Documentation of Chemical Reactions. I. A Faceted Classification

M. OSINGA and A. A. VERRIJN STUART*
Centraal Rekeninstituut der Rijksuniversiteit Leiden, Wassenaarseweg 80, Leiden, Holland

Received September 8, 1972

Existing methods for coding chemical compounds are discussed and evaluated as to their suitability for documentation of chemical reactions, a new classification for chemical reactions is presented, and possibilities of automatic encoding are studied.

Over the years, a lot of time and effort has been spent on the development of systems for the documentation of organic chemical structures. Much less has been done on systems for chemical reactions.

This paper is concerned with a system for the documentation of chemical reactions. For the purpose of this paper, the system is defined as a mechanism able to answer questions which are not expressed in terms of the starting material or the end-product. For brevity, the expression "the compounds" will be used for one or more starting materials and one or more end-products and "reaction-sites" for indicating the part of the starting material that is being changed and the part of the end-product that has been changed.

An important reason for the slow development of systems for the documentation of chemical reactions must be that, unlike chemical compounds, chemical reactions are not static entities, but dynamic processes and as such are not as easy to describe as a compound.

First, the possibilities for developing the documentation of chemical reactions from the different methods of coding chemical compounds will be reviewed. Second, a more detailed account will be given of how a new version of one of these developments was created.

POSSIBILITIES FOR DESCRIBING A CHEMICAL REACTION

There are two possible ways of describing a chemical reaction: by coding the compounds and by coding the reaction sites.

If the compounds are coded, three levels of possibilities for correlating them can be distinguished:

1. Little or no correlation.
   This is found in indexes, etc. Of course, correlation *a postiori* can be done, but this is not supplied by the system. This level will be referred to as *manual*.
2. Limited possibilities for correlation.
   This is found in *mechanical* systems, like those using machine sorting, edge-punch cards systems, etc.
3. Extensive possibilities of correlation.
   This can only be provided by a *computer*.

The medium, by which the information is carried, is not always decisive in this respect. A system on microfilm could be of each of the levels.

A manual approach to coding systems, which gives only the compounds, is not a system as previously defined, so it will be ignored.

The mechanical approach produces a useful system only when applied to a fragmentation code in which fragments are associated with fixed position in the cards.

This approach has given rise to a few applications—e.g., the Reaktionskartei[1] and the Ziegler system.[2,3]

The Gremas system[4,5] for coding reactions is also based on a fragmentation code, but cannot be used mechanically, because the fragments do not have fixed positions.

Eight possible methods of coding compounds can be distinguished. The degree to which the combination of coding the compounds with a specific code and computer-comparison of the results which lead to usable systems will be discussed. In addition, the possibilities for coding the reaction-sites are explored.

1. Systematic Nomenclature. It is possible to code the compounds with systematic names, but a systematic name is not very accessible for computer manipulation owing to the complicated structure of the chemical nomenclature. As a means of coding reaction sites, a kind of systematic nomenclature is used for simple reactions—e.g., decarboxylation, bromination, acetylation, etc.

Patterson[6] created a more extended system based on this nomenclature. Although only designed for simple reactions, it was not generally accepted.

2. Molecular Formulas. It is simple to "subtract" molecular formulas, but the indication of a reaction site so produced is so ambiguous that no system can be based on this principle.

The same is true for coding the reaction site in the form of molecular formulas, but it might be sufficient for simple cases. No system based on this principle is known to us.

3. Line Notations. Line notations have in general a quite consistent and not so complicated set of rules, which make them readily accessible for computer manipulations.

One of the more simple manipulations possible with a line notation is the production of a permuted KWIC-index, which can be used as a kind of fragmentation code. A system based on this principle was described by Gelberg.[7]

A line notation could also be used for coding the reaction site using the specific fragment notation, although a slight change can make quite a difference to the notation. No system based on this principle has yet been reported.

4. Fragmentation Codes. As stated before, fragmentation codes are suitable for comparison of the compounds, because each fragment can be searched for separately. The codes previously mentioned can be used ei-

*To whom correspondence should be addressed.

ther mechanically or on a computer, but the Gremas system[4,5] can be exploited only by the latter.

Fragmentation codes are also very suitable for coding the reaction sites, as has been done in the Ziegler system[2,3] and the Gremas system. A disadvantage of these systems, however, is the ambiguity of fragmentation codes, especially when overcoding is used—for example, in the Reaktionskartei.

5. Topological Codes. Because topological codes denote every detail of a structure, they are very suitable for comparison of the compounds by computer. However, processing topological codes uses a great deal of computer time, and this is a serious objection to their use as such.[8] Encoding of complex structures can also create difficulties unless one has available the type of special equipment used by large institutions such as CAS and IDC.

No system for the retrieval of reactions by means of a topological code is known to the authors, although the IDC system might serve this purpose.

The topological code can be used for coding reaction sites, but up to now this has been done only by Lynch[9-11] and coworkers. Using a computer, they compared the topological coding of starting material and end-product of a reaction and arrived at the reaction site from the difference in code.

Although not completely concerned with the same subject, the work of Corey[12-18] should be mentioned here. He derived a possible precursor from the end-product, which was coded in the form of a topological code, with aid of a computer program.

After that the precursor of this precursor can be derived, etc. Usually there are several precursors possible for one end-product, giving a "tree" of precursors. A chemist then selects the pathway he thinks is the best for the synthesis of the end-product.

6. Classifications. There are several classifications for the coding of reaction sites. Such a classification forms a part of each general classification scheme—e.g., UDC, where it is located at 542, mainly 542.9.

There are also separate systems for coding chemical reactions, based on this principle, like Weygand,[19] Theilheimer,[20] Sugasawa,[21] Vleduts.[22] The Theilheimer classification also forms a part of the code of the Reaktionskartei.

7. Trivial Nomenclature. Trivial nomenclature is not suitable for coding the compounds, unless one uses a kind of dictionary, but this can hardly be called a system.

Trivial nomenclature is often used for coding the reaction sites (see Mishchenko[23] for a review). This is usually done by using the name of the chemist who first performed the reaction. This system is very useful for those chemists who know the name reactions, but it is difficult to find the trivial name of a reaction, if you do not know it. The disadvantage is that not all reactions have a name, and moreover, some reactions are known under different names, especially in different countries.

8. Numbers. Numbers, such as registry numbers, are not suitable for coding the compounds, unless a dictionary is used. Neither are they suitable for coding the reaction sites.

A NEW CLASSIFICATION

We decided to study classifications in greater detail, and considered it worthwhile to develop a new one. The basic principle involved in the new classification is that a reaction can have different facets, each of which should be retrievable separately. For example, the formation of a lactone is a ring closure as well as an ester formation.

Therefore, we developed a faceted classification consisting of five parts. In this classification, the notation for a reaction can be built from numbers from each of these five parts and a '+' sign is used for a connection.

These five parts are:

1. Elimination, Substitution, and Addition Facets
   These consist of four classes:
   1. The carbon skeleton is not influenced.
   2. The degree of saturation of the carbon skeleton is changed.
   3. C—C bonds are formed.
   4. C—C bonds are broken.
2. Ring Aspects of the Reaction
   It consists of:
   5. Ring formation.
   6. Ring opening.
3. Rearrangements
   This part forms class 7.
4. Unusual Elements Are Involved in the Reaction
   These are elements other than C, H, N, O, S, P, F, Cl, Br, I. This part forms class 8.
5. Other Facets
   This part forms class 9.

Class 0 is reserved for *reaction conditions and related phenomena*. Each of the classes has been further subdivided.

For classes 1-4, 7 and 9 only the first subdivision will be presented here. As far as the subclasses are not self-explanatory, a few examples will be added.

A more detailed description is available from the authors. For class 1 the following notation is used for brevity: HF, standing for hetero-functions, followed by a number representing the number of bonds from carbon to a hetero-atom. Thus HF 1 means, only one bond between a hetero-atom and a certain carbon-atom—e.g., in an alcohol.

Class 1 is subdivided as:

10. Reactions in which no carbon-bonds are involved.
    Nitrobenzene gives aniline.
11. Reactions from HF 0 to HF 1, 2, or 3, and the reverse.
    Methane gives methyl chloride.
12. Reactions within HF 1.
    Methyl chloride gives methanol.
13. Reactions from HF 1 to HF 2, or 3.
    Methanol gives formaldehyde or formic acid.
14. Reactions from HF 2 and 3 to HF 1.
    Benzoic acid gives benzyl alcohol.
15. Reactions within HF 2.
    A ketone gives an oxime.
16. Reactions from HF 2 to HF 3.
    Aldehyde gives carboxylic acid.
17. Reactions from HF 3 to HF 2.
    Ester gives aldehyde.
18. Reactions within HF 3.
    Carboxylic acid gives ester.
19. Reactions from, to, and within HF 4, without breaking C—C bonds.
    Phosgene gives urea.

Class 2:

20. Formation of a double bond from a single bond.
21. Formation of a triple bond from a single bond.
22. Formation of a single bond from a double bond.
23. Formation of a triple bond from a double bond.
24. Formation of a single bond from a triple bond.
25. Formation of a double bond from a triple bond.

Class 3:
30. Substitution of RX with R-Metal or RX + Metal.
Wurtz-Fittig
31. Substitution of RX to double or triple bond.
Friedel-Craft reaction.
32. Addition of olefin to diene.
Diels-Alder reaction.
33. Addition of R-Metal to double or triple bond, one atom of which is not carbon.
Grignard reaction.
34. Addition of RH to double or triple bond, one atom of which is not carbon.
Claisen condensation.
35. Addition of isonitriles and ylides.
Wittig reaction.
36. Addition of carbon monoxide.
37. Decarboxylative formation of C—C bonds.
Kolbe electrochemical synthesis.
38. Forming of C—C bond under elimination of N_2.
Graebe-Ullmann synthesis.
39. Other reactions forming C—C bonds.
Passerini reaction.

Class 4:
40. Decarboxylation.
41. Retroaldol synthesis.
42. Ozonolysis and other oxidations of a double bond.
43. Oxidation of carbon chains, with formation of carboxylic acids and carbon dioxide.
44. Other reactions.

Classes 5 and 6 are subdivided in the same way: The first digit gives the size of the ring. If the size of the ring is 10 or bigger, a zero is used. The second has the following meaning:
0 = no double bonds in the ring.
1 = one double bond.
2 = two double bonds, conjugated.
3 = two double bonds, not conjugated.
4 = three double bonds, all conjugated.
5 = three double bonds, not all conjugated.
6 = more than three double bonds.
The third digit defines the heterocyclic character of the ring.
0 = No hetero-atoms in the ring.
1 = 1 N in the ring.
2 = 2 or more N's in the ring.
3 = The ring contains O, but not N.
4 = The ring contains O and N.
5 = The ring contains S, with or without N and O.
6 = The ring contains P, with or without N, O and S.
7 = The ring contains any other element.

Thus, the formation of a five-numbered ring lactone without double bonds in the ring and with one O and no N gives 5.503. Total code: 18 + 5.503 (18 = ester formation).

Class 7:
70. Shift of double and triple bonds.
Allylic rearrangement.
71. Shift of C or H from C to C.
Camphor rearrangement.
72. Shift of C or H from N to C.
Stevens rearrangement.
73. Shift of C or H from C to N.
Lossen rearrangement.
74. Shift of C or H from C to other atoms.
75. Shift of C or H from other atoms to C.
Fries reaction.
76. Other rearrangements.

Class 8:
The 8 is followed by a full stop and two figures, indicating the number of the element in the periodic table. Accordingly, 8.03 indicates a reaction in which lithium is involved.

Class 9.
90. Polymerization
91. Pi-complexes
92. Other facets.

FURTHER RESEARCH

The main purpose of our research is to study conversions, in particular those where a classification of chemical reactions is derived from the chemical codes of starting material and end-products. This classification has been outlined in the preceding paragraphs.

Out of the different possibilities, the Wiswesser Line Notation[24] was selected in preference to a systematic nomenclature or a topological code. A systematic nomenclature was not chosen because it is too complicated for our purpose. A topological code was not chosen because it would probably mean more or less duplicating the work of Lynch.[9-11]

The Wiswesser Line Notation was selected because it is used so much that possible results might find practical application. A data-base of chemical reactions has been created.

The next step is to form from the Wiswesser Line Notations pairs of symbols which are attached to each other. It is expected that comparison of the list of pairs obtained from starting material and end-product will give sufficient information for encoding the reactions.

LITERATURE CITED

(1) Schier, O., Nuebling, W., Steidle, W., and Valls, J., "Ein System zur Dokumentation chemischer Reaktionen," *Angew. Chem.* **82,** No. 15, 622-8 (1970) (Ger. ed.).

(2) Ziegler, H. J., "Reactiones Organicae, une nouvelle technique de documentation des rėactions organique," *Inform. Chim.* 41, 22-4, 27-8, 31-4 (1966).

(3) Ziegler, H. J., "A New Information System for Organic Reactions," *J. Chem. Doc.* **6,** 81-9 (1966).

(4) Fugmann, R., Braun, W., and Vaupel, W., "GREMAS—ein Weg zur Klassifikation und Dokumentation in der organische Chemie," *Nachr. Dok.* **14,** 179-90 (1963).

(5) Lobeck, M. A., "Recherchen mit dem IDC-system," *Angew. Chem.* **82,** 598-605 (1970) (Ger. ed.).

(6) Patterson, A. M., "Systematic Names for Substitution Reactions," *Chem. Eng. News* **32,** 4019 (1954).

(7) Gelberg, A., "Rapid Structure Searches Via Permuted Chemical Line Notation. IV. A Reactant Index," *J. Chem. Doc.* **6,** 60-1 (1966).

(8) Ming, T. K., and Tauber, S. J., "Chemical Structure and Substructure Search by Set Reduction," *J. Chem. Doc.* **11,** 47-51 (1971).

(9) Armitage, J. E., and Lynch, M. F., "Automatic Detection of Structural Similarities among Chemical Compounds," *J. Chem. Soc. C* **1967,** 521-8.

(10) Armitage, J. E., Crowe, J. E., Evans, P. N., Lynch, M. F., and McGuirk, J. A., "Documentation of Chemical Reactions by Computer. Analysis of Natural Changes," *J. Chem. Doc.* **7,** 209-15 (1967).

(11) Harrison, J. M., and Lynch, M. F., "Computer Analysis of Chemical Reactions for Storage and Retrieval," *J. Chem. Soc.* **1970,** 2082–7.

(12) Corey, E. J., and Wipke, W. T., "Computer-assisted Design of Complex Organic Synthesis," *Science* **166,** 178–92 (1969).

(13) Corey, E. J., "General Methods for the Construction on Complex Molecules," *Pure Appl. Chem.* **14,** 19–37 (1967).

(14) Corey, E. J., "Computer-assisted Analysis of Complex Synthetic Problems" (Centenary lecture), *Quart. Rev.* **25,** 455–82 (1971).

(15) Corey, E. J., Wipke, W. T., Cramer, R. D., III, and Howe, W. J., "Computer-assisted Synthetic Analysis. Facile Man-machine Communication of Chemical Structure by Interactive Computer Graphics," *J. Amer. Chem. Soc.* **94,** 421–30 (1972).

(16) Corey, E. J., Wipke, W. T., Cramer, R. D., III, and Howe, W. J., "Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules," *J. Amer. Chem. Soc.* **94,** 431–9 (1972).

(17) Corey, E. J., Cramer, R. D., III, and Howe, W. J., "Computer-assisted Synthetic Analysis of Complex Molecules. Methods and Procedures for Machine Generation of Synthetic Intermediates," *J. Amer. Chem. Soc.* **94,** 440–59 (1972).

(18) Corey, E. J., and Petersson, G. A., "An Algorithm for Machine Perception of Synthetically Significant Rings in Complex Cyclic Organic Structures," *J. Amer. Chem. Soc.* **94,** 460–5 (1972).

(19) Weygand, C., "Organic Preparations," VII-XIII, 1-3, Interscience, New York, 1945.

(20) Theilheimer, W., "Synthetische Methoden der Organischen Chemie," **1,** VI-VII S. Karger, Basel, New York, 1946.

(21) Sugasawa, S., and Nakai, S., "Reaction Index of Organic Synthesis," Wiley, New York, 1967.

(22) Vleduts, G. E., "Concerning One System of Classification and Codification of Organic Reactions," *Inf. Stor. Retr.* **1,** 117-46 (1963).

(23) Mishchenko, G. L., "Information Retrieval in the Field of Chemistry," 117-36, *Nat. Techn. Inf. Serv. JPRS-53523.* From *Zh. Vses. Khim. Obshch.* **16,** 55-63 (1971).

(24) Smith, E. G., "The Wiswesser Line-formula Chemical Notation," McGraw-Hill, New York, 1968.

Error Checking Digit for Nonconventional Chemical Codes*

K. SUBRAMANIAM and P. V. SANKAR**

Department of Electrical Communication Engineering and Molecular Biophysics Unit,
Indian Institute of Science, Bangalore 560012, India

Received September 22, 1972

A method of constructing a check digit for nonconventional chemical codes is described.

In handling large volumes of data such as chemical notations, serial numbers for books, etc., it is always advisable to provide checking methods which would indicate the presence of errors. The entire new discipline of coding theory is devoted to the study of the construction of codes which provide such error-detecting and correcting means.[1] Although these codes are very powerful, they are highly sophisticated from the point of view of practical implementation. With this in view, several inexpensive, but fairly effective means to guard against errors resulting from improper data transcription have been evolved. Some of these checks are already available for use in documentation and libraries—in the International Standard Serial Numbering (ISSN) for books and journals, and CODEN used for periodic Titles and in the Chemical Abstracts Service (CAS).

The ISSN system[2] uses a seven digit decimal code $\alpha\beta\gamma\delta\epsilon\phi\psi$ (where $\alpha,\beta,\gamma,\delta,\epsilon,\phi,\psi,\ = 0, 1, 2, \ldots\ldots 9$) and derives a check digit σ from these seven digits by using the equation

$$\sigma \equiv [11 - (8\alpha + 7\beta + 6\gamma + 5\delta + 4\epsilon + 3\phi + 2\psi) \bmod 11]$$

and sets for σ = X for $\sigma \equiv 10 \bmod 11$

The Coden System,[3] on the other hand, uses only a five alphanumeric character code $\alpha\beta\gamma\delta\epsilon$ and derives the check character by the equation

$$\sigma \equiv (11\alpha + 7\beta + 5\gamma + 3\delta + \epsilon) \bmod 34$$

by assigning integral weights as follows:

CODEN: A, B, ... Y, Z, 1, 2, ... 9, 0
Equivalent: 1, 2, ... 25, 26, 27, 28, ... 35, 36.

The remainder σ is converted into a check character by the following set of equivalents:

Remainder: 1, 2, ... 25, 26, ... 33, 34 (or zero)
Check character: A, B, ... Y, Z, 2, 3, ... 8, 9

The numeric check characters one (1) and zero (0) have been eliminated to avoid confusion with the alphabetic characters I and O.

Since the above methods of constructing check digits are applicable only for fixed length codes, it becomes necessary to consider a different scheme for constructing the check digits for a variable length code like WLN, etc.

Fortunately, the code suggested by Black[4] for decimal systems can be easily extended to a general base. This code detects a single error or an adjacent transposition error.

PRINCIPLE OF BLACK'S CODE

Given an N digit number, a check digit is constructed as a function of the N digits using a multiplication operation '*' defined below:

$$A * B = A \oplus (-1)^A B \qquad (1)$$

* Contribution No. 26 from the Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.

** To whom correspondence should be addressed.