A Screen Set Generation Algorithm

PETER WILLETT

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

Received November 28, 1978

A general procedure is outlined for the description of chemical substructures by strings of integers. These strings may be manipulated to produce sets of approximately equifrequently occurring fragment screens for use in chemical substructure search systems.

THEORETICAL CONSIDERATIONS IN THE DESIGN OF SCREEN SETS

Atom-by-atom matching of a substructural query against a file of connection tables, a subgraph isomorphism search, belongs to the class of problems known as NP-complete[1] for which no efficient algorithms are known and thus may require excessive amounts of computer time if many trial structures need to be compared with the query. Such searches are accordingly feasible only if the number of matches can be reduced by the rapid and inexpensive elimination of that large portion of the file not satisfying certain minimal requirements in the query.[2] We use the term screen set to describe the group of structural characteristics which is used to carry out this initial partitioning.

For even quite small files of compounds the total number of potential screens is very large indeed,[3] and hence strict criteria must be used to determine which features should be selected for use in the screen set. Lynch and his co-workers showed that the highly variable frequencies of the fragments of a given type, e.g., augmented atom or bonded pair, may be compensated for by employing several levels of description, the frequently occurring characteristics being delineated in some detail in the final screen set while the less common features are described in more general terms.[2,4,5] In this way, a balance may be achieved between the proliferation of low incidence fragments of superfluous specificity and the small number of high-incidence, low-precision fragments. Also, the occurrences of the resultant screen set members will become much less disparate than if a single level of description were to be employed, in accordance with simple considerations of information theory.[6] The move toward screen equifrequency may, however, be lessened by the need to describe frequent characteristics at the more general levels, as well as in detail, to allow easy query encoding since, otherwise, the union of many highly specific features may be required in order to describe a more general feature common to all of these. This suggests the need for a hierarchically ordered screen set in which there are well-marked relationships between the fragments at different levels of description.

A second problem is that the screenout performance of a screen set cannot be predicted accurately from fragment incidence data since it is found that the incidences of the screen set members are not independent of one another. Thus an analysis of the co-assignment frequencies of pairs of screens showed that the association between fragments of a given type increased with the type size; however, the study concluded that, in practice, no consideration need be given to such fragment associations as long as the screen set members were not too large.[7] Additionally, iterative fragmentation procedures introduce very strong associations between a fragment and its immediate parent, i.e., the fragment obtained in the previous iteration of the algorithm from which the new fragment has been derived. It is clear that if the two incidences are not dissimilar the filial fragment is redundant and should not be included in the screen set. A theoretical description of such associations has been developed by Hodes[8] and applied to the formation of a screen set in use at the Walter Reed Army Institute of Research (WRAIR).[3]

DESCRIPTION OF CIRCULAR SUBSTRUCTURES BY INTEGER STRINGS

The investigations outlined in the previous section employed an iterative fragment generation procedure whereby fragments were initially considered at a very simple level of description, e.g., element type for atom-centered fragments, and then more detailed descriptions were produced for the more commonly occurring features in subsequent processing of the file; thus in the Sheffield studies a frequent simple pair would be considered for inclusion at the augmented pair level so that the hierarchial nature of the fragment types studied was reflected in the method of screen selection.[4] The fragments considered at WRAIR covered a much wider range of substructural sizes, but a similar procedure was adopted in that any fragments occurring in more than 1% of the structures in the file were used to generate additional filial fragments in the subsequent iteration of the fragmentation algorithm.[3]

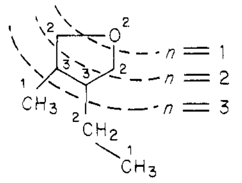
The work described here represents an alternative method for algorithmic screen set generation in that it is possible to produce a screen set from a single pass of the structure file. The screens may be atom- or bond-centered, are symmetric (in the sense that they extend outward equally in all directions from the center of the fragment), and form a strong hierarchy. They are also inexpensive to generate and to assign since no path tracing is required even for the largest substructures that need to be considered for inclusion in the screen set. The procedure involves three stages, these being the generation of all possible fragments at the most specific level of description, cumulation of the individual fragment occurrences (this being combined with the measurement of the frequencies of the more general features from which they are derived), and then selection of certain of the fragments for inclusion in the screen set upon the basis of the frequency and association considerations outlined above.

It is clear that a large amount of sorting will be required to obtain the cumulated fragment frequencies although the computational requirements are much reduced by the indirect calculation of the frequencies of all but the most specific fragments so that only these latter substructural types need be sorted prior to cumulation. Even so, the overall process will be most efficient if the fragment representations are chosen to be as compact as possible to allow rapid sorting. Such an approach will also bear fruit in the initial fragment generation since simple fragment representatives are likely to be simple to produce. While it is relatively easy to encode small fragments, the exact description of a large substructure, such as a nongeneralized octuplet, requires some form of connection table as the representative; this implies that the record must be converted into a canonical format with the minimum of effort[9] and that the total volume of data to be sorted will be large owing to the bulk of each record. Hence, since we wish

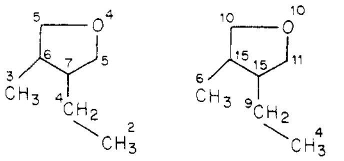
0095-2338/79/1619-0159$01.00/0

to generate fragment representatives only at the most detailed level of description, the main problem is one of producing a canonical, compact fragment while retaining easy access to more generic fragments contained within the substructure.

The fragment descriptors we have used are strings of integers, each successive integer representing a more precise definition of the environment of the feature described by the initial member of the string. The integers were obtained by an adaption of the Morgan algorithm.[10] This was developed to discriminate between the atoms within a molecule upon the basis of their extended connectivity values where the *n*th order connectivity of an atom is calculated by summing the $(n - 1)$th order connectivities of all immediately adjacent atoms. Consider the molecule shown below where the numbers attached to each atom represent the initial connectivity, i.e., the number of attached atoms.



The second and third order connectivities are shown below.



If we consider the oxygen atom we may describe it by the string of integers (2, 4, 10) where the *n*th integer represents a generalized description of the topological environment within $(n - 1)$ bonds of the central atom. As mentioned in earlier work using connectivity-derived descriptors,[11,12] the *n*th integer may be thought of as a hash of that portion of the molecular connection table which describes the corresponding circular substructures of radius $(n - 1)$ bonds; the three integers thus represent the features contained within the concentric radii illustrated above.

Thus far, the descriptors are of purely topological origin, but their discriminatory power may be increased by the use of additional characteristics, specifically atom type and bond order, in the calculation of the initial property values; such information can be included by various means and a specific implementation is given in the fourth section of this paper.

Given the initial set of property values, higher order descriptions may be generated as many times as desired so that it is necessary to define the maximum level of structural detail that is required. Hence if the largest substructures to be represented are of radius 4 bonds, five-integer strings will be produced. The smaller, more generic fragments may then be easily obtained by successively replacing the right-hand-most integer by zero; thus the oxygen string mentioned earlier, (2,4,10), will yield the strings (2,4,0) and (2,0,0).

The advantages of this method of substructural description are the compactness of the representation, the possibility of extracting parent fragments with the minimum of difficulty, and the fact that no path tracing need be involved to characterize the larger substructures; all of these factors help to reduce the computational cost of fragment generation and manipulation. The procedure is also applicable to bond-centered descriptors by the use of the bond order as the initial property value with the second and subsequent values being calculated from the multiplication, addition, or some other manipulation, of the first and subsequent integers describing the atoms at each end of the bond. Two possible disadvantages should be noted. Firstly, it is possible that some of the longer strings may describe several parent substructures. This is characteristic of hashing techniques, but we have not experienced any difficulty in this respect to date, which implies that the number of collisions has been kept to a low level; a discussion of this point is given by Unger in the context of a graph isomorphism detection algorithm.[13] Secondly, the screens are intelligible only in machine terms insofar as it is not generally possible to reconstruct manually the substructure corresponding to a particular screen representation; instead the screens have been designed for automatic generation from an input connection table so that the user need know little or nothing of the characteristics of the screen set.

SCREEN SET GENERATION ALGORITHM

In this section we describe how the integer strings may be used to produce an approximately equifrequently occurring screen set. The first step involves the analysis of the connection tables in the structure file, or some subset thereof. For each molecule, atom- and/or bond-derived integer strings are built up to the maximum level of description that is required and then written out to tape. This implies that the subsequent cumulation, both of these detailed fragments and the more general ones from which they are derived, will be upon an occurrence, rather than an incidence, basis; an alternative procedure would be to generate the strings at all levels of description and then to write out a single occurrence of each string type, this course resulting in accurate incidence figures at the cost of a considerable increase in the number of records that must be sorted. The use of occurrence statistics results in a slight overemphasis on the shorter strings in the final screen set. While this may result in lower equifrequency properties for the assignment of screens to the structure file, it should mean that a wider range of small screens is available for assignment to substructural queries which will tend to be smaller than the compounds against which they are to be matched.

Once all the connection tables have been processed, the set of strings is sorted into order so that all descriptions of the same substructure appear together on the tape. These occurrences are then summed for each fragment and a simultaneous count made of the less specific fragments which may be derived from it; the resulting ordered fragment occurrence dictionary is used as a basis for the screen set selection program.

The derivation of sets of approximately equally frequently occurring sets of attributes has been thoroughly investigated in the context of bibliographical information systems where the objects are textural in nature, e.g., document index terms or author names, and the attributes are strings of alphabetic characters.[14-18] In one algorithm, character strings are generated from the text by moving along it one character at a time and producing a fixed-length string at each point, the length being equal to the largest string that may be included in the final set of attributes. The string occurrences are summed together with the frequencies of the substrings that may be obtained by successive right-hand character truncation; thus the string "COMPUTE" will yield "COMPUT", "COMPU", etc., down to "C". The technique is clearly analogous to the generation of integer string substructural descriptors, and hence these may be processed using one of the text-oriented algorithms referenced above with the minimum of modification.

The procedure makes use of a threshold frequency, *T*, below which strings will not be considered for inclusion in the final screen set. Knowing the total number of atoms, *T* is calculated from the desired screen set size, this being the sole program parameter. The size is usually one less than a multiple of 24; the computer used for this work has a 24-bit word, and a single

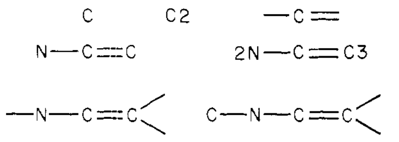


Figure 1. Typical substructures described by *n*-integer strings (n = 1–7).

bit is reserved for use as a conflated screen, i.e., one which may be assigned if no match can be obtained for a substructure with any of the other screens in the set. Alternatively, one could ensure that all the single integer types are included in the final set.

A trial screen set is obtained by including in the set all those single integer strings with associated frequencies $\geq T$, and during subsequent iterations, the sets of strings of a given size are considered for inclusion in the set to improve its equifrequency properties. Consider a general string, S_i, of the length n whose parent fragment S, of length $(n - 1)$, was included in the set created at the end of the previous iteration. S_i will be stored as a potential new screen if both its frequency, f_{S_i}, and the difference in frequency between it and its parent are not less than T; i.e., $f_{S_i} \geq T$ and $f_S - f_{S_i} \geq T$. The presence of the parent is dictated by the need for a strict fragment hierarchy while the latter requirement is to remove the strong parent–filial associations described by Hodes.[8] At the end of each iteration, i.e., at the end of the processing of the strings of size n, the set of potential new screens emanating from a given parent string are added one by one in decreasing frequency order so long as their immediate parent's frequency remains above T, i.e., while $f_S - \sum f_{S_i} \geq T$. After all the strings have been inspected in this manner, a note is made of the size of the current set and, if greater than that required, the least frequent strings are removed until the desired size is achieved. A more detailed description of the procedure is given in ref 19.

IMPLEMENTATION OF THE PROCEDURE

We shall exemplify the technique by a screen set generated from CROSSBOW connection tables for use in an experimental chemical reactions retrieval system.[19] The heart of the CROSSBOW record is the units section which consists of a string of symbols, each of which is associated with one of the nonhydrogen atoms in the molecule.[20] These symbols describe the type of an atom and the number and types of the surrounding bonds, i.e., a bonded atom, and we have used the binary representation of the symbol corresponding to the atom as its initial property value.

Successive integers in a string increase the area of substructural description outward by one bond in all directions, this implying a large increase in fragment size and a concomitant proliferation in fragment variety.[21] This is undesirable since, apart from the potential overlap with other fragments,[7] only a very few screens may be assignable to small query substructures. Accordingly, two property values are inserted in the string before the CROSSBOW-derived integers, these initial values corresponding to the elemental type and the type plus degree of coordination. Higher order values for $n > 3$ are then calculated by summing the $(n - 3)$th values of all adjacent atoms. The features delineated are hence in the regular order shown in Figure 1, the first four corresponding to atom type, coordinated atom, bonded atom, and augmented atom. If these four atomic values are then used to obtain bond-centered descriptors, the progression simple pair, augmented pair, bonded pair, and octuplet is obtained. We have studied only the substructures shown in the figure, but related subspecies, such as the various fragments which may be obtained from an augmented atom by deletion of one or more of the pendant atoms, could be described by summing only over certain of the attached atoms when calculating higher order property values.

In the present implementation, higher order values up to $n = 7$ are calculated, but only the five integer strings corresponding to $n = 3$ to $n = 7$ are written out to tape for subsequent sorting so that the minimal level of description in the final screen set is the bonded atom.

A simple measure of the effectiveness of the selection procedure may be obtained from the relative entropy, which describes the degree of the equifrequency of the screens in the set.[17] Analysis of a sample file of 8078 compounds yielded a total of 160 294 atom-centered strings, and these were used to produce a 240-member screen set. This had a relative entropy of 0.953 which may be compared with the relative entropy of the original different unit symbols, i.e., the bonded atom set, which was 0.798. This increase is at least comparable to that observed in textural studies, but it should be noted that a large amount of a priori selection has already been carried out in the design of the CROSSBOW symbol set since frequently occurring atom types may be assigned several different symbols. This is done to reflect the variety of bond surroundings that need to be taken into account for adequate discrimination while rarer features are generalized by the use of a single symbol; consequently, the degree of equifrequency is already quite high. To illustrate the point, the corresponding 240-member bond-centered set had an initial relative entropy of 0.603 and a final one of 0.967, a much larger increase. It is clear that the screen selection procedure can remove a high proportion of the frequency variation in the data at little cost since an Algol68 implementation required less than 30 seconds cpu time using the University of Sheffield ICL 1906S computer to generate both screen sets from the cumulated fragment dictionaries.

Query encoding uses a connection table as input, and this is processed as above to produce integer strings which may then be matched against the screen set. If a match is not obtained for a query string at the maximum level of description, the string has its final, nonzero integer replaced by zero and the assignment procedure called again. The process continues until a matching string is found, when the appropriate bit is set in the query bit string, or the conflated screen is assigned. An unsatisfied connection in a query substructure connection table is filled by the use of a dummy atom with an initial property value not corresponding to any of the symbols in the CROSSBOW set. This being so, not only will no match be obtained if we search such an atom against the screen set, but we shall also not assign screens corresponding to substructural features larger than that explicitly delineated by the query since the contribution of a dummy atom to the property values will reduce the chance of a match being obtained for a string describing a substructure which contains that atom.[19]

CONCLUSIONS

In this paper we have outlined a general method for describing circular chemical substructures[22] by strings of integers. These strings may then be manipulated, using well-established methods, to produce sets of approximately equifrequently occurring screens for use in chemical substructure search systems. The screens may be used to describe either atom- or bond-centered features, are symmetrical, and form a strongly linked hierarchy. They are also extremely cheap to generate and to assign since no path tracing is required for even the largest substructures.

ACKNOWLEDGMENT

Thanks are due to ICI Ltd. (Pharmaceutical Division) for CROSSBOW software, to David Cooper and Michael Lynch for helpful discussions, and to the Department of Education and Science for the award of an Information Science Research Studentship.

REFERENCES AND NOTES

(1) R. E. Tarjan, "Graphic Algorithms in Chemical Computation", *Am. Chem. Soc. Symp. Ser.*, No. 46, 1–19 (1977).

(2) M. F. Lynch, "Screening Large Chemical Files", in "Chemical Information Systems", J. E. Ash and E. Hyde, Eds., Ellis Horwood, Chichester, 1975.

(3) A. Feldman and L. Hodes, "An Efficient Design for Chemical Structure Searching. I. The Screens", *J. Chem. Inf. Comput. Sci.*, **15**, 147–152 (1975).

(4) G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, and A. M. Yapp, "Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files", *J. Chem. Doc.*, **13**, 153–157 (1973).

(5) M. F. Lynch, "The Microstructure of Chemical Data-Bases and the Choice of Representation for Retrieval", in "Computer Representation and Manipulation of Chemical Information", W. T. Wipke, S. R. Heller, R. J. Feldman, and E. Hyde, Eds., Wiley, New York, 1973.

(6) C. E. Shannon, "A Mathematical Theory of Communication", *Bell Syst. Tech. J.*, **27**, 379–423, 623–656 (1948).

(7) G. W. Adamson, D. R. Lambourne, and M. F. Lynch, "Analysis of Structural Characteristics of Chemical Compounds in a Large Computer-Based File. Part III. Statistical Association of Fragment Incidence", *J. Chem. Soc. C*, 2428–2433 (1972).

(8) L. Hodes, "Selection of Descriptors According to Discrimination and Redundancy. Application to Chemical Structure Searching", *J. Chem. Inf. Comput. Sci.*, **16**, 88–93 (1976).

(9) M. Bersohn, "Rapid Generation of Reactants in Organic Synthesis Programs", *Am. Chem. Soc. Symp. Ser.*, No. 61, 128–147 (1977).

(10) 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).

(11) L. A. Evans, M. F. Lynch, and P. Willett, "Structural Search Codes for On-Line Compound Registration", *J. Chem. Inf. Comput. Sci.*, **18**, 146–149 (1978).

(12) M. F. Lynch and P. Willett, "The Automatic Detection of Chemical Reaction Sites", *J. Chem. Inf. Comput. Sci.*, **18**, 154–159 (1978).

(13) S. H. Unger, "GIT—a Heuristic Program for Testing Pairs of Directed Line Graphs for Ismorphism", *Commun. ACM.*, **7**, 26–34 (1964).

(14) D. Cooper and M. F. Lynch, "The Compression of Wiswesser Line Notations Using Variety Generation", paper published in this issue.

(15) P. W. Williams, "Criteria for Choosing Subsets to Obtain Maximum Relative Entropy", *Comput. J.*, **21**, 57–62 (1978).

(16) E. J. Schuegraf and H. S. Heaps, "Selection of Equifrequent Word Fragments for Information Retrieval", *Inf. Storage Retr.*, **9**, 697–711 (1973).

(17) M. F. Lynch, "Variety Generation—a Reinterpretation of Shannon's Mathematical Theory of Communication and Its Implications for Information Science", *J. Am. Soc. Inf. Sci.*, **28**, 19–25 (1977).

(18) M. F. Lynch, Principal Investigator, "Comparison of the Efficiency of Bibliographic Search Codes", British Library, Research and Development Department Report No. 5422, 1978.

(19) P. Willett, "Computer Analysis of Chemical Reaction Information for Storage and Retrieval", unpublished Ph.D. thesis, University of Sheffield, 1978.

(20) J. E. Ash, "Connection Tables and Their Role in a System", in J. E. Ash and E. Hyde, Eds., ref 2.

(21) G. W. Adamson, M. F. Lynch, and W. G. Town, "Analysis of Structural Characteristics of Chemical Compounds in a Computer-Based File. Part II. Atom-Centred Fragments", *J. Chem. Soc. C*, 3702–3706 (1971).

(22) Alternative descriptions of circular substructures have been described by several authors, but the methods of feature description and manipulation are very different; see, e.g., J. E. Dubois, "Ordered Chromatic Graphs and Limited Environment Concepts" in "Chemical Applications of Graph Theory", A. T. Balaban, Ed., Academic Press, London, 1976; W. Schubert and I. Ugi, "Constitutional Symmetry and Unique Descriptors of Molecules", *J. Am. Chem. Soc.*, **100**, 37–41 (1978); M. Randic, "Fragment Search in Acyclic Structures", *J. Chem. Inf. Comput. Sci.*, **18**, 101–107 (1978).

Plausible Paths in the Rearrangement Reaction of Polycyclic Hydrocarbons Searched by the Graph-Theoretical Method and Computer Techniques[†]

NOBUHIDE TANAKA and TADAYOSHI KAN

Department of Physics, Faculty of Science, Gakushuin University, Mejiro, Tokyo, 171, Japan

TAKESHI IIZUKA*

Department of Chemistry, Faculty of Education, Gunma University, Maebashi, Gunma, 371, Japan

Received August 11, 1978

We have formulated a graph-theoretical concept "transmutation" which corresponds to the change of the skeleton of a molecule. In order to see the relations between two given graphs, we have devised two algorithms, mono-source and di-source propagation algorithms. We applied these algorithms to a transmutation process corresponding to adamantane and diamantane rearrangements and obtained relationships between the transmuted graphs. In the di-source propagation algorithm, we show the "shortest paths" which correspond to the plausible paths of the rearrangement reaction.

We have studied graph-theoretical and computational isomer enumeration,[1] representation of molecular structures,[2] and analysis of rearrangement paths of polycyclic hydrocarbons.[3] In this paper we focus on methods of studying the rearrangement paths of polycyclic hydrocarbons by graph theory and computer methods.

Balaban et al.[4] first studied graph theoretically 1,2 shifts of acyclic hydrocarbons and tried to classify the rearrangement reactions.[5] A study of the cyclic hydrocarbon adamantane, based on graph theory and a computer method, was reported by Whitlock and Siefken.[6] They found the rearrangement paths from twistane to adamantane by graph-theoretical predictions. Using calculations by molecular mechanics, Schleyer et al. challenged the diamantane rearrangement.[7] We think, however, that the rearrangement reaction paths become more complicated as the number of carbons and/or rings increases. The complexity is far beyond the limit of manual work. Thus it is essential to devise efficient computer methods for this problem.

We define the graph-theoretical concept, which we call "transmutation", as corresponding to chemical rearrangement, and devise computer methods to obtain plausible[13] information without imposing physicochemical conditions. We have established two algorithms to find relationships between isomers

[†] A part of this paper was presented by N. T. in the 26th IUPAC Congress.[3d]

0095-2338/79/1619-0162$01.00/0