common DEAN's and the other does not. These bond pairs are involved in the reaction. A table is composed, consisting of the common DEAN, the DEAN present only in the starting material, and the DEAN present only in the end product, which replaces the former (Table I). Sometimes the last two represent the same atom, sometimes not. This is caused by the fact that e.g., exchange of a heteroatom for a carbon atom and the reverse does change the

Table I

REACTION	COMMON Pairs NUMBER	MEANING	DEAN OF STARTING MATERIAL NUMBER	MEANING	DEAN OF END PRODUCT NUMBER	MEANING
$R-CH_2OH \rightarrow R-CH_2Cl$	1200	(C)-CH ₂ -(X)	14200	(C)-CH	18600	(C)-(C)
	100	CH ₃ -(C)	1300	(C)-C=O (C)-C=O	3200	(C)-C-(X) (C)-C-(X)
	1100	(C)-CH ₂ -(C)	1400	(C)-CH=C(C) (C)-CH=C(C)	3200	(X) (C)-CH=C(C) (X) (C)-CH=C(C)
$CH_3CHO + CH_3COCH_3 \rightarrow$ $CH_3-C(OH)(CH_3)-CO-CH_3$	100	CH ₃ -(C)	1400	(C)-CH=O	1400	(C)-C=C(C)
	1200	(C)-C=C(C)	100	(C)-CH ₂	1400	(C)-CH=C(C)
	1100	(C)-CH ₂ -(C)	1400	(C)-CH=C(C) (C)-CH=C(C)	1400	(C)-CH=O (C)-CH=O

DEAN of the carbon atom, but exchange of a heteroatom for another heteroatom does not. A horizontal row in this table will be referred to as a "half-reaction". To be more clear, not only the DEAN's themselves are given, but also the atomic environment they represent. From such a table a reaction description of the type of Hendrickson³ could be easily derived.

In reactions of simple compounds this procedure does not always work. Sometimes there are so few bond pairs, that starting material and end product do not have one in common. In that case the half-reactions are derived directly from the bond pairs. One has to be very careful especially when a big and a small compound react, because by comparing the bond pairs of starting material and end product no bond pairs of the small compound will be found present in the end product, but many of the big one will be. Thus one has to check whether "representatives" of each compound in starting materials and in the end products are present. Formation of methyl ethers from methanol is an example. Each complete description leads to two half-reactions, each partial description to one.

As mentioned above, not all reactions are described completely. Usually inorganic reagents are not defined. This affects especially class I. This is also reflected in the structure of the classification. Classification in class I can be done on the basis of one half-reaction, and class II could be classified by a half-reaction as deep as the classification described in the first article of this series.

Using the more detailed subdivision, the second half of the reaction is also needed. Classes III and IV need the second half of the reaction. If carbon atoms are missing from the description, they are supposed to be in the form of CO₂.

THE RING FACET

The ring facet is expressed by coding the rings that are present in the end product but not in the starting material, and the ones that are present in the starting material but not in the end product. As the analysis starts with decoding of a WLN, it would not be very sensible to select the rings that build up the ring system in another way than the WLN rules do. Thus the ring selection of the WLN rules is followed.

During the analysis of a ring system, an intermediate table is produced in which a row corresponds to an individual ring of which the ring system is built. In this row, the topological numbers of the atoms, which form the ring, are given (see Figure 6). This table has ten columns. If the tenth contains a one, the ring composition is continued in the next row. If it contains a zero, it is not.

From the translation table, as given in Figure 1, the exact composition of the composing rings could easily be derived.

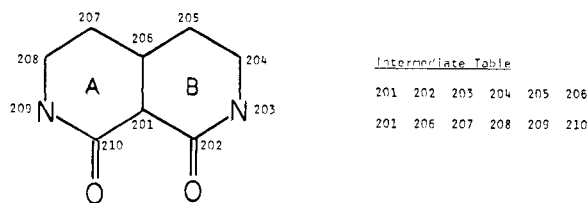


Figure 6. Diagram and intermediate table of a compound, showing the topological numbers for the rings A and B; structurally, the latter are equivalent.

COMPOSITION						
RING 1	C _{unsat}	N _{sat}	C _{sat}	C _{sat}	C _{sat}	C _{sat}
RING 2	C _{sat}	C _{sat}	C _{sat}	C _{sat}	N _{sat}	C _{unsat}

Figure 7. Ring composition of the individual rings in the compound of Figure 6; a simple rotation of the composition of ring 1 will not reveal the equivalence with ring 2.

However, the DEAN's contain too much information, because the ring facet deals only with what happens with this ring and not with what happens to atoms attached to the ring atoms. For comparison of the rings, the only thing we need to know is the degree of unsaturation and the heteroatomic character of the atom. So to each DEAN a number is assigned which reflects only the type of ring atom and its degree of unsaturation. For carbon, only four types are possible: (1) saturated, (2) with one double bond attached, (3) with one triple bond attached, (4) with two double bonds attached (allene). In an analogous way the other elements are divided.

The problems arising by comparing two rings can be illustrated by Figures 6 and 7. Ring A and ring B are obviously equal because they contain the same atoms in the same sequence. Nevertheless, the ring compositions are different. It is easy to see that rotating the composition of one of the rings will never give the ring composition of the other ring, because of the inverted order of N_{sat} and C_{unsat} in the WLN. Therefore not only rotation is necessary, but also the composition has to be inverted and the inverted order rotated. By this procedure, equal rings will always be found.

The same problems occur in comparing the composing rings of two ring systems. Because the compound with the fewest rings is rotated, any new formed ring and opened rings are in the second compound. Since those rings which were found during the comparison have been registered, the others are new or opened.

The classification for ring opening and formation (as discussed before¹) consists of three digits representing: (1) the size of the ring, (2) degree of unsaturation, and (3) heteroatomic character of the ring.

Size of the Ring. The size of the ring in general does not present problems. This digit is directly derived from the above-mentioned intermediate table.

State of Unsaturation. For coding the digit for unsaturation, one needs to know the number of double bonds in the ring. If both atoms of a double bond are part of the same ring system, there is no problem, but a decision is needed if only one of the atoms is part of the ring under consideration. For the purpose of this program such a bond counts for a half double bond. The degree of unsaturation is the sum of the number of double bonds plus half the number of the half double bonds, rounded to a whole number. This also applied when the double bond is not between two carbon atoms. Thus cyclohexanone is a completely saturated ring, and cyclohexadione is a ring with one double bond, as far as the ring facet is concerned. If two or more double bonds are present, the kind of conjugation is included in the digit.

The Heteroatomic Character of the Ring. As stated before, the combination of the intermediate table and the translation table with the simplified DEAN's gives the composition of the

C A R B O N					N I T R O G E N			
Substituted		Potential bonds			Substituted		Potential bonds	
Type of carbon	WLN symbol	Type of carbon	WLN official rules	WLN our rules	Type of nitrogen	WLN symbol	Type of nitrogen	WLN symbol
$\text{CH}_3\text{---R}$	number	CH_3^+	number	number	$\text{NH}_2\text{---R}$	Z	NH_2	Z
$\text{CH}_2=\text{R}$	number	$\text{CH}_2:$	number	number	$\text{NH}=\text{R}$	M	NH:	M
$\text{CH}\equiv\text{R}$	number				$\text{N}\equiv\text{R}$	N		
$\text{---CH}_2\text{---R}$	number	---CH_2^+	number	number	---NH---R	M	---NH^+	M
$=\text{CH---R}$	number	$=\text{CH}^+$	number	number	$=\text{N---R}$	N	$=\text{N}^+$	N
$=\text{C}=\text{R}$	C	$=\text{C:}$	C^{W}	C	$\equiv\text{N}\rightarrow\text{O}$	N^{W}	$\equiv\text{N:}$	N^{W}
$\text{C}\equiv\text{C---R}$	number	$\equiv\text{C}^+$	number	number				
$\text{H}\equiv\text{C---R}$	C	$\text{H}\equiv\text{C}^+$	C	C				
$>\text{CH---R}$	Y	$>\text{CH}^+$	number	Y	$>\text{N---R}$	N	$>\text{N}^+$	N
$>\text{C}^{\text{---}}$	Y	$>\text{C}^+$	number	Y	$>\text{N}\rightarrow\text{O}$	N	$>\text{N}^+$	N
							$\text{---N}\equiv$	N^{W}
							---N^+	N^{W}
---C---	X	---C^+	Y	X	---N---O	K	---N:	N^{W}
							---N^+	K
							$>\text{N}^+$	K

^W strictly speaking not defined

Figure 8.

ring. From this the third digit can be derived without problems. This may be illustrated by the three digits for a few well-known rings (see first article of the series).

cyclohexene	610 (size is 6 atoms, double bond, no heteroatoms)
pyridine	641 (size is 6 atoms, 3 conjugated double bonds, one nitrogen)
tetrahydrofuran	503 (size is 5 atoms, no double bonds, one oxygen)

THE REARRANGEMENT FACET

Definition of a Rearrangement. A rearrangement is a reaction within a reaction. To be called a rearrangement it is necessary that it can be described according to the following scheme: a reactive intermediate originating from the starting material, such as an anion, a cation, a radical, or a carbene, forms another reactive intermediate, which then gives the end product. The two successive reactive intermediates should consist of the same atoms. More complicated versions of this reaction scheme are of course also allowed. Thus at least two intermediates must be given at the input stage. If possible intermediates cannot be given, the rearrangement part of the reaction is considered to be of doubtful value for documentation. In a reactive intermediate there is by definition always an atom present which is capable of forming one bond (anion, cation, radical) or two bonds (carbene, etc.). Such bonds will be referred to as "potential bonds".

The first thing that had to be considered was whether all types of potential bonds could be encoded in WLN. During this consideration one of the basic "inconsistencies" of the WLN was again noticed, viz., the rule to code nitrogen atoms according to the number of nonhydrogens attached to it, but carbon according to the number of attached hydrogens. Quaternary nitrogen is an exception to this rule. The change in the use of the symbols is not congruent. Although the rule

STRUCTURE	WLN OFFICIAL RULES	WLN OUR RULES
	3R &1/0	3R &1/0
	3R &1/0	1Y1R &2/0
	3R &1/0	2YR &2/0

Figure 9.

does not lead to ambiguities, it is not elegant. When a potential bond is present, the same type of discrepancy arises. For symbol designation of carbon atoms the number of nonhydrogen substituents determines the symbol, for nitrogen the number of hydrogens (see Figure 8).

However, this rule leads to ambiguities when coding an aliphatic chain indicated by a number (see Figure 9). According to the rules, a change in the "structural" part of the notation cannot be indicated, so the string of symbols used to indicate the potential bond does not indicate which of the carbon atoms of the chain carries the potential bond.

There is no rule for coding carbenes, etc. In the WLN coding practice they are treated as diradicals, which is unfortunate from the chemical point of view, but it is unambiguous. In our program all potential bonds are treated like those of nitrogen. They are thus coded as ordinary substituents (see Figures 8 and 9). A proposal suggesting a change in the WLN rules in this respect has been sent to the C.N.A.⁴ The following DEAN's were designated for these "substituents":

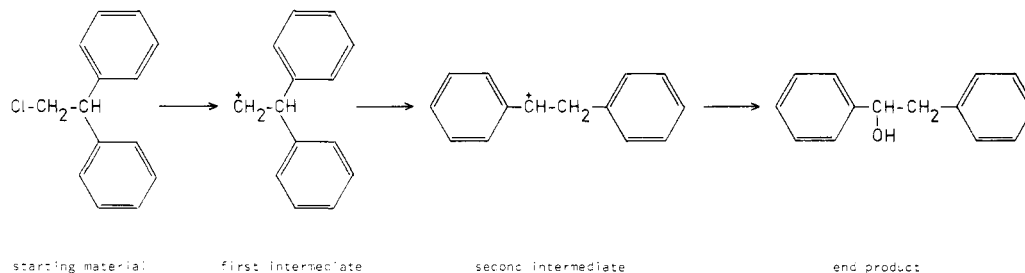


Figure 10. Example of a rearrangement.

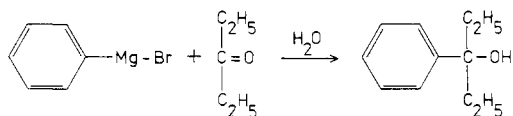


Figure 11. Example of a reaction in which the unusual element facet is involved.

Radical	-100
carbene, etc.	-200
anion, monovalent	-300
anion, divalent	-400
anion, trivalent	-500
cation, monovalent	-600
cation, divalent	-700
cation, trivalent	-800
cation, tetravalent	-900

This is in accordance with the use of coding a carbene as a diradical.

The Coding of Rearrangements. The fact of the rearrangement is stated at the input stage, by indicating two intermediates. The change between these two intermediates is the actual rearrangement (see Figure 10). As soon as the change is found between these two compounds, as expressed in the chain facet, finding the code number is fairly simple. The only problem which caused some programming difficulties was to distinguish between allyl rearrangements and other carbon-carbon rearrangements.

THE UNUSUAL ELEMENT FACET

The DEAN's of the unusual elements are derived from their place in the Periodic System. The code number of the reaction facet of unusual elements is also derived from the number of the element in the periodic system, so after having found which unusual element is not present in both starting material and end product the code number can easily be derived. A reaction with an unusual element is given in Figure 11. As the number of magnesium in the Periodic System is 12, the code number indicating the unusual element facet is 8.12.

DISCUSSION

In this paper a description is given of a computer program performing automatic encoding of chemical reactions. In Figure 13 the output of the computer program is shown for the seven reactions of Table I, Figures 10, 11, and 12. Comparing example 3 with example 5 shows that in classes 1 and 2 each nonhydrogen atom is classified separately. However, in classes 3 and 4 the carbon atoms are classified in pairs of atoms. Example 3 shows that the concepts "ring formation" and "ring opening" are interpreted very widely. In example 4 the second starting material was missed in the first round. This problem was solved in the way described elsewhere in the article.

Example 6 is a rearrangement. A rearrangement has to be coded in three steps. However, the end product of the first and the second step need not be reformulated as starting materials for the second and the third step. This is done by

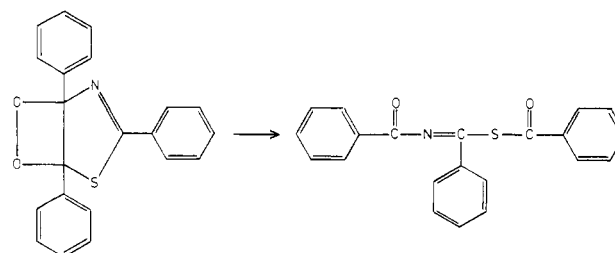


Figure 12.

1. STARTING MATERIAL(S) C1 END-PRODUCT(S) C1 REACTION-CODE IS 12.	2. STARTING MATERIAL(S) 1V1 END-PRODUCT(S) QY1&1 REACTION-CODE IS 14. REACTION-CODE IS 14.
3. STARTING MATERIAL(S) L&U END-PRODUCT(S) L&U A1 B1 RING-CLASSIFICATION IS 6611 RING-CLASSIFICATION IS 6601 REACTION-CODE IS 22. REACTION-CODE IS 22.	4. STARTING MATERIAL(S) 1V1 END-PRODUCT(S) 1V1U2 NUMBER ABOVE 1000 MISSING IN STARTING MATERIAL REACTION-CODE IS 34.
5. STARTING MATERIAL(S) L&U END-PRODUCT(S) V&V RING-CLASSIFICATION IS 6611 REACTION IS 40	6. STARTING MATERIAL(S) G1YR&R END-PRODUCT(S) 1YR&R REACTION-CODE IS 12. STARTING MATERIAL(S) R1YR REARRANGEMENT, CENTRAL PART REACTION-CODE IS 71 END-PRODUCT(S) QYR&R REACTION-CODE IS 12.
7. STARTING MATERIAL(S) D-WG-R END-PRODUCT(S) QX&R REACTION-CODE IS 8.12 REACTION-CODE IS 33.	8. STARTING MATERIAL(S) T45 B00 EN GS DUT0 AR& DR& FR END-PRODUCT(S) FV&VR&UNVR RING-CLASSIFICATION IS 6403 RING-CLASSIFICATION IS 6515 SOMETHING WRONG REACTION IS 42

Figure 13. Some examples of reactions encoded by the program.

the program. However, because the computer has to retain the first and the second end product as starting material for the next reaction, the indication: "starting material" was used erroneously in the second step.

The eighth reaction was taken from the report by Lynch.⁵ It is a reaction in which the normal procedure leads to inconsistent equivalences, as indicated by: "something wrong". This time it was caused by the fact that inside rings the opposite of rule 2 has to be applied. Again the problem was

solved by applying the fact that the atoms which formed the ring still have to form a chain.

Up to now only a limited number of reactions was encoded by the program. It is expected that more inconsistency checks and guide lines may be necessary when more reactions are processed. The program encodes only as deep as the description of the classification.¹ A greater depth of encoding will most likely not give rise to serious problems.

ACKNOWLEDGMENT

The authors wish to express their gratitude to the management of Gist-Brocades Research for the use of their

IBM-1130.

REFERENCES AND NOTES

- (1) M. Osinga and A. A. Verrijn Stuart, "Documentation of Chemical Reactions. I. A Faceted Classification", *J. Chem. Doc.*, **13**, 36-39 (1973).
- (2) M. Osinga and A. A. Verrijn Stuart, "Documentation of Chemical Reactions. II. Analysis of the Wiswesser Line Notation", *J. Chem. Doc.*, **14**, 194-8 (1974).
- (3) J. B. Hendrickson, "A Systematic Characterization of Structures and Reactions for Use in Organic Synthesis", *J. Am. Chem. Soc.*, **93**, 6847-54 (1971).
- (4) Letter to W. J. Wiswesser, October, 8, 1974.
- (5) M. F. Lynch, "Development and Assessment of an Automatic System for Analysing Chemical Reactions," Final report to the British Library, Research and Development, July 1975.

Computer Recognition and Segmentation of Chemically Significant Words for KWIC Indexing[†]

DAVID R. HEYM, HERBERT SIEGEL, MARGARITE C. STEENSLAND,* and HAO V. VO

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received April 2, 1976

This paper describes an algorithm which recognizes and segments chemically significant words and thus provides additional index entry points for Keyword-In-Context (KWIC) Indexing. The words are those appearing in titles of documents selected for coverage in *Chemical Titles*. The procedure begins by matching strings of characters in the title word with roots, or stems, of words stored in computer memory. When a match occurs, parts of the recognized words are compared with other fragment lists, and a matrix is formed to select the fragments to be indexed. For example, dichloronitrobenzene is segmented as *di chloro nitro benzene*. Types of words recognized and segmented include names and classes of chemical substances, complex substituents, and reaction processes. The technique can be extended to words in other fields of science by appropriate modification of the lists and dictionaries stored in the computer. Implementation of machine segmentation improves consistency in the index production and eliminates most of the intellectual effort involved in the process.

INTRODUCTION

Chemical Titles (CT) is a biweekly current-awareness service which reports the titles and bibliographic information of those documents of chemical and chemical engineering interest which have recently been published in selected journals. The production of CT is computer-based and results in a machine-readable version as well as the printed issue.

Although the printed issue of CT contains a bibliographic section and an author index, the basic access tool for the information is the Keyword-In-Context (KWIC) Index (see Figure 1). Each line of the index contains up to 60 characters and spaces, plus a code identifying the specific article in a journal.

The access point for each line is the keyword (lipid, lipids, etc., in Figure 1). Keywords appear in their normal context in the title and are listed alphabetically in a fixed, columnar arrangement.

Each significant word in the title is used as the keyword or entry point for a new index line. A significant word is defined to be any technical term describing a process, chemical substance name, class of chemical substances, animal name, organ, chemical reaction, operation, apparatus or equipment, theory or hypothesis, or scientific law. A nonsignificant word is one which is not in one of the above classes or which is too general for index purposes. The "CT Stopword List", printed

in each issue of CT, alerts users to nonsignificant words.

Words are frequently made up of several fragments or segments, some of which are themselves words that have meaning and therefore have value as access points in a KWIC Index. When this is true of chemically significant words (chemical names, chemical processes) in the titles processed for the KWIC Index in CT, those words are segmented to yield additional index entries. In Figure 1, the next to last entry is indexed at the segment *lipo* of the keyword *apolipoprotein*. Later in the KWIC, the same entry is indexed at the segment *protein*.

The object of the work described in this paper was to develop a computer-based algorithm which would provide segmentation for indexing purposes, i.e., find and indicate access points within chemically significant words. Since segments derive their significance from their technical meaning, not all are suitable index terms. Examples of acceptable segmentation are shown in Figure 2; the apostrophes indicate the segmentation points.

The segmentation algorithm must recognize chemically significant words, distinguish between general terms and chemical names, and indicate proper segmentation points. It must also recognize words which are candidates for segmentation but which cannot always be segmented because of ambiguity. These words must be listed for review and proper handling by chemists.

The advantages of such an algorithm include elimination of much of the intellectual effort otherwise required and an increased consistency in the keyword index. In the Chemical

[†] Presented before the Division of Chemical Information, 170th National Meeting of the American Chemical Society, Chicago, Ill., Aug 27, 1975.

* Author to whom correspondence should be addressed.