mathematical models for computer-assisted solutions of non-numerical problems in chemistry

b.jerman-blažič

UDK 681.3:54

J. Stefan Institute, University of Ljubljana,
61000 Ljubljana, Yugoslavia

MATEMATIČNI MODELI ZA REŠEVANJE NENUMERIČNIH PROBLEMOV IZ KEMIJE S POMOČJO RAČUNALNIKA. V članku je dan pregled uporabe molekularne topologije pri računalniškem reševanju nekaterih nenumeričnih problemov iz kemije. Obdelani so naslednji problemi: identifikacija spojin v sistemih za iskanje in shranjevanje informacij, enumeracija in generiranje strukturnih izomerov s ciljem izločevanja molekularnih struktur na podlagi eksperimentalnih podatkov, predstavitev molekularnih struktur in procesa načrtovanja sintez v računalniškem načrtovanju organskih sintez.

The article gives a review of the applications of the molecular topology in computer-aided solving of some non-numerical chemical problems. The problems are: identification of a compound in chemical information retrieval systems, enumeration and generation of structural isomers for the purpose of special chemical studies and in computer-aided elucidation of molecular structure on the basis of experimental data, representation of molecular structures and synthesis-design process in computer-aided planning of organic syntheses.

1. Introduction

There seems to be hardly any concept in natural sciences which is closer to the notion of a graph than the molecular structure of chemical compounds. A molecular structure may be viewed as a graph composed of nodes (atoms) linked with edges (chemical bonds). In fact there is no essential difference between a graph and a structural formula. A graph is a mathematical structure which can be used to represent the topology of given molecule. The advantage of using graphs in the representation of molecular structure lies in the possibility of applying directly the mathematical apparatus of the graph theory for solving special chemical problems. The idea that metric characteristics of the molecules (that is bond lengths and bond angles) can be neglected in chemical studies is more and more popular. The molecular topology allows the non-metric relationships of the molecular structures and the totality of information contained in the molecular graphs to be investigated and applied in a very simple manner.

Concepts of topology and graph theory though not always recognized as such, are nowdays analysed and applied to various branches of chemical science: photochemistry (1), stereochemistry (2), transition metals chemistry (3), boron hydrid chemistry (4), saturated (5) and unsaturated (6) hydrocarbon chemistry, etc. Furthermore, basic concepts of chemistry such as configuration, isomerism, valency etc. are shown to have a topological basis (7).

In the present article we wish to review mainly the application of the molecular topology and the topology of a space of molecules in computer-aided solving of some non-numerical chemical problems. The problems are: identification of a compounds in chemical information retrieval systems, enumeration and generation of structural isomers for the purpose of special chemical studies and in computer-aided elucidation of molecular structure on the basis of experimental data, representation of molecular structures and synthesis design process in computer-aided planning of organic syntheses.

Let us at first to define a chemical graph and the associated notions: A chemical graph is (8) a graph consisting of nodes associated with atom names, and edges which correspond to chemical bonds. The degree of a node in the chemical graph has its usual meaning, i.e. the number of (non-hydrogen) edges connected to it. The valence of each atom determines its maximum degree in the graph. A special kind of chemical graphs are vertex-graphs. Vertex graphs are cyclic chemical graphs (8), from which nodes of degree less than three have been deleted.

2. Identification of a structure of chemical compound in an Information retrieval system

There is probably no science in greater need of mechanized information retrieval than chemistry. Millions of chemical compounds are known; new ones are produced at an even faster rate. The chemist has two main problems: first, he wants to find out whether the substance in his test tube is already known; second, given a substance, he wants to know the properties of similar substances.

Both problems can be reduced to a matching process: a description of the given compound has to be matched against descriptions of the compounds that make up the data base of the retrieval system. To assure a complete identification of the compound structure, a detailed atom-by-atom comparison is usually needed between the compound in the query and the compounds in the data base. If chemical compounds are represented as chemical graphs, the problem of matching the query item with the library item becomes identical to the problem of isomorphism of graphs, considerably simplified by the labels carried by the chemical graph nodes.

The problem of isomorphism of graphs received little attention in the literature until late 1950's (9). Research into this problem was stimulated by the development of chemical information retrieval systems, with chemical structure representation in the system's files. An approach that was implemented in several computer programs was the procedure of a node-to-node matching in search of coincidence. The nodes of two chemical graphs are matched one at time until either a valid correspondence is found or until incompatibility arises (10). In the later, it is necessary to backtrack to a point of former coincidence and start again with a different

choice of nodes. Large amount of backtracking is required in this technique due to the lack of any criteria in the decision-making step. From the computational point of view, the unavoidable backtracking is time wasting as only in rare instance is the correct choice made at each point of decision. This technique also requires information to be saved in order to restart from the last point of agreement.

The use of a standard numbering procedure for the nodes in the chemical graphs makes the problem of establishing isomorphism in graphs trivial. Many attempts have been made in order to develop standard numbering procedures. Bouman (11) has suggested ordering of the nodes in chemical graphs based on the examination of the degree of a node and the degree of the nodes to which it is connected. Randić (12) proposed a very interesting procedure for labelling the atoms in the graph by considering the rows of the adjacenty matrix of the graph as digits coded in a linear code. The search for matrices corresponding to a complete graph or to a fragment of a graph is to be carried out by ordering of the matrices according to decreasing values of the numbers representing the row vectors. A year later, the same author suggested another solution to the unique labelling of the graph nodes. The sequencing of atoms is performed according to the relative magnitudes of the coefficient of the largest eigenvector of the adjacency matric (13). Similar approach to the Bouman scheme is the Morgan algorithm (15) which exploits the concept of extended connectivity. The Morgan's algorithm was implemented in the information system of Chemical Abstract Service. The classification of the atoms is obtained by adding the initial connectivity values of nearest neighbours and assigning the sum to the node considered. As (16) recognized, this method does not always allow the maximum possible differentiation, although it generally allows the atoms to be divided into several classes depending on the number of non-hydrogen attachements to each atom.

Bart and Giordano proposed (14) a new graph matching procedure in which for the one-to-one matching the Fourier maxima of specific entities of the known chemical structure was used.

A completely different and most computer-oriented approach to graph identification was advanced by Sussenguth (17). The procedure that he suggested is based on two principles. First, if graphs G and G* are isomorphic, then the subset of nodes of G that exhibit some property must correspond to the subset of nodes of G* that exhibit this same property. Second, if the subsets of the nodes of G and G* that are characterized by some property do not have the same number of elements, then the two graphs cannot be isomorphic. The matching procedure starts with generation of subsets of nodes that represent the same type of atom.

The purpose of generation of subsets of nodes is to reduce the number of nodes of G* to which a node of G can correspond. The purpose is achieved by taking intersection of the subsets and by matching the resulting nodes. If some nodes are not matched, new subsets must be generated. The algorithm terminates when every node in G is paired off with a node in G*, or when two corresponding subsets of nodes of G and G* are found to differ in the number of nodes they contain. If the former is the case, graphs G and G* are isomorphic, if the later, isomorphism is impossible. Occasionally, the algorithm exhausts all subset generating properties before one of the two conditions is satisfied. This happens when more than one isomorphism is possible between the two graphs, or when the subset generating properties are incomplete in the sense that some property that would establish isomorphism or the lack of it has been neglected in the design of the algorithm. An improvement of the described algorithm was suggested by Ming and Tauber (18). They separated the structure search and substructure search into a distinct part of the algorithm and included first order degree (17) and second order degree in the control vector for use in structure search. A short cut of the atom-by-atom search technique and set generation procedure is the connectivity code developed by Penny (19). The connectivity code although it is not a solution in itself, can be a useful tool when used in conjunction with the two general techniques (atom-by-atom search and the set generation algorithm) as it is done in the computer program of Tauber and Ming (18).

3. Computer-aided generation and enumeration of structural isomers

Problems of structural isomerism in chemistry have received much attention for a long time, but only occasional attempts have been made toward a systematic solution of the underlying graph theoretical problems of structural isomerism. Graph theoreticians have frequently considered various aspects of this topic, but not necessarily in the context of organic molecules. Polya presented a theorem (20) which permits calculation of the number of structural isomers for a given ring system. Hill (21) and Taylor (22) pointed out that Polya's theorem permitted enumeration of geometrical and optical isomers in addition to structural isomers. More recent formulas for the enumeration of isomers of monocyclic aromatic compounds based on the graph theory, permutation groups and Polya's theorem were presented (23). Although the number of isomers may be interesting, these methods do not display the structure of each isomer. Even in simple cases, the task of specifying each structure by hand without duplication is an enormous one. Balaban published a series of papers (16) addressed in part to the problem of specification of isomeric structures. Although his method represents an important contribution to the problem of isomeric structures, it does not contain a mechanism for avoiding a duplicate structures. Most successful in solving this problem were the works based on the Dendral algorithm (24). The algorithm permits an enumeration and representation of all possible molecular structures with a given empirical formula, i.e. a given set of atoms. Chemical structures of all possible isomers are obtained by mathematical permutation of acyclic and cyclic graphs representing appropriate ring systems and attached acyclic chains of atoms. The Dendral algorithms was implemented in a computer program called Structure Generator (8). The list of the structural isomers generated by the program is in the form of a special kind of graph -AND/OR tree (25).

The ring systems in the program are constructed from vertex graphs (8), which are defined in a given problem by a series of calculations. The first level of the tree, after the specification of the initial collection of atoms, is the set of all possible partitions of the initial set of nodes. Each partition consists of the cyclic subunit and the remaining set of nodes. The cyclic subunits are a collection of atoms from which all possible ring systems can be constructed on the basis of the appropriate vertex graph. The atoms in the remaining set form acyclic parts of final structures, combined in all possible ways with the ring structures from the corresponding initial partition. The second level of the tree specifies all possible ring systems that can be constructed from the vertex graph corresponding to the cyclic subunit in the first level of the tree. The next level of the tree just beyond the node specifying a possible ring system, specifies the possible ways in which the remaining atoms can be linked to the unfilled links of the system. After the three first levels of the tree generation, the program becomes recursive. Each set of unstructured nodes is taken up as a fresh problem until there are no more unstructured nodes. The Structure Generator represents a part of a very complex and sophisticated computer program - Heuristic Dendral Program (25,8) for elucidation of molecular structure based on structural features of unknown molecules derived from chemical, physical and spectroscopic data.

Recently, a similar approach to the problem of exhaustive enumeration and generation of chemical structures was published by Sasaki and Kudo (26). Their system successfully deduce all logically valid structures, acyclic and cyclic on the basis of previously settled proposition according to the input information about the structure, of a given compound.

4. Mathematical models in the computer-aided-planning of organic syntheses

The problems of isomorphism of chemical graphs and generation and enumeration of structural isomers are closely related to the works connected with the design of chemical structure information systems. The justification of the chemical structure information systems is the assistance in the research process. It happens very often that it is not only the structure of the compound which intersets the chemist, but also the properties of the compound which the structure represents. For instance, the chemist may be interested in the synthesis of the compound, in some of its physical properties, in its behavior in a living system etc. For these reasons greater attention should be paid to the problem of collection, evaluation and correlation of the data associated with a particular compound. Closer to these desired objectives are the studies carried out in the field of the application of machine computation to the generation of chemical pathways for the synthesis of complex organic molecules. Mathematical models in the form of graphs and the associated theory in the computer-aided design of chemical synthesis are involved in two ways: first as a tool in the representation of structures stored into the programs; and second, as a model in the computer representation of synthesis pathways. Synthesis pathways can be viewed as a tree (27) in which the root is the synthetic target, the intermediates in the synthesis process are the nodes, and the chemical reactions are the edges linking the nodes of the tree.

Several alternatives to the computer-aided design have been attempted. The strongest attention from the chemists has evinced the work of the groups of Princeton and Harvard (28). Their works were based on the interactive computation with an on-line guidance by a chemist. This feature has enabled them to solve interesting chemical problems by pooling the resources of a computer with a chemist as a source of information on reaction and on strategic design decisions.

The other most significant approach is based on the Heuristic Search Paradigm of Artificial Intelligence (30) research. The program developed at Stony Brook (29) designes synthesis without the chemist's intervention using the reactions from the program reaction library and programmed design strategies.

A similar approach can be found in the works of Whitlock on the heuristic solutions of the functional group synthesis problem (32). The reaction library in his program represents the implementation of the transition graph of a finite automata, wherein the nodes are functional groups and edges are the reactions that transform one functional group into another.

The most mathematical scheme of synthesis-planning problem was suggested by a group of prominent chemists and mathematicians (31). The scheme is based on the recognition that all chemical reactions correspond to interconversions of isomeric ensembles of molecules (IEM) within a family of isomeric ensembles of molecules (FIEM) (31). Distinguishable IEM of FIEM is represented by a family of be-matrices (bond and electron matrices) $F = (M_0, M_1 \ldots, M_f)$. The be-matrix M_i of an ensemble of molecules EM_i consisting of a set A, which contain n atoms, $A = (A_1, \ldots, A_n)$ is an n x n matrix as shown below:

$$M_i = \begin{pmatrix} a_1 & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & & \cdots & a_{nn} \end{pmatrix}$$

were the entries a_{ij} $(i \neq j)$ are the formal bond orders of the bonds between pairs of atoms A_i and A_j, the diagonal entries a_{ii} are numerically equivalent to the number of free valence electrons belonging to atom A_i in EM_i.

A slightly modified version of this mathematical model of chemical systems and their relations was used as a basis for the construction of algorithm, which generates multistep syntheses of a given chemical compound. The algorithm and the former mathematical model were implemented in an organic-synthesis-planning program called HEDOS (33). The program is confined to systems consisting of the benzene ring and functional groups attached to it. Specially designed heuristic rules governing the generation of the best synthesis were incorporated into the program in view of the complexity of the synthesis pathway and the number of steps involved.

In this approach, the be-matrices of a FIEM defining metric topology, were embedded as elements of state space, that we called the state space of ensembles of molecules (33). The associated set of operators of the state space was defined as a set of reaction matrices $D(n)$; $D(n) = |-R|R$ is the reaction matrix which fits (31) the elements of FIEM|. The set of fitting matrices $F(n)$ for some state EM_i in the FIEM represented by a be-matrix B_i is obtained by the mapping γ:

$$\gamma: D(n) \times FIEM \longrightarrow F(n);\ \gamma(D(n), B_i) = |F \quad F - B_i \leq 0|$$

The target molecule Z is contained in an initial EM, denoted with EM_Z. Final states are all possible EM assigned as EM_L provided that the chemical species in EM_L are contained in the list of available chemical compounds $\mathfrak{L}$, meaning that they can be easily synthetisized or found in the commercial catalogue of the world-known suppliers of fine chemicals. The molecules in EM_L are starting materials for the synthesis of molecule Z and compose the list L, $L \subseteq \mathfrak{L}$. Thus, the problem of synthesis design for a compound Z can be reduced to the problem of finding a path K, $K = |R_1, \ldots, R_n|$ into a space of FIEM which transforms EM_Z in EM_L. The search for a path through a state space is equivalent to the travel through a directed graph, in which the nodes correspond to various EM from FIEM and edges correspond to the set of reactions D(n). The root in the directed graph which is a tree in this case is the ensemble EM_Z. The implemented synthesis-planning-algorithm in the program HEDOS generates the space of FIEM and searches for a minimal length path which leads from EM to some EM_L. This minimal path meets some prescribed criteria, which garantee the feasibility of the proposed reactions and the validity of generated structures. Information contained in the be-matrices of the EM is usually insufficient for the evaluation of the proposed reactions and intermediate structures, so additional information concerning other molecular properties was stored in the second symmetric triangle of the be-matrix. The new matrices were defined as bei matrices:

$$M_1 = \begin{pmatrix} a_{i,j} & \text{for } i \leq j & \\ i_{i,j} & \text{for } i > j & \text{additional information} \end{pmatrix}$$

The program is not interactive, i.e. the chemist cannot interrupt the program to assist in the search for synthetic intermediates or in the evaluation of the synthetic path. The program must make all decisions by itself and is strictly experimental. It was designed for the purpose of developing and testing artificial intelligence mechanism with the aid of a strongly defined topology of molecular structures and related reactions.

5. Conclusion

The recent approaches to computer-aided solving of non-numerical chemical problems have been reviewed and the merits and drawback of implemented mathematical models outlined. It seems that the use of graphs as mathematical structure in the representation of chemical compounds, as they provide a form suitable for computer manipulation, becomes more and more popular. Best results in this field was achieved in application of graph theory and permutation groups in computer programs which generate and enumerate all possible structural isomers of a given set of atoms. Thus, the problem of exhaustive isomer generation can in general be considered as solved.

The other problem, identification of chemical compounds in the information retrieval system, for the solution of which the same mathematical model was used, was not so successfully solved. The majority of information chemical systems still performe the structure and substructure searches by using logical combinations of the structure fragments, as the compounds in the system's files are presented in one of the linear notations (WLN, JUPAC-DYSON) or by different fragment codes (Mechanical Chemical Code, KWIC indexes). Both forms of representations are simple to operate. As the volume and the interdisciplinary needs of chemistry, especially in the research process have increased, and the need for fully explicit structure representation of molecules becomes essential, various chemical information systems (CAS, DARC, TOSAR) have included graphs as a form of representation of chemical notions, but only as a supplement to the files with standard records. The great deficiency of this form of representation is the time - consuming identification of compounds. The chemist and information scientists still work on the development of fast and effective graph matching techniques, as the problem is not only a chemical problem, but also a computing problem. The task is large and difficult and should require common efforts for its solution.

The computer-assisted planning of organic syntheses is just beginning. The applied mathematical models and artificial intelligence methods have exhibited many deficiences, but they can be overcome. The success of the implemented computer programs justifies the expectations that the use of the computer-assisted-synthesis analysis will become a routine in the near future. The described mathematical model of the space of ensembles of molecules, provides a useful basis for the construction of a synthesis-planning algorithm. At the same time it offers possibilities for further uses, as it is the first attempt toward a systematization of procedures for storing and handling of the vast quantity of chemical information that is currently available.

6. References

1. H.E. Zimmermann, Angew.Chem.Internat.Edit, 8 (1969) 1.
2. V. Prelog, Chem. Britain, 4 (1968) 382.
3. H.A. Schmidtke, Coord.Chem.Revs. 2 (1967); J.Chem. Phys. 48 (1968) 970.
4. S.F.A. Kettle and V. Tomlinson, J.Chem.Soc. 91 (1969) 2002.
5. H. Hosoya, Bull.Chem.Soc. Japan.44 (1971) 2332.
6. K. Ruedenberg, J.Chem.Phys. 22 (1954) 1878.
7. I. Ugi, D. Marquarding, H. Klusacek, G. Gokel and P. Gillespie, Angew.Chem. 82 (1971) 741.
8. L.M. Masinter, N.S. Sridharan, J. Lederberg, D.H. Smith, J.Am.Chem.Soc. 11 (1974) 7702; L.M. Masinter, N.S. Sridharan, D.H. Smith, J.Am.Chem.Soc. 11 (1974) 7715; R. Carhart, D. Smith, A. Brown, N.S. Sridharan, J.Chem.Infor.Comp.Sci. 15 (1975) 2.
9. A. Sachs, Publ.Math.Debrecen (1962) 270; 11 (1963) 119; L. Spialter, J.Chem.Doc. 4 (1964) 269; F. Harary, J.Math.Phys. 38 (1959) 104.
10. L.C. Ray, R.A. Kirsch, Science, 126 (1957) 814.
11. C.M. Bouman, J.Chem.Doc. 3 (1965) 92.
12. M. Randić, J.Chem.Phys. 60 (1974) 3920.
13. M. Randić, J.Chem.Infor.Comp.Sci. 15 (1975) 105.
14. J.C. Bart, N. Giordano, Proceedings of the International Conference on Computers in Chemical Research and Education, Ljubljana-Zagreb, 1973, Paper 3/1
15. H.L. Morgan, J.Chem.Doc. 5 (1965) 107.
16. W.T. Wipke and T.M. Dyott, J.Am.Chem.Soc (1974) 4825.
17. E.A. Sussenguth, Jr., J.Chem.Doc. 5 (1965) 36; E.A. Sussenguth, Jr., "Structure Matching in Information Processing, Thesis Harvard Univ., 1964.
18. T.K. Ming, S.J. Tauber, J.Chem.Doc., 11 (1971) 47.
19. R.A. Penny, J.Chem.Doc. 5 (1965) 113.
20. G. Polya, C.R. Acad.Sci. 201 (1935) 1167.
21. T.L. Hill, J.Phys.Chem. 47 (1943) 253; T.L. Hill, J.Phys.Chem. 11 (1943) 294.
22. W.J. Taylor, J.Phys.Chem. 11 (1943) 532.
23. A.T. Balaban, F. Harary, Rev.Roum.Chim. 12 (1967) 1511; ibid., 11 (1966) 1097; ibid. 12 (1967) 103.
24. J. Lederberg, "DENDRAL-64, Part I; Notational Algorithm for Tree Structures", NASA Star. No. N-65-13 158, NASA CR-57029; "Part II, Topology of Cyclic Graphs" NASA Star. No. N 66-19074, NASA Cr-68898; "Part III Complete Chemical Graphs: Embedding Rings in Trees" NASA Star. No. N 71-76061, NASA CR-123176.
25. B. Buchanan, J. Lederberg, Proceedings of IFIP Congress 1971, Ljubljana, paper TA-2-91; B. Buchanan, G. Sutherland, E.A. Feigenbaum, In Machine Inteligence 4, 5 (1969) ed. Meltzer and Michie, 209, 253; J. Lederberg, G.L. Sutherland, B.G. Buchanan, E.A. Feigenbaum, A.V. Robertson, A.M. Duffield, C. Djerassi, J.Amer.Chem.Soc. 91 (1969) 11.
26. Y. Kudo, S. Sasaki, J.Chem.Infor.Comp.Sci. 16 (1976) 1.
27. E.J. Corey, W.T. Wipke, Science 166 (1969) 178.
28. E.J. Corey, Quart.Rev. (London) 25 (1971) 455; E.J. Corey, W.T. Wipke, R.D. Cramer, W.J. Howe, J.Am.Chem.Soc. 94 (1972) 421; ibid. 94 (1972) 431.
29. N.S. Sridharan, Proceedings of IFIP Congress, 1974, Stockholm, 828-837.
30. J.R. Slagle, Artificial intelligence: The Heuristic Programming Approach, Mc Graw-Hill, New York, 1971.
31. J. Dugundji, I. Ugi, Topics in Current Chemistry, 39 (1973) 21, Springer Verlag; I. Ugi, P. Gillespie, C. Gillespie, Trans.New York Ac.Sci. 34 (1972) 416.
32. E. Blower, Jr., W. Whitlock, Jr., J.Am.Chem.Soc. 98 (1976) 1499, W. Whitlock, ibid., 98 (1976) 3225.
33. B. Džonova-Jerman-Blažič, I. Bratko, Proceedings of AFCET Congress, Gif-Sur-Yvette, 1976, 283; ibid., Proceedings of the International Conference on Information Sciecnes and Systems, Patras, 1976, Hemisphere Publishing Corporation, Washington, 461.