Note Added In Proof. Another method for symmetry perception (atoms only) came to our knowledge recently: Davis, M. I.; Ellzey, M. L., Jr. *J. Comput. Chem.* **1983**, *4*, 267.

ACKNOWLEDGMENT

We thank Prof. M. Randić for information on endospectral graphs and on the Cayley-Hamilton theorem and Dr. A. Bömelburg (IBM Deutschland) for a useful programming hint.

REFERENCES AND NOTES

(1) Rücker, G.; Rücker, Ch. *Chimia*, in press.
(2) Shelley, C. A.; Munk, M. E. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 110; **1979**, *19*, 247.
(3) Jochum, C.; Gasteiger, J. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 113; **1979**, *19*, 49.
(4) Schubert, W.; Ugi, I. *J. Am. Chem. Soc.* **1978**, *100*, 37. Schubert, W. *MATCH* **1979**, *6*, 213.
(5) (a) Randić, M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 36. (b) Randić, M.; Brissey, G. M.; Wilkins, C. L. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52.
(6) Hendrickson, J. B.; Toczko, A. G. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 171.
(7) Balaban, A. T.; Mekenyan, O.; Bonchev, D. *J. Comput. Chem.* **1985**, *6*, 538, and references cited therein.
(8) Bersohn, M. *Comput. Chem.* **1987**, *11*, 67.
(9) Ihlenfeldt, W. D.; Gasteiger, J. In *Software-Entwicklung in der Chemie 2*; Gasteiger, J.; Ed.; Springer-Verlag: Berlin, 1988; pp 13-33.
(10) Gray, N. A. B. *Computer-assisted structure elucidation*; Wiley: New York, 1986; Chapter 9.
(11) For the kind of symmetry discussed in this paper (which is clearly not the usual geometric symmetry in three-dimensional space) the terms "constitutional symmetry" and "topological symmetry" have both been used in the literature[2-5] since the information contained in the constitution (not configuration or conformation) is considered exclusively; i.e., geometric properties like bond lengths and angles or cis/trans relationships are disregarded. However, the term topological symmetry may be misunderstood since topological *isomers* [as defined earlier, e.g., a simple macrocycle and its knotted isomer (Frisch, H. L.; Wasserman, E. *J. Am. Chem. Soc.* **1961**, *83*, 3789) or the pair **28**/**29** in Figure 3] are *indistinguishable* in terms of the symmetry under discussion here (the two connectivity matrices of such a pair are identical).
(12) The higher powers of the adjacency (or a similar) matrix were used earlier, but their full potential for symmetry recognition was not exploited: Randić, M. *J. Comput. Chem.* **1980**, *1*, 386. Uchino, M. *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 116. Golender, V. E.; Drboglav, V. V.; Rosenblit, A. B. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 196. Razinger, M. *Theor. Chim. Acta* **1982**, *61*, 581. Randić, M.; Woodworth, W. L.; Graovac, A. *Int. J. Quantum Chem.* **1983**, *24*, 435.
(13) A walk is a pathway with repetition: The entry 6 in element 1,2 in C^3 in our example cuneane means there are six different walks of length 3 bonds from atom 1 to atom 2, namely, 1-2-1-2, 1-4-1-2, 1-7-1-2, 1-2-3-2, 1-2-5-2, 1-4-5-2.
(14) As a rule of thumb, all classes of atoms are found at this stage. In fact, in most cases considerably fewer steps are required. We know of only one exception to this rule (Figure 5 in ref 7).
(15) The Cayley-Hamilton theorem in this connection states that if two entries are not differentiated in all matrices up to the nth power, then they will not be differentiated in higher matrices either. While this is theoretically pleasing, the stop criterion suggested thereby (stop after the nth power matrix has been evaluated) is of little practical value: In the majority of graphs the diameter is considerably less than n.
(16) Randić, M. *SIAM J. Alg. Disc. Meth.* **1985**, *6*, 145. Knop, J. V.; Müller, W. R.; Szymanski, K.; Trinajstić, N.; Kleiner, A. F.; Randić, M. *J. Math. Phys.* **1986**, *27*, 2601. Randić, M.; Kleiner, A. F. *Ann. N. Y. Acad. Sci.* **1989**, *555*, 320.
(17) Though the critical vertices in endospectral graphs obviously are not segregated by procedure i, they are differentiated either by procedure ii (the graphs in ref 16, which exhibit different patterns of entries in the rows corresponding to the critical vertices in low power matrices already) or by procedure iii, e.g., graphs **30** and **31** in Figure 3.
(18) The only known case where our algorithm stops prematurely is a graph of 18 equivalent vertices of degree 3 with diameter 4 (Figure 1.1 in: Coxeter, H. S. M.; Frucht, R.; Powers, D. L. *Zero-Symmetric Graphs*; Academic Press: New York, 1981) where in the fifth step only 9 different pairs are perceived (procedure v obviously cannot trigger further steps here). Forcing the program to do two further steps results in the correct perception of 13 different pairs.
(19) A factor of ca. 30 is to be expected (information given by IBM Deutschland).
(20) A plot of the tenth root of the CPU time vs n, however, looks almost the same and results in a least-squares straight line with $r^2 = 0.95$. Thus, it is by no means clear, either theoretically[5b,21] or experimentally, whether the symmetry perception effort increases exponentially or polynominally with increasing n.
(21) Wirth, K. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 242.
(22) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 172.

Substructure Search Systems. 1. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 Substructure Search Systems

MARTIN G. HICKS* and CLEMENS JOCHUM

Beilstein Institute, Varrentrappstrasse 40-42, 6000 Frankfurt 90, West Germany

Received December 1, 1989

A comparison of the performance of the substructure search systems MACCS, DARC, HTSS, and S4 has been carried out in-house at the Beilstein Institute, and that of the CAS Registry MVSSS system on STN International at FIZ Karlsruhe was carried out on-line. Included in the comparison were the hit sets, screening efficiency, task times, and elapse times. The results showed that all systems gave similar results in terms of retrieved hit sets, but S4 dramatically out-performed the other systems in terms of task and elapse times. A subsequent test of S4 with a very much larger file showed the search time/file size relationship to be very much less than linear.

Effective management of the information associated with the ca. 10 million chemical compounds known to date is of major importance to chemists in industry and at universities alike.[1-5] The ability of the computer to handle vast amounts of information has brought it to center stage in chemical information management. Recent advances in computer technology, fast mainframes with large storage capacities and cheap personal computers, have led to dramatic changes in chemical information handling. The user friendly interfaces, made possible by the graphical capabilities of the PC, have given easy access to this information to the lay chemist.

A chemical compound can be described and defined in various ways; irrespective of the method adopted, effective searching can only be achieved if a unique compound has a unique description. Moreover, effective storage of the information for a compound is only possible if the description of

0095-2338/90/1630-0191$02.50/0

any one compound is always the same. It must be normalized according to a set of rules, and the numbering of the atoms in the structure must be canonicalized. Chemical nomenclature, while providing a method to communicate verbally or in writing a description of a chemical compound, is unsuited both for the role of a unique descriptor and for further processing involving the manipulation of structural information.[6-8]

The key to chemical information is the chemical structure, and the ability to store, search, and retrieve computer representations of chemical structures is central to any chemical database. Recent decades have seen the continual development of systems to handle chemical structures. These systems operated initially on fragment codes, subsequently on line notation, and finally on connection tables. The advent of computer graphics, particularly cheaply available PC graphics, has given the chemist the tools to access the information by simply drawing the chemical structure in the user interface. Greater flexibility combined with graphical input has made line notation obsolete and connection tables the de factor standard.

The Beilstein Institute is nearing the completion of a project to create the world's largest structurally oriented factual database of organic compounds based on the *Beilstein Handbook of Organic Chemistry*. In December 1988, the first release of the database with 350K heterocyclic compounds became available on-line, and by the time the database is fully up-to-date in 1992, the figure will have reached ca. 4.5M. It is therefore of paramount importance to have a substructure search system which can confortably accommodate upward of 5M compounds. Such a system is not only required internally by Beilstein to assist with database production, administration, and handbook production, it is also a requirement of our database hosts who need to offer users a well-tailored system.

To benchmark the result of our own system and to determine whether any currently available system could better meet our needs, performance comparisons have been carried out at the Beilstein Institute.

Due to restrictions imposed by the system software owners, the tests could not all be carried out at the same time; thus, in the first phase DARC, HTSS, and S4 were tested and in the second DARC, MACCS, and S4. Due to storage and processing limitations on the Beilstein computer, the files available for testing always had to be the currently supported work files. Thus, for phase 1 only a test file was available and for phase 2 the initial on-line file. Within each of the phases the systems were tested on the same files and under, as far as possible, identical conditions.

The systems tested in-house, MACCS, DARC, HTSS, and S4, can be compared and contrasted according to their basic design. As part of this comparison the CAS Registry Substructure Search system will be discussed. Although the system was not available for us to test in-house, the host MVSSS implementation at STN/FIZ Karlsruhe had the same phase 2 file loaded and the same queries were tested. The results from the searches (number of hits, number of candidates, and elapse times), which are available to every STN user, are still very much of interest and will be discussed.

SYSTEMS

In its simplest form searching for a compound within a database is carried out by comparing directly the code of the query (fragment, line notation, or connection table) with that of each of the compounds in the database in turn. For a connection table based substructure search system this is achieved by the so called atom-by-atom-mapping (ABAM). Here the connection table of the query is mapped, i.e., compared, with that of a structure in the database to see whether they are absolutely identical (full structure search) of whether the query can be found as a subset within a database compound (substructure search).

Comparison of the query with every structure in the database is very time consuming and therefore only feasible for small databases. This is how the very successful ChemBase PC system from MDL operates. However, this method becomes impractical when the size of the database rises above a few thousand structures. A solution to this problem was to impose a screening step before the time-consuming ABAM.[9-12] In this step the molecules that, because of a particular structural characteristic, are clearly not going to be hits are discarded. This reduces the number of ABAMs that need to be carried out. The implementation of the screening step varies slightly from system to system, but they all operate on the same principle.

When the database is built, the molecules are analyzed in terms of the structural elements that they contain. For example, CAS uses the following structural screen types:[12] augmented atom, hydrogen augmented atom, twin augmented atom, atom sequence, bond sequence, connectivity sequence, ring count, type of ring, atom count, degree of connectivity, element composition, and graph modifier. There are over 2000 screens in the CAS screen library.

These screens are coded and stored in lists which are searched as the first phase of a structure search. The screens do not form a unique description of the molecule but do provide a means of ordering it in terms of structural characteristics. The ways the screens are stored and searched vary for the different systems.

The second phase of the search involves carrying out an ABAM on the candidates selected by the screening stage. It is essential that the screens are highly discriminating and that they are well balanced, producing good results for all classes of compounds.

CAS.[13] **(a) Search Machines.** The first implementation of the CAS search system with parallel architecture—the so-called search machines—consisted at the last count of 13 pairs of PDP-11 minicomputers. The file was split in two sections, the screens and the connection tables. The screens were stored as sequential lists on half of the pairs of PDP-11s, the file was read through sequentially, and the candidates were then transferred to the other half of the pairs of PDP-11s for ABAM. The advantages of this system were that the multiprocessor environment overcame the relatively primitive system design and produced good search times. The search times were fairly constant, the screens could be easily added, but the software was not transportable.

(b) Search Engines. In the recent years CAS has been developing a new version of their search software—the search engines. The search engine is again a multiprocessor system, this time using 11 Unisys 5000/95 minicomputers. The main difference is that the screens are now stored as inverted lists. The new implementation is faster than the old one and also portable to single-processor environments (MVSSS on STN International at FIZ Karlsruhe).

The CAS system is fairly sophisticated in its search options. Full generic search options are available which allow highly discriminating queries to be designed. The main shortcoming is that stereochemistry is not stored in the connection tables and is therefore not searchable. Correction of this will require an extension of the registry format and re-registration of existing compounds.

MACCS.[14,15] MACCS, the Molecular Access System developed by MDL, has been available since 1979 and is the most popular in-house system. MACCS uses a set of screens similar to but smaller than that of CAS. This is tailored to the smaller sizes of MACCS in-house files. There are about 1000 screens in the MACCS system. The screens are stored in inverted lists for optimal searching.

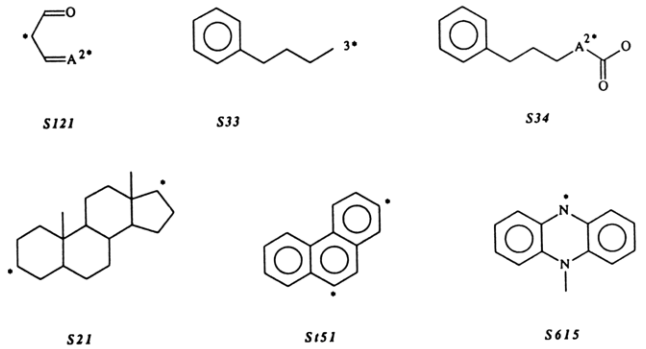

Figure 1. Phase 1 query structures.

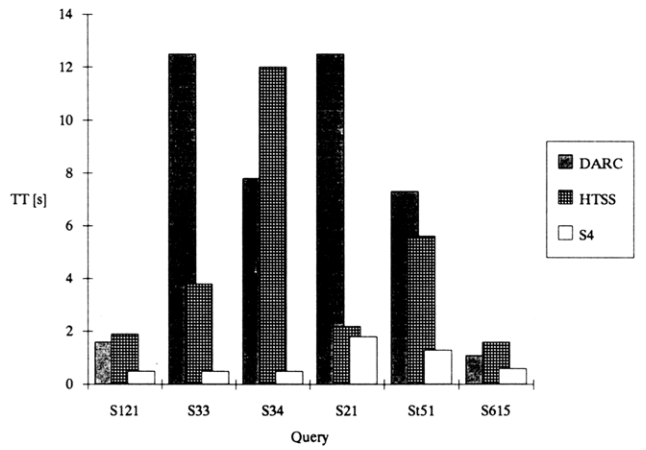

Figure 2. Phase 1 task times (seconds).

Full structure searches are carried out efficiently via a hash code. This is a special canonicalized encoding of the connection table which, while not allowing any substructure comparisons, provides a very fast method for full structure searches.

In its newest release MACCS can search for stereochemistry and also generic structures.

DARC.[16,17] This system has been under development by Dubois since 1963 and has been available since the mid 1970s as an in-house system. Since 1981 it has been used on-line to search the CAS Registry file (EURECAS). This is again a two-phase system of screening and ABAM. DARC does not use a screen library but analyzes molecules in terms of FRELs (Fragments Reduced to an Environment that is Limited). These FRELs, which consist of two sphere fragments built around a central focus, are stored in a tree form for searching. The focus can be an atom or a bond; the more highly connected the focus is, the more discriminating the FREL. Thus, the screens are file dependent and optimal for the file. It was found to be a disadvantage that the FRELs lacked the discrimination of some screen libraries in terms of extrafragment graph information such as ring information. To remove this problem, a small library of screens that contained these features was added.

The design of DARC allows files of millions of structures to be searched on a single processor machine. For full structure searches DARC also uses hash codes. Searches for generic structures and substructures and for stereochemically defined structures are also possible.

HTSS.[18] The Hierarchical Tree Substructure Search System was developed by Bruck and co-workers. In some ways it can be viewed as an extension of the techniques developed in DARC. In this system, all the atoms of a molecule are coded to three connectivity spheres. The encoding includes descriptors for rings and chains. An iterative process further characterizes an atom by the "color" of its neighbors with

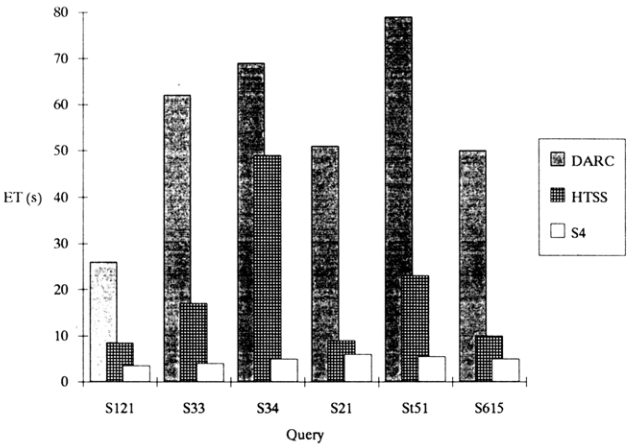

Figure 3. Phase 1 elapse times (seconds).

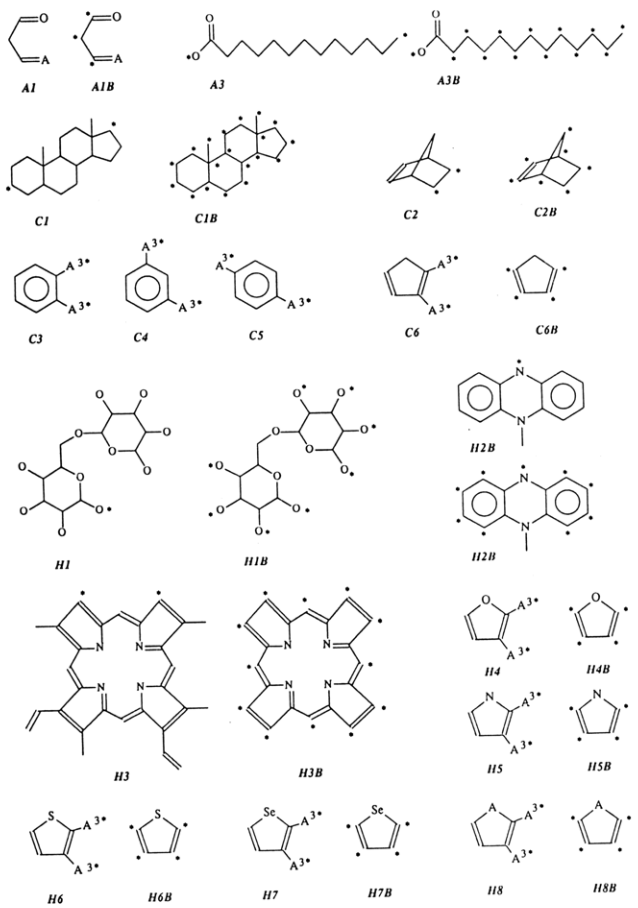

Figure 4. Phase 2 query structures.

subsequent iterations leading to a definition of an atom that is highly discriminating in terms of its extended connectivity and environment. This information is stored in the form of a tree. A search is carried out by conducting a tree walk for every atom in the query structure. Hits occur when the tree walk ends at a leaf for every atom in the query.

S4. This system has been under development by Beilstein–Softron over the past 3 years. The system is based on a compact code developed initially for use as a means of carrying out a fast full structure search. All molecules are encoded by using each atom in turn as the starting atom; thus, for a molecule of N atoms there are N connection tables. The codes are so compact that for an average molecule of 20 atoms they require only 15 bytes. These codes are sorted, stored, and indexed in a file. This highly redundant storage is the

Table I. Numbers of Hits for Phase 1 Queries

query	hits	query	hits
S121	87	S21	524
S33	364	St51	12
S34	!	S615	7

Table II. Average Task Times for Phase 1

system	time, s
DARC	7.1
HTSS	4.5
S4	0.9

means to achieve retrieval of all hits with very few (sometimes only one) sequential reads, carried out on a part of the file determined by the index. This eliminates the need for many time-consuming random accesses; in other systems there is one random access per candidate compound. The coding is so discriminating that in most cases the part of the file read contains exclusively hits; thus, no atom-by-atom match is required.

The present version of S4 can be used to search for full structures, substructures, and generic structures, and in its final version users will be able to search for stereochemical and tautomeric isomers.

The compact nature of the coding and the fast search times make this system ideally suited for very large files. Furthermore, the design of the system, in particular the fact that searching requires only very few sequential reads, makes this system ideally suited for transfer to a CD-ROM with its very long access times. This has been successfully carried out and has further demonstrated the power and flexibility of S4.[19]

COMPARISONS

Experimental Conditions. The system tests were carried out on an IBM 3090/150 at the Beilstein Institute.

System Versions. Phase 1: Generic DARC, HTSS 3.31, and S4 0.34. Phase 2: Generic DARC, MACCS II 1.41, S4 0.35, and CAS Registry MVSSS at STN/FIZ Karlsruhe.

The task time is defined as the total CPU time used by the machine to carry out the search process. The elapse time is the total actual elapsed time from the pressing of the key to start the search to the receipt of the message stating that the search has been completed.

Files. Phase 1: The file contained 581 619 structures taken from throughout the *Beilstein Handbook*. Phase 2: This time the first Beilstein on-line file was used. This file contained 350 418 structures, that is, all the single component handbook heterocyclic compounds from 1830 to 1960. Use of this file allowed the results obtained from the CAS STN system to be included.

Queries. Phase 1: At the time of the test the systems HTSS and S4 were under development and not all search options had been implemented. The queries used were, therefore, of a simple type. The common subset of queries is shown in Figure 1.

Phase 2: A new test set of queries was developed, which was designed to test the different systems as fully as possible with simple queries. It was also designed to be applicable not only to the file of heterocyclic compounds but also to subsequent Beilstein files.

Sixteen different structures were each posed in two different forms, set 1 having in total two free sites, to represent a fairly well defined query which a chemist might put, and set 2 having many free sites to show the performance with less well defined queries.

The queries, shown in Figure 4, have been drawn here in normalized form for clarity. Thus, a "star" represents one free site in set 1 and maximum free sites in the fully fuzzy set 2.

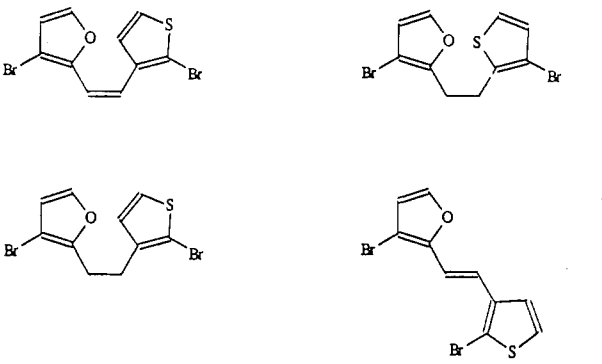

Figure 5. Molecules containing two substructures for searches H4–H7.

Table III. Numbers of Hits for Phase 2 Queries

query	hits	query	hits
A1	4	A1B	2744
A3	150	A3B	187
C1	117	C1B	679
C2	54	C2B	57
C3	10440		
C4	5918		
C5	43753		
C6	0	C6B	3
H1	95	H1B	230
H2	7	H2B	14
H3	6	H3B	625
H4	101	H4B	8720
H5	40	H5B	4230
H6	277	H6B	7395
H7	23	H7B	367
H8	437	H8B	20641

This does not mean, however, that the queries were phrased for each system by literally (as is possible in some systems) setting a star on the relevant atom. Each system has different conventions, and the queries were phrased accordingly to give the equivalent (or closest to it) of this normalized form as depicted.

Queries C3–C5, which would provide an inordinately high number of hits when fully fuzzy and could therefore only be carried out with S4, were not included in set 2.

RESULTS

Phase 1. In this phase DARC, HTSS, and S4 were tested.

(a) Hit Sets. The results showed that the accuracy and consistency of all systems was good. The numbers of hits, which were identical for all systems, for each query are listed in Table I.

(b) Search Times. The task and elapse times are plotted in Figures 2 and 3, respectively. Inspection of the results shows that there was a marked difference in the search times. S4 had by far the fastest search times (task and elapse times), which were relatively unaffected by molecular characteristics. The average search times are listed in Table II. HTSS was somewhat slower than S4 and had a much greater degree of variation in search times. Fragmented queries such as S34 and those with very many free sites would be theoretically slower in HTSS, and this supposition was borne out in the results. DARC was the slowest of the three.

Phase 2. In this phase the systems DARC, MACCS, and S4 were tested in-house, and the CAS Registry (STN/FIZ Karlsruhe) MVSSS was tested on-line. The CAS STN system imposes such a low system limit on the hit lists relative to the file size that it is often not possible to pose fuzzy queries without resorting to splitting up the query into ranges—a workable but time-consuming practice. With DARC the system limits could not be bypassed, and queries C3, C4, C5, and H8B could not be set. All comparisons of average search

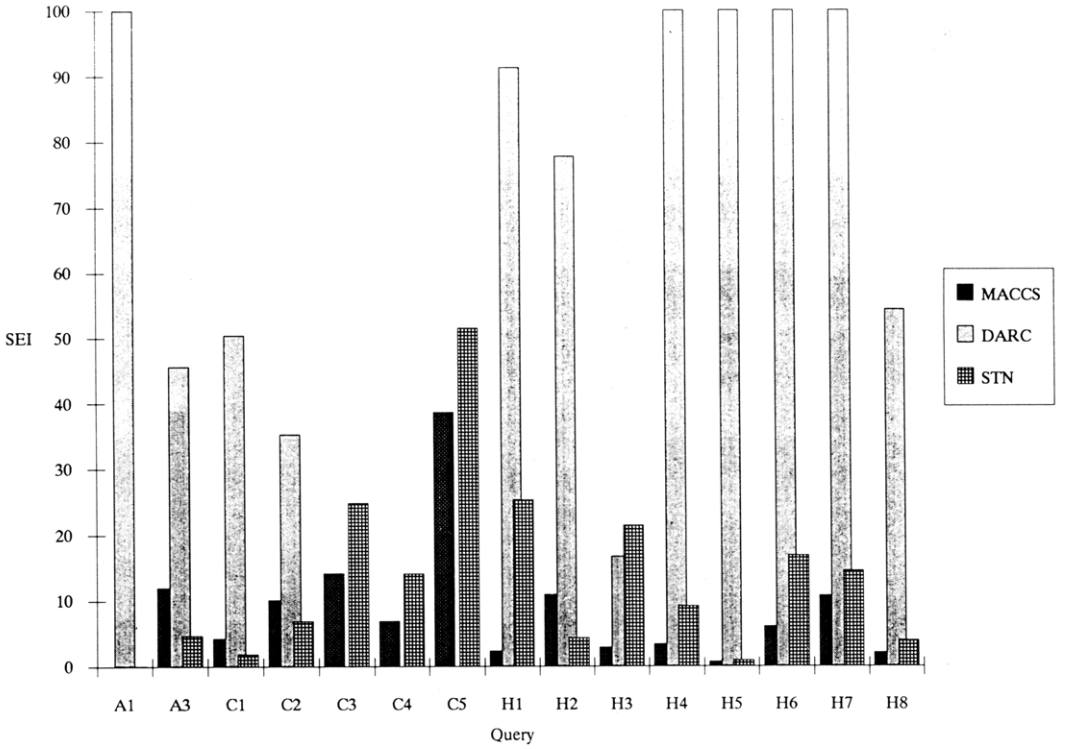

Figure 6. Phase 2 screen efficiency indices for well-defined queries.

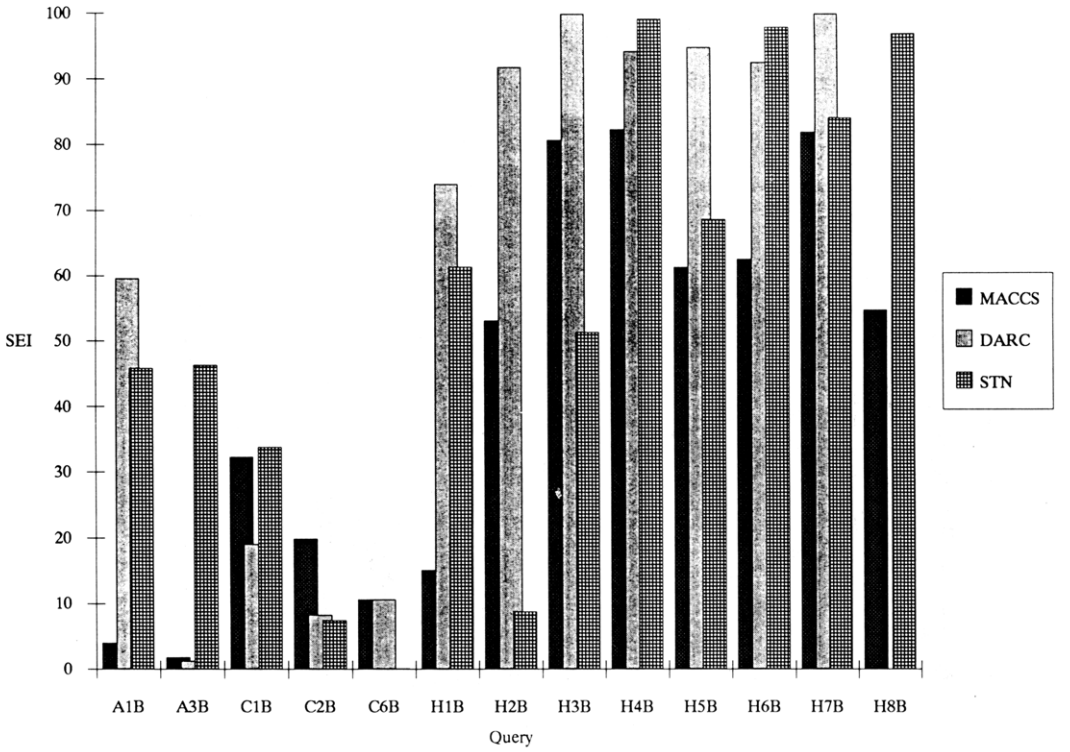

Figure 7. Phase 2 screen efficiency indices for fuzzy queries.

times were carried out on the common set of queries for which all systems produced an answer.

(a) Hit Sets. The number of hits for each query obtained by S4 and CAS STN are listed in Table III. The other systems gave answers with minor variations (a few hits) due to different input conventions and search algorithms.

The queries C6 and H4–8 were designed to further test the internal accuracy of S4. Although there was always excellent agreement between the different systems, these queries provided an intrasystem check. Simple addition of the numbers of hits from queries C6 and H4–7 and subsequent comparison with the results from the generalized query H8 lead to an apparent discrepancy. There are four less hits retrieved with query H8. This discrepancy can be easily explained by examining these four hits. These are shown in Figure 5 and it can be seen that each of these hits contains two structures posed in queries H4–7. H8 is effectively an OR of the hit lists, and redundancies are removed. The same holds for the B set, whereby a few tellurium-containing structures are also recovered. This simple test demonstrates the self-consistency and accuracy of S4.

(b) Screening Efficiency. This comparison can only be made

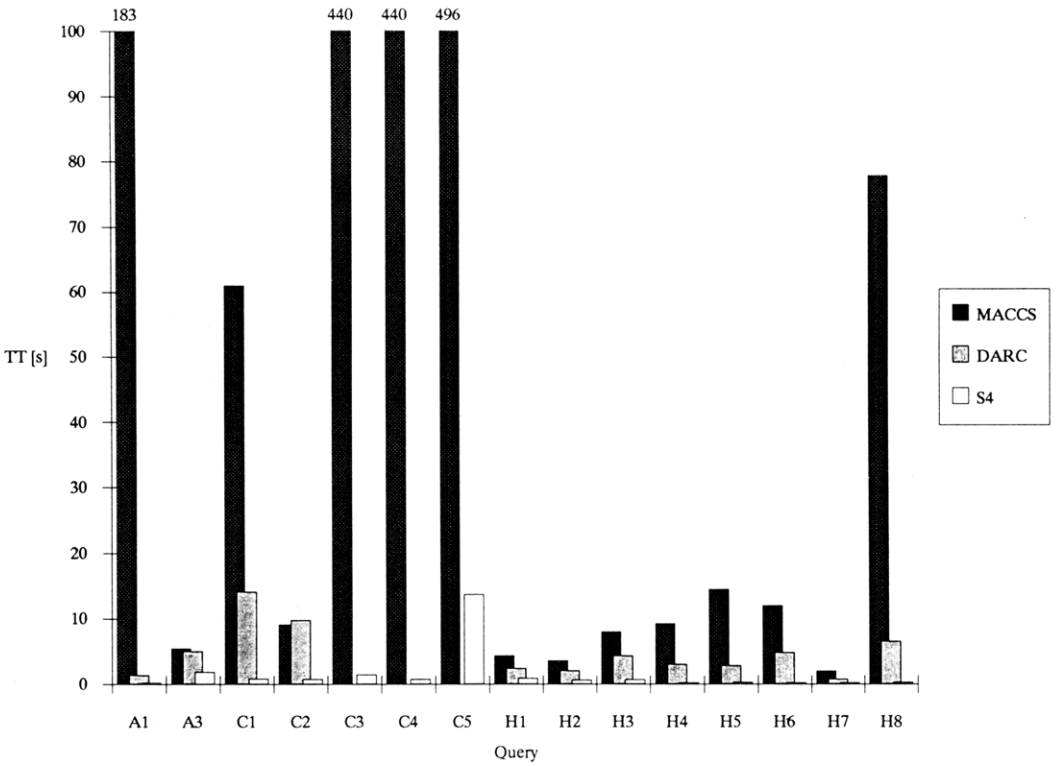

Figure 8. Phase 2 task times (seconds) for well-defined queries.

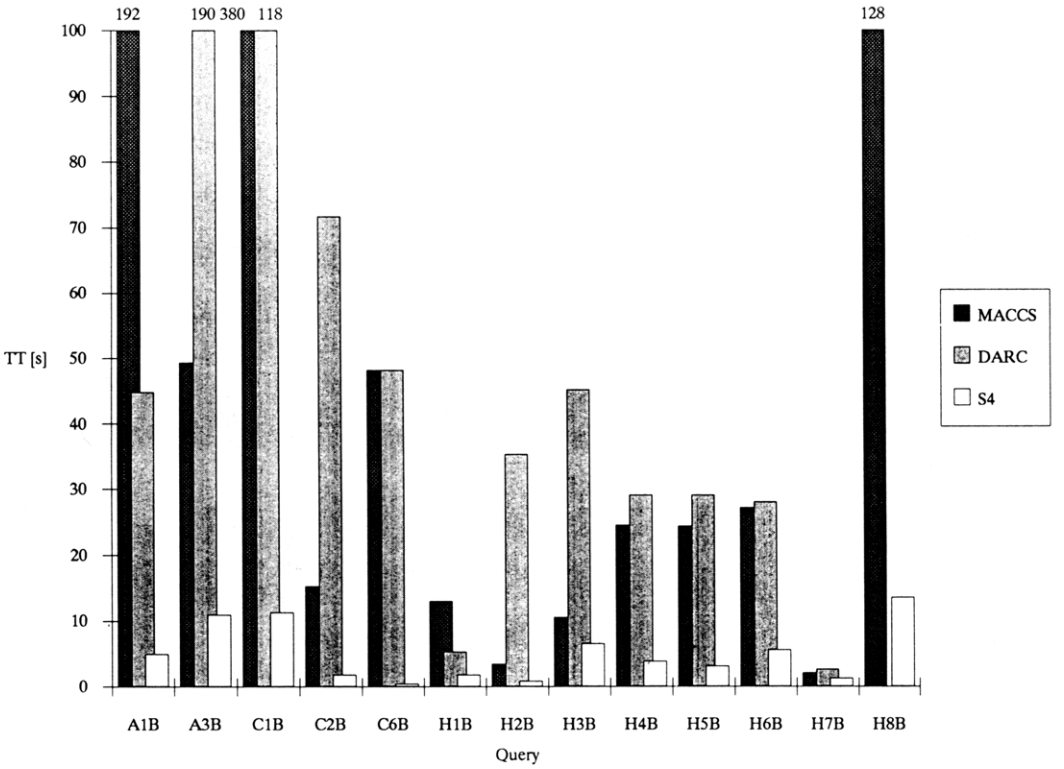

Figure 9. Phase 2 task times (seconds) for fuzzy queries.

between systems with a conventional screening phase; thus, the systems MACCS, DARC, and CAS STN are compared. Comparison is made by using the screening efficiency index (SEI), which is derived from the formula

$$\text{SEI} = (\text{no. of hits}/\text{no. of screens}) \times 100$$

This is measured for each query, and the results are shown in Figures 6 and 7. The most efficient screening has an SEI of 100.

The more efficient the screening, the shorter the search times. With very efficient screening the system does not waste time on too many unfruitful ABAMs. High efficiency not only gives short times but also has an influence on the largest practical file size, since with inefficient screening the system limit of candidates will be reached at an earlier stage.

As can be seen from Figure 6, DARC has by far the most efficient screening for well-defined queries. As would be expected, the screening is most efficient for queries containing

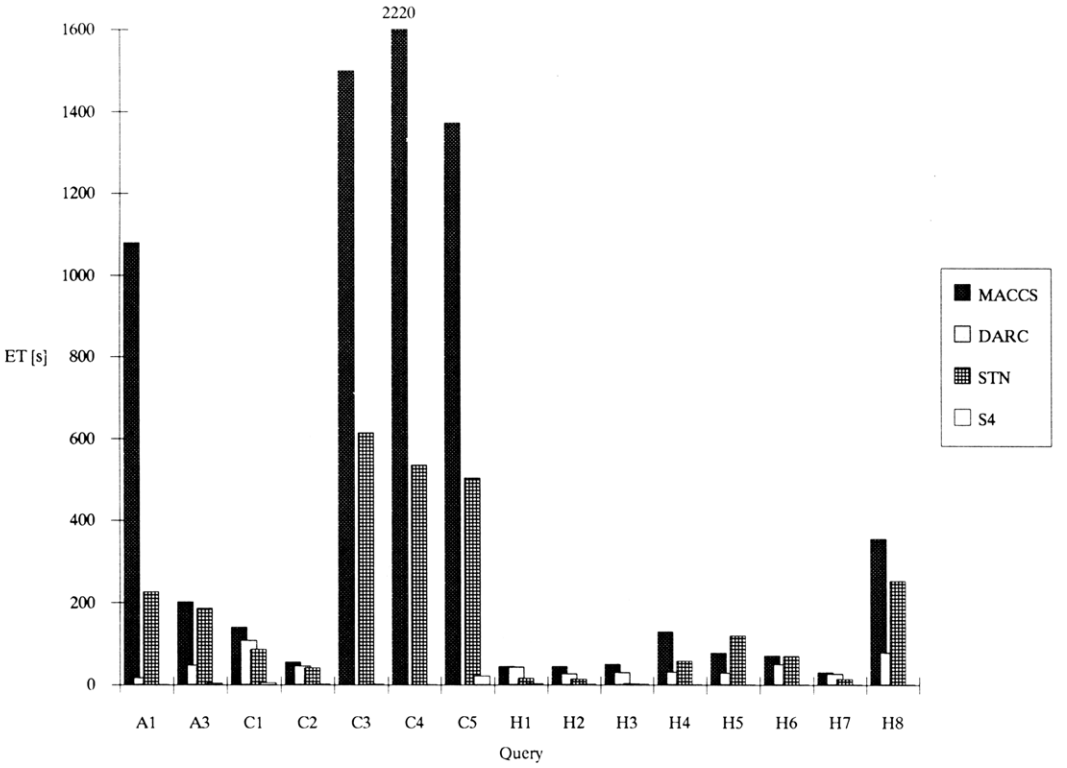

Figure 10. Phase 2 elapse times (seconds) for well-defined queries.

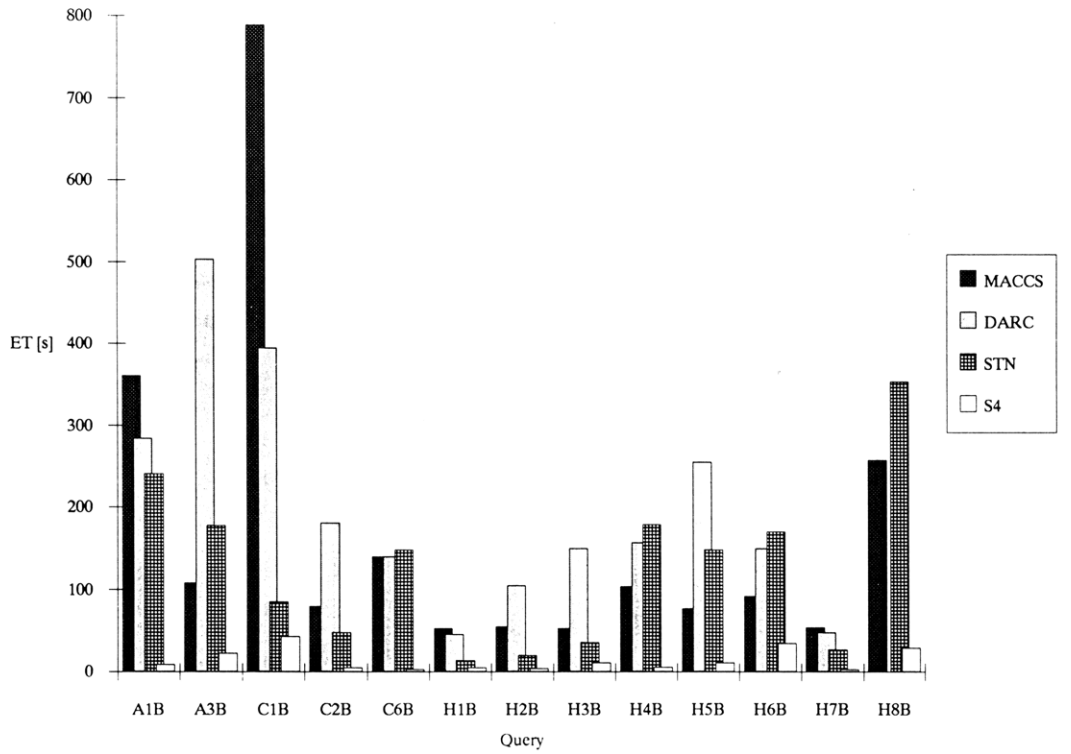

Figure 11. Phase 2 elapse times (seconds) for fuzzy queries.

heteroatoms. CAS STN had the next best results, followed by MACCS.

For the fuzzy set, shown in Figure 7, DARCs screening, although still the best, was slightly less efficient. CAS STN and MACCS had much improved efficiencies relative to the first set.

The effects of bad screening are not always noticed by the user. Query H5 had an SEI of 100 for DARC and less than 1 for MACCS, but the task and elapse times show little difference. This is because for this query MACCS retrieves relatively few candidates (ca. 5K) and the ABAM does not take long; this, combined with the fact that DARC has a time-intensive screening system, compensates for the inefficient screening. The situation is much different for A1. The SEIs are still very different, but the actual number of candidates retrieved by MACCS (95K) requires so much time to process that the search took over 3 min.

The similar screening efficiency for MACCS and DARC in the fuzzy set combined with the longer screening times for DARC virtually removes the advantages of the FRELs for

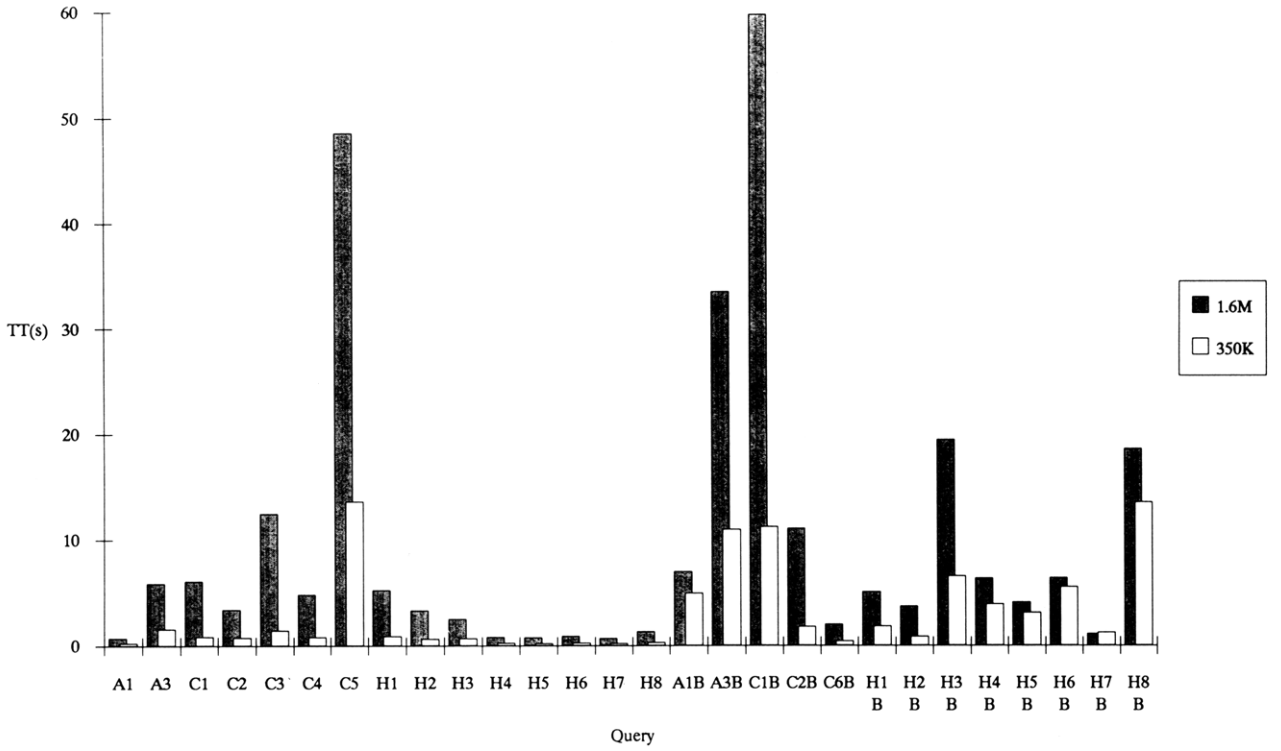

Figure 12. S4 task times (seconds) for 350K and 1.6M structure files.

Table IV. Overall Screen Efficiency Indices

system	SEI
MACCS	25
CAS	32
DARC	71

Table V. Average Task Times (Seconds) for Phase 2

system	defined set	fuzzy set	all
MACCS	32.6	65.8	49.2
DARC	4.7	53.9	29.3
S4	0.6	4.4	2.5

such queries. For cases such as the heterocycles, which are screened efficiently, MACCS is indeed faster. The overall screen efficiency indexes calculated as the average SEI from all common queries are shown in Table IV.

(c) Search Times. The task times for the well-defined set are plotted in Figure 8 and for the fuzzy set in Figure 9. The respective elapse times are shown in Figures 10 and 11. Since the elapse times generally reflect the task times, the discussions will be limited to the task times. The average task times are listed in Table V.

The task times of MACCS increase slowly with the fuzziness of the query. This slow increase is primarily due to inefficient screening, which does not provide much discrimination with well-defined queries. Thus, in going from well-defined to fuzzy queries, the number of ABAMs required rises more slowly than the number of hits. MACCS has by far the slowest task times. DARC with similar SEI for both well-defined and fuzzy queries exhibits accordingly a larger rise in the task times. The task times for S4 rise also with the fuzziness of query but not so steeply as with DARC. S4 has by far the fastest task times, being an order of magnitude faster than DARC or MACCS, and when compared directly with MACCS including queries C3–5, the difference approaches 2 orders of magnitude.

This large advantage of very short task times for S4 is even more noticeable to the user when the file size increases. S4 enables the user to pose queries which, because they either take too long or give too many hits, are just not possible with the other systems. The behavior of task times with file size is discussed later.

TASK TIME RESULTS

MACCS. The slowest queries can be divided into three groups: cases where the screening was bad (A1, A1B, and C4), cases where there were a great many hits (C3 and C5), and cases where the ABAM took inordinately long (C1 and C1B). There is little to be done for C3 and C5; many hits give rise inevetibly to long search times. However, improvements to the screen library would help with queries like A1, and it appears that the ABAM algorithm could be in need of refinement for complicated ring systems, such as C1, which have very long search times.

CAS STN. The screen library for the CAS system seemed much better balanced than that used by MACCS and was able to cope generally well with most queries. The elapse times were measured on a different (faster) machine and cannot therefore be compared to those of the other systems, but are generally only longer when there are a great many hits. The elapse times were measured at a time when the file was only available for testing and not nearly so heavily used as at present.

DARC. The set of well-defined queries presented little problem to the DARC system. The slowest queries were A3B and C1B; these queries provide few screens, and the long chain of A3B is in any case not well-defined in terms of FRELs. With the exception of A3B the screening stage in DARC always took longer than the ABAM. Clearly, with all systems there is a balance to be struck between the work in the screening and the work in the ABAM. Whether the optimal balance has been struck is not clear, but it would be interesting to see if there is any difference with a very much larger file.

S4. There is not much variation in the time taken to process a query within each set. This desirable feature can be traced to the well-balanced coding of the search file and to the efficiency of the search algorithm. This brings many advantages to users in that they do not need to be concerned with whether a particular search will take too long or exceed certain limits.

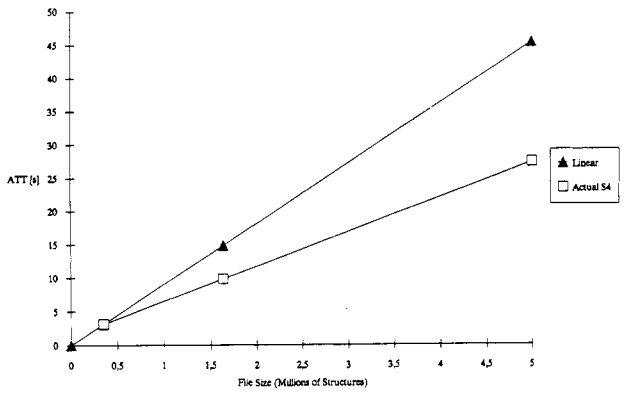

Figure 13. Average task time for S4 extrapolated for 5M structures.

Relationship between File Size and Task Times for S4. It was only possible to measure this relationship for S4. The updated on-line file which contained the file card heterocyclics in addition had 1 635 258 compounds. This file was used to determine the relationship. Basing an analysis on only two points is obviously not the best statistical practice, but nevertheless the information is very much of interest. Further results will be published as soon as possible. The same queries as for phase 2 were tested, and the task times are plotted in Figure 12. As expected there is a rise in the task times. The average task times have been plotted in Figure 13, and it can be seen that the rate of increase in search times is very much less than linear. Additionally, the average task times for S4 have themselves been linearly extrapolated to 5M compounds (the actual results are again expected to be less than linear). As can easily be seen, the extrapolation provides estimates of results which still give an excellent average task time for 5M compounds, which even with the overestimation in the linear extrapolation is still less than that for MACCS or DARC with 350K compounds.

The coding of S4 is such that there is no theoretical limit to the number of structures that can be encoded. Further tests will be carried out on future files, and it is expected that this linear extrapolation for 5M greatly overestimates the increase.

CONCLUSIONS

HTSS. With much shorter retreival times this system provides, with certain exceptions, a significant improvement on previous systems. Some classes of queries seem to pose problems. The results when used with small or medium size files (up to 600K) are quite acceptable; it remains to be seen what the performance is like with very large files.

MACCS. MDL has developed a generally excellent software system that provides chemists with most of the tools they need. The results show that MACCS is clearly limited to smaller sized files (ca. 350K).

CAS STN. At the time of testing the on-line system gave good responses. The new multiprocessor system in Columbus is now on-line and giving excellent results. The performance of the MVSSS implementation with very large files in Karlsruhe is being eagerly awaited. The system itself allows sophisticated generic searching lacking only the stereochemical capabilities.

DARC. This system has good search options including generic and stereochemical isomer searching and, in Markush DARC, markush searching. The performance with large files was good. It has been proven to be able to search very large files such as EURECAS.

S4. In terms of search times, this system gave by far the best performance of those tested, and it is able to search for all of those queries that can be posed by other systems. S4 is the system that Beilstein uses in-house and is operated by Dialog on the Beilstein on-line file. The implementation at Dialog demonstrates the excellent behavior of S4 in a multiuser environment. Other hosts are currently assessing S4 for on-line implementation; at the time of writing no final decisions have been reached. S4 is also available as an in-house system.

A large advantage of the S4 system is that the encoding makes it ideally suitable for CD applications. This will bring the databases to the chemist's bench, starting a new era in chemical information technology.

ACKNOWLEDGMENT

The Beilstein Institute thanks wholeheartedly the firms and their representatives who helped with the installation of software for the carrying out of this benchmark.

REFERENCES AND NOTES

(1) Willett, P. A Review of Chemical Structure Retrieval Systems. *J. Chemom.* **1987**, *1*, 139–155.
(2) Willett, P. *Similarity and Clustering in Chemical Information Systems.* Research Studies Press Ltd.: Letchworth, Hertfordshire, England, 1987; pp 10–18.
(3) Ash, J. E.; Chubb, P. A.; Ward, S. E.; Welford, S. M.; Willett, P. Chemical Structure Search Systems and Services. In *Communication, Storage and Retrieval of Chemical Information*; Ellis Horwood: Chichester, U.K., 1985.
(4) Stobaugh, R. E. Chemical Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 271–275.
(5) Barnard, J. M. Problems of Substructure Searching and Their Solution. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 113–126.
(6) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113.
(7) Wipke, W. T.; Dyott, T. M. Stereochemically Unique Naming Algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4825–4834.
(8) Petrarca, A. E.; Lynch, M. F.; Rush, J. E. A Method for Generating Unique Computer Structural Representations of Stereoisomers. *J. Chem. Doc.* **1967**, *7*, 154–165.
(9) Lynch, M. F. The Microstructure of Chemical Databases and the Choice of Representations for Retrieval. In *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldmann, R. J., Hyde, E., Eds.; Wiley: New York, 1974.
(10) Lynch, M. F. Screening Large Chemical Files. In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, U.K., 1974; pp 177–194.
(11) Graf, W.; Kaindl, H. K.; Kniess, H.; Warszawski, R. The Third Basic Fragment Search Dictionary. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 177–181.
(12) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS Online Search System. 1. General Design and Selection, Generation and Use of Search Screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(13) Farmer, N.; Amoss, J.; Farel, W.; Fehribach, J.; Zeidner, C. The Evolution of the CAS Parallel Structure Searching Architecture. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 283–295.
(14) Ahrens, E. K. F. Customisation for Chemical Database Applications. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 97–111.
(15) Shlevin, H. H.; Graham, M. M.; Pennington, D. F.; von Wartburg, W. Integration of Chemical Structures with Information in Support of Business Needs. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 79–90.
(16) Dubois, J.-E.; Panaye, A.; Attias, R. DARC System: Notions of Defined and Generic Substructures. Filiation and Coding of FREL Substructure (SS) Classes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74–82.
(17) Attias, R. DARC Substructure Search System: A New Approach to Chemical Information. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 102–108.
(18) Nagy, M. Z.; Kozics, S.; Veszpremi, T.; Bruck, P. Substructure Search on Very Large Files Using Tree-Structured Databases. In *Chemical Structures*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 127–130.
(19) Hicks, M. G.; Jochum, C.; Maier, H. Substructure Search System for Large Chemical Databases. Proceedings of the IXth ICCCRE. *Anal. Chim. Acta* (in press).