Messenger and S⁴: A Comparison of Structure Search Systems

Andreas Barth,* Ulla Westermann, and Beate Pasucha

STN International, c/o FIZ Karlsruhe, D-76012 Karlsruhe, FRG

Received December 14, 1993®

In the past few years several large structure oriented databases have become available through on-line hosts. These databases need special software in order to perform substructure searching. STN International has been using the Messenger software successfully for a number of years to search the Chemical Abstracts Service Registry file, and for some years this software has also been used to search the Beilstein and Gmelin on-line databases. The Beilstein database is also available on DIALOG using the S⁴ structure search software which has been developed jointly by the Beilstein Institute of Organic Chemistry together with the software company Softron Ltd. The two structure search systems have been compared by FIZ Karlsruhe with respect to the architecture, functionality, and performance of the systems. In our analysis the applicability of the two systems is discussed with respect to the structure searchable databases available on the two on-line hosts.

INTRODUCTION

The retrieval of information from databases has evolved from the access of simple bibliographic and patent data to the searching of more complex entities like numeric data (e.g. physical entities) and chemical structures. In the areas of chemistry and patent databases, substructure searching has become a very important tool both for in-house and for on-line chemical information systems (for an overview see refs. 1-3 and references cited therein). In the area of in-house systems there are very powerful database management systems available which can handle chemical structures in two or three dimensions, stereochemistry, and chemical reaction information. The major in-house systems which are commercially available are MACCS-II (Molecular ACCess System) and REACCS (REaction ACCess System; MACCS and REACCS are chemical reaction database management systems from Molecular Design Ltd.),[4] DARC In-house from Questel,[5,6] and S⁴ from Softron Ltd.[7,8].

In the on-line domain there are only a few hosts which offer a capability for structure searching based on the topology of the chemical substances. Major hosts offering this feature are CIS with the SANSS System,[9,10] Questel with its DARC software,[11] DIALOG with S⁴,[12] and STN International with its Messenger software.[13,14]

In this paper the structure search systems of STN International (Messenger) and of DIALOG (S⁴) are compared with respect to architecture, functionality, and performance. [For technical reasons and access restrictions to the DIALOG software, an in-house version of S⁴ has been used for the analysis. The difference in the performance of the systems is assumed to be neglectable.] In a previous study performed by Hicks et al.,[7] four different structure search systems have been analyzed, but the comparison has been restricted to the performance aspect of structure searching.

ARCHITECTURE AND QUERY PROCESSING

Messenger System. The Messenger system is a general purpose retrieval system which supports all kinds of on-line databases like bibliographic, full-text, thesaurus, topological, and graphical files. It has been developed by Chemical Abstracts Service (CAS), and it is used by the on-line service STN International, the *S*cientific and *T*echnical *N*etwork which is operated jointly by CAS (Columbus, OH), FIZ Karlsruhe (Germany), and JICST (Tokyo, Japan). CAS as a major producer of chemical databases is offering several large topological files to the chemical community, and hence they have a strong need for a sophisticated topological (structure) search system.

The architecture of the Messenger system supports both mainframe and workstation based structure file handling. The latter uses a parallel searching algorithm that allows the distribution of the search file on a set of parallel processing search engines running under the UNIX operating system.[14] Due to this architecture it was possible to enhance the performance of structure searching considerably with respect to mainframe files. This was especially important for the CAS Registry file with more than 12 million substances. The search engine architecture is very flexible for structure files since it is possible to keep the performance of the system at the same rate by adding more search engines when the file size increases. Text and numeric files are also supported by this architecture, and the CA file has already been moved from the mainframe computer to search engines. Since the price/performance ratio for these machines is improving quite fast, it can be assumed that the response times can be reduced stepwise by purchasing new computer equipment every few years.

Substructure searching is equivalent to a subgraph isomorphism, and all algorithms for this problem are Np complete.[15] This means that the time which is required for a substructure search is increasing exponentially with the number of nodes in the structures to be compared. In order to obtain acceptable search times, the search process is normally divided into two parts. In the first part the structure search file is screened using a filter, and in the second part the actual comparison of the structures is performed. Chemical Abstracts Service developed a structure search system where the structure queries were processed in two steps, i.e. screening and atom-by-atom matching (ABAM).

The basis for the screening is a dictionary of approximately 2000 structure fragments (screens) which have been chosen by the BASIC group in Switzerland[16,17] for their in-house screen dictionary, and it has been adopted by CAS for the Messenger structure search system.

Each search engine has a copy of a subset of the structure search file, and the structure search is performed parallel on all search engines. A detailed description of this system has

® Abstract published in *Advance ACS Abstracts*, May 15, 1994.

0095-2338/94/1634-0714$04.50/0

been published elsewhere.[14] At the time of this analysis, the CAS Registry file is accessible on search engines and the Beilstein file is available on both the mainframe computer and the search engines. The Beilstein file has been implemented originally on the MVS based mainframe computer, but due to performance considerations it has been transferred recently to search engines, and it has been released at the end of 1993.

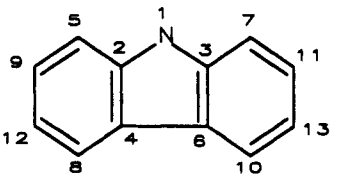
The structure queries can be specified in a graphical form which is converted into a connection table by the system. Following this, the query is compiled and the structure fragments (screens) are generated and sent to the screen process. The screens are indexed in an inverted file, and the screening process is an inverted index search. [It is not necessary to store the screens in an inverted file. An alternative approach is the use of bit vectors. In this case a bit vector is constructed for each structure where each position in the vector is associated with a given screen. A binary "1" means the screen is present in this structure, and a binary "0" means it is absent. The bit vector of the input structure (substructure) is compared against the table of the bit vectors of all structures stored in the structure file.] The result of the screening is an answer set of potential hits, i.e. a superset of the final answer set from the structure search.

After completion of the screening process, the individual answer sets are sent to the associated iteration engines. The iteration engines pick up the connection tables of all candidates which have been found in the screening process and are stored on this specific search engine. At this stage the algorithm must compare all atoms and bonds of the query structure and the file structures. It is clear that this part is very CPU time consuming since this is an Np complete problem, especially if the stored structures are complex and the structure file is very large. For the CAS Registry file with more than 12 million substances, it is very important that the structure search algorithm is able to perform the atom-by-atom and bond-by-bond matching in an acceptable time frame, e.g. in less than a few seconds for the average search query.

An important capability of the Messenger software is the ability to perform Markush structure searches.[18] A Markush structure is a structure with generic parts, variable points of attachment, and/or parts which cannot be expressed in a purely graphical form. While most structure search systems are able to deal with generic structures as user input, there are only two systems available which can match generic queries against generic file structures. These two systems are Markush DARC from Questel[19] and Messenger[20-22] from STN International. A search of a generic structure query in a file of specific substances like Beilstein or Registry is actually not a Markush structure search.

The Messenger software also allows one to manage chemical reactions and to retrieve those reactions on the basis of a search of the reaction sites. Currently, there are three reaction files on-line available on STN International (CASREACT, ChemInformRX, and CHEMREACT). The handling of chemical reactions is an extension of substructure searching, and it is based on the same type of algorithm. In the meantime, most features required for a reaction retrieval system have been implemented in Messenger.[23]

S⁴ System. *S*oftron's *s*ubstructure *s*ystem (S⁴) is a pure structure search system (search engine). It may run independently (e.g. under TSO from IBM) or it may be integrated into a retrieval system environment (e.g. DIALOG). There is also a CD-ROM version available, and a completely re-engineered version for workstation computers is under development. For the analysis we have used the TSO version of S⁴, and, for comparison, the software was also integrated in Messenger and accessible through the Messenger RUN command interface. Both versions of the software were implemented to search the Beilstein database with 3.5 million substances. The algorithm of the S⁴ software has been described in detail in a recent paper[24] and in a summarized form in ref 25.



$B0_1$	$B1_1$	$B2_1$	B
$B3_1$	$B4_1$	$B5_1$	R
$B6_1$	$B7_1$	....	N

Figure 1. Example molecule and its corresponding bit string. Here, atom 1 has been chosen as the starting atom.

$B0_1$	$B1_1$	$B2_1$	$B3_1$	$B4_1$	....	BRN
$B0_1$	$B1_1$	$B2_1$	$B3_2$	$B4_2$	....	BRN
$B0_1$	$B1_1$	$B2_1$	$B3_2$	$B4_3$	....	BRN
$B0_1$	$B1_1$	$B2_4$	$B3_4$	$B4_4$	....	BRN
$B0_5$	$B1_5$	$B2_5$	$B3_5$	$B4_5$		BRN

Figure 2. Example of a search file for the S⁴ software.

In order to perform searches with the S⁴ software, it is necessary to set up the files in a special way. At first, the atoms in the connection tables are numbered according to the extended Morgan algorithm. The whole molecule is divided into spheres beginning with the first atom (starting atom). All atoms which are connected to the same node build a bundle. In this way, there is a hierarchy of screens describing an atom and its environment, i.e. the environment of a given atom up to the *n*th sphere. Beginning with the first atom the complete structure is coded as a bit string which contains the complete description of the substance (see Figure 1) plus the Beilstein Registry Number. In the next step, the atoms are renumbered and atom 2 becomes the starting atom. Now, this atom becomes the center for the generation of the bit string. This procedure is continued until all non-hydrogen atoms and their environments have been encoded. It is clear that this procedure generates a large number of connection tables. However, these connection tables are stored in a compressed form, and the required storage is of the same size as that for the Messenger search file. Thus, for each substance the structure has been encoded as often as there are non-hydrogen atoms in the compound. The bit strings are sorted and stored in a compressed form. Finally, the search file may look like that shown in Figure 2.

In the next step the bundle files are generated from the search file. The starting atom belongs always to the *zero*th bundle. The setup of the bundle files is done in such a way that the first bit string of the search file is completely stored in the bundle files. According to our illustration (see Figure 3), the bundles $B0_1$, $B1_1$, $B2_1$, $B3_1$, and $B4_1$ are stored in the bundle files 0–4. Each bundle has a unique pointer to its corresponding bundle in the next bundle file. When the second structure is loaded, the bundles $B0_1$, $B1_1$, and $B2_1$ are already present and it is not necessary to store them again. The next bundle is $B3_2$ which is unequal to $B3_1$, and therefore it must

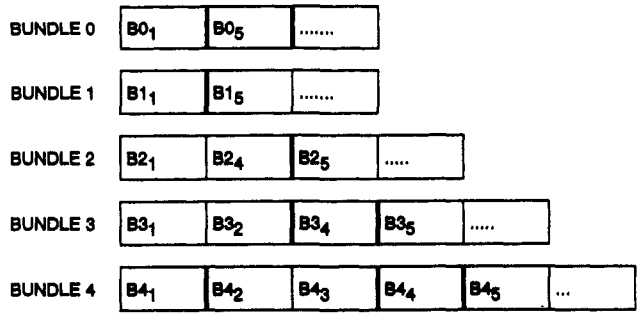


Figure 3. Structure of the bundle files 0–4.

be stored in the third bundle file. It is stored in the same interval as $B3_1$, and all intervals in the higher bundle files are closed. Bundle $B4_2$ leads to the opening of a new interval in the fourth bundle file, and $B3_2$ must have a pointer to this bundle. In this manner all the bit strings are stored in the bundle files. Only those bundles are stored which are not yet present in any previous open intervals. As soon as a bundle has been loaded, all the intervals of the higher bundle files are closed and the subsequent bundles for this bit string (substance) are loaded. This is necessary in order to keep unique pointers between the bundle files. In this way all the bitstrings of the search file can be reproduced from the bundle files up to the tenth bundle (sphere) with respect to the starting atom. For practical reasons only 10 spheres are allowed, and thus the software generates 10 bundle files plus a zeroth bundle file which contains only the starting atoms.

The structure searching consists of a screening part with 11 "bundle" files (0–10) and an atom-by-atom search called ABAS. The structure query is entered as a so-called ROSDAL string [ROSDAL is the abbreviation for *r*epresentation of *o*rganic *s*tructure *d*iagrams *a*rranged *l*inearly].[26] which is converted into a bit string by the S[4] software. The starting atom for this bit string is chosen in such a way that it is as selective as possible. A careful choice of the starting atom is important for the performance of the structure searching. The screening phase starts with a comparison of the zeroth bundles of the query structure and the file structures (see Figure 3). As a result one obtains the matching zeroth bundles with their pointers to all the entries in the first bundle file, which may match the corresponding bundle of the query structure. Now, the set of entries in the first bundle file are matched against the first bundle of the query structure, and the result is a set of matching first bundles with their pointers to the entries in the second bundle file. The matching bundle from the first bundle file is appended to the matching bundle of the zeroth bundle. In this way the entry points to the search file are constructed successively. The procedure continues until either the query structure has been completely reconstructed from the matching bundles in the bundle files or until all the bundle files have been compared.

In many cases the screening is sufficient to determine whether a given structure is an answer to the query. This is a difference from that in the Messenger software, which actually requires the completion of the atom–by–atom matching until it is able to decide whether or not the structure belongs to the answer set. The results of the screening phase in S[4] are the entry points to the search file for the atom-by-atom match. If the query structure consists of less than 11 bundles, then the answers can be constructed completely from the bundle files. In this case no atom-by-atom search is required. In all the other cases the bit string which results from the screening phase is the access key to the structures in the search file which have to be compared in an atom-by-atom search with the query structure. All structures which start with the same bit string resulting from the screening phase are potential hits (if the query structure consists of more than 11 bundles). For each entry point only one random access is necessary. Then the file is read sequentially until the first part of the codes changes. The file search algorithm is equivalent to a search of a truncated text term in an inverted file (see Figure 4).

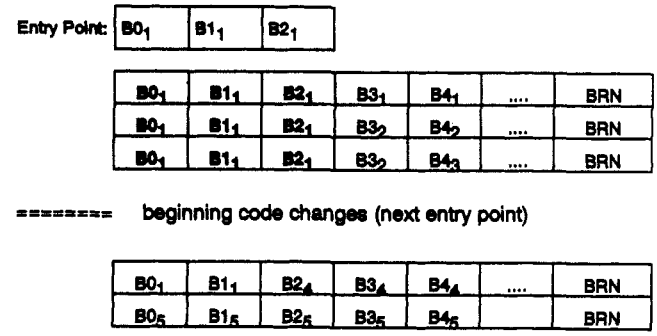


Figure 4. Atom-by-atom search with S[4].

Comparison of the Algorithms. The structure search systems S[4] and Messenger have several features in common, but the architecture, the query processing, and the search algorithms differ considerably. Both systems consist of a two-phased searching algorithm, i.e. a screening phase and an atom-by-atom match. The result of the screening in Messenger is a list of potential hits—in other words the screen answer set is a superset of the final answer set. A dictionary with a fixed number of screens forms the basis for this algorithm.

In S[4] the screening algorithm is based on a hierarchical graph-type approach using a limited number of spheres for the screens or bundles. The Messenger screening algorithm uses a random access to the inverted screen files, and it compares and merges large data sets while S[4] reads only certain bundles of information which is stored at the same physical location on the disk (sequential read). In many cases S[4] can decide after the screening which structures are hits without performing an atom-by-atom match. The atom-by-atom match algorithm of S[4] requires only a few random read accesses, and the rest are sequential read processes. Since the hit structures are partially constructed from the bundles, the atom-by-atom match needs to compare only the rest of the structures. As a result of these features the S[4] software has a very good performance, and it is especially suitable to very large structure files, i.e. with several million structures.

FUNCTIONALITY

The functionality of a software system can be defined as the set of features or capabilities of the given software system. It can be applied on different conceptual levels; e.g. on a very broad level the functionality of a structure search system includes functions like Markush or substructure searching. In our analysis we are focusing on a medium conceptual level; i.e. features like chemical bonding conventions or the different possibilities to search for stereochemistry are discussed.

A chemical structure is described by the nodes (atom or groups of atoms), the bonds between the nodes, the type of nodes and bonds, and a set of additional information about structures like the three-dimensional configuration in space. In general, this information is stored in a connection table. The following sections describe both Messenger[27,28] and S[4] [29] with respect to the functionality and give a brief comparison of the two systems.

Table 1. Node Types for Structures in Messenger

symbol	node type
element symbol	standard chemical element
A	any standard element except hydrogen
X	any halogen
Q	any element including H, except carbon
M	any metal (all elements except Ar, As, At, B, Br, C, Cl, F, H, He, J, Kr, N, Ne, O, P, Rn, S, Se, Si, Te, Xe)

Table 2. Bond Types for Structures in Messenger

symbol	bond type	symbol	bond type
SE	single exact bond	S	single or normalized bond
DE	double exact bond	D	double or normalized bond
T	triple bond	U	unspecified bond (SE, DE, T, or N)
N	normalized bond		

Messenger System. The nodes of chemical structures can be drawn using the standard chemical element symbols or any other symbol given in Table 1. It is also possible to exclude certain types of nodes using the symbol "-", e.g. "-X" means that no halogen may be present. For structure fragments which are frequently present, there is also a list of shortcut symbols available (see ref 27). Generic queries can be formulated using generic group symbols or the generic group definition. There are four generic group symbols available in Messenger: AK (acyclic carbon chain), CY (cyclic group), CB (carbocyclic group), and HY (heterocyclic group). It is also possible to assign an attribute to these generic groups:

BRA/LIN	branched/linear chain
HIC/LOC	high/low (≥7/≤6) C number
HIQ/LOQ	high/low (≥2/1) number of heteroatoms
MCY/PCY	monocyclic/polycyclic
SAT/UNS	saturated/unsaturated

In addition to the specification of a single node, the user may also define a generic group (Gk group) through a list of atom fragments. Such a Gk group is defined separately according to the following rules: (i) 20 different Gk groups are allowed per structure, and (ii) 20 node symbols are allowed per Gk group, including standard element symbols, variable atom symbols, generic groups, shortcut symbols, and user defined fragments. A user defined structure fragment could be another Gk group and according to the above rules a maximum of 19 Gk groups can be nested in one Gk group structure.

A node may be further characterized by its attributes. Such attributes include the assignment as ring or chain, an abnormal mass, a charge, or a delocalized charge. Without any further specification it is assumed that all possible attribute values are allowed.

A structure may contain specific or variable bond types (see Table 2). If the user does not define the type of bond, then the software assumes an unspecified bond (type U). The normalized bond type is a formal bond type which does not correspond to an aromatic bond. It could be used to search for structures with alternating single or double bonds in ring systems, certain types of tautomers, and structures with delocalized bonds. The use of this bond type requires some experience from the user, since it is not always clear what kind of structures will be retrieved when normalized bonds are included in the structure. In analogy to the nodes the bonds may also have certain attributes like ring or chain.

A node may also be characterized by the number and type of its substituents. If nothing is specified, the software assumes that all free sites may be substituted according to the valency of this node. It is possible to define the number of hydrogen atoms (as an exact or minimum value) and the number of non-hydrogen substituents (as an exact, minimum, or maximum value) and to specify the valency of a node (as an exact value between 1 and 16 or as an abnormal valency). An interesting feature is the possibility of allowing variable bond sites for a substituent in a ring system. In Messenger there may be between 2 and 20 variable bond sites in a ring for non-hydrogen atoms; i.e. the substituent may be present exactly once at one of these positions.

It is possible to search for generic structures either in a file of specific (e.g. Registry) or a file of generic substances (Marpat) (see ref 30).

At the time when the analysis was done, the Messenger system could not cope with stereochemical information in the connection tables. Since the end of 1993, stereosearch is on-line available. This feature can handle absolute and relative stereochemistry[31,32] and allows one to distinguish geometric isomers (handling of geometric double bonds) and to distinguish optical isomers (handling of tetrahedral chiral centers). There are several possibilities for the specification of the search query which cannot be described here.

Chemical reactions can be handled by the Messenger software using a structure- or a reaction-oriented search approach. The first type of search is based on a substructure search of the individual reaction partners, and the latter search takes the reaction sites into account. The bonds in a reaction can be declared as completely or partially broken.

A search for multicomponent structures can be performed in two different ways in Messenger. Firstly, it is possible to input two separate fragments in a single query. However, this search results in structures which contain both fragments in the same structure with no overlapping, and it does not result in any multicomponent structures. Secondly, the fragments can be defined separately, and the two queries can be combined by AND logic. The results of this substructure search comprise structures where both components are in the same component or where each fragment occurs in a different component. Both query fragments may or may not overlap in the resulting structures.

In Messenger four types of structure searches and four different search scopes are allowed. The structure search types are

EXA for an exact search (no further substitutions or multifragments are allowed)
FAM for a family search (as EXA plus salts and multifragments)
CSS for closed substructure search (substitutions are only allowed on free sites which have been declared explicitly)
SSS for substructure search (any substitution is allowed; this is the system default)

A structure search could be run against a whole file (scope: FULL), against part of the file (scope: RANGE or SUBSET), or against a sample set (scope: SAMPLE). The current limits for FULL, RANGE, and SUBSET on-line searches are 100 000 iterations and 100 000 answers for Registry and 200 000 iterations and 100 000 answers for Beilstein. For a batch search the corresponding limits are 250 000 iterations and 250 000 answers for Registry, and 300 000 iterations and 250 000 answers for Beilstein. SAMPLE searches are restricted to 1000 iterations and 50 answers.

S⁴ System. The substructure search system S⁴ supports a rich variety of different node types (see Table 3). In addition

Table 3. Node Types for Structures in S⁴

symbol	node type
element symbol	standard chemical element
A	any standard element except hydrogen
AH	any standard element including hydrogen or no substituent present
Ai	user defined atom list
A(n,m)	any standard element with an atomic number between *n* and *m*
AH(n,m)	any standard element with an atomic number between *n* and *m*, including hydrogen or no substituent present
X	any halogen
XH	any halogen, including hydrogen or no substituent present
Q	any element, including H, except carbon
QH	any element, including H, except carbon, or no substituent present
M	any metal (all elements except Ar, As, At, B, Br, C, Cl, F, H, He, J, Kr, N, Ne, O, P, Rn, S, Se, Si, Te, Xe)
MH	any metal or hydrogen or no substituent present

to the node types available in Messenger there is also the possibility of defining generic node ranges, e.g. A<6,8>. This means standard elements with an atomic number between 6 and 8, i.e. carbon, nitrogen, or oxygen. It is also possible to exclude certain node types by using the prefix "-" before the node type, e.g. "-C" means exclusion of carbon. The exclusion prefix may be used in conjunction with standard elements, variable atoms (A, M, X, and Q), user defined atom lists, and a range of atomic numbers. Shortcuts are not supported by S⁴.

Generic groups can be specified either by selecting a predefined group or by defining a Gk group. Predefined generic groups are divided in three classes (any, acyclic, and cyclic), and there are more than 30 different groups available, including a group symbol for rings without carbon which is very important for organometallics. The definition of Gk groups is more flexible than in Messenger since it allows 99 Gk groups in a single structure and any number of nodes in a given Gk group. The maximum nesting depth for these groups is 98.

A node may be further characterized by a set of attributes like the type of node (ring or chain), an abnormal mass, a charge, a delocalized charge, a radical, a valency, or a stereo center. In addition to these node attributes it is also possible to assign attributes to the molecule as a whole.

The S⁴ software requires the specification of the molecular substitution when the structure is built. If no substitutions are given, the software assumes that all free sites are filled with hydrogen atoms. Each node may have free sites according to its valency. It is also possible to specify variable bonding positions at a set of nodes. The declaration of free sites at the time of structure building is a difference to Messenger, where the search command allows the user to restrict the query to a certain search type (EXA, FAM, SSS, or CSS).

The bond types of S⁴ are quite similar to those of Messenger. It is possible to specify exact bonds (single, double, or triple) and variable bonds (undefined, automatic, or multiple). Since the bond type must be specified in any case in S⁴, there is no default bond type. If the "automatic" bond type is applied, all multiple bonds will be converted to aromatic or tautomeric bonds if necessary. Attributes may be associated with each bond according to the list given in Table 4.

S⁴ has an option to include tautomeric compounds in the search. This concept takes proton, deuterium, and tritium atoms into account. There are some cases, however, where the search results in algorithmic tautomers instead of chemical tautomers. Since the latter is a subset of the former class, no information will be lost for the user.

Table 4. Bond Attributes for Structures in S⁴

attribute	attribute value
type	chain
	ring
	ring/chain (default)
orientation	in the plane (default)
	above the plane
	below the plane
	out of plane
stereo	active
	inactive

Table 5. Polyhedral Geometries for Structures in S⁴

polyhedron geometry	symbol
trigonal	P-3
tetrahedral	T-4
square planar	SP-4
trigonal bipyramid	TB-5
square pyramid	SP-5
octahedral	OC-6
trigonal prism	TP-6
pentagonal bipyramid	PB-7
cube	CU-8
antiprism	SA-8
dodecahedral	DD-8
hexagonal bipyramid	HB-8
tricapped trigonal bipyramid	TPS-9
heptagonal bipyramid	HB-9
trigonal with π-ligands	P-3-PI
tetrahedral with π-ligands	T-4-PI
square planar with π-ligands	SP-4-PI
trigonal bipyramid with π-ligands	TB-5-PI
square pyramid with π-ligands	SP-5-PI
octahedral with π-ligands	OC-6-PI

While S⁴ supports a broad variety of node types, including generic nodes in the definition of the query structure, it does not support files with generic (Markush) structures.

The software can handle stereochemistry at the double bond, at the tetrahedral geometry, and at 19 additional polyhedral geometries (see Table 5). It is possible to formulate the query for a stereochemical search using the orientation of the bonds at a stereo center or the coordinates of the ligands. A stereo center could be all regular and variable atoms, atom lists, and ranges of atoms. For a stereo double bond it is necessary to assign the attribute "stereo active" to the bond and to draw the configuration in an unambigous way. The following levels of a stereo search are supported:

stereo searching is off (S0)
No checking for stereochemistry is done.

exact stereo search (S1)
The configuration of the stereo double bond and of the steric tetrahedron is checked. The racemic structures are discarded.

mirror image stereo search (S2)
Stereo double bonds and stereo tetrahedra are checked. In addition, all structures which have all the parities of the requested tetrahedral centers inverted are included as hits.

Independent of the stereo level, it is possible to search for structures in inorganic chemistry using a flag for the specification of the mirror images. The three levels are (i) the mirror image flag is not checked, (ii) the mirror image flag must be present in the hit structures, and (iii) the mirror image flag must not be present in the hit structures.

A management of chemical reactions as provided by Messenger for in-house reaction database management systems is currently not implemented in S4.

A structure query can be composed of two fragments. S4 finds all structures with non-overlapping fragments either in a single structure or in different components.

The structure search in S4 is based on the input structure, and it is not possible to specify a search type as in Messenger. The free sites of a structure determine implicitly the search type. This means that a clear distinction between the types of structure searching does not exist in S4 and the user may define any level of substructure search.

Comparison of the Capabilities. Both systems support a large class of different nodes, both specific and general. The number of generic nodes and of Gk groups is much larger in S4, and it is possible to specify more complex generic search queries. Concerning the bond types, both systems are equivalent and they differ mainly in the definition of a normalized (Messenger) and an aromatic bond (S4).

The concept of free sites is complementary in S4 and Messenger. In the S4 system the default is a search for all unsubstituted compounds; i.e. there are no free sites unless this is explicitly stated by the user. All free sites have to be specified when the structure is built. Messenger has no default for free sites when the structure is composed. This will be specified when the search query is performed. At this stage, the customer may use the search default substructure search and the software assumes free sites on all nodes. However, it is also possible to perform other types of searches like a closed substructure search which corresponds to the default search type in S4.

Both systems understand the concept of generic nodes in the search query, e.g. Gk groups or generic groups. The handling of generic structures in file structures and the matching of various stages of generality of Markush structures is only possible with Messenger. In other words both systems can run a generic structure query against a file of specific structures, but only Messenger is able to run these queries against a file of generic structures.

At the time of the analysis the Messenger system could handle stereochemistry only in the form of text descriptors. In the meantime the search system has been enhanced considerably to handle stereochemistry at tetrahedral geometry and at double bonds (organic compounds). S4 on the other side is able to handle stereochemistry for both organic and inorganic compounds. There is a set of 20 polyhedra geometries (including the tetrahedra) which are supported by S4 and which where developed especially for the needs of the Gmelin database of inorganic and organometallic substances.

The concept of tautomer searching is different for the two systems. In Messenger proton tautomers of the general form M=Q—ZH ↔ HM—Q=Z are found as a total according to the concept of normalized bonds. Valence and ring/chain tautomers are retrieved as individual tautomers and not as a total set. S4 has an option for tautomer searching which results in the set of all proton tautomers, but valence and ring/chain tautomers are not found.

The retrieval of multicomponent substances is more flexible in Messenger since it allows one to formulate the query with two different fragments in one or more structures. In neither system is it possible to enforce the search to yield only substances where all fragments occur in different components.

A true reaction searching is only possible with the Messenger software since it can both store and search for reaction sites.

Messenger allows the user to decide the search type at the time when the query is actually run, while in S4 this has to be specified when the structure is built. In addition, Messenger can perform the search on different sizes of the file by specifying the search scope.

USER INTERACTION

The user interaction of a structure search system requires additional features with respect to a bibliographic retrieval system. It is especially convenient if graphical structure input is supported or even integrated into the system. In the Messenger system a structure may be input using the STRUCTURE command which is a line-based editor for chemical structures. The terminal mode type 2 allows also a graphical input of structures using terminals with a Tectronics Plot10 emulation. In addition, there are a number of PC software products which have an integrated graphical structure editor, e.g. STN Express, and they can be used in conjunction with the access to STN International. Another possibility for the generation of structure queries is the modeling of an existing structure using a registry number as input (the registry number modeling depends on the registry number which is the primary key for this file; in the case of the Registry file it is the CAS Registry Number, and in the case of the Beilstein file it is the Beilstein Registry Number). In this case, a structure is retrieved from the database, the customer modifies the structure according to his or her requirements, and the result is stored for later use in a search query.

In S4 structures can be input as a so-called ROSDAL string, which is a linear representation of chemical structures. Alternatively, structures could be input using the PC software MOLKICK, which has been developed for the special needs of the Beilstein database. A difference from the Messenger system is the fact that S4 does not distinguish between different search modes as Messenger does, e.g. substructure or exact search. When the structure is built, the user has to decide which search mode he wants to use, and this mode is incorporated into the structure query.

The input of structures for Messenger is quite flexible. Besides the simple (and fast) line-oriented structure editor (STRUCTURE command), graphical emulation software could be used if it supports the Tectronics Plot10 format, e.g. like STN Express. Since the S4 software is only a search engine, there is additional software required for the input of chemical structures. Although the PC software MOLKICK is quite convenient to use, the integration of this software in the DIALOG system is not very satisfactory.

PERFORMANCE

Substructure searching belongs to the class of problems which are known as Np complete problems. The time for solving such mathematical problems is depending on the number of objects involved (N), i.e. the number of nodes to be compared, and the time for the matching increases exponentially with N. Therefore, the problem is generally split into two tasks where the first task is a kind of filter (screening) reducing the number of substances which have to be compared in the second task against the query structure (atom-by-atom match). The overall time required for the substructure searching is dependent on the screening method and the algorithm for the atom-by-atom match. Since S4 and Messenger are using different search methods and are based on different architectures, it is interesting to analyze the performance of substructure searching for both systems.

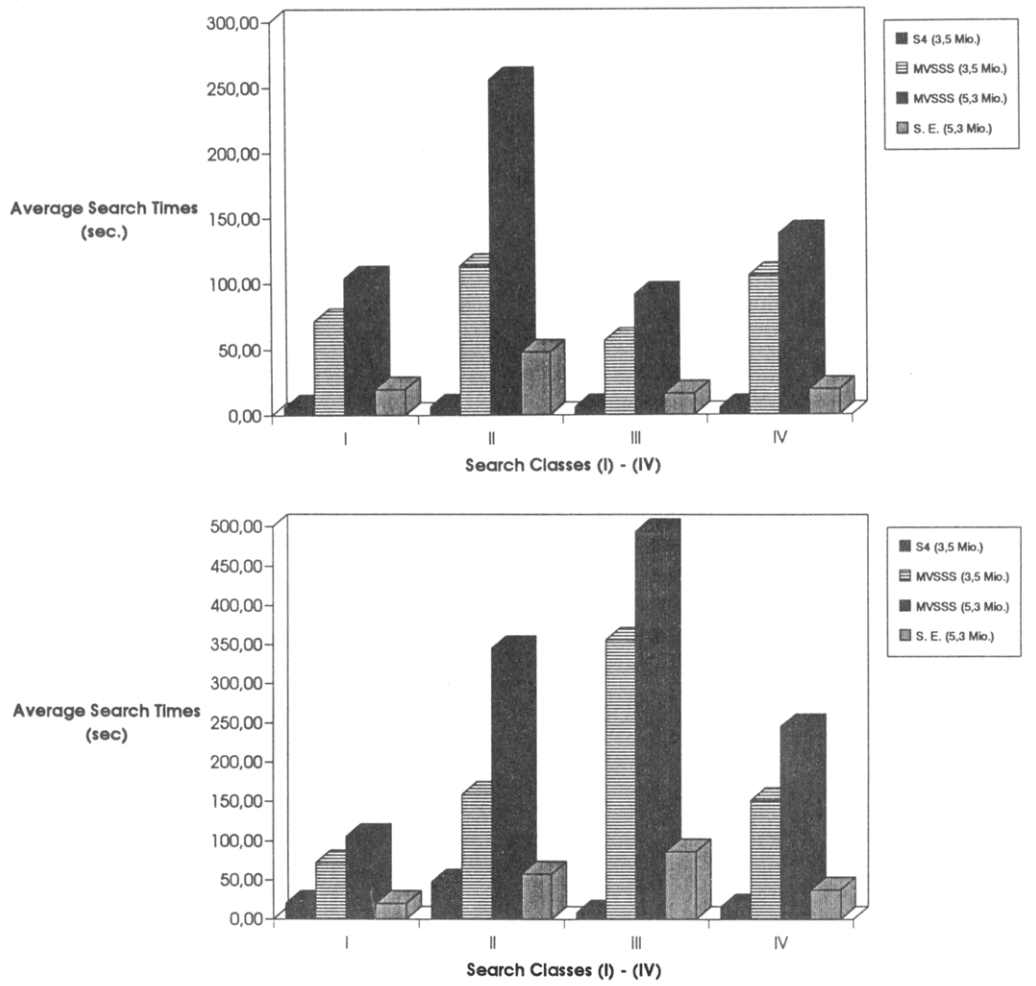


Figure 5. Average search times per query in seconds: (a, top) for CSS; (b, bottom) for SSS.

The performance has been tested with four different classes of structures, and we have measured the average search times for the Messenger search engines, for the Messenger mainframe system, and for the S4 system. The four classes of structures can be described as follows:

(I) acyclic compounds
seven structures with heteroatoms in the chain, i.e. alcohols, carbonic acids, ethers, and esters

(II) isocyclic compounds
14 structures both with and without heteroatoms in the side chains

(III) bicyclic compounds
4 structures without any heteroatom in the rings

(IV) heterocyclic compounds
25 structures with and without side chains, fusioned rings, etc.

Each structure has been searched with and without free sites, i.e. as a substructure search (SSS) and as a closed substructure search (CSS).

In the initial analysis the test was performed with the Beilstein file of approximately 3.5 million substances on the Messenger mainframe search system and on the S4 mainframe system. The search with S4 was done on both an independent system under TSO and in an integrated environment in Messenger using the RUN command interface. Recently, the Beilstein file has been loaded on search engines, and the performance test has been repeated with the Beilstein file with 5.3 million substances on both the search engines and the mainframe.

Table 6. Total Number of Hits Found in Substructure Searches of the Search Classes I–IV

search class	S4 3.5 million	MVSSS 3.5 million	MVSSS/SE 5.3 million
	(a) For Closed Substructure Searches		
I	59	59	106
II	53	57	229
III	13	14	22
IV	132	160	256
	(b) For Substructure Searches		
I	3 772	3 878	5 791
II	2 570	8 489	9 672
III	72	80	119
IV	21 780	23 590	37 370

The Beilstein file is loaded on both a mainframe computer and a set of workstations. The mainframe computer at STN Karlsruhe is an IBM 3090 Model 400E with four processors, and it performs with 58 MIPS (MIPS = mega-instructions per second). The search engines for the Beilstein file are IBM risc 6000 machines Model 530H, and they show a SPECint92 value of 28.5. The performance measures for the different machines cannot be compared directly. Actually, there is no measure available which can be used for different machine architectures. The complete file of 5.3 million substances is currently distributed over 10 machines. For safety reasons a backup file is also attached to the search engines.

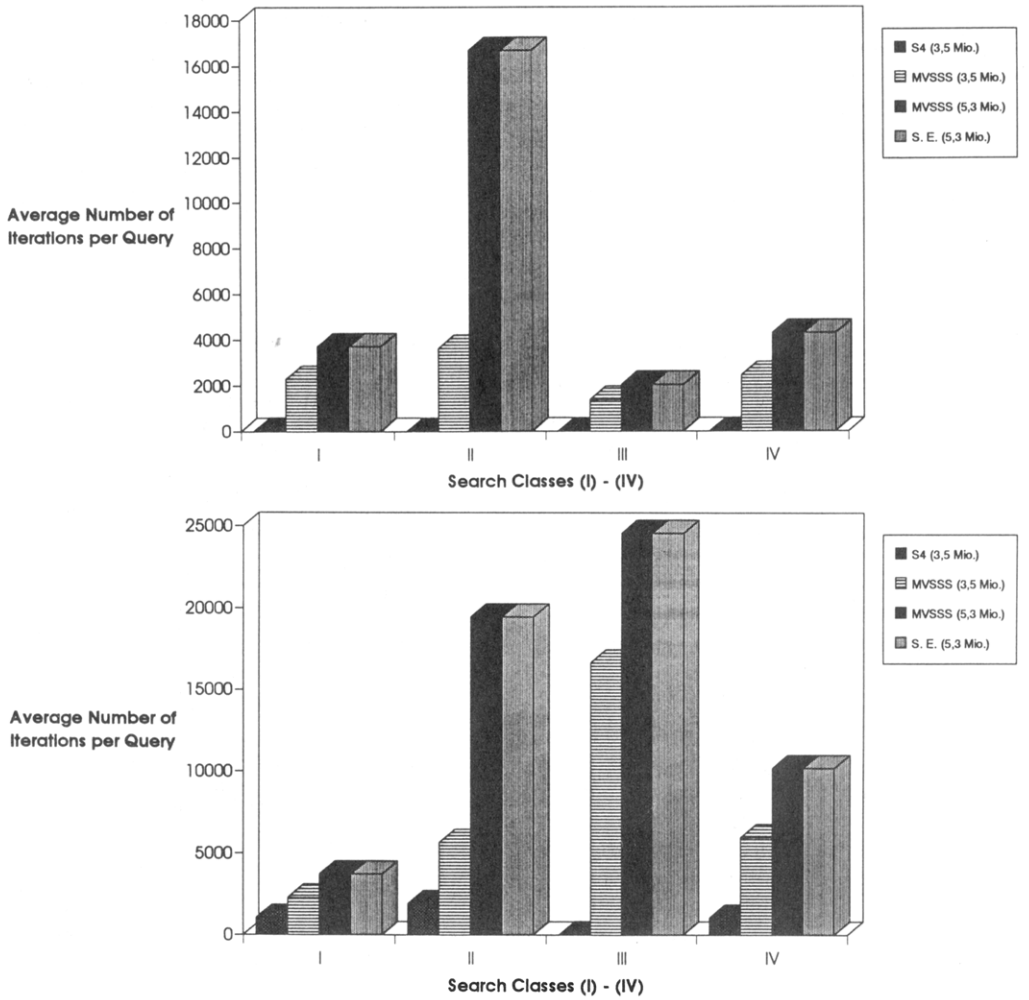


Figure 6. Average number of iterations per search query summed over all four classes of substances: (a, top) for CSS; (b, bottom) for SSS.

In order to obtain realistic figures, all the tests have been run at different times during the day, and the results reported in this paper are the averages of the original values which have been measured at different times. It is clear that it makes no sense to report the exact values since the performance depends strongly on the search query of the user and it changes with the increase of the structure files. For the tests of the Beilstein file the system limits have been increased to 200 000 iterations in order to avoid incomplete iterations.

Each structure class is treated as one single query, although the structures were searched independently. Since the databases were of different sizes and three different search systems were used, the number of hits found by the systems were different (see Table 6). However, there is also a small difference in the number of hits between S4 and Messenger which is due to the different algorithms used. The treatment of normalized/aromatic bonds or of tautomers, for example, is different in the two systems. From the two parts of Table 6 it can be deduced that the increase in the number of hits is not increasing linearly due to the special procedure of the Beilstein Institute to add new substances.

In Figure 5a,b the average search times are shown for the closed substructure (CSS) and the substructure search (SSS), respectively. Correspondingly, parts a and b of Figure 6 illustrate the average number of iterations per search query for both search types. The first value in each group refers to the Beilstein database searched with S4 (with 3.5 million substances), the two values in the middle result from the mainframe Messenger search system (with 3.5 and 5.3 million substances.), and the last value stems from the Messenger search engines (5.3 million substances).

In general, the S4 software performs faster than both Messenger systems (mainframe and search engines). On an average of all four search classes, the S4 software is about 6 times as fast as the mainframe Messenger structure search system (MVSSS) for substructure searches of type SSS and it is approximately 16 times faster for closed substructure searching (CSS). In some special cases this ratio may be even higher. There is a considerable decrease in the performance of the mainframe Messenger system when the file size is increased from 3.5 to 5.3 million substances. Using the Messenger search engine technology the search times for all classes are reduced considerably and in some cases they are even in the same range as the values for the S4 software (comparing 5.3 million to 3.5 million substances). As a result of this analysis it can be stated that both the S4 software and the search engine technology are able to cope fairly well with very large structure files.

In Figure 6a,b the average number of iterations per search query is given for the four search classes. The number of iterations per search query is given for the four search classes. The number of iterations for the file of 5.3 million substances is actually identical for the Messenger mainframe and the search engine architecture. It can also be seen from the figure that the number of iterations is considerably lower for the S4 software than for the Messenger system (both architectures). This fact accounts for the large differences in the performance of S4 and Messenger. As has been described in a previous section, in many cases the S4 software can decide already after the completion of the screening step whether a given

substance belongs to the answer set. Thus there are considerably less iterations for S⁴ to perform, and therefore the S⁴ software is very fast. It is clear that the number of iterations is higher for a substructure search type SSS than for the closed substructure search type CSS. In the case of S⁴ the number of iterations for the search type SSS increases for all search classes except for class III. It seems that these structures (bicyclic compounds) can be handled by S⁴ in a very effective way.

SUMMARY AND CONCLUSIONS

The two structure search systems are very similar with respect to their search capabilities. A large number of features is actually almost identical in both systems. The number of hits for a given query differs slightly due to the different treatment of some special chemical effects like tautomeric bonds. In general, the Messenger structure search system provides more functionalities, e.g. a user interface and a Markush and a reaction search capability. S⁴ on the other hand is very well suited for the needs of the Beilstein and Gmelin databases. Especially the inorganic chemistry is supported on a very high level of sophistication. Stereochemical information is included in the connection table, and it can be searched for about 20 different polyhedra, not only for the tetrahedron and the double bond.

An important difference between the two systems is the search time required for structure searching. Since Messenger is designed to perform structure searching for large files on a set of parallel search engines, the search times are dependent on the hardware equipment and the number of search engines. In our test case, the response times were very long on the mainframe computer but they reduced to reasonable times on the search engines. S⁴ is a single processor search engine, and it can be run on both mainframes and workstations. In both cases the search times are very short due to the extremely fast algorithm.

As a summary, it can be stated that the average structure search times are faster for the S⁴ structure search system than for the Messenger structure search system. However, the Messenger system can be adopted to large files by using the search engine technology. In addition, the search engines can easily be extended by using more or faster machines. Concerning the functionality of the two systems, it can be stated that the S⁴ software is very well suited for the Beilstein and Gmelin databases, while the Messenger search system provides more general capabilities, especially for Markush and reaction databases.

ACKNOWLEDGMENT

The funding of this work by the Federal German Ministry for Research and Technology is greatly acknowledged. Softron Ltd. provided a copy of the in-house software for testing, and we kindly acknowledge their support of this work.

REFERENCES AND NOTES

(1) Wiggins, G. *Chemical Information Sources*; McGraw-Hill: New York, 1991; p 352.
(2) Warr, W.; Suhr, C. *Chemical* Information Management; VCH Verlagsgesellschaft: Weinheim, 1992, p 261.
(3) Barnard, J. M. (1993). Substructure Searching Methods: Old and New. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.
(4) French, S. E. Our Reaction Access System. *CHEMTECH* **1987**, *17* (2), 106–111.
(5) Dubois, J. E. (1973). Computer Representation of Numerical and Graphic Data. *Proceedings of the Third International CODATA Conference on Generation, Compilation, Evaluation and Dissemination of Data for Science and Technology*, Le Creusot, France, June 26–29, 1972; CODATA Central Office: Frankfurt/Main, Germany, 1973; pp 58–66.
(6) Gay, J. P.; Alardo, H. Integrating Standard DBMSs Functionalities and Structures Handling Capabilities: The DARC Approach. In *Chemical Information, Proceedings of the International Conference*; Collier, H. R., Ed.; Springer: Berlin, Federal Republic of Germany, 1989; pp 221–236.
(7) Hicks, M. G.; Jochum, C. Substructure Search Systems. 1. Performance Comparison of the MACCS, DARC, CAS Registry MVSSS and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191–199.
(8) Welford, S. M. (1990). Chemical Structure Searching. Using S4/MOLKICK on DIALOG. In *The Beilstein Online Database: Implementation, Content and Retrieval*; Heller, S. R., Ed., ACS Symposium Series 436. American Chemical Society: Washington, D.C., 1990; pp 64–79.
(9) Milne, G. W. A.; Heller, S. R.; Heller, R. S.; Martinsen, D. P. The NIH/EPA Chemica Information System. *Adv. Mass Spectrom.* **1990**, *8B*, 1578–1581.
(10) Feldmann, R. J.; Milne, G. W.; Heller, S. R.; Fein, A.; Miller, J. A.; Koch, B. An Interactive Substructure Search System. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 157–163.
(11) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.
(12) Hartwell, I. O.; Haglund, K. A. (1990). An overview of DIALOG. In The Beilstein Online Database: Implementation, Content and Retrieval; Heller, S. R., Ed.; ACS Symposium Series 436; American Chemical Society: Washington, D.C., 1990; pp 42–63.
(13) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS ONLINE Search System. 1. General System Design and Selection, Generation, and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(14) Farmer, N. A.; Amoss, J.; Farel, W.; Fehribach, J.; Zeidner, C. The Evolution of the CAS Parallel Structure Searching Architecture. In: Chemical Structures: The International Language of Chemistry. *Proceedings of the Chemical Structure Association Conference*, Noordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988; pp 283–295.
(15) Barnard, J. M. Problems of Substructure Search and their Solution. In: Chemical Structures: The International Language of Chemistry. *Proceedings of the Chemical Structure Association Conference*, Noordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988; pp 113–126.
(16) Graf, W.; Kaindl, H. K.; Kniess, H.; Schmidt, B.; Warszawski, R. Substructure Retrieval by Means of the BASIC Fragment Search Dictionary Based on the Chemical Abstracts Service Chemical Registry III System. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 51–55.
(17) Graf, W.; Kaindl, H. K.; Kniess, H.; Warszawski, R. The Third BASIC Fragment Search Dictionary. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 177–181.
(18) Barnard, J. M. Computer Handling of Generic Chemical Structures. *Proceedings of a Conference at the University of Sheffield, U.K.*; Gower: Aldershot, U.K., 1984; p 230.
(19) Shenton, K. E.; Norton, P.; Ferns, E. A. Generic Searching of Patent Information. In *Chemical Structures: The International Language of Chemistry*, Proceedings of the Chemical Structure Association Conference, Noordwijkerhout, The Netherlands, June 1987; Springer-Verlag: Berlin, 1988; pp 169–178.
(20) Fisanick, W. Requirements for a System for Storage and Search of Markush Structures. In *Computer Handling of Generic Chemical Structures, Proceedings of a Conference at the University of Sheffield*, U.K.; Barnard, J. M., Ed.; Gower: Aldershot, U.K., 1984; pp 106–129.
(21) Fisanick, W. Storage and Retrieval of Generic Chemical Structure Representation by Computer; American Chemical Society, U.S. Patent 4,642,762, 1987.
(22) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145–154.
(23) Barth, A. Status and Future Developments of Reaction Databases and Online Retrieval Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 384–393.
(24) Bartmann, A.; Maier, H.; Walkowiak, D.; Roth, B.; Hicks, M. G. Substructure Searching on Very Large Files by Using Multiple Storage Techniques. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 539–541.
(25) Hicks, M. G.; Jochum, C.; Maier, H. Substructure Search Systems for Large Chemical Data Bases. *Anal. Chim. Acta* **1990**, *235* (1), 87–92.
(26) Rohbeck, H. G. *Representation of Structure Description Arranged Linearly*; Beilstein Institut: Frankfurt/Main, Germany, 1991.
(27) *STN Manual: Building and Searching Structures on STN*; Chemical Abstracts Service: Columbus, OH, 1992.
(28) Kasparek, S. V. Computer Graphics and Chemical Structures: Database Management Systems, CAS Registry, Chembase, REACCS, MACCS-II, Chemtalk; Wiley: New York, 1990; p 798.
(29) Heller, S. R.; Milne, G. W. A. Online Searching on DIALOG. *Beilstein Reference Manual*; Springer-Verlag: New York, 1989.
(30) *STN Manual: MARPAT User Guide*; Chemical Abstracts Service: Columbus, OH, 1990.
(31) Petrarca, A. E.; Lynch, M. F.; Rush, J. E. A Method for Generating Unique Structural Representations of Stereoisomers. *J. Chem. Doc.* **1967**, *7* (3), 154–165.
(32) Blackwood, J. E.; Blower, P. E., Jr.; Layten, S. W.; Lillie, D. H. Lipkus, A. H.; Peer, J. P.; Qian, C.; Staggenborg, L. M.; Watson, C. E. Chemical Abstracts Service Chemical Registry System. 13. Enhanced Handling of Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 204–212.