20 000 compounds had been described, in 1899 about 75 000, in 1910 about 140 000. The further rapid increase is illustrated by the following features: In 1940 about 400 000 compounds had been described, 1960 about 1 200 000, 1980 about 5 000 000 compounds, of which an estimated 90% belong to the realm of organic chemistry. These figures provide impressive evidence of the quantitative increase in research results produced by the scientific discipline of "organic chemistry" over the last decades. Despite all efforts of the Beilstein editorial staff to present the handbook user with a comprehensive "concentrate" of checked, reliable, and reproducible data, by critical appraisal of all published data and facts, the increased volume of primary publications has unavoidably had its impact on Beilstein. Up to the end of 1980, about 220 subvolumes of the fourth edition (main series plus four supplementary series), begun in 1918, have been published. With completion of Supplementary Series IV in 1984/1985 the total will have reached about 280 subvolumes. Supplementary Series V, which will make its appearance in 1984/1985—in English—will deal with the chemical literature between 1960 and 1980, i.e., two decades which belong to the most fruitful of all in organic chemistry. In processing the literature of this period, as other periods, the Beilstein editorial staff have to keep foremost in their minds the objectives defined by Beilstein himself in compiling the first edition, i.e., only to allow data and facts known to be reliable in terms of current scientific knowledge to appear in the handbook and so provide Beilstein users with a comprehensive concentrate of original literature, free of erroneous results and trivial information. Scientists working in the field of organic chemistry throughout the world may rest assured that in its second century, as in its first, "Beilstein" will remain a competent and dependable aid in dealing with day-to-day research problems, able to save many hours of painstaking literature searching through its systematic organization and critical assessment of the known facts.

## REFERENCES AND NOTES


(1) "How to Use Beilstein"; Springer-Verlag: Berlin, 1979 (free of charge, obtainable in English, German, and Japanese).
(2) "Beilstein Reference Chart"; Springer-Verlag: Berlin (free of charge, also obtainable in brochure form).
(3) "Beilstein-Outline", Springer-Verlag: Berlin (free of charge).
(4) "The Short Cut to Compound-Location in Beilstein"; Springer-Verlag: Berlin (poster free of charge).
(5) Luckenbach, R. "Der Beilstein", *CHEMTECH* **1979** (*10*), 612.
(6) Sunkel, J.; Hoffmann, E.; Luckenbach, R. "A Straightforward Procedure for locating Chemical Compounds in the Beilstein Handbook", *J. Chem. Educ.*, in press.
(7) Luckenbach, R.; Sunkel, J. "Das wissenschaftliche Handbuch: Ein Konzept zur Bewältigung der Literaturflut in der Chemie", *Naturwissenschaften* **1981**, *68*, 53.
(8) Walentowski, R. "Unique, Unambiguous Representation of Chemical Structures by Computerization of a Simple Notation", *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 181.
(9) The following sources may also be consulted on the history of Beilstein: (a) Hjelt, E. *Ber. Dtsch. Chem. Ges.* **1907**, *40*, 5041. (b) Adams, R. *Chem. Eng. News* **1956**, 6310. (c) Krätz, O. *Chem. Ztg.* **1970**, *94*, 115.


# Evaluation and Implementation of Topological Codes for Online Compound Search and Registration


DAVID BAWDEN,* J. TREVOR CATLOW, and TREVOR K. DEVON

Pfizer Central Research, Sandwich, Kent, CT13 9NJ, England

JUDITH M. DALTON,† MICHAEL F. LYNCH, and PETER WILLETT

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England


Received September 30, 1980

A topological search code has been found to have high discriminatory power within large sets of disparate structures. The technique has been implemented in a pharmaceutical company's computerized chemical information system, for interactive registration and structure search.

## INTRODUCTION

Structure search, the matching of a compound against a machine-readable file of chemical structures to see whether it is already present in the file, is one of the most common and essential functions in a computer-based chemical information system.[1] A routine application of this procedure is compound registration in which compounds not found by structure search, and thus identified as being novel, are incorporated into the system. If structure search is not to involve a time-consuming serial search through the whole file, some form of initial file partitioning must be carried out, such as the use of a molecular formula check in the "isomer sort" approach to registration.[2] Structures occurring in the same partition as the query compound are then compared with it by direct comparison of unique structure representations, e.g., canonical connection tables or WLNs, atom-by-atom matching in the case of non-unique representations, or by eye. It is obviously desirable for purposes of efficiency that the numbers of compounds in each partition should be as small as possible: molecular formula groups are not ideal in this respect since their sizes are highly disparate.[3]

An online system for compound registration and full structure search offers many advantages both to the operators and to the users of such systems. One approach to the provision of such facilities involves the calculation of a numeric search key for a structure representation. The search key for the query structure may then be compared with the keys for structures already in the file by using standard search techniques such as hashing or binary search. A detailed comparison is then carried out when the keys match. Howe and Hagadone described a design for such a system which involved the calculation of a hash address from a canonical connection table and subsequent visual inspection of possible matching

† Beechams Medicinals Research Centre, Harlow, Essex, CM19 5AD, England.

0095-2338/81/1621-0083$01.25/0 

structures,[4] while Freeland et al. used a similar procedure but with atom-by-atom searching for the final matching.[5] Evans et al. studied the use of topological indexes as search keys for nonunique structure representations.[2] The basis of their work was an index devised by Randic to describe the degree of branching in alkanes and was based on the first-order connectivities of pairs of adjacent atoms in a molecule.[6] This index was extended to include higher order connectivities by using an adaption of the Morgan algorithm[7] and to encompass different atom and bond types.

This paper reports the evaluation of this index, with various parametrizations, on large sets of diverse structures, incorporating subsets of structurally similar compounds, e.g., β-lactams, from the Pfizer Central Research (UK) internal files, and its subsequent incorporation into a computer system for interactive registration and structure search.

## EXPERIMENTAL DETAILS

The index originally proposed by Randic, $I$, is given by

$$I = [({}^1d_i \times {}^1d_j)]^{-1/2}$$

where ${}^1d_i$ and ${}^1d_j$ represent the first-order connectivities of the connected pair of atoms $i$ and $j$ and where the summation is over all of the bonds in the molecule.[6] The elaborations on this index developed by Kier and co-workers for structure–activity correlation[9] are not appropriate for structural codes of the sort tested here.[2] The two most effective indexes tested by Evans et al. were

$$I_{2AB} = [({}^2d_i a_i \times {}^2d_j a_j)b]^{-1/2}$$

and

$$I_{3AB} = [({}^3d_i a_i + {}^3d_j a_j)b]^{-1/2}$$

where $a_i$ and $a_j$ are integers describing the atomic types of $i$ and $j$, ${}^2d_i$, ${}^2d_j$, ${}^3d_i$ and ${}^3d_j$ are the second- and third-order connectivities of $i$ and $j$, and $b$ is the order of the bond connecting them.[2] Note that the third-order index is a sum, whereas the second order is a product.[2] Evans et al. arbitrarily chose 3, 5, and 7 as parameters for the atomic types C, N, and O, respectively, these being the only atomic types considered. In this work, three types of parameterization were tested for the representation of the much wider range of atomic types: these were the atomic numbers in the range 1–100, prime numbers in the range 1–523, and prime numbers in the range 2–541. Prime numbers were tested, as potentially more discriminating, i.e., avoiding fortuitous "false clashes", than sequential integers. Two sets of prime numbers were used to investigate the sensitivity of discrimination to alternative parametrizations. The connection tables used by Evans et al. used bond orders of 1, 2, and 3 for acyclic single, double, and triple bonds, respectively, with all ring bonds being given an order of 1. This was found to cause problems in the partitioning of the molecular formula groups tested. The connection table used in this work gave bond orders of 1, 2, and 3 to single, double, and triple bonds, respectively (both cyclic and acyclic), and 4 to aromatic bonds, and these values were used for the bond parameters.

Three kinds of data set were tested. Initial experiments used the $C_8H_8N_2O_3$ and $C_{10}H_{10}O_2$ molecular formula groups from the Formula Index of the Chemical Abstracts 8th Collective Index (1967–1971), omitting polymers, stereoisomers, salts, indefinite compounds, dimers, and addends: the groups contained 80 and 192 structures, respectively. Secondly, three sets of 500 structures each were selected from the beginning, middle, and end of the Pfizer compound file to typify small sets of compounds of diverse structure. Finally a larger proportion of the Pfizer compound file was used for the evaluation of two particular search keys. The total numbers of structures involved were 30 538 and 30 563, respectively.

**Table I.** Summary of Results for the Molecular Formula Groups: Numbers of Structures with Identical Codes[a]

| | | $C_8H_8N_2O_3$ | | $C_{10}H_{10}O_2$ | |
|---|---|---|---|---|---|
| bond values | atom values | $I_2$ | $I_3$ | $I_2$ | $I_3$ |
| 1-3 (4 = 1) | C = 3, N = 5, O = 7 | 1 pair | none | 3 pairs | none |
| 1-4 | C = 3, N = 5, O = 7 | 1 pair | none | 2 pairs | none |
| 1-4 | C = 11, N = 13, O = 17 | 1 pair | none | 3 pairs | none |
| 1-4 | C = 13, N = 17, O = 19 | 1 pair | none | 2 pairs | none |
| 1-4 | C = 6, N = 7, O = 8 | 1 pair | none | 2 pairs | none |

[a] $C_8H_8N_2O_3$–80 structures, $C_{10}H_{10}O_2$–192 structures.

For all the data sets, the structures were input to the computer as Wiswesser Line Notations (WLNs), which were then converted automatically to redundant, noncanonical connection tables. These tables were used for the generation of the search codes which were then sorted into order so as to identify the occurrence of duplicate codes, with six significant figures used for comparison. The connection tables generated did not take account of salt forms, stereoisomerism, and other structural factors encoded after the "space-hyphen" or "space-ampersand" of the WLN, and compounds differing only in these factors were therefore assigned identical indexes. For the sets of 500 Pfizer structures, these "valid clashes" were distinguished from those clashes arising from assignment of the same code to distinct connection tables. This was done automatically in the case of the salt forms and otherwise by visual inspection. In the case of the larger Pfizer sets, only salt form matches were eliminated. All processing was carried out by using FORTRAN programs on a PDP 11/45 computer, with double precision arithmetic. Fuller details of the experiments are given by Dalton.[8]

## RESULTS AND DISCUSSION

Initial experiments were carried out with the two molecular formula groups to test the effectiveness of the various types of atom and bond parameters. The summarized results are shown in Table I. The WLN to connection table routine used did not deal with all the structures in these groups, and hence the numbers of structures used are smaller than for these molecular formula groups in earlier work.[2] Good partitioning is seen in both groups, with $I_3$ superior to $I_2$, as found originally. The bond parameters used here are at least as effective as the original. The $I_{3AB}$ code was found to be quite insensitive to the alternative parametrizations used, and no two compounds in either of the groups were found to have equal index values. Index clashes with the $I_{2AB}$ index were due to positional isomers. The use of bond parameters compared with the other two types of parameters.

Bond parameters 1–4 were used for experiments with the sets of 500 Pfizer compounds. If valid matches for differing salts, stereoisomers, etc., were discounted, very few index clashes occurred at the six decimal places comparison.

In the first group of 500 structures, one pair of different structures was not separated by the $I_2$ index (with the 1–523 parameter set) only. A further pair were given identical values for all six index/parameter combinations. These structures were $R_1O(CH_2)_3NH(CH_2)_2OR_2$ and $R_1O(CH_2)_2NH(CH_2)_3OR_2$, respectively. In the second group of 500 structures, one pair of structures was falsely matched by the $I_3$ index (with atomic number parameters) only. In the third group of 500 structures, one pair of structures was falsely matched by the $I_2$ index (with the 1–523 parameter set) only. The high discriminatory power of these indexes, regardless of atom parametrization, on sets of diverse structures is demonstrated by these results.

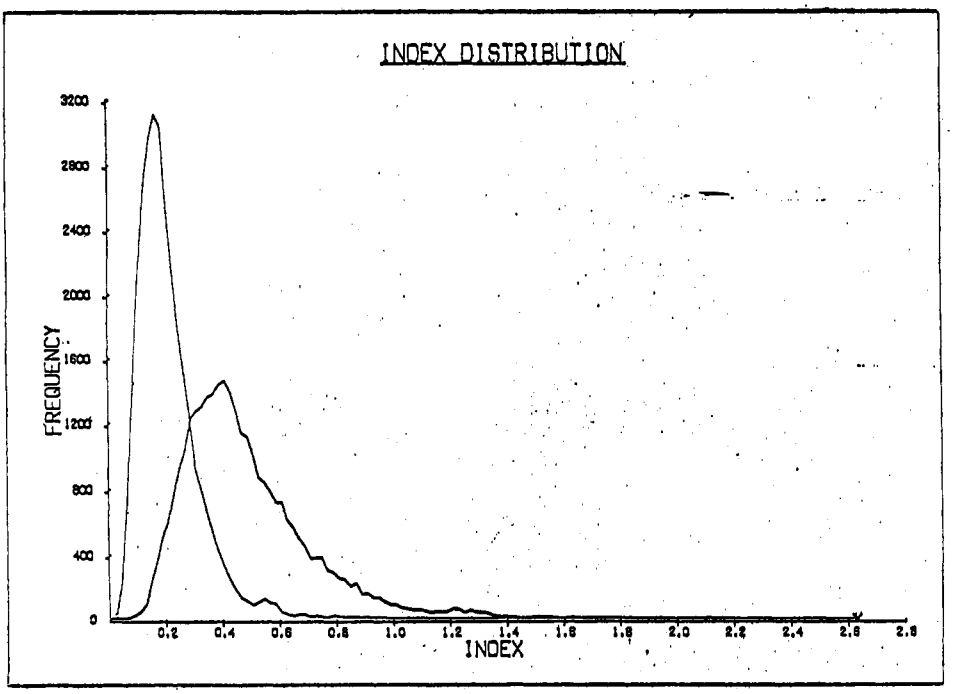


**Figure 1.** Distribution of search keys for the $I_{3AB}$ code by using atomic number (right-hand curve) and prime number (left-hand curve) atomic parameters.

**Table II.** Large Pfizer Structures Sets: Numbers of Groups of Structures with Identical Codes ($I_3$)

| no. of structures | atom parameters | nos. of groups of particular size | | | |
|---|---|---|---|---|---|
| | | 2 | 3 | 4 | 7 |
| 30 563 | 2-541 | 270 | 19 | 2 | 1 |
| 30 538 | 1-100 | 274 | 19 | 2 | 1 |

Evans et al. had proposed the use of a topological index as an adjunct to the molecular formula so as to differentiate between molecules in the same molecular formula group: the results with the three sample sets suggest that the indexes were sufficient to discriminate between quite diverse structures without the use of an initial molecular formula check.

Only the $I_{3AB}$ index was used for generating search keys for the large Pfizer compound sets, this index being based on bond orders 1-4 and atom parameters 1-100 and 2-541. The difference between the two parametrizations was small: the results for the prime numbers are presented in Table II. In this table, the groups of structures with matching indexes include both valid matches and erroneous ones where different structures have equal index values. The figures demonstrate the high resolving power of the index with about 98% of the compounds in the file possessing a unique index value and with only a minimal occurrence of large groups of compounds with equal index values.

The frequency distributions of the second- and third-order indexes, with all parametrizations, were virtually identical over all groups of structures. The prime number sets gave "peaked" distributions, with the great majority of indexes falling between 0 and 1. The atomic number sets gave a flatter distribution with a long tail of infrequent high key values. The third-order distributions with atomic number and prime number (2-541) parameters for the larger Pfizer compound sets are shown in Figure 1.

## IMPLEMENTATION

As a result of the success of this evaluation, an interactive structure-matching system based on these indexes has been introduced as an integral part of the Pfizer Central Research (U.K.) internal data bank system. This system is currently operational on a VAX 11/780 computer.

Third-order indexes, with atomic number parametrization, have been generated for approximately 120 000 compounds in the Pfizer internal files for which a WLN can be assigned and a connection table generated.

For structure matching, an index is generated from a WLN entered interactively and compared against the sorted file by an efficient binary search technique. WLNs corresponding to matching indexes are displayed at the terminal. Response is virtually instantaneous, and the very low percentage of clashes make the procedure highly effective.

This technique has been incorporated within the daily interactive compound registration process, with the index used for novelty matching after checking and validation of the WLN. It is also utilized for specific compound searches, providing much more rapid access than previously possible. Because of the high discrimination of these indexes, a molecular formula check has not been found necessary in compound searches: this could be readily introduced, if required, with larger files. Since the indexes are not dependent on a canonical representation, the method is useful in allowing detection of WLN coding errors.

## CONCLUSIONS

The topological search codes discussed here were originally developed to discriminate only between molecules within a given molecular formula group. It has now been shown that the codes are sufficiently specific to discriminate between large numbers of disparate structures and between large subsets of compounds of similar structure, e.g., penicillins, present in the file. Their usefulness as search keys in online compound registry systems is thereby demonstrated. The advantage of the approach chosen in comparison with other registration methods is that the codes can be generated from any unambiguous chemical structure representation even though it may not be unique.

The successful implementation of this technique within a pharmaceutical company's chemical information system demonstrates its effectiveness for interactive registration and structure searching.

ACKNOWLEDGMENT

We thank C. J. Bridgeman, D. A. Faulkner, and S. R. Morgan for their assistance.

REFERENCES AND NOTES

(1) J. E. Ash and E. Hyde, Eds., "Chemical Information Systems", Chichester, Ellis Horwood, 1975.
(2) L. A. Evans, M. F. Lynch, and P. Willett, "Structural Search Codes for Online Compound Registration", *J. Chem. Inf. Comput. Sci.*, **18**, 146-149 (1978).
(3) J. H. R. Bragg, M. F. Lynch, and W. G. Town, "The Use of Molecular Formula Distribution Statistics in the Design of Chemical Structure Registry Systems", *J. Chem. Doc.*, **10**, 125-128 (1970).
(4) W. J. Howe and T. R. Hagadone, "Progress towards an Online Chemical and Biological Information System at the Upjohn Company" in "Retrieval of Medicinal Chemical Information", *A.C.S. Symp. Ser.*, **84**, 107-131 (1978).
(5) R. G. Freeland, S. A. Funk, L. J. O'Korn, and G. A. Wilson, "The Chemical Abstracts Service Registry System. II. Augmented Connectivity Molecular Formula", *J. Chem. Inf. Comput. Sci.*, **19**, 94-98 (1979).
(6) M. Randic, "On Characterization of Molecular Branching", *J. Am. Chem. Soc.*, **97**, 6609-6615 (1975).
(7) H. L. Morgan, "The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service", *J. Chem. Doc.*, **5**, 107-113 (1965).
(8) J. M. Dalton, "Evaluation of the Application of a Topological Index to Compound Registration at Pfizer Central Research", M.Sc. thesis, University of Sheffield, 1979.
(9) L. B. Kier and L. H. Hall, "Molecular Connectivity in Chemistry and Drug Research", Academic Press, New York, 1976.

# Quantum Chemistry Literature Data Base

Y. OSAMURA[†]

Department of Chemistry, Osaka City University, Sumiyoshi-ku, Osaka 558, Japan

S. YAMABE

Faculty of Education, Nara University of Education, Nara 630, Japan

F. HIROTA

Department of Chemistry, Shizuoka University, Ohya, Shizuoka 442, Japan

H. HOSOYA

Department of Chemistry, Ochanomizu University, Bunkyo-ku, Tokyo 112, Japan

S. IWATA

Institute of Physical and Chemical Research, Wako, Saitama 351, Japan

H. KASHIWAGI and K. MOROKUMA

Institute for Molecular Science, Myodaiji, Okazaki 444, Japan

M. TOGASI,[‡] S. OBARA,[§] K. TANAKA, and K. OHNO*

Department of Chemistry, Hokkaido University, Sapporo 060, Japan

Received August 1, 1980

The quantum chemistry literature data base (QCLDB) contains literature concerning ab initio computations of atomic and molecular electronic structures. Approximately 2000 literature references published from Jan 1977 to June 1979 have been collected from 19 internationally well-known core journals. Keys to references are computational methods, basis sets, and calculated properties and printout is by author and compound indexes.

## INTRODUCTION

With many ab initio computations of atomic and molecular electronic structure appearing in many journals, chemists, experimental and theoretical, who would like to know and utilize the results of such calculations often have a difficult time in finding proper references. Of late, computer-based information retrieval systems are becoming available, but it is not easy to satisfy both specialists and nonspecialists who want to make an overall survey as well as obtain some specific information on the available calculations for compounds of interest. Richards' famous book series[1-3] is quite useful in that each molecule is treated separately. Molecules are ordered according to their size, and by looking up a particular molecule, one can find all ab initio calculations published up to a certain date as well as additional information on the geometry, type of calculation, energy, and computed properties (except in the last book of the series[3]). In spite of these merits, it takes at least a year for the bibliography to be published in a book form, and therefore current references cannot be covered.

In this project of the Quantum Chemistry Literature Data Base (QCLDB), a major emphasis is placed on current awareness as well as exhaustiveness. The potential for computer searching is desirable. Therefore, the information in QCLDB should be created in a computer-readable form.

[†] Department of Chemistry, University of California, Berkeley, CA 94720.
[‡] Department of Physics.
[§] Institute for Molecular Science, Myodaiji, Okazaki 444, Japan

0095-2338/81/1621-0086$01.25/0 