

- 1986, 26, 212. (e) Fujita, S. "3. Classification of One-String Reactions Having an Odd-Membered Cyclic Reaction Graph". *J. Chem. Inf. Comput. Sci.* **1986**, 26, 224. (f) Fujita, S. "4. Three-Nodal and Four-Nodal Subgraphs for a Systematic Characterization of Reactions". *J. Chem. Inf. Comput. Sci.* **1986**, 26, 231. (g) Fujita, S. "5. Recombination of Reaction Strings in a Synthesis Space and Its Application to the Description of Synthetic Pathways". *J. Chem. Inf. Comput. Sci.* **1986**, 26, 238. (h) Fujita, S. "6. Classification and Enumeration of Two-String Reactions with One Common Node". *J. Chem. Inf. Comput. Sci.*, first of five papers in this issue. (i) Fujita, S. "7. Classification and Enumeration of Two-String Reactions with Two or More Common Nodes". *J. Chem. Inf. Comput. Sci.*, second of five papers in this issue. (j) Fujita, S. "8. Synthesis Space Attached by a Charge Space and Three-Dimensional Imaginary Transition Structures with Charges". *J. Chem. Inf. Comput. Sci.*, third of five papers in this issue. (k) Fujita, S. "9. Single-Access Perception of Rearrangement Reactions". *J. Chem. Inf. Comput. Sci.*, fourth of five papers in this issue. See also: *Chem. Eng. News* **1986**, 64(39), 75.
- (19) The term "imaginary" transition structure stems from the analogy of imaginary numbers which are counterparts of real numbers. The ITS may be a real transition structure when we consider a concerted reaction, and so, the term "complex" transition structures would be more suitable if we consider that complex numbers contain real and imaginary numbers. However, the term "complex" has been used widely in the chemical field. Therefore, we have adopted "imaginary" transition structures in our case.
- (20) After submission and acceptance of this paper, the author heard Vladutz's report: Vladutz, G. In "Modern Approaches to Chemical Reaction Searching". Willet, P., Ed., Gower: Aldershot, U.K., 1986; p 202. The author thanks Dr. Vladutz for sending his article. See also: *Chem. Eng. News* **1987**, 65(6), 2. The main differences between Vladutz's *superposed reaction graph* (SRG) and our ITS are as follows. (a) The SRG contains no catalysts even if they participate the reactions. Thus, the SRG counterpart of the reaction of Figure 14 does not contain hydrochloric acid. (b) The ionic character of a bond is represented by the concept of "charge space" in our ITS approach.^{18j} On the other hand, an ionic SRG and the corresponding nonionic SRG are not integrated.
- (21) Information on stereochemistry can be treated by three-dimensional ITS's, which will be discussed elsewhere.
- (22) Balaban, A. T., Ed. *Chemical Application of Graph Theory*; Academic: London, 1976.
- (23) Wilcox, C. S.; Levinson, R. A. In *Artificial Intelligence. Applications in Chemistry*; Pierce, T. H., Hohne, B. A., Eds.; ACS Symposium Series 306; American Chemical Society: Washington, DC, 1986; pp 209-230.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 8. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval[†]

VALERIE J. GILLET, GEOFFREY M. DOWNS, AI LING, MICHAEL F. LYNCH,*
PALLAPA VENKATARAM, and JENNIFER V. WOOD

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

WINFRIED DETHLEFSEN

BASF, Ludwigshafen am Rhein, Federal Republic of Germany

Received February 18, 1987

Reduced chemical graphs for specific chemical substances comprise summary descriptions of the gross structural features of these substances; an example is summarization in terms of only the ring and nonring components, giving a tree structure in which each node is either a cyclic or an acyclic component. Other bases for graph reduction are possible, including those in which the components are formed from the separate aggregates of carbon atoms and of heteroatoms. The reduced graph of a generic chemical structure is usually a multigraph, in which the identities of the variables are summarized in similar terms. The varieties and distributions of several types of reduced graphs created from almost 50 000 specific chemical substances in the Fine Chemicals Directory are characterized. The data provide qualitative guidance on the power of reduced graphs as retrieval keys when a database of generic structures described in this way is searched for queries that are complete specific or generic structures. The performance of several types of reduced graph, taken both singly and in combination, as retrieval keys for searches of generic chemical structures is reported. The test database is a small set of generic structures in which the variables are specific partial structures; the queries comprise both specific structures from patents and the generic structures of the test database itself. The results confirm the potential of reduced chemical graphs for high performance.

INTRODUCTION

Substantial progress has already been reported by us in the development of methods of representing generic chemical structures in machine-readable form for retrieval purposes.¹⁻¹⁰ It was evident from the beginning of our work that a powerful tool kit combining a variety of search representations, search algorithms, and high-performance hardware would be necessary in order to provide workable and economic solutions to searching in the most general sense. Our approaches to

search representations have, thus far, been "bottom-up" in design; i.e., we have sought to describe the generic structures in terms of atoms and bonds and their aggregates, in much the same way as the screens employed for substructure searching of databases of specific substances are derived and used. These and similar search screens have played and will continue to play an important role in our work since they provide the most generally applicable description of generic structures for the range of types of searches required.

Two important classes of searching in files of generic chemical structures are those that involve complete structures; i.e., the queries are either specific structures or generic structures. When the query is a specific structure, the purpose is to discover which file structures include the query. When

* Author to whom correspondence should be addressed.

[†] Paper presented at the Division of Chemical Information Symposium on Generic Chemical Structure Searching, 192nd National Meeting of the American Chemical Society, Anaheim, CA, Sept 9, 1986.

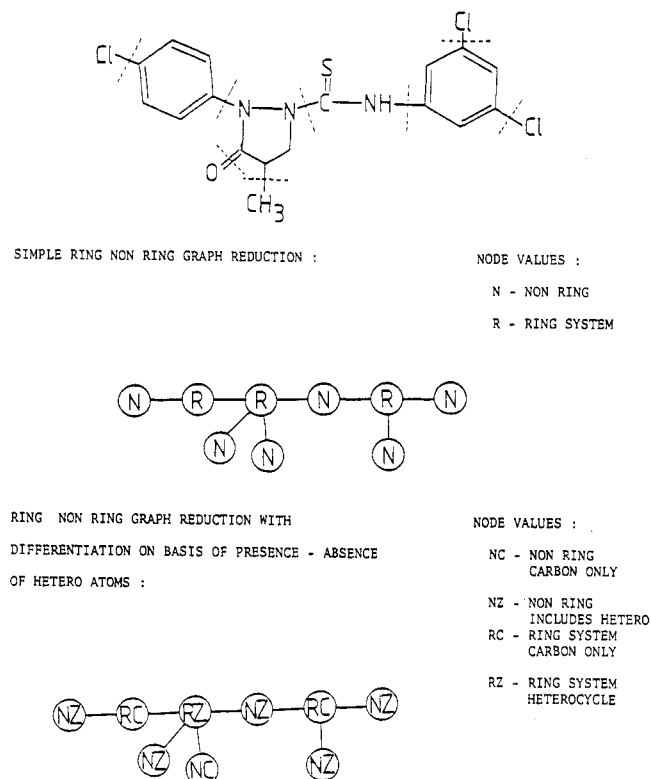


Figure 1. Graph reduction by categorizing components as ring or nonring.

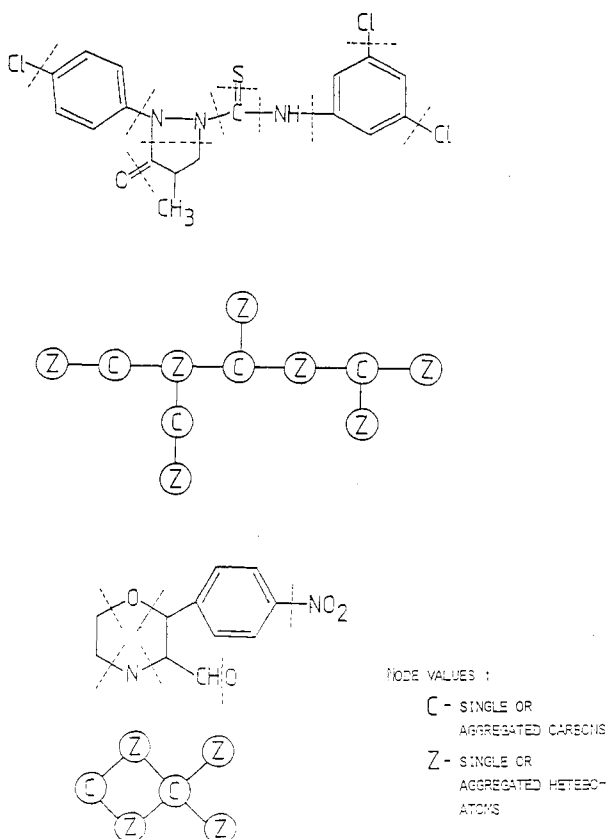


Figure 2. Graph reduction on the basis of aggregation of carbon atoms and heteroatoms.

the query is a generic structure, the purpose is to discover whether or not there is common membership, i.e., whether one or more file structures are identical with the query, include it or are included by it, or intersect with it. In currently

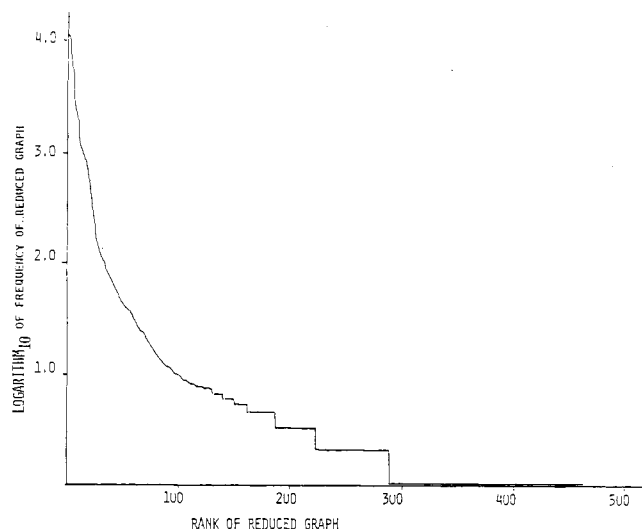


Figure 3. Frequency distribution of simple ring/nonring reduced graphs.

operating generic structure retrieval systems, these queries are not always readily handled.

It appeared important, therefore, to investigate search representations that would be well suited for these cases. This appeared to call for a "top-down" representation that would provide a hierarchy of simple search filters; at successive stages of application the mesh could be made finer and finer by bringing increasing levels of detail into play. Further, the notion is in keeping with our objective of seeking orthogonal searchable representations, i.e., those which depend on distinct and different characteristics of generic structures, so that the dependencies between them can be minimized and their power, individually and collectively, maximized. Accordingly, the simple notion of the reduced chemical graph was studied.

Graph reduction involves the generalization of certain features of chemical structures, whether specific or generic. Its purpose is to bring about a homeomorphic mapping from the structure onto some simpler graph, resulting, in general, in a smaller number of nodes than in the original graph. The simpler graphs can then be searched more rapidly against queries that have been subjected to the same process. The structures retrieved must then be subjected to a more detailed examination to determine whether an exact match exists in fact.

Two methods of graph reduction, with variations on them, are reported here. Many other variations on the themes are possible, some of them currently under investigation here. Those reported here involve (a) reduction on the basis of categorization of components of the chemical graph as ring or nonring and (b) reduction on the basis of aggregation of contiguous assemblages of one or more carbon atoms or one or more non-carbon atoms (other than hydrogen) into distinct and different node types.

Typical specific structures and their corresponding reduced chemical graphs, reduced, that is, on each of these bases, are shown in Figures 1 and 2. The reductions assume that (a) hydrogen atoms are suppressed, (b) ring systems comprise solely ring atoms and ring-internal bonds in the case of reduction to ring/nonring reduced graphs (diphenyl, for instance, is taken to be two connected cyclic nodes) and (c) heteroatom aggregates multiply connected to carbon-atom aggregates can form rings of size four or greater only.

The notion of a reduced graph is by no means novel, but little use appears to have been made of it as yet in the context of computer-based retrieval systems. Cyclic/acyclic reduction is implicit, for instance, in most systematic nomenclature, leaving aside radical names such as "cinnamyl" and "tolyl",


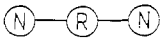

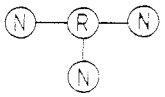
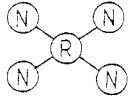
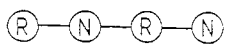
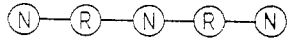

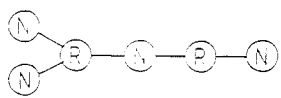
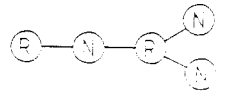
RANK	SPECIES	FREQUENCY %
1		22.3
2		15.6
3		11.2
4		11.2
5		4.9
6		3.9
7		3.6
8		3.2
9		2.2
10		1.9

Figure 4. Ranking of simple ring/nonring reduced chemical graphs.

which imply combinations of cyclic and acyclic components, as well as in chemical line notations. Again, it is implicit in the treatment of structure records in the Chemical Abstracts Service Chemical Registry System, in which ring systems and acyclic radicals are separated and treated differently.¹¹

At a different level of detail, Lederberg first made explicit one form of cyclic reduced graph, the vertex graph, used in the exhaustive generation of the cyclic isomers of a given molecular formula in the DENDRAL project.¹²⁻¹⁴ Here, the vertex graph is the cyclic graph from which all nodes of degree (or connectivity) of less than three have been eliminated. More recently, Balaban et al. have mentioned a similar form of graph reduction as the basis for an algorithm to find all possible rings in chemical graphs;¹⁵ in this, all ring nodes of degree less than three are again ignored, so as to give the homeomorphically reduced graph (although it must be noted that, as stated, the algorithm will fail with unsubstituted monocycles). As will be evident from the following, the notion of the reduced graph, based on different but equally simple principles, has much potential for searches of generic structures involving the query types mentioned above and may have useful applications in other respects too.

In order to gain estimates of the variety and distributional characteristics of these reduced graphs and assess their applicability for searches of these kinds, analyses of the Fine Chemicals Directory (FCD), a database of common specific chemical substances, were first undertaken. Then files of generic structures that contain only specific partial structures as variables were converted into reduced multigraphs at several levels of detail in order to determine their performance in retrieval when queries of the types mentioned above were used.

The complexity of the graphs that result from reduction of specific substances varies considerably, but, in general, the process results in a substantial and often large reduction in the number of nodes to be dealt with, a welcome characteristic if good keys for retrieval are to result. The reduction may produce a single node, depending on the characteristics of the

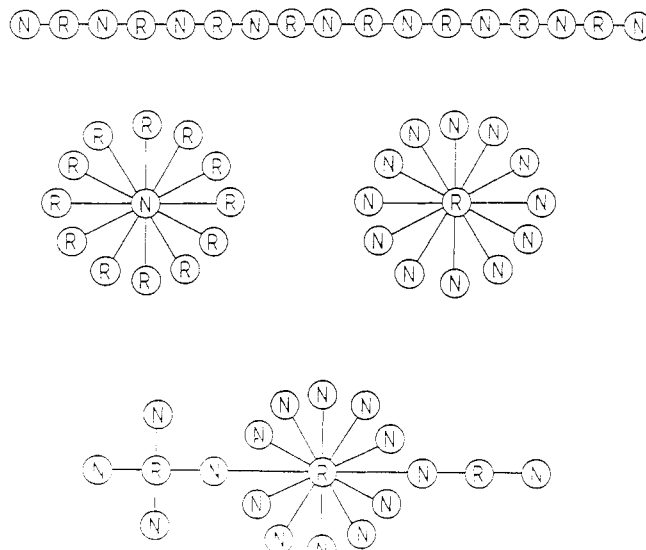


Figure 5. Extreme examples of simple ring/nonring reduced graphs.

substance and of the reduction method being applied. Thus in the case of ring/nonring reduction, all purely acyclic structures collapse to a single nonring node, and isolated ring systems reduce similarly to a single ring node. In the case of carbon/heteroatom reduction, all hydrocarbons reduce similarly to a single node. The proportion of such extreme cases, taken together, is small, and, overall, a substantial variety of reduced graphs can be expected.

These methods have now been investigated by applying simple algorithms to the 48 270 specific structures of the Fine Chemicals Directory that are successfully decoded by the DARING software and which consist of nondisjoint graphs. A reduced graph representation of each specific structure is produced and is then canonicalized by applying the Morgan algorithm.¹⁶ The resultant representations are then sorted, first to count the number of "tokens" for each "type" and then to order these in decreasing rank/frequency order. In each instance, for the 48 270 structures from the FCD, the total variety of types and the inverse frequency rankings are obtained.

RING/NONRING GRAPH REDUCTION

The simplest form of reduction involves the representation of structures as graphs that indicate solely the relationships of cyclic and acyclic components. The symbols for the nodes are "R" (for ring component) and "N" (for nonring component). The degree of reduction is substantial; each node in the reduced graph corresponds to an average of 4.75 non-hydrogen atoms in the FCD database. The total variety of distinct reduced graphs produced in this way is 464. The distribution of these, as expected, is highly skewed; the most frequent of them, N, denoting wholly acyclic structures, accounts for over 22% of the file (although there is evidence from a file of specific structures from patents that the proportion of such completely acyclic structures is much lower here). As Figure 3 shows, this maximum frequency falls off very rapidly, with 38% of the reduced graphs occurring once only. Figure 4 shows the 10 top-ranking species, with their relative frequencies, while Figure 5 illustrates some of them, the extrema, so to speak. All of these reduced graphs are trees by definition.

A much wider variety of reduced graphs is obtained by stepping up the distinctions slightly so that the individual components are now distinguished according to whether they comprise solely carbon atoms or one or more heteroatoms; the node values are now designated "RC", "RZ", "NC", and

RANK	SPECIES	FREQUENCY %
1		21.1
2		9.0
3		6.8
4		4.9
5		3.5
6		2.4
7		2.4
8		2.1
9		2.1
10		2.1

Figure 6. Ranking of heteroatom differentiated ring/nonring reduced graphs.

RANK	SPECIES	FREQUENCY %
1		8.3
2		8.3
3		6.5
4		5.6
5		5.0
6		3.6
7		3.0
8		2.8
9		2.6
10		2.4
...		
14		0.95

Figure 7. Ranking of carbon/heteroatom aggregated reduced graphs.

"NZ". The variety of these reduced graphs now increases to 1405. The mean reduction, the ratio of atoms to nodes, remains as before, 4.75 to 1. Figure 6 shows the top 10 species, the highest-ranking being NZ with a slightly reduced frequency of 21%; this slight reduction follows from the fact that acyclic hydrocarbons, denoted NC, are very infrequent in this file.

It is evident that continuing this process of including greater detail about each component within the node will have the

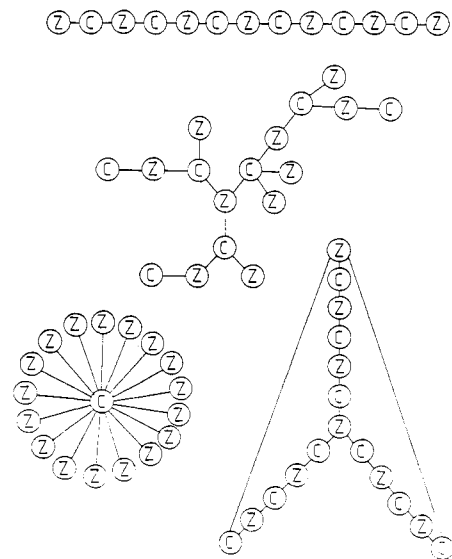


Figure 8. Extreme examples of reduced graphs resulting from carbon/heteroatom aggregation.

effect of reducing the maximum frequency of the most frequent types and will increase the variety of types. Ample precedents for such manipulations are to be found both in chemical and in other contexts, as Cooper and Lynch have shown.¹⁷ Thus a progression can be considered in which the addition of greater detail corresponds to making the meshes of a net finer and finer, as circumstances, which here include factors such as database size, query size, and structural complexity, dictate. In the limit, of course, we have the full representation of the structure itself. An illustration of the effect of this principle is given in the retrieval tests below and is particularly apparent when the nodes include the numbers of carbon and non-carbon atoms.

CARBON/HETEROATOM GRAPH REDUCTION

The variety of graph types resulting from aggregation of one or more connected carbon atoms to form nodes of one type, designated "C", and of one or more heteroatoms to form nodes of the other type, designated "Z", is found to total 1864, just over 4 times larger than that for ring/nonring graph reduction. The distribution is similarly strongly skewed, with 865 of the types appearing once only. Figure 7 details the 10 top-ranking species. These now include cyclic graphs of even ring sizes of four or greater, in which the node types alternate in the ring; Figure 8 shows some of the extreme examples. Ring formation contributes in part to the larger variety of types produced. A smaller reduction in the numbers of nodes as between the original structures and the reduced graphs, a ratio of 2.8 to 1, is found.

Further degrees of differentiation among node types, leading to larger varieties of species and, with that, an overall reduction of their mean frequency of occurrence, can obviously be applied. One such is the differentiation of the heteroatom type into oxygen and other, oxygen, nitrogen, and other, etc., following the rank order of heteroatoms.

REDUCED MULTIGRAPH REPRESENTATION

Generic structures of the type with which we are dealing, i.e., those with specific partial structures as variables, are readily represented by a simple extension of the methods developed for specific structures, as illustrated in Figure 9 for the case of the simple ring/nonring reduction. First, the components of the generic structure are identified and represented in the appropriate symbolism. Neighboring nonring

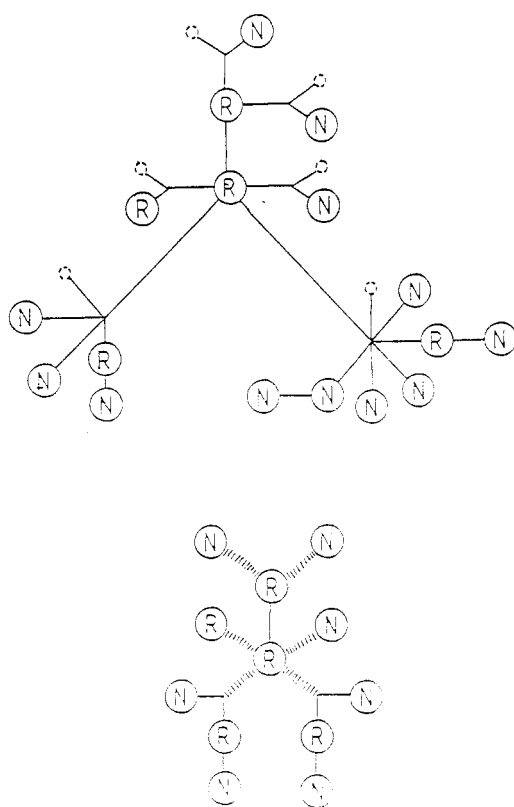
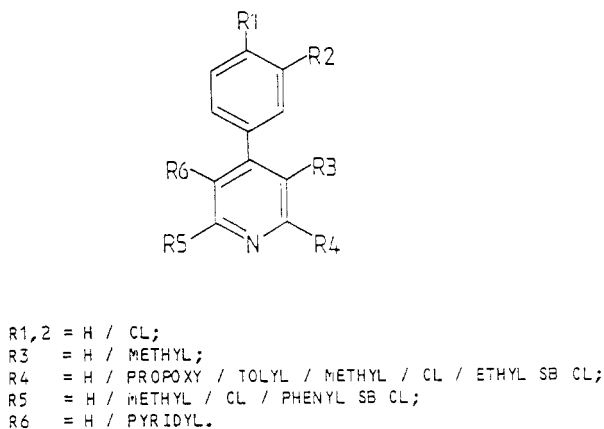


Figure 9. Formation of the ring/nonring reduced graph of a generic structure.

nodes are collapsed into a single node, carrying their further attachments, if any, with them. Next, components that are alternative to one another are conflated if they are of the same type. A multigraph, in the form of an AND/OR tree, will usually result. Since a frequent feature is optional substitution, i.e., hydrogen is listed as an alternative, the AND/OR tree is modified, and the dashed lines in the figure indicate this optional substitution. Nodes joined by solid lines in the center constitute the invariant parts of the generic structure.

The procedure is similar for the ring/nonring reduction with heteroatom differentiation; here, however, neighboring nonring nodes, when conflated, need examination to determine whether the conflated nodes should carry an indication of the presence, absence, or possible presence of heteroatoms, as noted below. A further stage of differentiation has also been studied. Here, the first stage of the reduction process associates a count of the number of carbon atoms and of the sum of the numbers of heteroatoms within the component. The next stage, con-

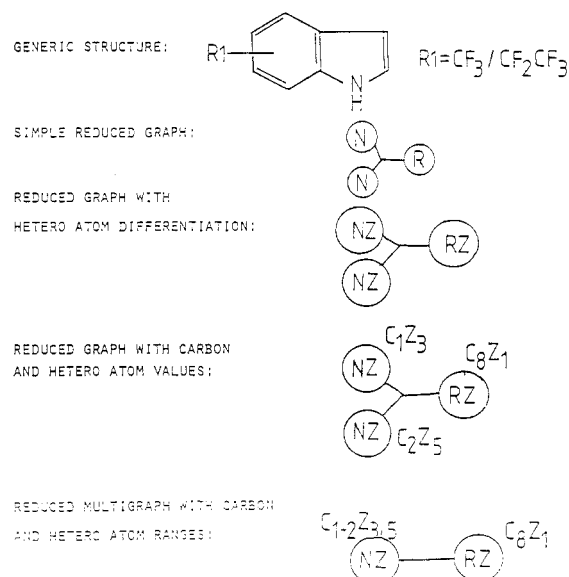
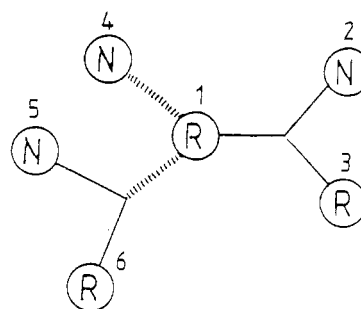


Figure 10. Ring/nonring graph reduction with carbon and heteroatom ranges.



NODE NO.	NODETYPE	CONGENERS
1	R	[2,3] [0,4] [0,5,6]
2	N	[1]
3	R	[1]
4	N	[1]
5	N	[1]
6	R	[1]

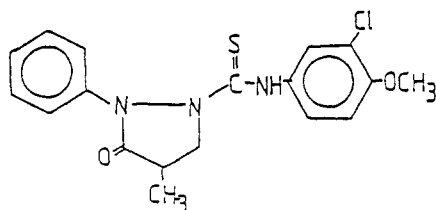
Figure 11. Representation of the reduced multigraph of a generic structure.

flation of neighboring nonring nodes, results in the addition of the number of carbon atoms, whereas the conflation of alternative nodes of the same types results in ranges of carbon atom and heteroatom counts, as illustrated in Figure 10. Subsequent node-matching processes must then use a more elaborate matching criterion.

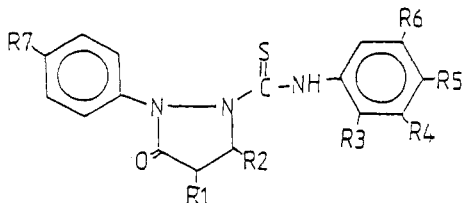
The information stored within nodes can be detailed further to improve on the effectiveness of the search, but this is at the expense of more complex representations and matching strategies. Consider the combining of information from alternative nodes of the same type: here the carbon and heteroatom counts are combined independently, e.g., combining the two nonring components methyl (C_1Z_0) and hydroxy (C_0Z_1) gives the atom ranges $C_{0-1}Z_{0-1}$; this component will therefore match either a methoxy or a hydroxymethyl substituent (C_1Z_1). An extension of the representation of atom ranges or the retention of alternative nodes as individual components would make the search at this level more effective.

von Scholley has already described the relaxation algorithm

SPECIFIC:



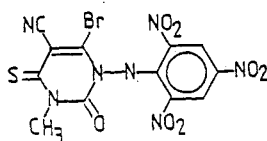
GENERIC:



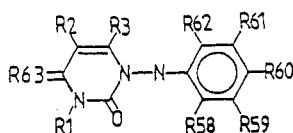
R1 = H / METHYL;
 R2 = H / METHYL;
 R3 = H / CL / BR / METHYL / I;
 R4 = H / CL / BR / F;
 R5 = CL / ETHOXY / H / BR / METHOXY;
 R6 = H / CL;
 R7 = H / METHYL / CL.

Figure 12. Example of a specific structure together with the generic structure from which it was derived.

SPECIFIC:



GENERIC:



R1 = H / METHYL OSB <3> CL / THIOL SB ETHYL;
 R2 = H / CYANO / CARBOXY OSB ETHYL / BR / CL;
 R3 = H / METHYL / ETHYL / AMINO / CL / F / CYANO / BR;
 R4 = H / METHYL SB (HYDROXY / <3> CL) / THIOL SB ETHYL;
 R58 = H / METHOXY / NITRO / METHYLTHIO / METHYL / F / CL / ETHYL / ACETYL / CF3;
 R59 = H / METHYL / F / CL / N-BUTYL / METHOXY;
 R60 = H / CL / CF3 / ETHOXY / METHOXY OSB <2> F / NITRO / I-PROPYL / BR / ACETYL / METHYL;
 R61 = H / CL / F;
 R62 = H / NITRO / CL / F / METHYL / METHYLTHIO;
 R63 = O / S.

Figure 13. Example of a specific structure together with the generic structure from which it was derived.

for searching generic chemical structures.⁸ This applies methods first developed in the context of image processing to substructure searching of specific and generic chemical structures and has been tested successfully on small numbers of both types.^{8,9} The representation developed for this purpose involves a modification of the standard connection table, in that substituents may be alternative to one another. Hence, in addition to those connections of an atom that are specific and certain and which are shown in the connection table in the usual manner, the groups of substituents that are alternative to one another are shown together within an extended cell in the connection table.

The representation of the reduced multigraph is similar to this, but where the entities in the original RELAX record are atoms, the entities are now nodes of the reduced multigraph, and attributes can be associated with each node. When the

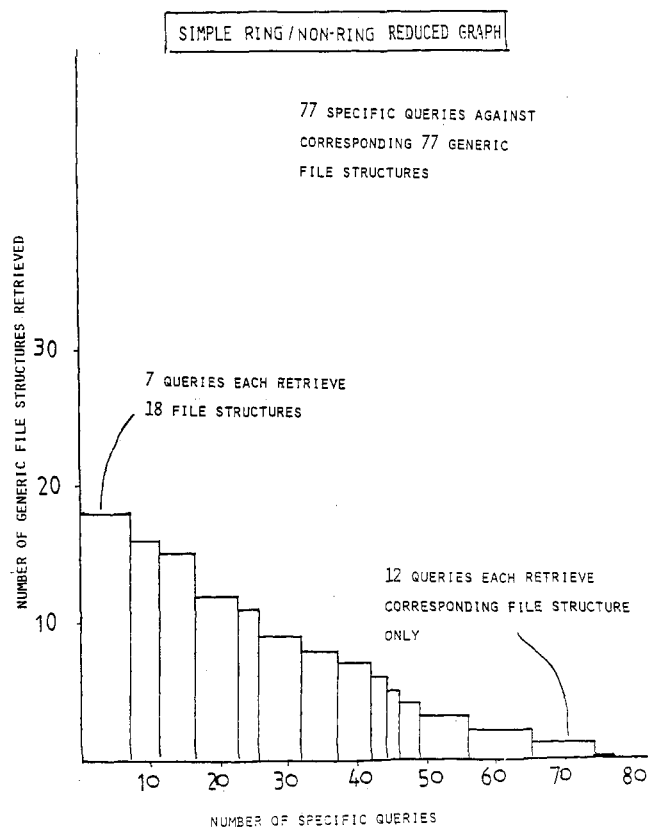


Figure 14. Distribution of results of searching 77 specific queries against 77 corresponding generic file structures for simple ring/nonring reduced graphs.

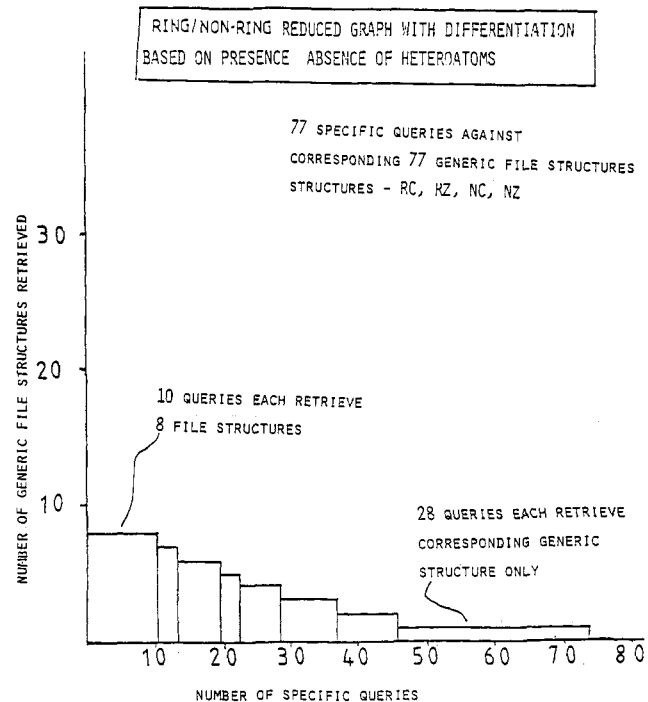


Figure 15. Distribution of results of searching 77 specific queries against 77 corresponding generic file structures for ring/nonring reduced graphs with heteroatom differentiation.

substitution is optional, this is shown by means of a zero entry in the cell for the congeners, as illustrated in Figure 11.

This adaptation of the RELAX record for representing reduced multigraphs has the great advantage that our RELAX algorithm is immediately applicable to searching these rep-

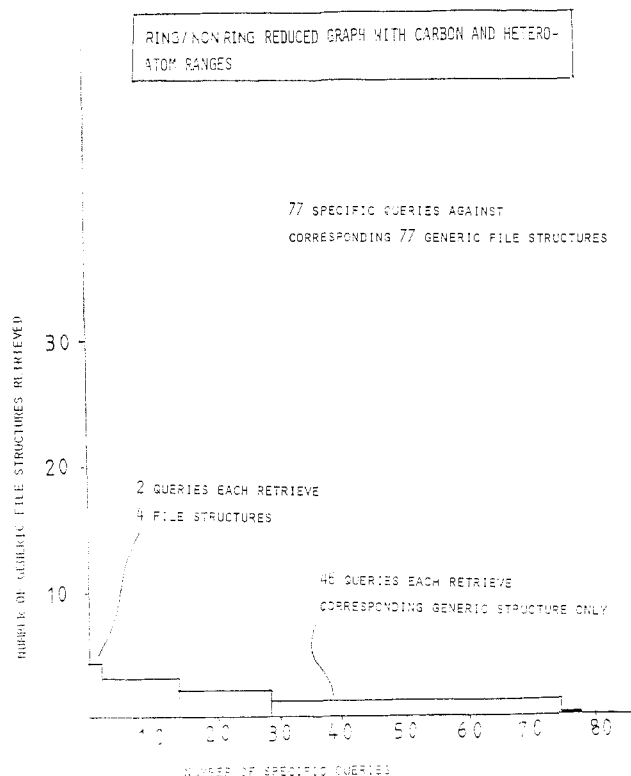


Figure 16. Distribution of results of searching 77 specific queries against 77 corresponding generic file structures for ring/nonring reduced graphs with carbon and heteroatom ranges.

QUERIES : 57 SPECIFIC STRUCTURES (not corresponding to the generic file structures)

FILE STRUCTURES : 77 GENERIC FILE STRUCTURES

RING/ NON RING REDUCTION

R-N ALONE	R-N HETERO DIFFERENTIATED	R-N WITH CARBON AND HETERO ATOM RANGES
2 X 17		
3 X 15		
4 X 12		
2 X 11		
3 X 10		
6 X 8	9 X 8	
4 X 7	1 X 7	
3 X 6	1 X 6	
1 X 5	1 X 5	
1 X 4	3 X 4	
3 X 3	3 X 3	1 X 3
11 X 2	7 X 2	2 X 2
6 X 1	7 X 1	1 X 1
10 X 0	25 X 0	55 X 0

TOTAL= 319

140

8

Figure 17. Results of searching 57 specific queries (which have no corresponding structures in the file) against 77 generic file structures at three levels of differentiation of ring/nonring reduced graphs.

representations. They fully represent the invariant structure and its substituents, including further nested levels of substituents; however, it does not yet deal with cases where a substituent on a parent may occur multiply, as designated by a range, e.g., (1-3). Again, a common form of expression in generic structures is the combined substituent, e.g., R_1 and R_2 combine to form a ring. This last case has been dealt with here by creating duplicate records with and without the added ring. None of these circumstances pose problems for a future, more comprehensive, implementation.

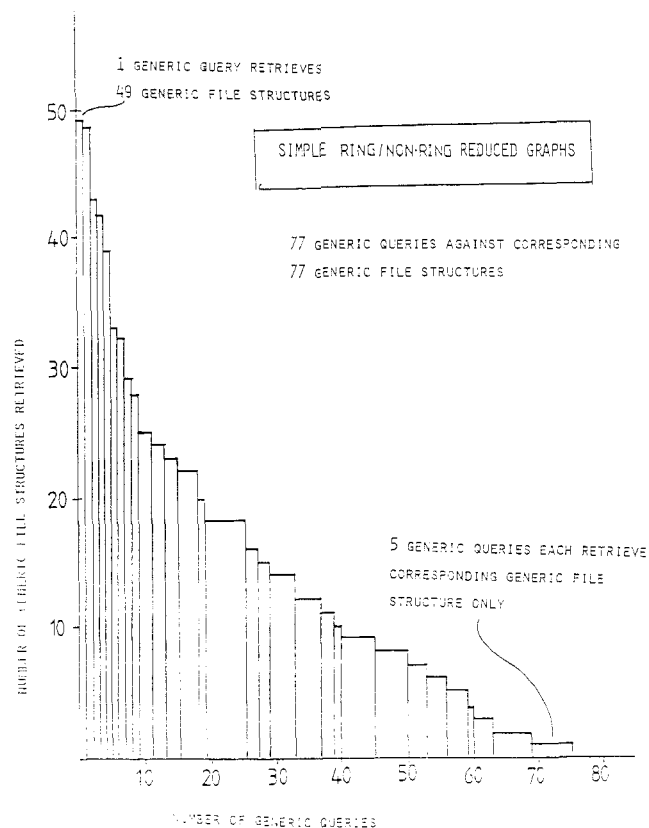


Figure 18. Distribution of results of searching 77 generic queries against 77 corresponding generic file structures for simple ring/nonring reduced graphs.

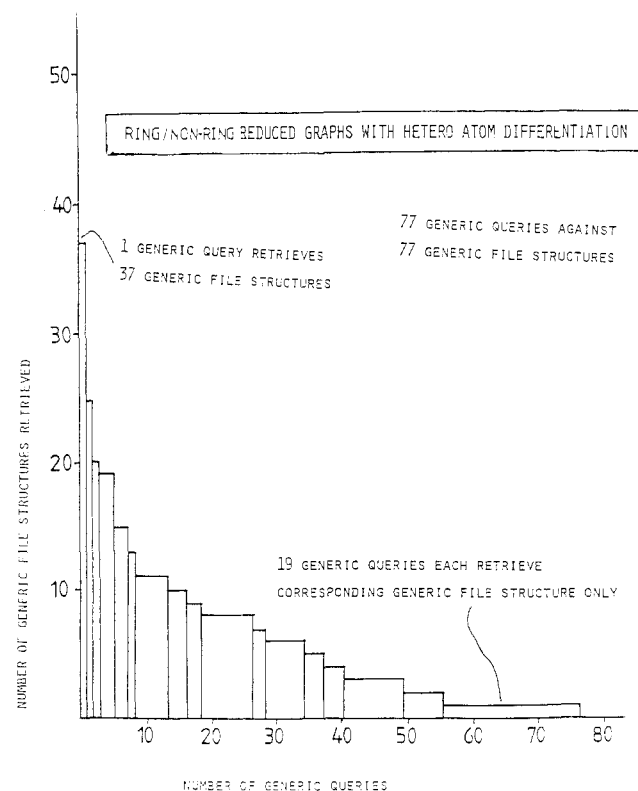


Figure 19. Distribution of results of searching 77 generic queries against 77 corresponding generic file structures for ring/nonring reduced graphs with heteroatom differentiation.

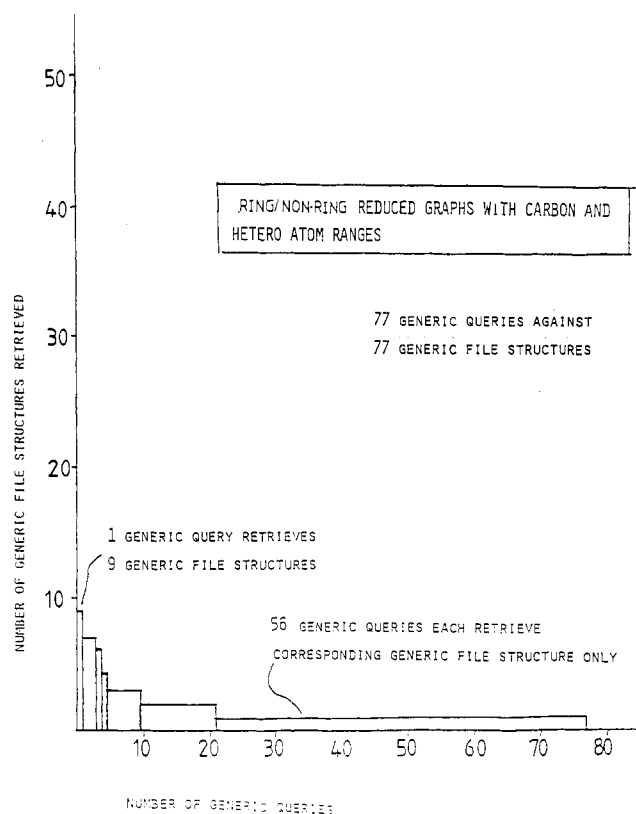


Figure 20. Distribution of results of searching 77 generic queries against 77 corresponding generic file structures for ring/nonring reduced graphs with carbon and heteroatom ranges.

RETRIEVAL TESTS AND EVALUATION

The main test database available to us at present comprises a file of over 2000 generic structures from patents, represented as GENSAL expressions. There is also a smaller database created as a systematic sample from this database, which comprises 77 generic structures representing the class of all specific substances instanced in the patent document. These generics involve no generic radical terms; all variables are specific partial structures. The sample is much smaller, because of the much greater effort involved in creating it. In addition, with each of the generic structures in the database of 2000, and hence with each of the 77 generics that include only specific variables, there is a single specific structure characteristic of that generic structure, recorded at the time of database creation in order to provide a set of simple queries for search evaluation purposes. Figures 12 and 13 provide examples of the generic and specific structures from the test database.

Ring/Nonring Reduction. The evaluation involving the ring/nonring reduction at three levels of detail is performed by transforming both query and file structures to their reduced graph forms, searching the queries against the file using the relaxation method, and determining the results at this level of search alone; i.e., no extension of the atom-by-atom method used with specific structures is applied as yet. The queries comprise both specific structures and generic structures. The first set of queries comprises the set of 77 specific structures that correspond one-to-one with the 77 generic structures; thus, each specific query should retrieve at least the corresponding generic. (It is known from other work in the project that, in searches conducted with fragment and ring type screens, each structure retrieves only its associated generic structure.) The second query set is the test database itself, with each generic structure run as a query against the database; once again, it

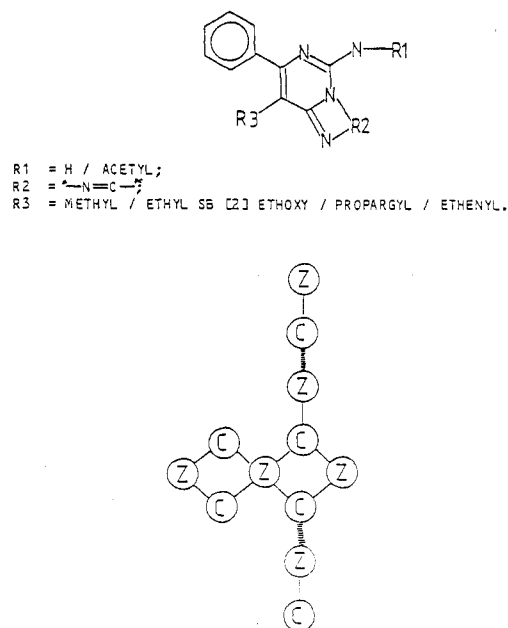


Figure 21. Example of a carbon/heteroatom aggregated reduced graph of a generic structure.

is known from other searches that each retrieves only itself. The third set of queries is a sample of 57 specific structures corresponding to other generic structures within the larger database, none of which corresponds to a generic file structure in this sample. Complete structure searching at the atom and bond level reveals that there are no hits in the database.

Noteworthy statistics relating to the test database of 77 structures are that the sample comprises a gross total of 4258 atoms, irrespective of which of these are invariant and which alternative to one another, i.e., an average of 55.3 non-hydrogen atoms per generic structure. The ring/nonring reduced graphs created from these entail a mean reduction of just over 7 to 1.

The basis on which the evaluation is performed is that the query, if a specific structure, should be included within the generic class defined by the file structure; if a generic structure, it should intersect with it, include it, be included within it, or be identical with it. The matching criterion, therefore, is that there is at least one specific reduced graph in common between query and file structure.

The three levels of description used for the ring/nonring reduced graph tests are as follows:

- The nodes of the reduced multigraph indicate solely whether the components are ring or nonring.
- The nodes of the reduced graph are differentiated according to whether the component comprises exclusively carbon atoms or includes one or more non-carbon atoms, i.e., heteroatom differentiation.
- The nodes of the reduced graph include carbon and heteroatom counts or ranges. These, rather than exact counts themselves, are used so as to take advantage of the collapsing between nodes of similar types.

Specific Structure Queries. When searches of the specific queries are performed against the test database, only three queries fail to retrieve their corresponding file structures; this is due to the fact that the file structures include variable numbers of substituents, designated by a multiplier in the GENSAL description. The representation currently in use treats this as a single substituent, so that query structures with two or more substituents covered by such a multiplier term fail to retrieve their corresponding file structures. Apart from these, each query retrieves at least its corresponding file structure, and, in addition, certain others that are matches at

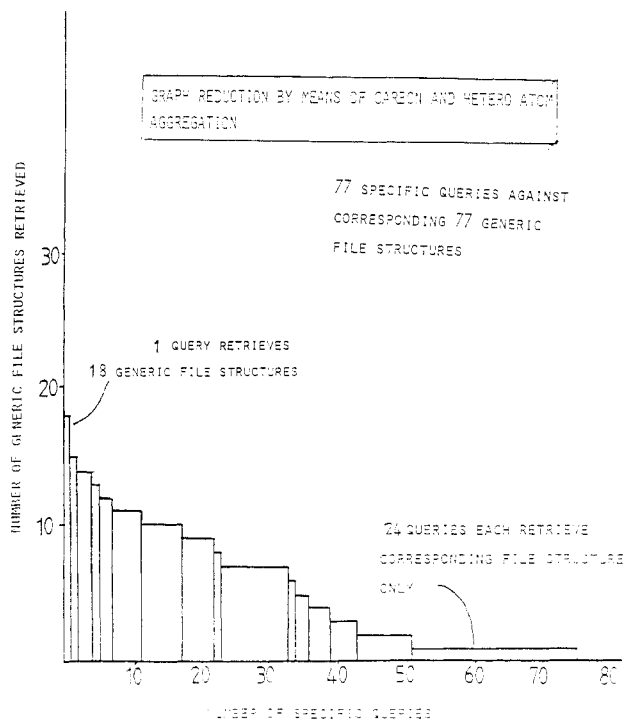


Figure 22. Distribution of results of searching 77 specific queries against 77 corresponding generic file structures for carbon/heteroatom aggregated reduced graphs.

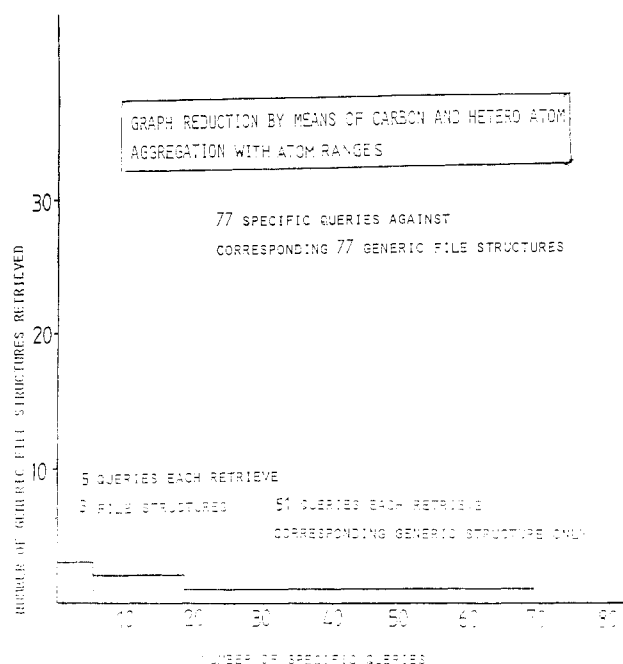


Figure 23. Distribution of results of searching 77 specific queries against 77 corresponding generic file structures for carbon/heteroatom aggregated reduced graphs with atom ranges.

the level of description of the particular reduced graph used but not when the structures are compared at the atom and bond levels. Thus, with the simple ring/nonring representation, i.e., option a, the 77 specific queries retrieve a total of 572 structures; 12 queries retrieve their associated structures only. The mean number of structures retrieved per query is thus 7.4. Figure 14 shows the rank-ordered figures for the numbers of structures retrieved by individual queries.

The representation in which the presence or absence of a heteroatom in the component is used to augment the description (case b) brings about an immediate improvement in

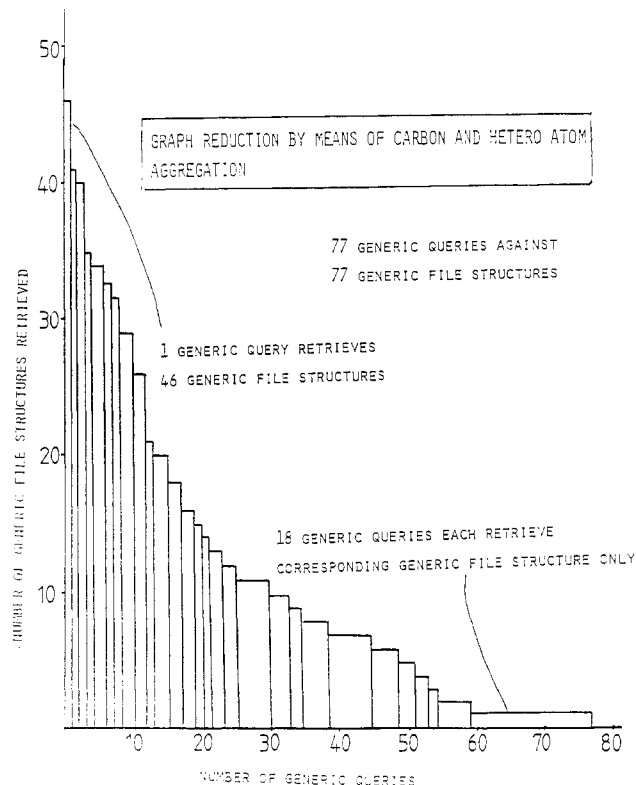


Figure 24. Distribution of results of searching 77 generic queries against 77 corresponding generic file structures for carbon/heteroatom aggregated reduced graphs.

that the mean number of structures retrieved falls at once to 3.2. The number of queries retrieving only their associated structures also increases to 28. The total number of matches at this level falls to 251. Figure 15 shows the detailed results; comparison with Figure 14 shows the improvement graphically.

Finally, the representation in which the nodes of the reduced graph also carry the number or ranges of carbon atoms and heteroatoms increases the number of queries retrieving only their associated structures to 46 and gives a total of 119 hits, giving an average per query of 1.5; the detailed results are shown in Figure 16.

Next, a set of 57 specific structures that do not correspond to the database structures was used as queries in order to determine how effective the representation is in excluding file structures which do not match queries. The numbers of file structures retrieved for this set are 319 for the simple ring/nonring reduced graph representation, 140 for that with differentiation on the basis of the presence or absence of heteroatoms, and 8 for that with counts or ranges of carbon atoms and heteroatoms. Figure 17 gives the details in tabular form. Thus, the representation can be seen to be highly effective both in detecting appropriate hits and in excluding nonrelevant items.

Generic Structure Queries. Each of the generic file structures in the database was used as a query to search the file. Complete recall was achieved in each case; however, some further structures are also retrieved, the numbers depending on the degree of detail in the representation. The ring/nonring representation results in the identification of a total of 1054 putative hits, a mean of over 13 per query, or of 12 additional structures per query. However, five of the queries retrieved only themselves. The distribution of the search results is shown more fully in Figure 18.

The number of file structures retrieved is reduced to 471 when the representation is augmented to include heteroatom differentiation, a mean of 6.1 per query, while the number of

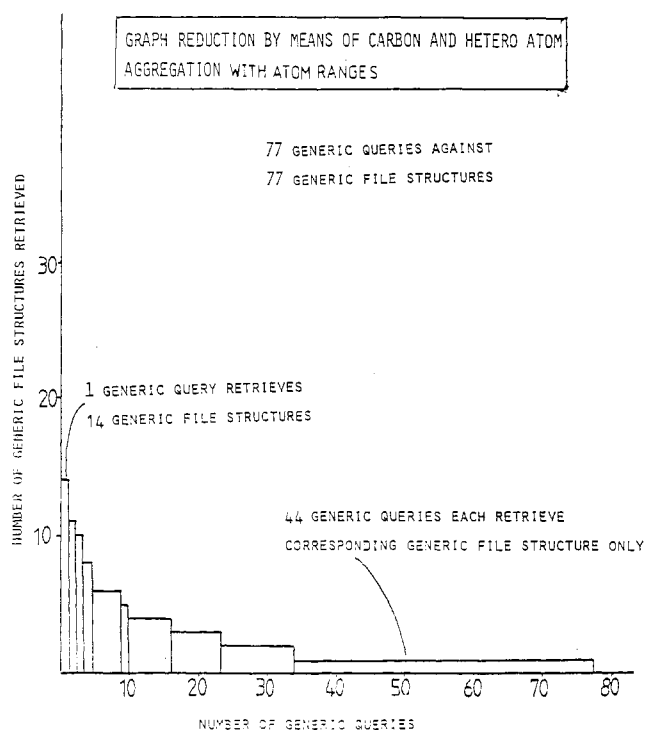


Figure 25. Distribution of results of searching 77 generic queries against 77 corresponding generic file structures for carbon/heteroatom aggregated reduced graphs with atom ranges.

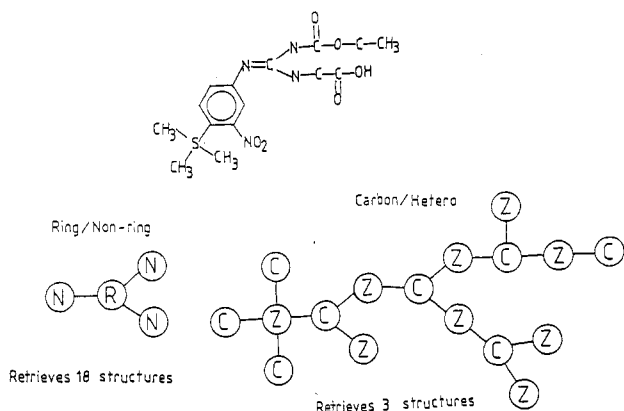
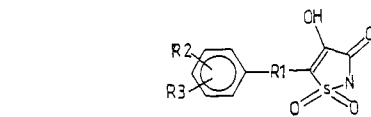


Figure 26. Example of specific structure that shows marked difference in performance when searched against the file of generics, depending on the method of graph reduction.

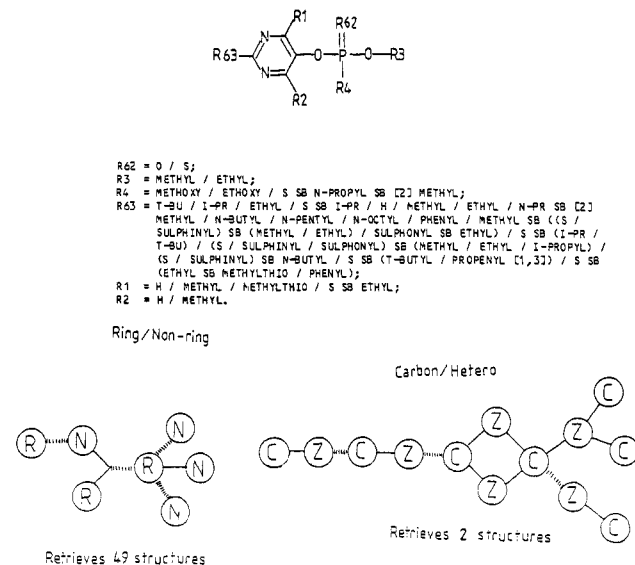
queries retrieving only themselves is now 19. The details are shown in Figure 19.

Finally, with the reduced graph representation including counts or ranges of carbon and non-carbon atoms, the total number of hits is reduced to 125, an average of only 1.6 structures per query. Moreover, 56 queries retrieve themselves alone, and the maximum number of additional hits per query is now 9. Details are given in Figure 20.

Carbon/Heteroatom Aggregation for Graph Reduction. The reduced multigraphs based on carbon/heteroatom aggregation are generated in a similar manner to the ring/nonring reduced graphs. The main difference here is that ring systems containing even numbers of nodes of sizes greater than or equal to four are possible. (An alternative strategy for their generation would allow rings of size two to be accommodated as well.) Quite complex ring systems can occur, as shown in Figure 21, and these are shown to be highly discriminating in search. Conflation of neighboring nodes now has to take into account whether or not the nodes form part of a ring system.



R1 = PHENYLENE;
R2 = H / [3] (METHYL / CL / ETHYL);
R3 = H / [4,5] CL / [4] METHYL / [5] ETHYL / [3] BR.



R62 = O / S;
R3 = METHYL / ETHYL;
R4 = METHOXY / ETHOXY / S SB N-PROPYL SB [2] METHYL;
R63 = T-BU / I-PR / ETHYL / S SB I-PR / H / METHYL / ETHYL / N-PR SB [2] METHYL / N-BUTYL / N-PENTYL / N-OCTYL / PHENYL / METHYL SB (S / SULPHINYL) SB (METHYL / ETHYL) / SULPHONYL SB ETHYL / S SB (I-PR / T-BU) / (S / SULPHINYL / SULPHONYL) SB (METHYL / ETHYL / I-PROPYL) / (S / SULPHINYL) SB N-BUTYL / S SB (T-BUTYL / PROPENYL [1,3]) / S SB (ETHYL SB METHYLTHIO / PHENYL);
R1 = H / METHYL / METHYLTHIO / S SB ETHYL;
R2 = H / METHYL.

Figure 28. Example of generic structure that shows marked difference in performance when searched against the file of generics, depending on the method of graph reduction.

Two levels of detail are included in this evaluation: first, only the node types are differentiated and, second, atom counts are included. Two sets of queries are applied, the set of 77 specific structures that correspond one-to-one with the generic file structures and the set of generic queries which are the test database itself.

Specific Structure Queries. The performance here is improved over the simple ring/nonring reduced graphs with a total of 405 structures retrieved, a mean of just over 5 per query, while 24 queries retrieve themselves alone. Five queries fail to retrieve their corresponding file structures due to the presence of multipliers, as explained earlier. The results are shown graphically in Figure 22.

When the number of atoms contained in each node is detailed, a total of 92 structures is retrieved, with 51 queries retrieving their associated structures alone (Figure 23). Eight queries now fail to retrieve their corresponding structures because of the presence of multipliers; the additional failures are due to multipliers that affect the atom ranges of nodes

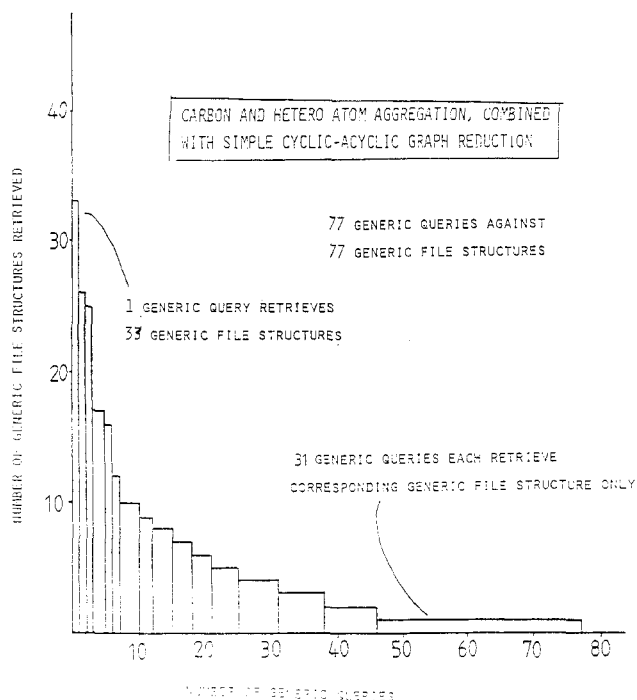


Figure 29. Results of combining ring/nonring graph reduction with carbon/heteroatom aggregation at the most general level of detail for generic queries.

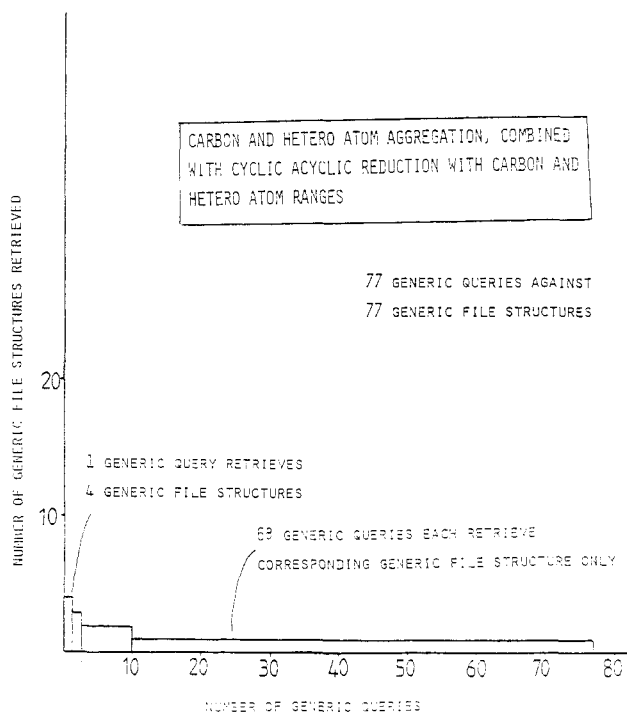


Figure 30. Results of combining ring/nonring graph reduction with carbon/heteroatom aggregation at the most specific level of detail for generic queries.

without affecting the total number of nodes, e.g., a repeating CH_2 group within a chain.

Generic Structure Queries. The performance of the simple carbon/heteroatom aggregated reduced graph shows a slight improvement over that of the corresponding ring/nonring reduction. A total of 862 structures are retrieved, a mean of 11 per query, and 18 queries retrieve only themselves, as shown in Figure 24.

The improved performance overall can be attributed to the greater complexity of the reduced graphs produced by this

method and to their consequentially greater variety. On the average, the generic structures are not as severely reduced as with their ring/nonring counterparts; the 77 carbon/heteroatom reduced graphs show a mean reduction of 2.75 to 1 in the number of nodes, and graphs with cycles are possible, adding a further discriminating feature.

The inclusion of counts of atoms within the nodes further improves performance, although not to the same degree as with the most detailed ring/nonring reduction, in that a total of 184 hits are now found, an average of 2.4 per query; Figure 25 illustrates that 44 structures retrieve themselves alone.

Combination of Both Graph Reduction Methods. Examples of specific and generic queries that show marked differences in performance depending on the nature of the graph reduction methods used are shown in Figures 26–28 together with their corresponding reduced graphs and the numbers of structures retrieved in each case. These examples demonstrate clearly that the resolving power of different graph reduction methods is dependent on the characteristics of both structures and queries.

When these two methods are combined for the 77 generic queries, an immediate increase in resolving power is apparent, as Figure 29 illustrates. Finally, the combination of carbon/heteroatom aggregation with ring/nonring reduction at the level of atom ranges shows the most powerful combination of retrieval keys yet demonstrated, as illustrated in Figure 30. A total of 89 structures is retrieved, with 68 queries retrieving themselves alone.

CONCLUSIONS

These results amply demonstrate the power of graph reduction methods in dealing with whole-structure queries, despite the small size of the sample of generics available to us. Moreover, the evaluation does not yet include the screen searching methods that have already been developed for this purpose, the addition of which should further increase the power of the methods.

The significance of the results, however, lies less in their relevance to these limited generic structure types; it has already been noted that these are expensive in labor to produce and record, since the patent document must be combed to find every mention of a specific structure and these must then be combined into a composite class representation. Rather, they give a strong indication that these methods, when suitably extended and accompanied by more elaborate matching criteria which take full account of the nature of generic radical terms, can contribute strongly to a more general solution to the problem of searching for both specific and generic whole-structure queries in generics databases.

The relevance of these characterizations to the generic radical terms of the more common generic structures is already apparent. Thus, terms such as "alkyl" and "hydrocarbyl" map directly onto C nodes in the carbon/heteroatom reduced graphs, while "cycloalkyl" and "heteroaryl" map onto the RC and RZ nodes of the ring/nonring reduced graphs. Work toward the implementation of these more general objectives is already in progress. Furthermore, it seems clear that these search methods may well prove powerful in providing further screening keys for substructure searches of databases of specific chemical structures, and work toward this end is now at hand.

ACKNOWLEDGMENT

We gratefully acknowledge funds in support of this work from International Documentation in Chemistry mbH and the provision of the Fine Chemicals Directory (FCD) by Frazier-Williams (Scientific Systems) Ltd. and of the DARING package by the SERC Daresbury Laboratory. We are also

grateful for helpful discussions with Drs. P. Willett, S. M. Welford, and J. M. Barnard.

REFERENCES AND NOTES

- (1) Lynch, M. F.; Barnard, J. M.; Welford, S. M. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 1. Introduction and General Strategy". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 148-150.
- (2) 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.
- (3) 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.
- (4) 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.
- (5) Barnard, J. M.; Lynch, M. F.; Welford, S. 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.
- (6) 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.
- (7) Welford, S. M.; Ash, S.; Barnard, J. M.; Carruthers, L.; Lynch, M. F.; von Scholley, A. "The Sheffield University Generic Chemical Structures Research Project". In *Computer Handling of Generic Chemical Structures*; Barnard, J. M., Ed.; Gower: Aldershot, 1984; pp 130-158.
- (8) von Scholley, A. "A Relaxation Algorithm for Generic Chemical Structure Screening". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 235-241.
- (9) Gillet, V. J.; Welford, S. M.; Lynch, M. F.; Willet, P.; Barnard, J. M.; Downs, G. M.; Manson, G. A.; Thompson, J. "Computer Storage and Retrieval of Generic Chemical Structures in Patents. 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 118-126.
- (10) Lynch, M. F. "Generic Chemical Structures in Patents (Markush Structures)—the Research Project at the University of Sheffield". *World Pat. Inf.* **1986**, *8*, 85-91.
- (11) 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.* **1979**, *19*, 51-55.
- (12) Lederberg, J. "DENDRAL-64. A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs. Part 2. Topology of Cyclic Graphs". NASA CR-68898, National Aeronautics and Space Administration Report No. N66-14074, 1966.
- (13) Lederberg, J. "DENDRAL-64. A System for Computer Construction, Enumeration and Notation of Organic Molecules as Tree Structures and Cyclic Graphs. Part 3. Complete Chemical Graphs: Embedding Rings in Trees". NASA CR-123176, National Aeronautics and Space Administration Report No. N71-76061, 1971.
- (14) Lederberg, J. "Topological Mapping of Organic Molecules". *Proc. Natl. Acad. Sci. U.S.A.* **1971**, *53*, 134-139.
- (15) Balaban, A. T.; Filip, P.; Balaban, T.-S. "Computer Program for Finding All Possible Cycles in Graphs". *J. Comput. Chem.* **1985**, *6*(4), 316-329.
- (16) Morgan, H. L. "Generation of a Unique Machine Description for Chemical Structures—a Technique Developed at Chemical Abstracts Service". *J. Chem. Doc.* **1965**, *5*, 107-113.
- (17) Cooper, D. C.; Lynch, M. F. "Review of Variety Generation Techniques". British Library R & D Report No. 5586, London, British Library, 1980.

ARTS: A Flexible Laboratory Instrument Control Language

W. A. SCHLIEPER and T. L. ISENHOUR*

Department of Chemistry and Biochemistry, Utah State University, Logan, Utah 84321-0300

J. C. MARSHALL

Department of Chemistry, Saint Olaf College, Northfield, Minnesota 55057

Received March 10, 1987

A generalized computer program, ARTS (Analytical Robot Telecommunications Software), has been developed to give the research scientist more flexible control of laboratory robots and instruments. As a stand-alone program, ARTS is a complete laboratory control language. ARTS can also be an extension of other software in either a master or slave mode. As master, ARTS can call on other software to perform certain tasks. In a slave mode, ARTS can act as a sensory extension of the calling software. ARTS is a flexible laboratory control language capable of adapting to changing laboratory requirements.

INTRODUCTION

Robots are important laboratory instruments, and as robot use increases, so does the need for improved external control by computers. Current laboratory robots are principally used to perform repetitive operations.¹⁻⁶ In research laboratories, the experiments performed are more varied and require diverse procedures⁷ including changing chemical or instrumental procedures on the basis of intermediate results. For laboratory robots to achieve maximum utilization in the research laboratory, it will be necessary to have external computer control designed for maximum flexibility.

Attempts have been made to improve communication between robot systems and external computers.^{8,9} These examples show the ability of the robots to interact outside their local environment. Individually these attempts address part of a bigger problem, which is the need for a more versatile robot control language to deal with changing laboratory requirements.

A generalized computer program, ARTS (Analytical Robot Telecommunications Software), has been developed to give the research scientist more flexible control of laboratory robots and instruments. Flexibility has been incorporated into the software design to promote greater software control. ARTS is capable of controlling laboratory robots and instruments on its own or in conjunction with other standard software packages and runs under the MS-DOS microcomputer operating system.

ANALYTICAL ROBOT TELECOMMUNICATIONS SOFTWARE

As a stand-alone program, ARTS is a complete laboratory control language that can interpret commands in either interactive or batch modes. The ARTS interpreter is written in the C programming language and uses Reverse Polish notation invented by Lukasiewicz.¹⁰ Each line of input is parsed according to precedence rules of the arithmetic operators