Beilstein Ring Search System. 1. General Design

László Domokos

Beilstein Institute, Varrentrappstrasse 40-42, 60486 Frankfurt/Main, Germany

Received March 16, 1993[⊗]

One of the most significant advantages of an online database is the possibility of accessing the data in many different ways using different types of search criteria. The Beilstein Online Database (Domokos, L. *Mikrochim. Acta* **1986**, *II*, 423–429), which is the largest factual data collection of organic chemistry, has more than 400 different search keys. It is available on two online hosts, Dialog and STN. The online experience has shown that the most frequently used access methods are those which select compounds by their structural properties, like full structure and substructure, chemical name, chemical name fragments, molecular formula, molecular weight, atom counts, and Lawson number. The Beilstein ring search system (BRSS) was developed to allow an easy and fast retrieval of compounds by exact or fuzzy definition of their embedded ring systems. This paper gives a general overview of the BRSS itself and of the new database fields supporting the BRSS. An early stage of the development has been reported previously (Domokos, L. In *Software Development in Chemistry*; Gasteiger, J., Ed.; Springer-Verlag: Berlin, 1990; pp 31–41).

INTRODUCTION

The Beilstein database of organic chemical compounds has been online since 1988. This rapidly growing database currently (October 1992) contains structural, chemical, physical, and bibliographical data of 5 million organic compounds. The data can be accessed via several hundred different searchable fields. The most frequently used fields in search terms are those of structural properties.

The Beilstein ring search system (BRSS) offers additional possibilities for accessing the data by using structural information, namely, by specifying the ring systems. The ring systems constitute an especially important class of substructures. The BRSS deals with the description and retrieval of the complete ring systems of organic compounds. It provides simple, well-defined, and fast retrieval. The user can describe a ring system in many ways: exact, by specifying its atom and bond types, and variable, by allowing any bond or any atom types, by giving only the formula of the ring system, etc. The ring system can be defined by a graphical structure input, or by specifying several of the numerical and text BRSS fields, or by combination of them. A significant aspect is the capability of browsing the BRSS fields, like the Beilstein ring index (BRIX), for example. The browsing gives a quick overview of the ring systems contained in the database, which can be helpful in formulating the proper BRSS or other type (e.g. substructure search) queries.

Important design requirements were simplicity and user friendliness. Therefore, in order to avoid the need for using and learning new interfaces, the online host's own structure input software can be used. Almost all BRSS fields are simple and self explanatory: *number of ring systems*, *ring system formula*, *number of multiple bonds*, etc. Parameters describing the elementary rings of the smallest set of smallest rings are not used. The reason is that in case of more complicated ring systems these descriptors are not easy to determine by hand.

In contrast to some other special purpose ring coding methods, e.g. see ref 3, the BRSS aims to handle all ring systems of the organic compounds in the Beilstein Information System. Recently, Nilakantan et.al.[4] have published a ring based query system which is able to handle large databases by using an easy-to-calculate hash code.

This paper focuses on the technical part of the BRSS. A detailed discussion of possible applications and examples are given in ref 5.

DISCUSSION

The BRSS consists of three major components: the *code generation* part, which includes the coding software; the generated *database field*; the *retrieval* part, i.e. the interface for query input, the query coding, and the retrieval software.

After releasing the BRSS, the retrieval part is embedded in the online host's own environment. Therefore, although the basic features are identical, the BRSS may have different looks and capabilities on the different systems.

Code Generation. The code generation creates the BRSS related database fields. This procedure uses the Beilstein description of the chemical structures, the SDF.[6] It is important to mention that only the three basic structural information are used: the connectivity, i.e. the "from" and "ring closure" SDF lists; the atom types, which are given by the "atom type" SDF list; the bond types which are coded in the "π-electron" SDF list. All other information, like stereochemistry, abnormal masses, charges, etc., is ignored. This simplifies the system and ensures a broader retrieval. Similarly to the Beilstein registry system,[6] there is no special consideration for tautomerism.

Extraction of the Ring Systems. The first step of the code generation is the recognition and extraction of complete ring systems from the structure. The procedure can be described as follows:

(a) Remove all "cutting bonds". A bond is called "*cutting bond*" if after removing it from the structure there is no path along the remaining bonds which connect the two end atoms of the removed bond.

(b) Remove the atoms which became isolated unconnected atoms after deleting all cutting bonds. These removed atoms are the atoms of the "*acyclic part*" of the structure. The remaining atoms constitute the "*cyclic part*". The connected

[⊗] Abstract published in *Advance ACS Abstracts*, August 15, 1993.

0095-2338/93/1633-0663$04.00/0

Chart I

Compound Level Fields.

Compound type (heterocyclic)

The compound type is either acyclic, isocyclic, or heterocyclic corresponding to the three large compound classes of the Beilstein system.

Strong fragment similarity, SFS (231768)

Weak fragment similarity, WFS (204509)

The definition of fragment similarities is given later.

Component Level Fields. The component fields are generated for each different component.

Acyclic part formula. ($C_2O_2S_2$)

Cyclic part formula. ($C_{22}N_4O_3$)

The cyclic and acyclic formulas are given without hydrogen atoms.

Component size. (29)

The component size is the number of non-hydrogen atoms.

Number of ring systems. (4)

The total number of complete ring systems of the component.

Number of different ring systems. (3)

Number of elementary rings in the component. (6)

The number of elementary rings is equal to the number of bonds minus number of atoms plus 1.

Number of hetero atoms in the cyclic part. (7)

Number of hetero atoms in the acyclic part. (4)

Ring System Fields. The ring system fields are generated for each different ring system of a component.

Size of the ring system, i.e. number of atoms (14)

Ring system formula (C_{14})

Similarly to the cyclic and acyclic formulas, the ring system formula is given without the attached hydrogen atoms.

Ring system code, RSCODE (binary string)

Beilstein ring index, BRIX (14.3.30-0.0-6.3)

The ring system code and index are discussed below.

Number of elementary rings in the ring system. (3)

Number of hetero atoms. (0)

Number of multiple bonds (6)

It is calculated as the number of pi-electrons divided by 2.

Multiplicity (1)

The number of occurrences within the component.

Number of residues (3)

It defines the number of the non-hydrogen atoms attached to the ring system. The BRSS does not allow to define the exact positions or types of the residues. The proper tool to do this is the substructure searching.

Number of residues attached to heteroatoms. (0)

subsets of the cyclic part constitute the ring systems of the compound.

The BRSS deals only with complete ring systems defined by the above algorithm. From this definition it follows, for example, that a spiro system is considered as a ring system. Subrings (elementary rings) of complete ring systems are not described separately. Hence, for example, there is no BRSS field which says, that the ring system of naphthalene has two 6-member subrings.

The ring systems are handled independently of their structural environment. The only field related to this environment is the number of attached residues.

In case of multicomponent compounds each component is processed separately. Identical components are considered only once.

BRSS Related Fields. Most of the generated fields are very simple. This enables the chemist to define search criteria using these fields easily.

The BRSS fields can be divided into three groups: (i) compound level fields, (ii) component level fields, and (iii) ring system level fields. The first two groups are "byproducts" of the ring system coding. Some of the fields are not strictly related to the ring systems of the compound.

A list of all generated BRSS fields is given in Chart I. The values given in parentheses are examples and correspond to the structure shown on Figure 1 and its largest ring system, respectively. The cutting bonds of the structure are drawn with dotted lines. The cyclic part consists of four ring systems. The acyclic part consists of four systems, namely, C, O, O, and S–S.

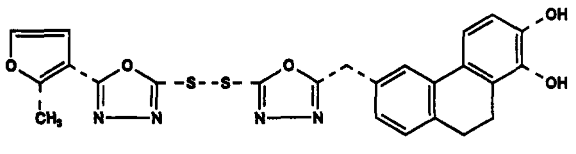

Figure 1.

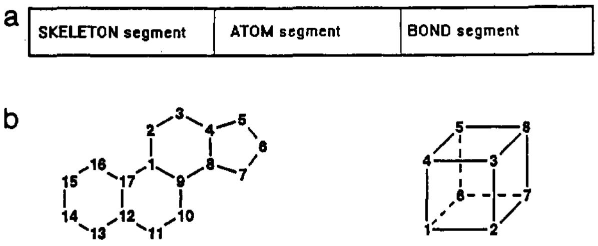

Figure 2. (a) Structure of the ring system code. (b) Examples for "chain numbering".

The BRSS fields, except for the ring system code (RSCODE), the Beilstein ring index (BRIX), and the fragment similarities are self-explanatory and will not be discussed here. The RSCODE and BRIX, which are the most powerful fields, are discussed below.

Both the RSCODE and the BRIX are unique codes of a particular ring system. There is a one to one correspondence between the RSCODE and the BRIX. The difference lies in the way they are generated, and used.

Ring System Code. The (RSCODE) is an algorithmically generated variable length binary string. The RSCODE gives a compact, on average 5–6 byte long, unique description of a particular ring system. The RSCODE consists of three segments, the *skeleton*, the *atom*, and the *bond segment*. The RSCODE is the concatenation of these segment strings (Figure 2a). The segments correspond to the three basic structural information used for BRSS, namely, the connectivity, the atom type, and the bond type. The skeleton segment describes the connectivity of the ring system. The skeleton segment code is not influenced by the atom and bond types of the ring system. For example, all 6-member single rings, e.g. benzene, pyridine, cyclohexane, etc., have the same skeleton segment code.

The atom segment describes the position and type of the atoms of a ring system within a particular skeleton. It is not affected by the bond types. The atom segment can only be used together with the skeleton segment. Ring systems with different skeletons might have the same atom segments with different meaning.

The third segment, the bond segment, identifies the bond types for ring systems with a particular skeleton and atom types. The bond segment is related to the skeleton and atom segments. It cannot be used without these two segments.

The coding of the ring systems is based on a unique numbering of the atoms. The numbers are assigned by analyzing the skeleton. If there are only symmetrical atoms in the skeleton relative to the previously assigned numbers, then the atom type information is used to decide to which atom the next number is assigned. If a unique candidate cannot be selected, then the bond type information is used. If even at this point no selection can be made, then the next number is assigned randomly to one of the candidates.

Unlike the well-known Morgan numbering,[7] the numbering need not be cooperative, which requires that the next number is assigned to one of the neighbors of the lowest numbered atom with unnumbered neighbors. This makes the algorithm considerably faster.

After the complete structure, in our case the ring system, has been numbered, it is very easy to renumber the atoms so that the resulting numbering is efficient for the further processing. It can be renumbered to a cooperative or to a canonical numbering, for example. However, in order to keep the skeleton code segment short, the BRSS renumbering algorithm tries to find a "chain numbering". A numbering is called a chain numbering if the subsequent numberings are along a spanning chain of the skeleton. A chain numbering can be found in 95% of all cases. Even relatively complicated ring systems, like cubane and steroid skeletons, can be chain numbered (Figure 2b). An example where a chain numbering is not possible is the skeleton of porphyrin. The advantage of chain numbering is obvious. The skeleton coding can be reduced to code the ring system size and the ring closures. If no complete chain numbering is found, then the longest possible chain numbered subchain is used. The single rings are treated separately, their skeletons coded simply by the number of atoms.

The atom and bond codes are created such that the more general information is coded before the details. This can be exploited later to build a powerful retrieval using the proper leading substrings of the code. The atom segment contains the information in the following sequence: number of heteroatoms, number of different heteroatom types, list of heteroatoms, counters of heteroatoms, and position of heteroatoms. As a result it gives identical heads of the heterocode segments for a particular ring skeleton and formula, for example.

The resulting RSCODE is unique for the ring system. Furthermore, the code is reversible; i.e. the ring system can be reconstructed from its code. There are only a few exceptions of very large ring systems where the reversibility was sacrificed in order to reduce the maximal possible code length to 64 bytes.

The code is compact; the average code length is 5.6 bytes. With the sacrifice of some properties, a further reduction to an average of 3.7 bytes could be achieved. One of the properties is that the segments end on a byte end. This allows the leading substrings of RSCODE to be used to retrieve classes of ring systems, e.g. all ring systems with a particular skeleton. The RSCODE supports the following three basic searches: (i) specified skeleton with any atom and any bond types, (ii) specified skeleton and atom types with any bond types, and (iii) specified skeleton, atom, and bond types, i.e. a fully defined ring system. Using more complicated bit substrings offers further options. For example, ring systems with specified skeleton and ring system formula with any positions of heteroatoms and with any bond types can be retrieved.

The advantage of the RSCODE is that it can be regenerated from the query ring system and it, or its leading substrings, can be used to retrieve the specified ring system. The RSCODE is completely transparent to the user, who can define it implicitly by defining the ring system structure by graphical structure input or directly by its ROSDAL string.[8]

Due to the algorithm and to the fact that there are only relatively few complicated ring systems (70% of all ring systems are single rings and 19% consist of two elementary rings), the code generation is very fast. The processing of 2 million compounds (195 304 acyclic and 1 804 696 cyclic structures), including the ring system extraction, generation of all BRSS fields, 3.45 million ring codes and additionally 6.23 million codes of acyclic parts, needed 9.5 h on an IBM 9121-480 machine.

Beilstein Ring Index. The BRIX is the counterpart of the RSCODE. In one aspect the BRIX is similar to the RSCODE

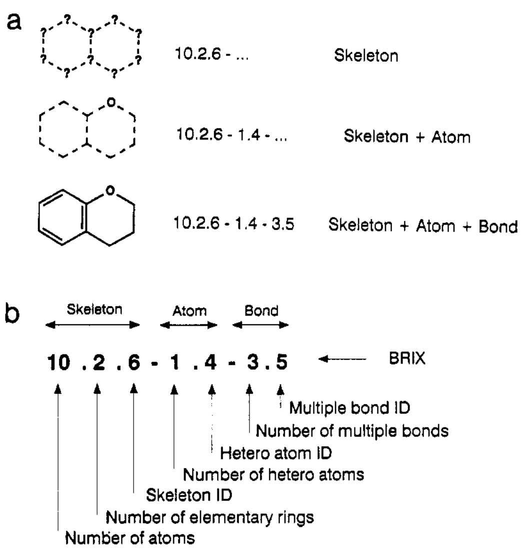

Figure 3. (a) Example for the BRIX segments ↔ structure relationship. (b) Structure of the Beilstein ring index, BRIX.

Table I. Statistics on the Distribution of Ring Systems in the Beilstein Structure File

feature	no.
no. of structures	5 325 850
no. of acyclic structures	507 399
no. of isocyclic structures	2 089 501
no. of heterocyclic structures	2 728 950
no. of all ring systems	8 817 792
no. of all ring systems without multiplicity	7 026 417
no. of different ring systems	113 079
no. of different skeletons	19 801
no. of different skeletons contained in more than	
1 structure	12 496
10 structures	3 531
100 structures	726
1000 structures	135

Table II. Most Frequent Ring Systems, Number of Compounds Containing Them, and Their Ring Indices

ring system	compound	BRIX
benzene	2694 220	6.1.0-0.0-3.1
cyclohexane	166 389	6.1.0-0.0-0.0
pyridine	160 087	6.1.0-1.1-3.1
piperidine	117 822	6.1.0-1.2-0.0
tetrahydropyran	114 445	6.1.0-1.2-0.0
naphthalene	113 457	10.2.6-0.0-5.1
tetrahydrofuran	82 606	5.1.0-1.2-0.0
pyrrolidine	70 583	5.1.0-1.1-0.0
chinoline	64 569	10.2.6-1.2-5.1
morpholine	58 968	6.1.0-2.11-0.0
cyclohexene	57 846	6.1.6-0.0-1.2
steroid system with single bonds		
hexadecahydrocyclopenta {*a*}phenanthrene	57 160	17.4.32-0.0-0.0

Table III. Number of Different Ring Systems Belonging to the Same Skeleton

skeleton of	no.	BRIX
benzofuran	4917	9.2.5-
naphthalene	2957	10.2.6-
	2529	13.3.26-
	1940	14.3.30-
	1861	14.3.12-
bicyclo{3.3.0}octane	1698	8.2.5-
	1582	13.3.13-
cyclohexane	1576	6.1.0-
	1563	13.3.33-
cyclopentane	1364	5.1.0-
	1105	12.3.31-
steroids	1078	17.4.32-
bicyclo{5.4.0}undecane	1037	11.2.6-

because the BRIX (i) is a variable length string, (ii) is unique for each particular ring system, and (iii) consists of skeleton, atom, and bond segments. On the other hand they are different because the BRIX (i) is a readable string, (ii) cannot be generated algorithmically from the ring structure, (iii) is not reversible, and (iv) is directly available to the user as a search and display field.

The BRIX is the result of a registration of the RSCODEs. As the examples of Figure 3a,b show, the three segments of the BRIX are separated by hyphens. Each segment is divided further by one or two dots. The skeleton segment consists of the number of non-hydrogen atoms and the number of elementary rings and of a skeleton identification number. The atom segment contains the number of heteroatoms and an atom identification number. Finally, the bond segment contains the number of multiple bonds and the bond identification number. Hence the BRIX consists of seven numbers, two hyphens and four dots. At first glance it seems to be a rather complicated string. In fact, it is fairly simple. The segmentation facilitates the effective use of the BRIX.

The number of atoms, of elementary rings, of heteroatoms, and of multiple bonds can be determined easily. The skeleton, atom and bond identification numbers are assigned by the ring registration process of the BRSS. Therefore these numbers cannot be generated algorithmically from the structure. The user must know them to use the complete BRIX. The BRIX can be determined easily by displaying a compound and the corresponding BRIXs. If the structure contains more than one ring system, the known elements of the BRIX, i.e. number of atoms, elementary rings, heteroatoms, and multiple bonds, help to assign the correct BRIX to the corresponding ring system. After the BRIX of a ring system is obtained, it can be used very easily in searches.

The BRIX can be used effectively both for retrieval and for browsing. Similarly to the RSCODE, the leading substrings, i.e. the right truncated form of BRIX, can be used for searching the database. For example, BRIX = "10.2.6-4.*" would find all compounds containing at least one ring system with a naphthalene skeleton with four non-carbon atoms at any position and with any bond. The asterisk stands for right truncation.

Most of the systems dealing with rings try to arrange the ring systems according to some similarity measure, e.g. ring ID numbers[9] or measure of complexity[4]. The BRIX works ideally for browsing, (expand command in the online databases) through the ring systems of the database. This is a remarkable tool, because except for browsing the molecular formula, Lawson number, or chemical name segment fields, it is not possible to get a quick overview of the database from a structural point of view. The BRIX offers a completely new way. For example, to get information about the large ring systems, larger than 100 atoms, of the database, it is enough to browse the BRIX starting at "100". The number of heteroatoms, multiple bonds, etc., given in the BRIX help to get a finer overview. This might be useful for formulating a proper substructure, BRSS, or other type of queries. Detailed examples are given in ref 5.

The evaluation of the generated BRSS fields provides an interesting overview of the database. Tables I–III show several statistics on the distribution of ring systems in the October 1992 version of the Beilstein structure file. The file contained 5 325 850 structures.

The tables show, that the 4 818 451 cyclic structures (90.5%) of the database contain a total of 8 817 792 ring systems, or

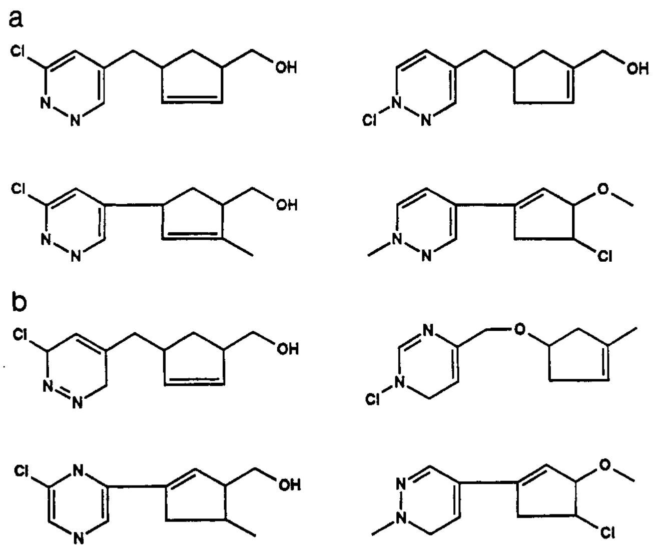

Figure 4. (a) Strongly fragment similar structures. (b) Weakly fragment similar structures.

7 026 417 if each ring system is counted only once per component. There are 113 079 different ring systems belonging to 19 801 different ring skeletons. It is interesting to note that 7305 (36.9%) ring skeletons are only in one compound, and only 726 (3.7%) skeletons are in more than 100 compounds. The skeleton of benzofurane (6-5 ring) has the most variations; 4917 different ring systems have this skeleton. Even the very simple and most common skeleton of a single 6-member ring is represented with 1576 different atom–bond combinations.

A comparison with the results published by Stobaugh[10] (Tables III, VI, and VII) shows good agreement regarding the distributions of acyclic and cyclic compounds, the number of ring systems per compounds, the most frequent ring system skeletons, and the most frequent ring systems.

The complete file was processed in several portions corresponding different literature periods. The comparison of the results shows clearly the obvious trend, namely, that the recent publications contain more and larger ring systems than the older ones. This trend underlines also the increasing importance of ring systems and the need for their effective handling.

Acyclic Systems, Weak and Strong Fragment Similarities. As mentioned above, the code generation divides the structure into a cyclic and an acyclic part. The connected subsets of the cyclic part are the ring systems. Similarly, the connected subsets of the acyclic part are called acyclic systems. The acyclic systems can be coded by almost the same algorithm as the cyclic system. Hence, fields characterizing the acyclic systems, e.g. acyclic system code ASCODE, acyclic system index BAIX, can be derived. The only difference is, that instead of the number of elementary rings the number of branching atoms is used. An atom is called a branching atom if it is connected to more than two non-hydrogen atoms.

Analogously to the retrieval of ring systems, the generated fields can be used for retrieving structures containing the specified acyclic systems. The generated cyclic and acyclic codes, or indices, provide an easy way to create useful clusterings of the structures of the database. The procedure can be outlined as follows:

(a) The cyclic and acyclic codes (or indices), describe the ring and acyclic systems of the structure uniquely. Hence, by concatenating the codes in ascending order the resulting string provides a unique description of the embedded systems (fragments) of the structure disregarding the connection between these systems.

(b) The strings obtained can be rather complicated and long. Therefore it is reasonable to replace them with simple integer numbers. It is done by a kind of registration process which assigns identical numbers to identical strings and the next unused number to a new string. The assigned integer number is called a "strong fragment similarity" number, SFS.

(c) The SFS defines classes of compounds. Compounds with identical SFS number are called strongly fragment similar. Positional isomers are strongly fragment similar, for example.

A similar procedure which uses instead of the complete code only the skeleton code and the molecular formula of the embedded systems results in the "weak fragment similarity" number, WFS. The WFS is invariant to moving the hetero atoms and multiple bonds within an acyclic or ring system.

Figure 4a shows several strongly fragment similar, and Figure 4b shows weakly fragment similar compounds, respectively.

SUMMARY

The BRSS is a powerful system for coding and retrieving complete ring systems effectively. The system is designed with the Beilstein database; however, it is not limited to it. It could be applied to any database containing structural information in some form of atom–bond connectivity.

REFERENCES AND NOTES

(1) Domokos, L. Data in Beilstein-Online. *Mikrochim. Acta* **1986,** *II,* 423–429.

(2) Domokos, L. Keys to the Beilstein Database (A Ring Searching Algorithm) In *Software Development in Chemistry 4*; Gasteiger, J., Ed.; Springer Verlag: Berlin, 1990; pp 31–41.

(3) Klingebiel, U.; Specht, K. Automatic Generation of the Chemical Ringcode from a Connectivity Table. *J. Chem. Inf. Comput. Sci.* **1980,** *20,* 113–116.

(4) Nilakantan, Ramaswamy; Bauman, N.; Haraki, K. S.; Venkataraghvan, R. A Ring-Based Chemical Structural Query System: Use of a Novel Ring Complexity Heuristic. *J. Chem. Inf. Comput. Sci.* **1990,** *30,* 65–68.

(5) Sunkel, J. The Beilstein Ring Search System, BRSS II. Applications. Manuscript in preparation.

(6) Domokos, L. The Beilstein Registry System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1991,** *31,* 320–326.

(7) Morgan, H. L. The Generation of a Unique Description for Chemical Structures. *J. Chem. Doc.* **1965,** *5,* 105–113.

(8) ROSDAL, Representation of Structure Description Arranged Linearly, internal documentation, Beilstein Institute.

(9) Randic, M. Ring ID Numbers. *J. Chem. Inf. Comput. Sci.* **1988,** *28,* 142–147.

(10) Stobaugh, R. E. Chemical Abstracts Service Chemical Registry System. 11. Substance-Related Statistics: Update and Additions. *J. Chem. Inf. Comput. Sci.* **1988,** *28,* 180–187.