Computer Storage and Retrieval of Generic Chemical Structures in Patents. 17. Evaluation of the Refined Search

J. D. Holliday and M. F. Lynch*

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

Received October 31, 1994[⊗]

Results are presented for the atom-level search, called the refined search, for matching components of generic chemical structures. The refined search is the last and most discriminating search strategy used by the Sheffield generic structures system and is performed after the faster screening stages, bit screening, and reduced graph screening. It operates on the real atom representations of components of the generic structure which are defined by reduced graph nodes, nodes which represent aggregates of atoms of the original chemical graph which are structurally similar. The nature of generic structures means that parts of a structure may be expressed in terms of real atoms and bonds and parts in terms of homologous series identifiers such as *alkyl*. The fundamental problem concerning the refined search is determining equivalences between these two types of representation. Search results are presented for four query database searches against one file database. The methodology and results of the screening searches are also briefly described.

1. INTRODUCTION

The preceding paper in this series[1] describes the operation of the refined search. It uses a depth-first backtracking process adapted from the subgraph isomorphism algorithm developed by Ullmann.[2] The refined search is the most discriminating search strategy and is carried out after the faster screening stages.[3-6] This paper presents the results of several searches using the refined search and discusses its suitability in dealing with the types of structure variation found in generic chemical structures. Four types of structure variation have been identified by Dethlefsen;[7] these are as follows:

Substituent variation. A list of alternative substituents on a ring or chain, defined by a substituent identifier, e.g., "R1 = methyl, ethyl, or propyl".

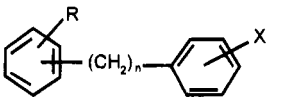
Position variation. A choice of atoms through which parts of the structure are connected, e.g., "monochlorophenyl".

Frequency variation. A variation in the frequency of occurrence of part of the structure, in a chain for example, e.g., "$-(CH_2)_n-$, n is 0, 1, or 2".

Homology variation. The use of standard nomenclature, a *homologous series identifier*, to denote a series of compounds with common structural features, e.g., "alkyl".

An example of a generic structure is shown in Figure 1 illustrating its obvious partitioning into "partial structures", each of which is related to a single expression in the patent document such as a structure diagram, a line formula, or a radical term. What is also apparent from the structure is that it reflects a tree-based heirarchy, the nonvariant partial structure being expressed at the root, further branches emanating from subsequent assignments.

The size and complexity of generic structures means that an atom level search on a file of full structure representations would be too time consuming to consider as an initial search strategy. Instead, screening searches are carried out on less detailed representations of the structure; these are faster but less discriminating than an atom level strategy. The two



n is 1, 2 or 3;
R is H, F or Cl;
X is F, Cl, Br or Alkyl.

Figure 1. Example of a generic chemical structure.

representations used for screening purposes, the bit screens[3,4] and the reduced graph,[5,6] are described in section 2.

The final search strategy, the atom level search (or "refined search"), compares regions of the full structure which have been mapped to each other during the reduced graph search strategy, using the atom and bond definitions of the partial structures. Some partial structures are not defined in terms of atoms and bonds, however, but emanate from homologous series identifiers. These generic partial structures represent a possibly infinite series of compounds with common structural features expressed in terms of numerical values or ranges of values, e.g., "alkyl 1−4C". The refined search has been designed to overcome this problem and its operation is summarized in section 3.

The strategies used at each stage of search are described in section 4, and results of several searches are shown in section 5.

2. SCREENING STRATEGIES

There are two stages of screening in our work: fragment screening,[3,4] implemented as a bitstring, the elements of which represent the presence or absence of a structural feature, and reduced graph screening,[5,6] 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 3000 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) at various levels of description.

[⊗] Abstract published in *Advance ACS Abstracts,* May 15, 1995.

0095-2338/95/1635-0659$09.00/0

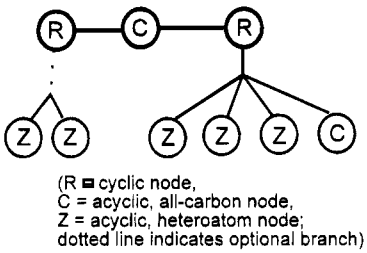


Figure 2. Reduced graph of the structure shown in Figure 1.

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.

In addition to fragment screens, ring screens provide information about the cyclic parts of the structure.[8] Ring screens categorize each ring of the structure according to size, composition, and degree of fusion. Again, each category 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.

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 they indicate the logical 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 full generic structure.

A reduced chemical graph represents a simple view of the structure whilst retaining the gross topological and logical relationships between components. Nodes of the graph represent regions of the structure which are chemically or structurally similar, depending on the criteria for reduction of the original chemical graph to the reduced chemical graph. Several types of reduction have been investigated and reported by us,[5,6,9,10] 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 original structure is retained in the reduced graph, a tree-structured graph which contains AND and OR branches; optional nodes arise due to the occurrence of hydrogen in a list of alternative substituents. The reduced graph of Figure 2 illustrates 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 numerical values or ranges of values which describe certain structural features, e.g., "number of rings".

3. THE REFINED SEARCH

The result of a successful match of two reduced graph representations, one representing the query structure and one representing the file structure, is a list of pairs of nodes, one query and one file node each, which are mapped onto one another. The refined search uses this list by comparing the two nodes of each item in the list in greater detail, i.e., at the level of atoms and bonds. The presence of homology variation within reduced graph nodes means, however, that the atom-level matching process must be able to translate between the real atom-bond representations of the specifically-expressed parts of the generic structure and those expressed as homologous series identifiers. The latter are represented by a series of 13 parameters each of which defines a structural feature, such as "number of carbon atoms" or "number of rings", and associated numerical values or ranges of values which describe that structural feature.[11]

The backtracking algorithm developed by Ullmann[2] for subgraph isomorphism has been adapted to deal with this translation problem by relaxing the 1:1 mapping between atoms of the two structures to allow 1:*N* relationships. This means that an atom can be used to represent a generic group in order that the group may map against more than one real atom. Indeed, *N*:*N* mappings may take place due to the existence of generic groups as part of the query node and as part of the file node. The problems posed by position variation have also been solved, and solutions are included in the algorithm. Frequency variation continues to be a problem for all our search strategies and will be dealt with in the future.

4. VALIDATION AND TESTING

An earlier report[12] documented the results of screening searches using five databases of generic structures of various levels of complexity. The same five databases have been used in this study; these are described here together with the search methodology.

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

Db1. A database of 2025 generic structures, i.e., structures which exhibit position-, substituent-, frequency-, and homology variation, created from the patent abstracts in sections B and C of the Basic Abstracts Journal published by Derwent Publications Ltd. during weeks 8340 to 8419 inclusive. Several of these structures failed at various stages of processing due to their size and complexity. The total number of structures which was 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 20th Db1 structure and altering the definition 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 from Db1. Of the 77 structures in Db3, two structures could not be processed during reduced graph generation due to

program failure. Results are reported for 75 structures which were successfully processed.

DbS. A database of 1205 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.

Db7. A database of 43 generic structures, exhibiting all four types of structure variation, taken from a random sample of non-English language patents from the Central Patent Index of 1984 published by Derwent Publications Ltd. Of the 43 structures in Db7, one structure could not be processed during reduced graph generation due to encoding errors at the earlier stage of database creation. Results are reported for the 42 structures successfully processed.

Db8. A database of 28 generic queries, exhibiting all four types of structure variation, was obtained from several industrial sources. Of the 28 structures in Db8, one structure could not be processed during reduced graph generation due to encoding errors at the earlier stage of database creation. Results are reported for the 27 structures successfully processed.

Search Methodology. The result of a successful match between two reduced graph representations is a list of pairs of reduced graph nodes. This list represents a "matching path" through the two reduced graphs, i.e., each item in the list indicates those query and file structure nodes which have been mapped onto one another. This list is used as the basis for the refined search. For each query structure-file structure match, the list of matching nodes is examined and node pairs are sent to the refined search stage. Each reduced graph node is mapped onto the region of the generic structure from which it was generated; this is then examined in the refined search stage. If every node pairing in the list results in a match at the refined search level then a structure match results, otherwise a structure mismatch is encountered.

More than one matching path may result from a reduced graph search, since there may be more than one instance of overlap between the two graphs. Only one structure match need be encountered from all of these possible matching paths for a hit to result.

Full structure searches were carried out using the five databases described above. Four database combinations were searched first by bit screening and reduced graph screening and then by the refined search. The combinations were as follows: (1) Db3 was used as a query database to search Db1, i.e., 75 structurally explicit generic queries against 1957 generic file structures; (2) DbS was used as a query database to search Db1, i.e., 1205 specific structure queries against 1957 generic file structures; (3) Db7 was used as a query database to search Db1, i.e., 42 randomly−selected generic queries against 1957 generic file structures; (4) Db8 was used as a query database to search Db1, i.e., 27 generic industrial queries against 1957 generic file structures.

Table 1 gives an overview of the results.

5. SEARCH RESULTS

Searches of Db3 against Db1. An average of 98.5% of the database was screened out in the bitscreening stage of search when using Db3 as a query database. Under normal search procedure the resulting hits would be sent to the reduced graph search stage. In order to achieve a more comprehensive evaluation, however, all Db3 structures were sent to the reduced graph search stage. Two Db3 structures failed to produce reduced graphs, as explained; the remaining 75 were used and produced a mean reduced graph screenout value of 99.95% of the database (an average of 0.95 hits per query). This figure is further reduced to 0.93 hits per query when both forms of screening, bitscreening, and reduced graph screening are used. This reflects the advantage of using two complementary screening stages. The resulting reduced graph hits were then sent to the refined search stage. Several structures failed to retrieve their parent Db1 structure due to encoding errors at the earlier stage of database generation and due to problems in dealing with frequency variation. Frequency variation, implemented as a multiplier label associated with the respective atom or substituent, is not yet handled by any search procedures and is the main cause of recall failures.

Table 1. Summary of Search Results

query database	Db3	DbS	Db7	Db8
bit screening screenout (%)	98.5	98.83	97.6	97.76
reduced graph matches	71	532	8	2
reduced graph parents retrieved	51	368	N/A	N/A
refined search matches	59	472	3	0
refined search parents retrieved	46	320	N/A	N/A

Using the 75 reduced graph queries, a total of 71 matches was found between the query structures and file structures at the reduced graph search level. Of these, 51 were matches between a query structure and its parent file structure. These figures reduce to 59 and 46, respectively, after the refined search. The 12 pairs of structures which did not match at the refined search level were examined, giving various reasons for elimination. All of these were correct with the exception of one pair in which the file structure has an ambiguous definition and caused an incorrect adjacency matrix to be produced by the refined search program.

The common representation of halogens in the reduced graph led to several structure mismatches since the explicit representation of the atom type is examined in the refined search. Further mismatches were due to the more detailed bond definitions in the structure representation, particularly in the case of aromatic rings. The five eliminated pairs of structures which should have retrieved their parent failed to do so due to incorrect coding of the query or file structure. These cases have also been identified in the fragment screening search stage and occur when a structure has been incorrectly encoded at the database generation stage. One such error is the incorrect assignment of bonds, in which double bonds have been missed or wrongly placed. In other cases, unsuitable specific definitions have been chosen as examples of generic radical expressions in the parent structures of Db1.

The maximum number of file structures retrieved by a single query structure was four; the average result was 0.77 hits per query.

Searches of DbS against Db1. Bitscreening produced a mean screenout value of 98.83% when using DbS as a query database. After reduced graph screening an average of 0.42 hits per query resulted. The low recall value is not only due in part to incorrect encoding during DbS generation but also due to problems in dealing with frequency variation. These problems have already been alluded to.[12]

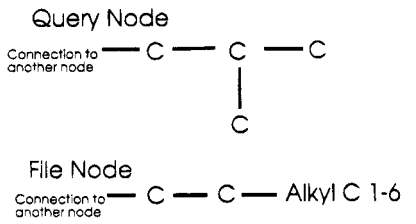


Figure 3. Example mismatch.

A total of 532 matches was determined between query structures and file structures at the reduced graph search level. Of these, 368 were matches between a query structure and its parent file structure. These figures reduced to 472 and 320, respectively, after the refined search. There were two errors in the refined search processing stage due to program failures.

The large number of structures tested makes examination of each elimination difficult. Several were examined and were found to be correct eliminations for reasons similar to those given above. The maximum number of file structures retrieved by a single query structure was 12; the average was 0.39 hits per query.

Searches of Db7 against Db1. Screenout values for bitscreening and reduced graph screening are 97.6% and 99.99%, respectively, the latter value giving an average of 0.18 hits per query.

A total of eight matches was found between query structures and file structures at the reduced graph search level; no parent structures exist in Db1. Five of these were eliminated during the refined search. Four were correct eliminations; the fifth was due to a processing error in a complex structure. Figure 3 illustrates an elimination in which two reduced graph carbon nodes which matched at the reduced graph level are an obvious mismatch at the refined level. Only three matches result after the refined search, making an average of 0.07 hits per query.

Searches of Db8 against Db1. The mean screenout value for bitscreening was 97.76%, and reduced graph screening resulted in an average of 0.074 hits per query. Of the 27 query structures tested, only one retrieved any database structures. This query retrieved two database structures.

Both structure matches resulting from the reduced graph search were eliminated during the refined search. These eliminations are correct, one being due to a coding error at the stage of database creation.

6. CONCLUSIONS

The atom-level search strategy presented here represents the most discriminating level of search in the Sheffield generic structure search system. This search stage is carried out after the two screening stages, bitscreening and reduced graph screening, which are faster but less discriminating. The high performance of the reduced graph screening stage has been reported in previous papers; the refined search examines in greater detail the resultant node mappings from the reduced graph search. The algorithm chosen, an adapted version of the Ullmann algorithm for subgraph isomorphism, appears to be a method suitable for dealing with the special problems posed by generic chemical structures; notably the establishment of equivalences between parts of the graph defined in terms of real atoms and those described in terms of structural feature attributes.

Performance has been reported in terms of recall for those query structures derived from parent database structures and in terms of screenout. Recall problems have arisen due to the system's inability to deal with frequency variation in the present search strategies and due to many encoding errors during the generation of the test databases. It is expected that a solution to the frequency problem would produce a system of very high performance. The second problem, that of database encoding errors, is not considered to be a problem of implementation and serves to support the high level of performance since the identification of such errors is a part of the performance evaluation.

The objectives of the refined search have thus been attained. It is not expected that they can readily be improved upon by further means. A very small number of structures did not produce the correct adjacency matrix, and this is an area where the processing can be improved upon.

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.

REFERENCES AND NOTES

(1) Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 16. The Refined Search: an Algorithm for Matching Components of Generic Chemical Structures at the Atom−Bond Level. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1−7.

(2) Ullmann, J. R. An Algorithm for Subgraph Isomorphism. *J. Assoc. Comput. Mach.* **1976**, *23*, 31−42.

(3) 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.

(4) 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.

(5) 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.

(6) 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.

(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) 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.

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

(10) 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.

(11) Welford, S. M.; Lynch, M. F.; Barnard, J. M. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 3. Chemical Grammars and Their Role in the Manipulation of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 161−168.

(12) Holliday, J. D.; Downs, G. M.; Gillet, V. J.; Lynch, M. F.; Dethlefsen, W. An Evaluation of the Screening Stages of the Sheffield Research Project on Computer Storage and Retrieval of Generic Chemical Structures in Patents. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 39−46.

CI940120O