Evaluation of the Screening Stages of the Sheffield Research Project on Computer Storage and Retrieval of Generic Chemical Structures in Patents

J. D. Holliday, G. M. Downs, V. J. Gillet, and M. F. Lynch*

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Winfried Dethlefsen

BASF AG, Ludwigshafen/Rhine, Germany

Received June 28, 1993®

An evaluation is given of the search strategies used in the screening stages of the Sheffield generic structure system. The results of several searches are presented; a detailed account is given of their performance together with a discussion of outstanding problems and weaknesses. The screening stages are of two types, a fast bitscreening stage using inverted files of bitstrings and a slower but more discriminating stage which operates on files of reduced graphs. Bitscreens are used to represent the presence, absence, or possible presence of features attributable to the structure. These features are of two types: fragment descriptors and ring descriptors. The results of searches using these screening stages are encouraging. The diversity of the methodology used is reflected in the results; the strengths of each method are shown to resolve the weaknesses of others. The final stage of search, the refined search, is introduced. This is a localized atom-level search which investigates, in greater detail, the mappings between reduced graph nodes which result from the reduced graph screening stage. This is the most discriminating search stage and is therefore expected to be highly efficient.

1. INTRODUCTION

The series of conferences held at Noordwijkerhout[5,6] has used the subtitle "The International Language of Chemistry". This is a particularly suitable subtitle when considering the field of generic chemical structures since it is the language used by the chemist, and by the patent agent, to describe the invention which is responsible for many of the problems inherent in dealing with generic chemical structures. This stems from patent applicants' need to disclose as little detail as possible about the product while complying with the legal requirements of a patent application. The generic description, as a result, usually represents a large or even infinite number of structures under a single disclosure.

The structure descriptions used by the chemist in a patent have been well documented[7,8] and reveal the use of several notations, such as structure diagrams, line formulas, specific or generic radical names or nonstructural, textual information, each of which describes a "partial structure". There is also much use of structural variation within and between the partial structures. Briefly, the types of structural variation are as follows:

Substituent Variation. This type is a list of alternative substituents on a ring, for example, any one of which is possible, e.g., "R1 is methyl, ethyl, or propyl."

Position Variation. This variation has alternative positions of attachment through which a substituent is connected to the parent partial structure, any one of which is possible, e.g., "monochlorophenyl".

Frequency Variation. This case is a variation in the frequencies of partial structures, e.g. "$-(CH_2)_n-$, n is 1, 2, or 3".

Homology Variation. Here the chemist is using a standard nomenclatural term to represent a series of compounds with common structural features; i.e., they are homologous to each other. The series may be finite or infinite and is often qualified by structural property descriptions defined in terms of numerical values of ranges, e.g., "alkyl 1–4C".

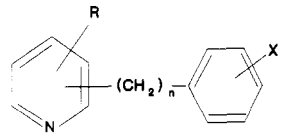
The interpretation of the language of the chemist,[7] its ambiguities, vagueness, and nonstandardization, and the sheer complexity of the structures which the chemist is disclosing pose great problems with regard to computer representation and searching. Indeed, size and complexity of generic structures are increasing all the time, making more difficult the tasks of defining an ideal representation and searching databases of such representations.

These problems have been the subject of the studies carried out at Sheffield for over a decade now. The aims of the project have been 3-fold: First, the design of a formal language which can readily be generated from the chemist's language and which can be interpreted by a computer program without ambiguity. The chemical patent description is characterized by an invariant partial structure to which variable partial structures are attached. The variable partial structures are designated by an assignment statement of the form "R is methyl or ethyl". The Sheffield language, GENSAL,[9,10] expresses these assignment statements in a formal grammar, for example, using the case above, "R1 = methyl/ethyl". Figure 1 shows an example generic structure from the patent literature. The GENSAL notation for this structure is shown in Figure 2. This example illustrates the partitioning of the generic structure, the diversity of notation used to describe each partial structure, and the types of structure variation described above.

The second aim of the project was to design a suitable data structure to represent the generic structure. The data structure developed at Sheffield is the extended connection table representation (ECTR).[11] The relationships and logic between partial structures are represented in the ECTR as an AND/

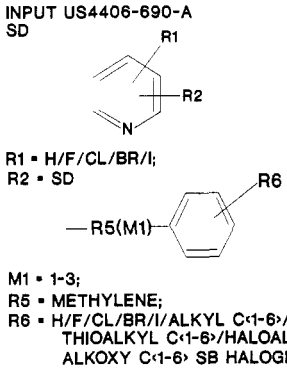
* Abstract published in *Advance ACS Abstracts*, January 15, 1994.

0095-2338/94/1634-0039$04.50/0



n is 1, 2 or 3;
R is H, F, Cl, Br, or I;
X is H, F, Cl, Br, I, or alkyl, alkoxy, thioalkyl, haloalkyl or haloalkoxy having 1-6 C.

Figure 1. Example patent structure.



M1 = 1-3;
R5 = METHYLENE;
R6 = H/F/CL/BR/I/ALKYL C<1-6>/ALKOXY C<1-6>/ THIOALKYL C<1-6>/HALOALKYL C<1-6>/ ALKOXY C<1-6> SB HALOGEN.

Figure 2. Example of the GENSAL input language.

Table 1. GENSAL Parameters

A	Total atom count
C	Total carbon count
T	Acyclic ternary branch count
Q	Acyclic quaternary branch count
E	Number of localised olefinic unsaturations
Y	Number of localised acetylenic unsaturations
RC	Ring count
RN	Ring atom count
RS	Ring substitution count
RF	Ring fusion count
RA	Normalised ring count
RZ	Ring heteroatom count
Z	Total heteroatom count

OR tree, the leaf nodes of which are representations of the partial structure. Partial structures which are expressed using specific notations, such as specific nomenclature or structure diagrams, are represented as partial connection tables in the ECTR. Generic nomenclatural expressions, those which exhibit homology variation, represent not one specific structure but one of a possibly infinite number of structures, all of which are related in terms of structural property. Generic partial structures cannot therefore be represented as a single connection table and are represented at Sheffield by means of a set of parameters, each of which describes a structural feature and is quantified by a value or range of values. The parameters were initially defined as shown in Table 1; an example list of values would be, for alkyl C 1–4, as follows:

A⟨1–4⟩C⟨1–4⟩T⟨0–⟩Q⟨0–⟩E⟨0⟩Y⟨0⟩RC⟨0⟩...RZ⟨0⟩Z⟨0⟩

The respective ECTR for the structure of Figure 2 is shown schematically in Figure 3.

The third aim of the project was to develop efficient search representations and routines in order that a file of generic structures may be interrogated using a query structure as a search key. This paper describes the search representations which have been investigated and reports on the results of several searches using these representations.

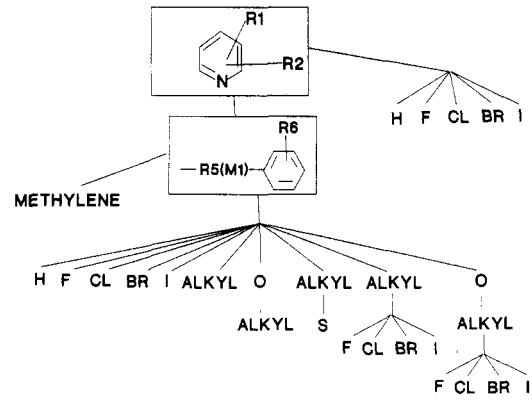


Figure 3. Schematic ECTR for the structure shown in Figure 2.

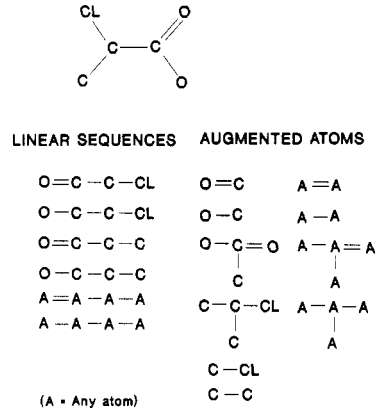


Figure 4. Example linear sequence and augmented atom fragments.

2. SCREENING STRATEGIES

The ECTR is the most detailed representation of the generic structure, describing all the structural information required as completely as possible. It describes partial structures in terms of their atoms and bonds or parameter lists, interrelations between partial structures in terms of points of attachment, and the logic between partial structures. The size and complexity of generic structures means that a search based on a representation of such detail would be far too computationally demanding for a complete file of structures. There is therefore a requirement for elaborate screening mechanisms based on less detailed representations of the structure. This reduces the number of candidate file structures which must be searched at an atom–bond level.

At Sheffield, there are two such stages of screening: fragment screening,[2,12–16] a bitstring, the elements of which represent the presence or absence of a structural feature; reduced graph screening,[1,17,18] a reduced representation of the generic structure in which graph nodes are defined in terms of aggregates of atoms.

Fragment Screening. A dictionary containing over three thousand specific fragment descriptors, each one having an associated bit vector in a bitstring, is used in the fragment screening stage. The fragment descriptors are of two types:

1. Sequence fragments describe linear sequences of atoms and bonds of various lengths, (4, 5, or 6 atoms) and various levels of description.

2. Augmented atoms describe a central atom, the atoms to which it is connected, and the bonds which connect them, again at various levels of description.

Figure 4 shows some example fragments which occur in the structure shown. The various levels of description mean that a single linear sequence of atoms can produce several fragments by a process of specificity reduction.

In addition to fragment screens, ring screens provide information about the cyclic parts of the structure.[3,19] Ring screens categorize each ring of the structure according to size, composition, and degree of fusion. Again, each categroy has a series of bit vectors in the bitstring, but in this case there is no need for a dictionary of descriptors as they are directly assigned.

In generic structures, the variability of parts of the structure means that those parts are optional to the overall structure; i.e., they *may* occur in the structure. The fragments which occur in these variable parts may not, therefore, be common to all of the specific structures which the generic structure describes. These fragments can be distinguished from those which are common to all the specifics by using two bitscreens to represent the fragments. The bitscreens used are then the MUST screens, representing the essential, or common, fragments, and the POSS screens, representing the union of the essential fragments and the optional, or noncommon, fragments.

In order to generate the fragments and differentiate between those which are common to all specifics and those which are not, it is not necessary to explicitly generate all the specific structures which the generic structure describes. Instead, the logical structure of the ECTR, the essence of which is an AND/OR tree, allows the accumulation of information to be carried out from the leaves of the structure, the partial structures themselves, to the root of the tree in a process called the *bubble-up*. Bubble-up is a two stage process starting with the identification of localized information within the partial structures during a top to bottom traverse of the ECTR. The process if fairly straightforward with specific partial structures and requires only a path trace of the partial connection table. For generic partial structures, however, there are no real atoms to trace, and graph theoretical techniques have been used to identify the subset of fragments[20,21] and rings[22] which may be contained within the possible partial structures denoted by the intensional description, the generic radical term, and determined by the intensional representation, the parameters. This information, be it fragment or ring descriptors or even molecular formulas, is accumulated using logical operation on the way back up to the root. In bitscreening, the result of the bubble-up is the pair of bitscreens which represent sequence fragments, augmented atom fragments, and rings.

Reduced Graph Screening. Fragment screens represent a detailed view of localized chemistry within the structure, i.e., at an atom–bond level or a ring level, but in no way reflect the structure's overall topology and indicate the logic relationships in simple terms only. There is therefore a requirement for a representation which reflects the logic and topology more fully and yet does not contain the detail of the ECTR.

A reduced chemical graph represents a simplisitc view of the structure while retaining the gross topological and logical relationships between components. Nodes of the graphs represent regions of the structure which are chemically or structurally similar dependent on the criterion for reduction of the original chemical graph to the reduced chemical graph. Several types of reduction have been investigated and reported by the Sheffield project,[1,17,19,23] the most notable of which uses nodes determined by aggregates of connected ring atoms, nonring carbon atoms, and nonring heteroatoms. The AND/OR logic of the ECTR is retained in the reduced graph using a tree-structured graph which contains AND and OR branches, and optional nodes arise due to occurrence of hydrogen in a list of alternative substituents. A further node

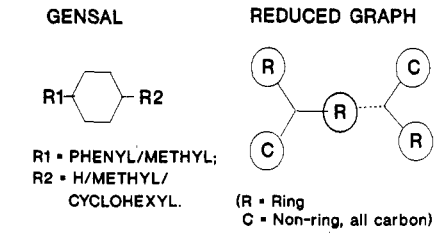


Figure 5. Example reduced graph.

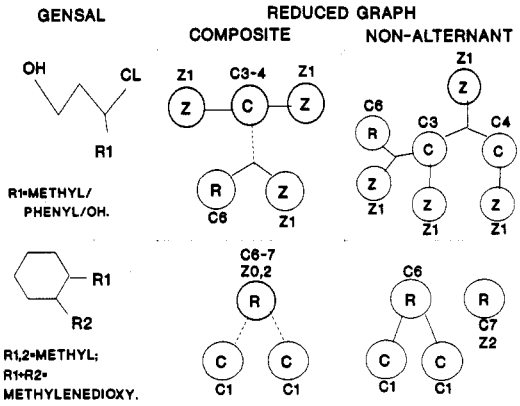


Figure 6. Composite and non-alternant reduced graphs.

type, "unsettled", has been added to deal with nodes emanating from generic partial structures whose cyclic or acyclic nature is not known, such as "radical" or "aliphatic". The ring parameters for these partial structures have values of zero or more, e.g., RC⟨0–⟩RN⟨0–6⟩.

The reduced graph of Figure 5 shows a simple structure in which the OR relationships are indicated by branches and an optional connection is shown by a broken line. The invariant nodes, those which are common to all alternative full structures, are shown in bold. In addition to the node label, nodes are also further colored by the inclusion of structural information in the form of a subset of the parameters used to represent generic partial structures in the ECTR. These parameters are quantified by values or ranges of values which are generated during a path trace of specific partial structures or directly from the parameters of generic partial structures.

The paper presented at the preceding conference at Noordwijkerhout[24] describes a type of reduced graph in which such nodes are collapsed into a single instance where two or more nodes of the same type are alternatives to each other. In this type of reduced graph, the "composite reduced graph", the parameter values are conflated into ranges; the conflated parameter ranges for the node representing the alternatives methyl or isopropyl would then be A⟨1,3⟩C⟨1,3⟩E⟨0⟩Y⟨0⟩T-⟨0–1⟩Q⟨0⟩Z⟨0⟩. There is a reduction in the number of nodes in this type of reduced graph which improves storage requirements and search times, but this is offset by a loss of topological information and mappings to the original structure.

In this paper, a more explicit form of reduced graph, the "non-alternant reduced graph", is used in which no node collapsing is carried out. The resulting graph retains a direct mapping onto the orignial ECTR structure and contains more precise structural information. This should result in higher search performance levels as it represents a more discriminating search; this is discussed in a later section. The graphs can, however, be very large, some in excess of 5000 nodes, and may contain much duplication of subtrees. Figure 6 illustrates several structures, in GENSAL notation, together with their associated composite reduced graphs and non-alternant reduced graphs (Z = nonring, all heteroatom). The discon-

nected non-alternant reduced graphs of the second example represent alternatives to each other.

A local ring screen has been added to the nodes which represent ring components of the structure. This ring screen has the same representation as the ring screen section of the bitscreens but is a localized representation of the rings within the node only.

The mappings of the reduced graph nodes to their relevant partial structures (note that this is not always a one-to-one relationship since a node may span several partial structures and a partial structure may reduce to more than one node) are necessary for the final search stage, the "refined search", which is an atom-level search strategy. This is described in a later section.

3. VALIDATION AND TESTING

The Databases. Five databases are used to test the performance of the search stages used by the Sheffield project. These are as follows.

Db1. A database of 2025 generic structures, i.e., structures which exhibit position, substituent, frequency, and homology variation, created from those patent abstracts in sections B and C of the *Basic Abstracts Journal* published by Derwent Publications Ltd. during weeks 8340–8419 inclusive. Several of these structures failed at various stages of processing due to their size and complexity.

Twenty structures failed to be processed by the fragment screen generation program, 30 failed to be processed by the ring screen generation program, and 54 structures failed to be processed by the reduced graph generation program. The latter problem was due to reduced graphs which exceeded a maximum of 5000 nodes in size. The total number of structures which were processed completely, and hence the number of structures reported in this study, is 1957.

Db3. A database of 77 structurally explicit generics, i.e., structures which exhibit all of the features of generics except homology variation. These structures have been created by selecting every twentieth Db1 structure and altering the definitition of substituent values, replacing generic radical terms with corresponding specific examples of the homologous series. A certain amount of repartitioning has also been introduced. In all cases, at least one of the specific structures described in the Db3 structure is an example of one of the specific structures described by the original (or parent) Db1 structure. As a result, each structure in Db3, when used as a query structure, should retrieve its corresponding parent structure in Db1.

Of the 77 structures in Db3, all were processed by the generation stages except for two structures which failed to produce reduced graphs. This is because they exceeded the limit of 5000 nodes. Results are reported for all structures in the bitscreening stage but thereafter are reported for the 75 structures successfully processed.

DbS. A database of 1205 example specific structures manually extracted from the first 1205 structures in Db1. Each DbS structure is a single specific instance of the class described by its corresponding parent structure in Db1 and should therefore retrieve that structure when used as a query.

All 1205 structures were successfully processed by all stages of generation.

Db7. A database of 43 generic structures, exhibiting all four types of structure variation, taken from a random sample of non-English language patents published in Derwent's Central Patent Index in 1984.

Of the 43 structures in Db7, one structure is not processed during reduced graph generation. Results are reported for all 43 structures in the bitscreening stage but thereafter are reported for the 42 structures successfully processed.

Db8. A database of 28 generic queries, exhibiting all types four of structure variation, which were obtained from several industrial sources.

Of the 28 structures in Db8, one structure is not processed during reduced graph generation. Results are reported for all 28 structures in the bitscreening stage but thereafter are reported for the 27 structures successfully processed.

Search Methodology. For the purpose of validation, searches are carried out using a complete file of query representations against a complete file of database representations using both forms of representation, reduced graph and bitscreen. Under normal circumstances the fast bitscreening stage would be used to screen out many of the database structures in order that fewer are sent to the slower reduced graph search stage.

Bitscreening is carried out using inverted files of bitscreens in order that, rather than matching every query bit with every file bit for every query structure against every file structure, the first bit is interrogated for the whole database, followed by the second, and so on. The set of candidate file structures is then reduced as each bit is interrogated, and the process may be halted at the desired stage. This process allows a possible reordering of bits to allow those which are highly discriminating to be used first.

Searches are carried out for full structure queries against the file full structures. The criteria for a match have been described;[2,25] briefly they are the inclusion of the MUST screen of the query structure in the POSS screen of the file structure and the inclusion of the MUST screen of the file structure in the POSS screen of the query structure.

The queries used in this study for the reduced graph search stage represent full structures. A graph matching algorithm is used to match nodes of the same type. Included within the node–node match stage is a process whereby the parameter information of the query nodes may be compared with those of the file nodes. This process results in a match where there is an overlap of at least one value in each of the parameter ranges of the query node with the corresponding ranges of the file node. For ring nodes there is also a process for matching the ring bitscreens.

4. SEARCH RESULTS

Four search combinations were carried out using the databases described above. These were as follows: (1) Db3

Table 2. Full Structure Search Results

query database	Db3 Db1	DbS Db1	Db7 Db1	Db8 Db1
FRAGS				
min	16	14	19	18
max	72	95	404	331
H/Q	35.5	28	52.5	50.7
median	32	27	33	33.5
RINGS				
min	8	0	104	102
max	885	362	1239	968
H/Q	253	194	304	370
median	203	188	227	350
BITS				
min	1	0	12	16
max	66	67	374	313
H/Q	29.5	22.8	47	43.8
median	28	23	29	31
R.G.				
min	0	0	0	0
max	4	4	2	2
H/Q	1.1	0.48	0.18	0.074
median	1			
overall result				
H/Q	**1.0**	**0.42**	**0.14**	**0.074**

used as a query database to search Db1, i.e., 77 structurally explicit generic queries against 1957 generic file structures; (2) DbS used as a query database to search Db1, i.e., 1205 specific structure queries against 1957 generic file structure; (3) Db7 used as a query database to search Db1, i.e., 43 randomly-selected generic queries against 1957 generic file structures; (4) Db8 used as a query database to search Db1, i.e., 28 generic industrial queries against 1957 generic file structures.

Table 2 gives a complete overview of the results obtained by all of the searches presented.

Searches of Db3 against Db1. Bitscreen searches of Db3 against Db1 give a fair indication of performance in terms of precision and recall, although the structures in Db3 are not good examples of queries. The average number of hits per query structure in a full structure search of Db1 was 29.5 (or 1.5% of the database; minimum number retrieved was 1; maximum number retrieved was 66). This represents a screenout of 98.5% of the database which is an encouraging figure. A more realistic measure which eliminates anomalous values is the median of the number of hits per query, 28 in this case. Seven of the Db3 structures failed to retrieve their parent structure which would seem to indicate a recall problem. The reason for these failures was, however, that the query structures were incorrectly encoded and cannot therefore be regarded as structurally explicit examples of their parent structure. Five of these failures were picked up by the fragment screens and two by the ring screens. The effectiveness of the ring screens is not as apparent as that of the fragment screens (253 hits per query for ring screens, 35.5 for fragment screens), although the ring screens do complement the fragment screens considerably. The graph in Figure 7 illustrates the relative performance of the bitscreens.

Reduced graph search results show, as expected, greatly improved performance over bitscreening. The 75 Db3 structures tested produced an average of 1.11 hits per query (0.057% of the database, a screenout of 99.94% of the database) with a maximum retrieval of four structures and a median value of 1. This represents a high level of performance and is an improvement on the composite reduced graph which produced an average of four hits per query (maximum of 7). Of the 75 structures tested, 12 fail to retrieve their parent Db1 structure. Five of these do not have a corresponding Db1 structure as they were not processed during the reduced graph

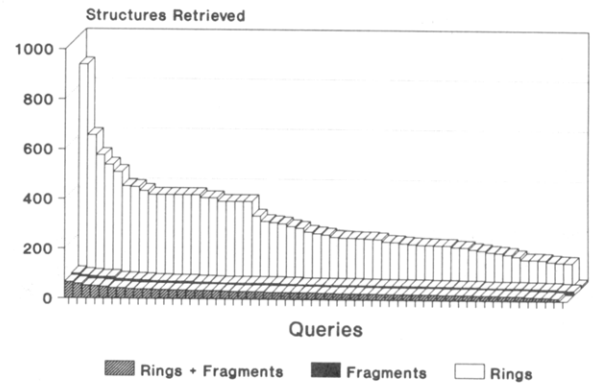


Figure 7. Full structure bitscreening: Db3 against Db1.

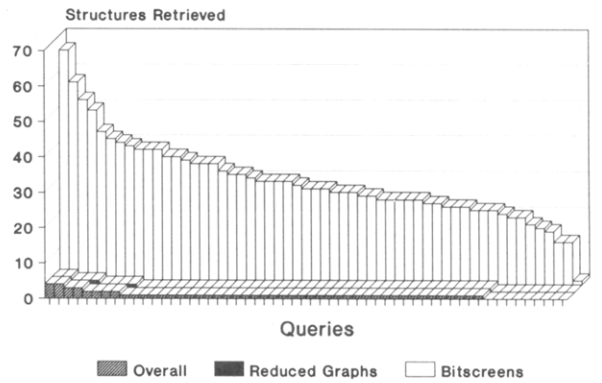


Figure 8. Overall screening performance: Db3 against Db1.

generation stage; 4 of the structures also failed to retrieve their parent in the bitscreening stage due to the reasons described above, and it is assumed that all recall failures were due to similar discrepancies. Figure 8 shows the relative performances of the search methodologies used.

A further seven file structures are screened out by combining the results of both screening methods. This represents a further reduction of 8% on the reduced graph search results to produce an overall result of 1.02 hits per query. This reduction reflects the advantage of using two different screening methods which complement each other.

Searches of DbS against Db1. Specific structures, when used as full structure queries against a file of generic structures, produced an average of 22.8 hits per query, or 1.17% of the database (maximum 67, minimum 0, median 23). The level of screenout, 98.83%, is again encouraging. Examination of the hit structures reveals that many of the file structures are retrieved consistently by a majority of the file structures. These file structures are all made up of a small root partial structure, often a single variable substituent identifier (such as R1), and at least one generic definition which denotes a wide extension, such as "radical". The result is a dense, or even black, POSS screen and a sparsely filled MUST screen producing constant recall. Again, fragment screens perform more effectively than ring screens; 28 hits per query for fragment screens (maximum 95, minimum 14, median 27), and 194 hits per query for ring screens (maximum 362, minimum 0, median 188).

Of the DbS queries, 30% do not retrieve their respective Db1 parent structure. This is mainly due to the selection of multiplier examples during the creation of DbS. The ECTR can represent multipliers, but these are a problem for all of the processing procedures: fragment generation, reduced graph generation, and ring screen generation. These procedures treat all multipliers as a value of 1, yet the DbS structures represent specific examples of any multiplier value described.

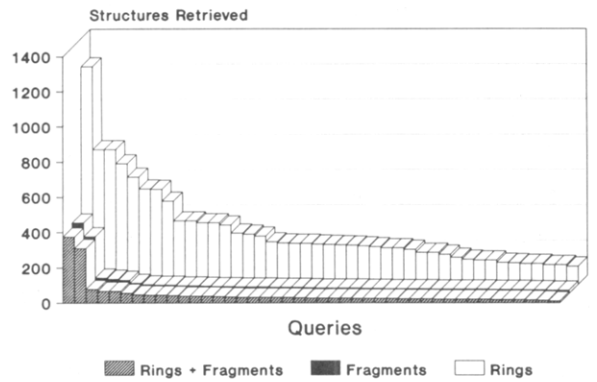


Figure 9. Full structure bitscreening: Db7 against Db1.

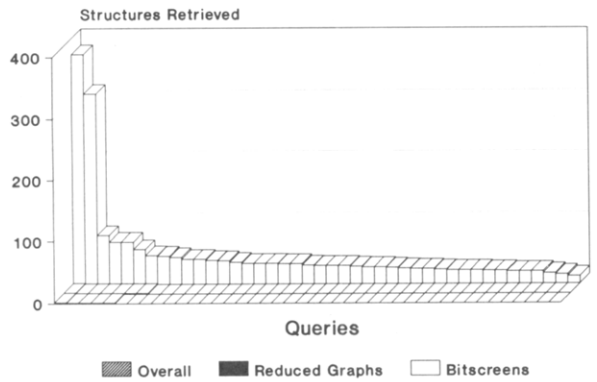


Figure 10. Overall screening performance: Db7 against Db1.

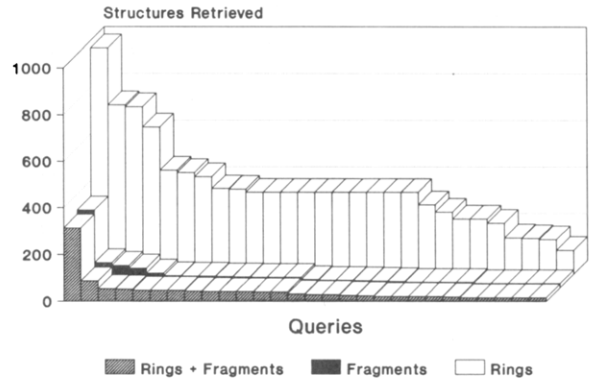


Figure 11. Full structure bitscreening: Db8 against Db1.

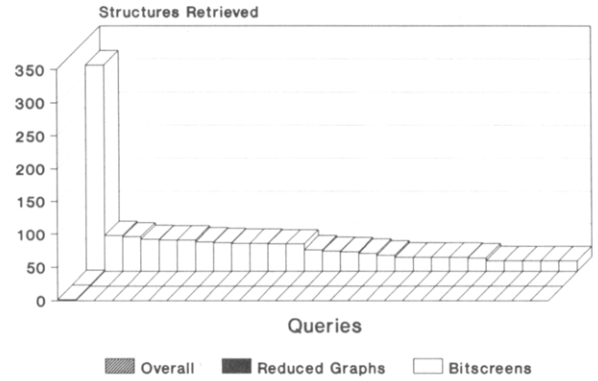


Figure 12. Overall screening performance: Db8 against Db1.

Multipliers remain as the last major problem in generic structure handling.

Reduced graph searches result in an average of 0.48 hits per query (maximum 4, minimum 0). Each DbS structure should retrieve its parent Db1 structure, but only 43.2% do so. Of those structures that do not retrieve their parent, 95% contain multipliers which are exemplified in DbS using values other than 1. This would explain the low value for recall.

Searches of Db7 against Db1. Full structure bitscreening of Db1 using Db7 as a query database produces an average of 47 hits per query (2.4% of the database, or 97.6% screenout). Two anomalous values (310 hits and 374 hits) show that a more accurate measure of performance is the median, 29 in this case. The maximum number of hits, disregarding these two values, is 79, and the minimum is 12. The results for the seprate screening stages are 52.5 hits per query for fragment screening (maximum 404, minimum 19, median 33) and 304 hits per query for ring screening (maximum 1239, minimum 104, median 227). The relative performance of the bitscreening stages is shown in Figure 9.

Reduced graph results represent a more accurate value for true screenout since the query structures in Db7 do not have an associated parent structure in Db1. The results are encouraging, an average of 0.18 hits per query (one query hits two database structures, six queries hit one database structure only, and the remaining queries retrieve no structures). The results represent a hit rate of one database structure in ten thousand or a screenout value of 99.99%. Two further structures are eliminated by combining both screening methods. The results of the reduced graph search are shown in Figure 10.

Searches of Db8 against Db1. Db8 represents the most realistic set of queries as they are examples of industrial queries which have been donated by several chemical information departments. Again, they do not have an associated parent structure in Db1. They have been used as both full structure and substructure queries for completeness. Full structure bitscreening of Db1 using Db8 produces an average of 43.8 hits per query (2.24% of the database, or 97.76% screenout). One anomalous value (313 hits) is reflected by a significantly lower median value of 31. The maximum number of hits, disregarding the anomalous value, is 87, and the minimum is 16. The results for the separate screening stages are 50.7 hits per query for fragment screening (maximum 331, minimum 18, median 33.5) and 370 hits per query for ring screening (maximum 968, minimum 102, median 350). The relative performance of the bitscreening stages is shown in Figure 11.

Reduced graph results are encouraging, an average of 0.074 hits per query (27 queries). Indeed, only one query retrieves database structures, two hits in all. No further structures are eliminated by combining both sets of results, as shown in Figure 12.

5. SUBSTRUCTURE SEARCH

Substructure searches were carried out for bitscreening alone. The results are encouraging with screenout values of 80–85%. The major problem with substructure searching is the use of expressions which denote wide extensions, such as radical. Such terms produce a POSS bitscreen which is densely filled or even black, i.e., all bit positions set to true, which means that all file structures which contain the term are retrieved in a substructure search. In Db1, 303 structures contain the term radical with no further structural qualification, such as number of atoms or number of rings. These structures are repeatedly retrieved in a substructure search. The screenout results should be improved by the use of reduced graph screening; this will be carried out in the near future. The results for bitscreening were as follows: Db3 against Db1, 362 hits per query (median 318); Db7 against Db1, 351

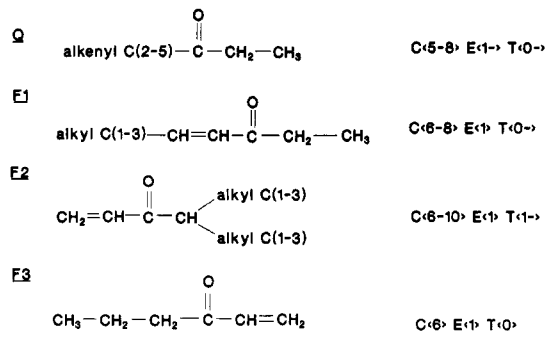


Figure 13. Refined search examples.

hits per query (median 317); Db8 against Db1, 402 hits per query (median 363).

6. THE REFINED SEARCH[4]

The result of the reduced graph search is a node mapping between nodes of the query file and those of the database file. This is in the form of a list of matched node pairs, one from the query and one from the file. At this stage, each pair of nodes has been matched at a basic level using the node labels and the local node parameter lists. The refined search is designed to compare the two nodes at a more detailed level, i.e., at the level of atoms and bonds in the case of specific components or parameters in the case of generic components.

Each node of the reduced graph has associated with it a list of those ECTR components from which it was generated; it is in effect mapped onto a part of the ECTR. The query node has a part of the query structure's ECTR to which it is mapped, and the database node has a part of the file structure's ECTR to which it is mapped. The process is then to match at a detailed level the respective areas of the two ECTRs represented by the two nodes. The simplest comparisons can be made when both parts of the ECTR are wholly specific, i.e., represented by atoms and bonds, or when both emanated from a single generic partial structure. The process is then either a local atom-by-atom match or a comparison of ECTR parameter values. In many cases, though, each reduced graph node may be mapped onto a part of the ECTR which contains a combination of generic components and specific components. The problem is therefore one of transparency between the generic and specific representations, and the matching mechanism must be able to translate between the two types of representation for the purpose of comparison.

A reduced graph node may represent any combination of ECTR components including the following:

- a single specific component contained within a specific partial structure (the atoms and bonds represented by the node may constitute all or part of the specific partial structure)
- a single generic component represented by a generic partial structure
- a single specific component which spans more than one specific partial structures
- a single generic component represented by more than one neighboring generic partial structures
- more than one component of mixed types which are represented in the ECTR by at least two partial structures (the generic parts are represented by generic partial structures and the specific parts are contained within specific partial structures).

Figure 13 illustrates the outcome of example refined searches on previoulsy matched reduced graph nodes. The query structure Q, containing two reduced graph nodes, matches the three file structures F1, F2, and F3. The operation on the acyclic carbon nodes is described since the acyclic heteroatom nodes are identical in all instances. These are made up of various components with the exception of F3 which is wholly specific. The relevant node parameters are given for the acyclic carbon nodes; the query node parameters overlap the file node parameters in all three cases.

F1 is a structure match since the alkenyl of the query matches the connected alkyl and ethenylene of the file structure. All other atoms in the node are specific and match exactly.

F2 is a structure mismatch. The alkenyl group of the query matches the left hand ethenyl of the file structure, but the ternary branch connecting the two alkyl groups of the file structure does not relate to the possible ternary branch point of the query node which emanates from the alkenyl group.

A structure match would be seen with F3 if the atoms of the node alone were examined. The position of attachment of the oxygen in the carbon chain, however, indicates a mismatch showing that the refined search must extend beyond the boundaries of the node and examine the immediate environment also.

7. CONCLUSION

The screenout results are encouraging, with overall screenout values above 99.9% for the combined screening search. The two methods used for screening, bitscreens and reduced graphs, seem to complement each other; the combination of the two strategies improves their individual performances. This is because the two representations are based on different aspects of the structure. Fragments and ring descriptors are a detailed view of the local topology but do not represent the global topology, whereas reduced graphs show more detail of the global topology but give a generalized representation of the local topology.

Multipliers have yet to be dealt with and will be the subject of future work. Their use reveals two distinct problems which must be cured at a practical level. These are the multiplier indicating a repeating group, such as $-(CH_2)_n-$, and the multiplier which indicates a choice of substitution frequency on, for example, a ring. The lack of treatment of multipliers in the database examples discussed means that the values given for recall are not representative of the true values. Indeed, in many cases the query examples in Db3 and DbS cannot be regarded as structure examples of Db1; those queries which can be regarded as such do, however, show complete recall.

Substructure bitscreening searches are not encouraging, but the results should be greatly improved by a reduced graph search stage.

The search speeds for reduced graph screening are slow, mainly due to the size of the reduced graphs themselves. This can be improved by a faster matching algorithm than that which was used, a rather primitive back-tracking algorithm, and would also be improved by adding a bitscreening mechanism. The kind of mechanism envisaged would use fragment descriptors representing sequences of reduced graph nodes, or augmented reduced graph nodes perhaps. The searches have been carried out for both screening strategies on the whole database. This would, of course, not be the case in practice, as the fast bitscreening stage would always precede the reduced graph search.

ACKNOWLEDGMENT

We gratefully acknowledge funding from International Documentation in Chemistry mbH, Derwent Publications Ltd., Questel SA, and the Department for Education in support of the research described here. We also thank John Barnard of Barnard Chemical Information Ltd. and Peter Willett for their advice, and Chemical Abstracts Service for provision of documentation on the screen sets. We also acknowledge the help of Beverley Jackson and Eluned Hall who were responsible for encoding Db7 and Db8 as part of their M.Sc. Dissertations at the University of Sheffield, and we thank the chemical information departments who provided us with test queries.

REFERENCES AND NOTES

(1) Gillet, V. J.; Downs, G. M.; Ling, A. I.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126–137.

(2) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 14. Fragment Generation from Generic Structures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 453–462.

(3) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 10. The Generation and Logical Bubble-Up of Ring Screens for Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 215–224.

(4) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 12. Principles of Search Operations Involving Parameter Lists: Matching-Relations, User-Defined Match Levels, and Transition from the Reduced Graph Search to the Refined Search. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 253–260.

(5) Warr, W. A. *Chemical Structures. The International Language of Chemistry*; Springer-Verlag: Berlin, 1988.

(6) Warr, W. A. *Chemical Structures. 2. The International Language of Chemistry*; Springer-Verlag: Berlin, 1993.

(7) Dethlefsen, W.; Lynch, M. F.; Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 233–253.

(8) Kaback, S. M. What's in a Patent? Information! But Can I Find It? *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 159–163.

(9) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 2. GENSAL. A Formal Language for the Description of Generic Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 151–161.

(10) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 6. An Interpreter Program for the Generic Structure Language GENSAL. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 66–70.

(11) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 4. An Extended Connection Table Representation for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160–164.

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

(13) *CAS Online Screen Dictionary for Substructure Search*, 2nd ed.; Chemical Abstracts Service: Columbus, OH, 1981.

(14) 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.

(15) Gannon, M. T.; Willett, P. Sampling Considerations in the Selection of Fragment Screens for Chemical Structure Search Systems. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 251–253.

(16) Dubois, J. E.; Panaye, A.; Attias, R. DARC System: Notions of Defined Generic Substructures. Filitation and Coding of FREL Substructure (SS) Classes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 74–82.

(17) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 260–270.

(18) Fowler, E. An Investigation into the Information Required for Improved Performance of Reduced Graphs as Retrieval Keys in Generic Chemical Structure Files. M.Sc. Dissertation, University of Sheffield, 1988.

(19) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. The Sheffield University Generic Chemical Structures Project-A Review of Progress and of Outstanding Problems. In *Chemical Structures. The International Language of Chemistry*; Warr, W. A., Ed.; Springer-Verlag: Berlin, 1988; pp 151–167.

(20) Holliday, J. D. Computer Storage and Retrieval of Generic Chemical Structures in Patents. Fragment Generation and Screening of Generic Chemical Structures. Ph.D. Thesis, University of Sheffield, 1991.

(21) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 15. Generation of Topological Fragment Descriptors from Nontopological Representations of Generic Structure Components. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 369–377.

(22) Downs, G. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. Ring Perception and Screening to Extend the Search Capabilities. Ph.D. Thesis, University of Sheffield, 1988.

(23) Holmes, M. R. GENERATE VRGS: A Computer Program for the Generation of Vertex Reduced Graph Representations of Generic Ring Structures. M.Sc. Dissertation, University of Sheffield, 1988.

(24) Gillet, V. J.; Downs, G. M.; Holliday, J. D.; Lynch, M. F.; Dethlefsen, W. Searching A Full Generics Database. In *Chemical Structures 2. The International Language of Chemistry*; Warr, W. A., Ed., Springer-Verlag: Berlin, 1993; pp 87–103.

(25) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 57–66.