structure $-CH(CH_3)_2$ in A and yields a structure match.

ACKNOWLEDGMENT

We gratefully acknowledge financial support by International Documentation in Chemistry—IDC GmbH, Chemical Abstracts Service, Derwent Publications, Ltd., and Questel S.A. We also thank the reviewers for useful comments, which added to the clarity of presentation.

REFERENCES AND NOTES

(1) Welford, S. M.; Barnard, J. M.; Lynch, M. F. 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.

(2) 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.

(3) 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.

(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. 11. Theoretical Aspects of the Use of Structure Languages in a Retrieval System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 000-000.

(5) Meyer, E. Topological Searches for Classes of Compounds in Large Files—even of Markush Formulas—at Reasonable Machine Cost. In *Computer Representation and Manipulation of Chemical Information*; Wipke, W. T., Heller, S. R., Feldman, R. J., Hyde, E., Eds.; John Wiley & Sons: New York, 1974; pp 105-122.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 13. Reduced Graph Generation

VALERIE J. GILLET, GEOFFREY M. DOWNS, JOHN D. HOLLIDAY, and MICHAEL F. LYNCH*

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, England

WINFRIED DETHLEFSEN

BASF, Ludwigshafen/Rhine, Germany

Received February 7, 1991

Criteria for creating reduced graph representations of full generic structures for screening in full structure and substructure searching are compared; ring/non-ring reduction is identified as the principal criterion. The current form of the Extended Connection Table Representation (ECTR), the internal representation of generic structures in the GENSAL system, is shown to be an AND/OR tree, in contrast with an earlier implementation, a logical graph. The role of the ECTR in facilitating the generation of reduced graphs from a wide range of generic structures, despite the complexity of their logical textures, is detailed. The structural descriptors associated with the nodes of the resultant graphs are detailed, together with their derivation. These descriptors are common, regardless of whether they are derived from specific or generic partial structures (PSs), thus ensuring the correctness of retrieval.

1. INTRODUCTION

Research efforts have been directed toward providing a topologically based system for the storage and retrieval of generic chemical structures for over a decade. The major contributors, to date, have been Sheffield University, Chemical Abstracts Service, International Documentation in Chemistry GmbH (IDC), and Derwent Publications Ltd., together with Questel SA and INPI (the French National Patent Office). The complexity of the problem is evidenced by the fact that the first publicly available system appeared on the market only in 1989, and the need for continuing research is evidenced by the comparison of this system with the fragmentation systems it is intended to replace.[1]

Many aspects of the procedures developed at Sheffield for the storage and retrieval of generic structures have already been presented in Downs et al.[2] and earlier papers. These include the definition of GENSAL, the formally defined language used to represent structures, and aspects of retrieval such as ring perception and screening, together with earlier studies on fragment screening. Fragment screening has now been extended to cover the full spectrum of structural variation found within generics and to provide a more accurate representation making full use of the logical relationships found within generics and will be the subject of a future paper. The concept of the reduced graph as a form of representation of generic structures has also been outlined in its earlier exploratory approach, first with regard to specific structures and then to generics of limited variability.[3] Reduced graphs have also been developed by Chemical Abstracts Service, although the criteria of reduction are different from those applied in the Sheffield approach.[4]

The techniques of graph reduction have now been substantially extended, and while one or two exceptions remain, the full variety of structures can be handled. The importance of reduced graphs both as screens and as an essential preparatory step toward the refined search, providing a gross correspondence between query and file structure prior to more detailed matching, is discussed in the previous paper in this issue.[5] This paper describes their generation from the internal representation of generic structures.

2. THE ECTR

Generic structures are of the internally as an Extended Connection Table Representation or ECTR.[6] The syntactic form of the ECTR has evolved since its earlier description, driven by the experience gained in developing software to translate this representation to other forms such as ring screens and reduced graphs.

0095-2338/91/1631-0260$02.50/0

The ECTR contains three types of structural information:

(1) Structural information on distinct partial structures. These are represented by simple partial expressions that are either connection tables for specific radicals or parameter lists[7] for generic radicals.

(2) Positional/multiplicity information. This defines the positions of attachment between partial structures and the frequency of occurrence of partial structures and is contained in the Parent and Child Gates of the ECTR.[6] Child Gates give access to partial structures occurring lower down the ECTR and contain information on the logical relationships described below. Parent Gates point directly back up the ECTR to the parent structure and duplicate information already available in Child Gates. Parent Gates facilitate path tracing from partial structures at lower levels of the ECTR to those at higher levels.

(3) Logical information. Partial structures within a generic structure exist in logical relationships to one another. These relationships are fully reflected within the ECTR. Partial structures may occur together, i.e., in AND relationship, or alternatively, i.e., in OR relationship.

The ECTR[6] is a logical **graph** in which the nodes of the graph represent partial structures and the branches represent the logical relationships between the partial structures. The ECTR is accessed by the root partial structure (which usually corresponds to the invariant part of a generic structure) to which variable groups may be attached. The substituent definitions of the variable groups are referred to as child partial structures and are accessed through the parent partial structure. There is, therefore, a hierarchical relationship between a parent structure and its children. The ECTR is recursive in that any partial structure may be a parent if it is further substituted and any level of nesting is allowed. A parent structure is **implicitly** in AND relationship with its children. Child partial structures arising from different variable groups in the parent are also in AND relationship, and this is represented **explicitly** in the Child Gates of the ECTR. The partial structures in a series of alternative definitions of a variable group are in OR relationship, also given explicitly.

In the original ECTR, it was possible for a partial structure to have more than one parent structure, i.e., partial structures were nonunique and could be reached by tracing more than one path through the ECTR. This situation arose where R groups were defined together and where a group of substituents was defined together as being further substituted, e.g., "(methyl/ethyl) SB Cl" [which, by default, has the same meaning as "(methyl/ethyl)SB(1)Cl", denoting substitution by **one** Cl, in current GENSAL]. In this expression, only one partial structure was generated in the ECTR for the "Cl" radical, and this radical had multiple Parent Gates: one Parent Gate led to the "methyl" radical and a second led to the "ethyl" radical. The occurrence of multiple Parent Gates in the ECTR complicated path-tracing, particularly where doubly connected R groups occurred. Therefore, the ECTR has now been modified by introducing redundancy into the representation so that each partial structure in the ECTR is unique and has at most a single Parent Gate. (The root partial structure has no Parent Gate.) Thus for the above expression, Cl appears as two separate partial structures; one is accessed via the Child Gate of the partial structure representing methyl and the other is accessed via the Child Gate of the partial structure representing ethyl. This revised version is therefore a logical **tree**.

The other major modification to the ECTR is the removal of locks. Locks were used primarily to represent combined substitution, i.e., the possibility of substituent groups having separate definitions or being combined together to form a ring.

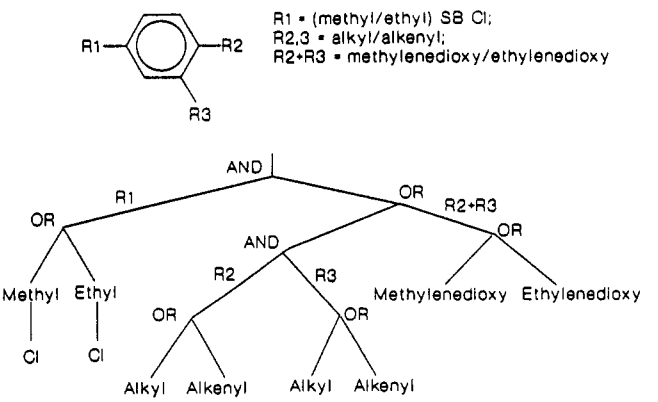


Figure 1. An example ECTR.

The logical dependencies are now made explicit in the ECTR. An example ECTR is given in Figure 1.

In summary, the original ECTR was a compact graph with minimal redundancy, e.g., partial structures could have more than one parent. It used locks to represent occurrences of combined substitutions. Conversion of this graph into a tree results in a trade-off between the additional storage requirements of a larger tree structure and the greater simplicity of algorithms required to search a tree rather than a graph.

2.1. THE ECTR AND AND/OR TREES

If the hierarchical relationships between parent and child partial structures are ignored and only the logical relationships are considered, then the ECTR can be formulated as an AND/OR tree;[8] the root node is an AND node, and all the internal nodes are AND or OR nodes. All the leaf nodes are now partial structures. This form permits data reduction within the ECTR—a process referred to as the bubble-up process. Bubble-up has been described previously[2] for the accumulation of ring screens for a generic structure; however, it is a general technique for data reduction in generic structures. It is a two-stage process in which information (e.g., ring screens or parameter values) is gathered locally with a partial structure as the ECTR is traversed top-to-bottom. The information is then combined logically at AND and OR nodes as processing recurses back up the ECTR. The result, when processing terminates at the root AND node, is information that describes the entire structure (e.g., ring screens representing the ring systems which occur within the structure either as invariant or as variable partial structures, or parameter values which represent the status of structural features within the generic structure). Data reduction, or bubble-up, can also be performed over selected branches of the ECTR, e.g., those branches and leaf nodes that correspond to a node in the reduced graph, as will be seen later.

In some cases, it is necessary to examine the context of a partial structure in order to determine information that is local to it, e.g., the connectivity of an atom can be determined only by examining connections to parent and child partial structures, and similarly for rings, since these may span partial structures (inter-PS rings[2]). The nature of the ring can be determined only by assembling the two parts and associating these with either the parent or the child. In fact, the identification of inter-PS rings is incorporated within the recursion back up the ECTR. Partial paths in the parent and child are identified as the ECTR is traversed top-to-bottom, and they are assembled in the parent as processing recurses bottom-to-top of the ECTR.

The hierarchical nature of the ECTR is important for path tracing from one atom to another within the ECTR, e.g., a path trace begins with the root partial structure and progresses to each child in a depth-first process. The AND/OR tree

character of the whole ECTR is **implicit** in the ECTR. However, an **explicit** AND/OR tree exists between a parent partial structure and its children; this tree is headed by the Child Gate field of the parent.

3. GRAPH REDUCTION

One of the requirements for a flexible system for handling generic structures is the ability to match a homologous series with one of the members of the series, e.g., "alkyl" with "methyl". In order to accommodate generic nomenclatural expressions within a conventional fragment screening system, all atom and bond centred fragments which are present in each member of the series must be identified. Thus a *bottom-up* approach is required where more *specific* information than is actually given must be derived for the generics. Generic radicals are typically characterized by the status of certain structural features, and detailed atom and bond information is not available. For example the homologous series named "alkyl" is represented in GENSAL by a parameter list indicating the absence of unsaturated bonds, rings, and heteroatoms and the presence of carbon atoms with the possibility of branching. Fragment generation must then use this generalized information to derive all possible atom and bond-centred fragments which exist in each member of the series, i.e., all fragments which fall within the limits imposed by the structural features.

The reduced graph approach to this problem is *top down* and involves generalizing specific structural features in such a way that specific and generic entities are reduced to the same form of representation. So rather than making the generic more specific, the specific is generalized to the same form as the generic.

As a very simplified example, consider matching "ethyl" against "alkyl C⟨1-6⟩". One possibility would be to generate all members of the alkyl series and to match each in turn against the ethyl group. An alternative approach involves analyzing the connection table for "ethyl" with respect to the structural features used to represent "alkyl"; a match exists if there is a match between all the structural features. In this case the end result is the same, but generalizing the specific requires much less computation.

Many different criteria can be considered for the reduction of specific structures, e.g., ring/non-ring reduction, reduction on the basis of aggregations of carbon and heteroatoms, or differentiation according to unsaturations. For generics, however, the aim is to provide a representation that is derivable both from connection tables and from parameter lists used in GENSAL, and so consideration must be given to the structural features used to represent generic structures in order to select an appropriate criterion for reduction.

Ring/non-ring reduction and carbon/hetero reduction were considered earlier.[3] Carbon/hetero (C/Z) reduction is inappropriate when full generic structures are considered; this is evident from the structural features used to represent generic radicals in patents, and consequently also in the parameter list chosen to represent those features in GENSAL. The reduction is dependent on the relative positions of carbon and heteroatoms; in many cases this information can be determined neither from the parameter list nor from the patent itself, particularly where heterocyclic ring systems are concerned. Consider the ring systems included within the expression: "R is a ring system having one or two component rings of variable size and containing two heteroatoms". This is represented in current GENSAL as heterocycle "RN⟨3-⟩RZ⟨2⟩RC⟨1-2⟩" (where RN denotes the number of cyclic atoms, RZ denotes the number of cyclic heteroatoms, and RC denotes the number of rings), and many different C/Z reduced graphs are possible, some of which are illustrated in Figure 2.

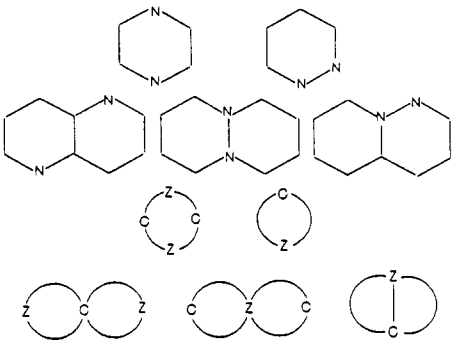


Figure 2. C/Z reduced graphs resulting from the GENSAL expression "heterocycle RN⟨3-⟩RZ⟨2⟩RC⟨1-2⟩".

The possibilities increase when the point of attachment to the parent structure is considered, e.g., the attachment could be via a carbon atom or a heteroatom (assuming that the valency would allow this). The position of attachment of a generic radical cannot be specified in the current parameter set, and so all possibilities must be assumed. (The parameter list originally specified in an earlier paper in the series is tentative, and a revised parameter list might include such a parameter; however, the problem still occurs when the attachment position is not specified in the patent.) Alternative partial reduced graphs also arise from variable substitution on heterocyclic ring systems, whether they are expressed specifically or generically. This is illustrated by the expression "morpholinyl SB Cl", where the chlorine may be attached to either a carbon atom or the nitrogen atom. A different reduced graph results in each case. Again, if the point of attachment to the parent is considered, the number of alternative partial reduced graphs increases.

Generic nomenclatural expressions are more amenable to ring/non-ring (R/N) reduction, and this is the method described here. Ring/non-ring reduction partitions the generic structures into **ring** (**R**) and **non-ring** (**N**) nodes and results in trees (unlike C/Z reduction which may result in cycles), since each ring system of the original structure reduces to a single node. A ring node represents an aggregate of atoms connected via ring bonds, i.e., a ring node represents a ring system rather than an individual cycle; thus both phenyl and naphthyl ring systems reduce to a single ring node. Connected ring nodes in the reduced graph arise from ring systems that are separated by an acyclic bond, e.g., biphenyl reduces to two connected ring nodes. Non-ring nodes are formed from aggregates of connected non-ring atoms, and connected non-ring nodes cannot exist in such a reduced graph.

The reduced graphs of specific structures may be seen, in analogy with the structure diagrams of specific structures, as graphs of "reduced structures", which can be represented as connection tables in analogy with the connection tables of the specific structures themselves, except that, as mentioned above, ring/non-ring reduced graphs are trees.

Generic structures pose more complex problems. Thus, as paper 11[9] indicates, several class-constituting mechanisms are evidenced by generic structures. These are **p-variation** (position-variation), **s-variation** (substituent-variation), **f-variation** (frequency-variation), and **h-variation** (homology-variation). The last of these designates variation within a (finite or infinite) series of specific structures that are homologous to one another in a very wide sense of homology. A non-h-variant generic structure then is one without h-variation but with p-, s-, and/or f-variation.

Reduced graphs generated from generic structures may have the same form as described for specifics, i.e., all nodes are in **AND** relationship; in addition they may contain **optional** nodes and nodes connected in **OR** relationship. In the latter case the graphs may represent classes of simpler reduced structures

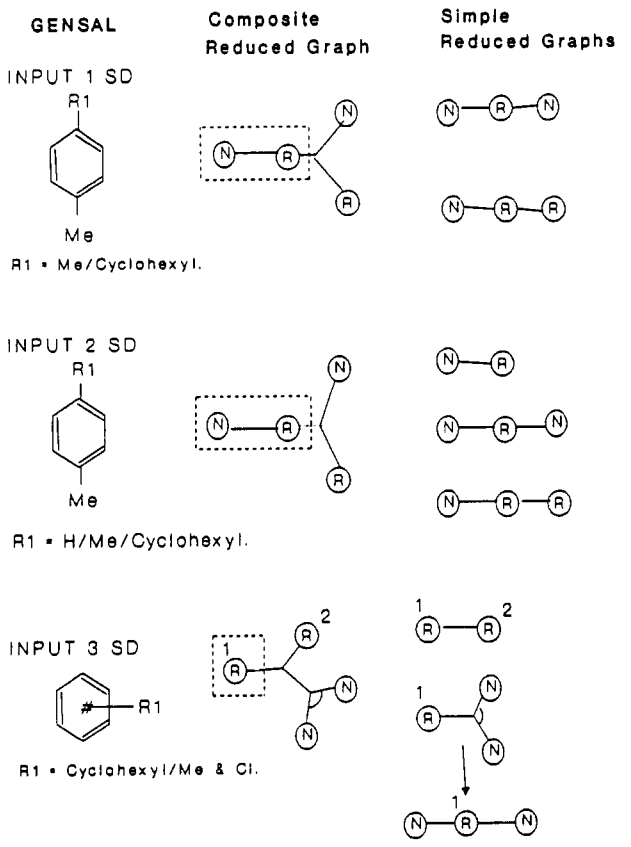


Figure 3. R/N reduced graphs of simple generic structures.

in much the same way as a non-h-variant generic expression[9] represents a class of specific structures. These reduced graphs are called **composite reduced graphs**, whereas reduced graphs consisting only of nodes in AND relationship are called **simple reduced graphs**. All specific structures reduce to simple reduced graphs. Simple reduced graphs may also be derived from generic structures, in that a single node may represent a series of alternatives, all of which reduce to nodes of the same type. This is illustrated by the expression "phenyl SB (Cl/Br/F)", which reduces to a ring node connected to a single non-ring node, the non-ring node representing the three halogen alternatives. Simple reduced graphs can be enumerated from composite reduced graphs. The nodes of a composite reduced graph that are present in each of the enumerable simple reduced graphs are referred to as the **root nodes** of the reduced graph and can be compared with the invariant part of a generic structure.[2]

Examples of reduced graphs of non-h-variant generic structures are given in Figure 3. The composite reduced graphs are shown together with the corresponding simple reduced graphs. In the simple reduced graphs all nodes are in AND relationship, shown by the solid lines connecting the nodes. The composite reduced graphs include more complex relationships: nodes in OR relationship to one another are indicated by branches; nodes in AND relationship are indicated either by solid lines with no branching or by branches connected by arcs; optional connections are indicated by **broken** lines. The nodes of a composite reduced graph which are present in each of its simple reduced graphs are enclosed by a **dashed** line.

The reduced graphs of generic structures are stored internally as extended connection tables with each row in the connection table representing a reduced graph node. The main difference between a connection table for a specific structure and a connection table for the reduced graph of a generic structure is the complex logical relationships which may exist between neighboring nodes. In this respect, the reduced graph representation can be compared with the ECTR. The neighbors of an atom in a specific structure are simply listed in an array, usually with a maximum connectivity (or array size) of 6 assumed, whereas the neighboring nodes in a reduced graph are represented as a dynamic structure that mimics the Child Gate structure of the ECTR and fully reflects the AND/OR relationships involved. The number of connections to a given reduced graph node is unlimited because a node generally represents many atoms, any of which has the possibility of forming connections across node boundaries. Since there is no limit to the number of atoms contained within a node, the number of connections to a node is also unlimited.

Each row in the reduced graph connection table contains a node identifier, the node type, i.e., ring, non-ring, or unsettled (described later), node descriptors, and also a list of all the atoms/partial structures in the ECTR which constitute that node.

4. REDUCED GRAPH GENERATION

The generation of reduced graphs from specific structures is trivial; however, generic structures require complex path-tracing routines within the ECTR, particularly where the more complex features of GENSAL are concerned. The algorithm is initially described for generic structures that are represented by a subset of the features of GENSAL, i.e., for non-h-variant structures containing only singly connected R groups and no combined substituents. The more complex features are introduced later.

Tracing of the ECTR is done at two levels. **Global** tree trace is a depth-first recursive trace through the ECTR which visits each partial structure once, and once only. Global tree tracing is used by all the algorithms developed for processing the ECTR, e.g., fragment screening, ring screening, and reduced graph generation; however, the actions to be performed on partial structures are obviously different for each algorithm. (When bubble-up is performed over an entire structure, it is incorporated within this global trace.) Nodes are generated during global tracing of the ECTR, and tracing of reduced graph nodes is initiated each time a partial structure is accessed by this global trace. However, nodes may span partial structures, i.e., a node generated in a partial structure may extend into its children and even its children's children, and so on. Therefore, **local** tracing is required to identify the boundaries to a node.

Graph reduction begins with the root partial structure, which is always represented by a partial connection table and contains substituents as R groups within the connection table. A start atom is selected in the root partial structure, its node type is determined, and an appropriate node is created. The partial structure in which the trace of a node begins is referred to the origin of the node. The operations that are performed are summarized as

(1) Path tracing within the partial structure to identify all connected atoms of the same type (intra-PS path tracing).
(2) Local tracing of all Child Gates to find the boundaries of the node.
(3) Identifying connections across the node boundaries. These can be either within the partial structure in which the node originates (internal or intra-PS neighbors) or in partial structures other than that in which the node originates (external neighbors), e.g., if a node boundary coincides with a partial structure boundary as is the case for "phenyl SB Cl", or if a node extends across partial structure boundaries as for "methyl SB benzyl".[7]
(4) The accumulation of node descriptors—described in a later section.

In practice, the identification of internal neighbors is incorporated within the intra-PS path trace, and the identification of external neighbors is incorporated within the local Child Gate tracing. As already discussed, neighboring nodes are represented as an AND/OR tree; however, internal neighbors are always in AND relationship with the node being traced, whereas the relationship of external neighbors is determined by the logical relationships that exist between the partial structures in the ECTR.

4.1. CHILD GATE TRACING

Nodes may extend from one partial structure into another via connections that are specified in the Gates of the ECTR, and local tracing is therefore required between partial structures. Node generation for singly connected substituents occurs during downwards tracing, i.e., a node originating at some partial structure in the ECTR can be extended only to its children, i.e., PSs which occur at lower levels in the ECTR.

Child partial structures are introduced in one of two ways in GENSAL: either by including R groups within a partial connection table, which may be in fixed or variable attachment positions on rings, or by the use of the SB and OSB operators ("substituted by" and "optionally substituted by", respectively). However, both cases are represented in the same way in the ECTR, i.e., part of the same AND/OR tree that is accessed via the Child Gate of the parent.

A partial structure may give rise to more than one node, e.g., the partial structure representing "benzyl" contains ring atoms and a non-ring atom. Therefore, as the Child Gates are traced for a given node, a check is made to see if the connection is to an atom contained in the current node and not to an atom in another node. The relative positions of attachment in the parent and child partial structures are contained in the Child Gates. When a connection to an atom of a child partial structure is established, one of four possible cases arises:

(1) The child partial structure is hydrogen.
(2) The connection is to an R group.
(3) The connecting atom forms a node boundary, e.g., the connecting atom in the child is a ring atom and the current node is non-ring, or vice versa.
(4) The connecting atom becomes included within the current node, i.e., the connecting atom is of the same type as the node being generated.

If the child partial structure is hydrogen, then this forms one of a series of alternative substituents and indicates that substitution is optional. Any nodes that are derived from the non-hydrogen alternatives are made optional neighbors. Local tree tracing through this branch of the ECTR is terminated, and path tracing proceeds to the remaining Child Gates, i.e., other alternative PSs or PSs in AND relationship. If the connection is to an R group, then the local trace must descend a further level in the ECTR to discover the connecting atom, and a recursive call is made to trace the Child Gates of the child partial structure. If the connecting atom forms a node boundary, it must be noted as an external connection. This atom will not have been reduced to a node at this stage (i.e., it will not have been accessed by the global tree trace), and therefore a temporary label is used to indicate the connection. This label will be converted to a node identifier at a later stage. Local tracing is terminated since a boundary to the node has been identified, and processing continues with other branches. If the connecting atom is of the same type as the node being traced, it becomes included within the node. The rest of this child partial structure must be examined in a manner completely analogous with the root partial structure (except that a new node is not created): the intra-PS connections of the child partial structure, if any, are noted, and the local trace

```
PROCEDURE Trace_Local_ChildGates(ChildGate : );

BEGIN
FOR Each_Child_Partial_Structure DO
  IF Connected_To_Atoms_In_Node(ChildGate) THEN
    CASE Connecting_Atom OF
      Hydrogen    : Assign_Neighbours_As_Optional;
      R-group     : Trace_Local_ChildGates(ChildGate^.ChildPs^.ChildGate);
      RingAtom    : Add_To_External_Neighbours;
      NonRingAtom : BEGIN
                      Trace_Intra_Ps_Path(ChildPs);
                      Trace_Local_ChildGates(ChildPs^.ChildGate);
                      Combine_Internal_And_External_Neighbours;
                    END
    END
END;

PROCEDURE Generate_NonRing_Node(Ps : );

BEGIN
Create_New_Node;
Trace_Intra_Ps_Path(Ps);
Trace_Local_ChildGates(Ps^.ChildGate);
Combine_Internal_And_External_Neighbours;
END;
```

Figure 4. Pseudocode description of the generation of a nonring node contained within the partial structure Ps.

is continued through the Child Gates of the child partial structure, exactly as already described, until eventually all the boundaries of the node have been identified. In this way all connected atoms of the same type become included within the node.

Once the extent of a node has been determined and local tracing has terminated for each branch of the ECTR which can be reached from the parent partial structure, processing returns to the root partial structure, the internal and external neighbors are combined, and the node is complete. If there are atoms remaining in the root partial structure that are not yet reduced, then a new node is created and generated as above. Once all atoms are incorporated within nodes, the processing of this partial structure is complete and the global tree trace continues to the next partial structure in the depth-first trace.

Ring and non-ring nodes are generated in a very similar manner, the only difference being the identification of atoms which belong to the node. Ring nodes, however, are extended into child partial structures only via doubly connected substituents, e.g., "phenyl SB methylenedioxy" or a ring containing R groups as ring members. The handling of doubly connected substituents is described later.

Processing of a non-ring node is described by the pseudocode in Figure 4.

When the global tree trace has progressed to partial structures at lower levels than the root partial structure, it is possible that some (or all) of the atoms have already become incorporated within a node, so it may be that no action is required, or that reduction is necessary only for some of the atoms. Thus, the connection to the parent must be examined for all partial structures lower than the root to see if atoms have already been reduced from above. Consider the expression "methyl SB (Cl/Br/I)"; all of the halogen atoms will have been included within the node initiated in the partial structure representing "methyl". Each time a node is completed, all previously generated nodes are examined to see if any contain temporary labels to atoms in the newly created node. If such labels are found they can now be replaced by an identifier of the newly created node.

At the completion of global tracing, a reduced graph has been produced; however, further processing of the reduced graph may be required.

4.2. FURTHER PROCESSING OF THE REDUCED GRAPH

When node boundaries coincide with partial structure boundaries, it is possible that further processing of the reduced

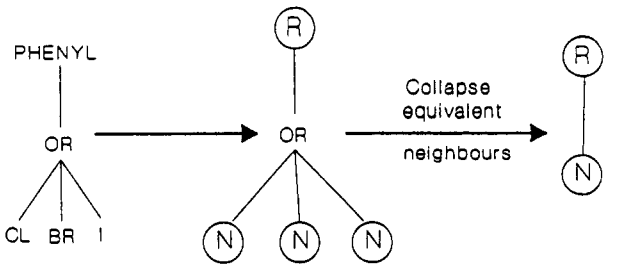


Figure 5. Collapsing equivalent node types.

graph is required. Consider the expression "phenyl SB (Cl/Br/F)"; the reduced graph generated initially consists of a ring node connected to three alternative non-ring nodes. The graph is then further processed to collapse equivalent alternative nodes of the same type. The process is illustrated in Figure 5. It is possible that the nodes which are collapsed together carry further neighboring nodes. The higher order neighbors of the collapsed nodes are transferred to the remaining node.

A worked example of graph reduction is illustrated in Figure 6.

5. H-VARIANT PARTIAL STRUCTURES

As previously noted, one of the aims of graph reduction is to produce a common representation for non-h-variant and h-variant generic full structures[9] as far as possible, in order that matches can readily be achieved between a homologous series and one of the specifics that is a member of that series.

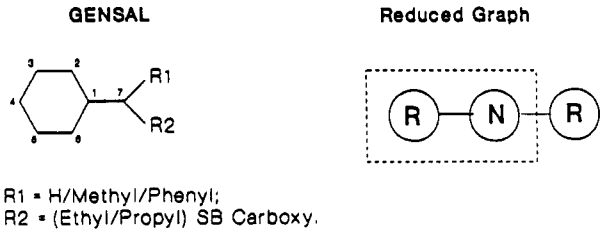
Non-h-variant generic full expressions (previously referred to as structurally explicit generics[10]) reduce to graphs where the nature of the nodes generated is settled with respect to the criteria of reduction, i.e., each atom in the structure is contained within either a ring node or a non-ring node. Generic structures containing h-variation **may** contain nodes which are no longer settled with respect to the criteria of reduction, e.g., hydrocarbyl is **unsettled** with respect to R and N. However, many generic nomenclatural expressions are settled with respect to R and N when considered in their narrowest conceivable extensions, e.g., "alkyl", "cycloalkyl", "aryl".[9] The only expressions which are unsettled with respect to R/N reduction within the current GENSAL system are "hydrocarbyl" and "radical", and any composite expression which includes these terms, e.g., "acyl". **Other Terms** in the ECTR, arising from expressions such as "electron-withdrawing group", also reduce to unsettled nodes. In h-variant generic structures both specific and h-variant partial structures are processed during node generation.

When a h-variant partial structure is accessed, the associated parameter list is analyzed to determine the node type as ring (**R**), non-ring (**N**), or unsettled (**U**). This process replaces the path trace within a specific partial structure to find connected atoms of the same type. The h-variant partial structures are then processed in the same way as specific partial structures: the parent connection is examined to see if the partial structure has already been incorporated within a node originating nearer the root of the ECTR; if so, the partial structure is skipped, otherwise its children are examined recursively to find the extent of the node and any neighbors external to the node. Each h-variant partial structure reduces to a single node, and there are therefore no internal neighbors.

h-Variant partial structures may be accessed during local tracing from specific partial structures. In this case, there are two possible actions only:

(1) The node is extended into the h-variant partial structure.

(2) The h-variant partial structure forms a boundary to the current node.



R1 = H/Methyl/Phenyl;
R2 = (Ethyl/Propyl) SB Carboxy.

```
Process PS 1: Create_New_Node(Ring Node 1);
              Trace_Intra_PS_Path - include atoms 1-6;
              Note internal connection to atom 7;
              Trace_Local_ChildGates - no connection to atoms in node;
            : Create_New_Node(Non-ring Node 2);
              Trace_Intra_PS_Path - atom 7;
              Note internal connection to Node 1;
              Trace_Local_ChildGates:
                 R1: Hydrogen:
                        Assign_Neighbours_As_Optional;
                     Methyl:
                        Trace_Intra_Ps_Path - include atom 1;
                        Trace_Local_ChildGates - none;
                        No internal neighbours;
     Phenyl:
Node boundary;
Note external neighbour - atom 1;
                 R2: Ethyl:
                        Trace_Intra_PS_Path - include atoms 1 and 2;
                        Trace_Local_ChildGates:
                           Carboxy:
                              Trace_Intra_PS_Path - include atoms 1-3;
                              Trace_Local_ChildGates - none;
                              No internal neighbours;
                     Propyl:
                        Trace_Intra_PS_Path - include atoms 1-3;
                        Trace_Local_ChildGates:
                           Carboxy:
                              Trace_Intra_PS_Path - include atoms 1-3;
                              Trace_Local_ChildGates - none;
                              No internal neighbours;
              Replace neighbour label in Node 1 by Node 2;
Process PS 2: Hydrogen:
                 No atoms to reduce;
Process PS 3: Methyl:
                 Atom already reduced;
Process PS 4: Create_New_Node(Ring);
                 Trace_Intra_PS_Path - include atoms 1-6;
                 Replace neighbour label in Node 2;
                 Trace_Local_Child_Gates - none;
Process PS 5: Ethyl:
                 Atoms already reduced;
Process PS 6: Carboxy:
                 Atoms already reduced;
Process PS 7: Propyl:
                 Atoms already reduced;
Process PS 8: Carboxy:
                 Atoms already reduced;
```

Figure 6. Worked example.

Processing allows a non-ring node originating in a specific partial structure to be extended into and through a h-variant partial structure, e.g., "carboxy SB alkyl SB Cl" collapses to a single non-ring node during local tracing, whereas for "carboxy SB heterocycle", the non-ring node traced for "carboxy" is bounded by the h-variant partial structure. h-Variant partial structures of an unsettled nature form boundaries to both ring and non-ring node generation.

6. DOUBLY CONNECTED SUBSTITUENTS

Doubly connected substituents are intrinsically more difficult to process, especially where path tracing is concerned. A complete path trace through doubly connected substituents at the atom-bond level begins in the parent partial structure, proceeds to the doubly connected children (possibly through multiple levels in the ECTR), and eventually must return to the parent partial structure, i.e., the ECTR must be traversed in both the upwards and downwards directions. It is possible to implement such tracing in the ECTR by using Parent Gates which lead from child partial structures back to their parent; however, the difficulty, as far as graph reduction is concerned,

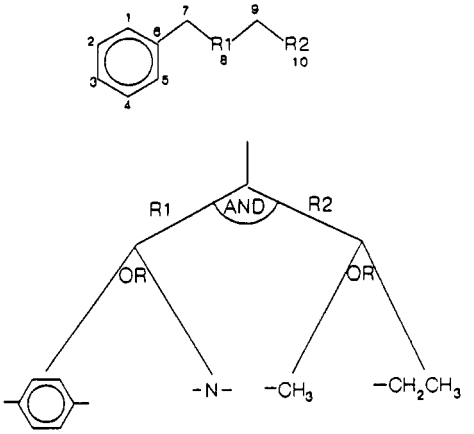


Figure 7. Doubly connected substituents.

is that the pattern of nodes in the parent structure is dependent on the nature of its children. This is illustrated in Figure 7. When R_1 is "-N-", the non-ring node is traced from the root partial structure, to the child, back to the root, and then on to the R_2 substituents. When R_1 is "phenylene", the non-ring node is terminated, and atom 9 in the root partial structure gives rise to a new non-ring node. The effect of this, as far as path tracing in the ECTR is concerned, is that atoms are no longer uniquely reduced to nodes, i.e., atom 9 of the root partial structure belongs to two different nodes in the reduced graph.

Doubly connected substituents, as with other substituents, may be entered in GENSAL either by the SB operator, e.g., "phenyl SB methylenedioxy", or by the inclusion of doubly connected R groups within a partial structure. Substituents added by the operators SB and OSB form rings, either by fusion with another ring, as in the above example, or by the closing of a chain, e.g., consider the GENSAL expression "*n*-butyl SB [1/4] (methylene/ethylene)", which means "*n*-butyl substituted through positions 1 and 4 by either methylene or ethylene" and results in a 5- or 6-membered, fully saturated carbocyclic ring. Doubly connected substituents represented by R groups can fall within chains or within rings.

Doubly connected substituents which are ring-forming and R groups within rings do not require any special handling. Local tracing in the ECTR need be only via Child Gates, i.e., downward through successive levels of the ECTR. If a path is traced from one of the connecting atoms to a child partial structure and back to its parent, it leads back to the same reduced graph node and does not change the pattern of nodes in the parent structure. However, doubly connected R groups within chains require special treatment.

Doubly connected substituents within chains are treated as forming boundaries to nodes during ECTR tracing, thus producing an intermediate reduced graph which contains *pseudorings*. The pseudorings are then expanded as necessary. The processes involved are illustrated in Figure 8 for the generic structure shown. The pseudorings are illustrated by **dotted** lines. Nodes derived from the partially reduced doubly connected substituents have the same parent node, and path tracing in the intermediate reduced graph leads to the same leaf nodes, i.e., leaf nodes of the reduced graph can be reached by tracing more than one path through the reduced graph. Chains of doubly connected R groups and nested doubly connected R groups give rise to connected pseudorings and nested pseudorings, respectively, as shown. Connected and nested pseudorings are grouped for expansion, e.g., the nodes 3–10 are grouped. The reduced graph is traced from the root node to the leaf nodes. When a group of nodes is found that requires expansion, the expansion begins from the nodes nearest the leaf nodes and works back toward the root, as

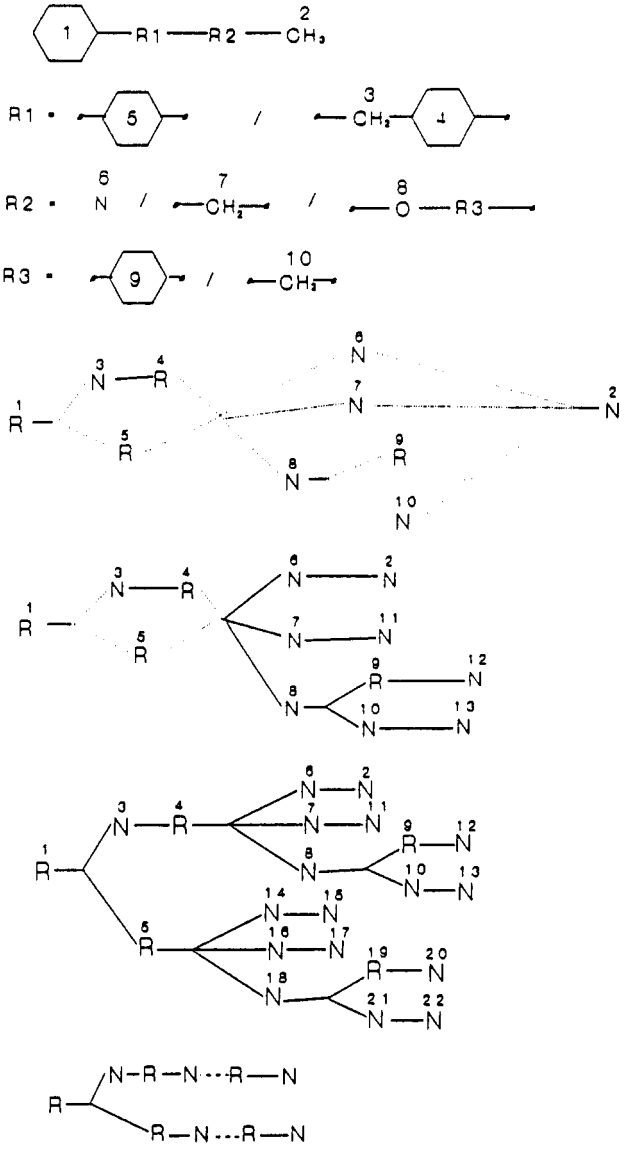


Figure 8. Intermediate reduced graph containing pseudorings.

shown. Forbidden node pairs may result from this process, e.g., N–N, and connected non-ring nodes are concatenated (this may involve combining neighbors, if the node carries further nodes.) This process is similar to that described by CAS.[4]

Although the reduction of nodes spanning doubly connected children is avoided by constructing artificial boundaries to the nodes, Parent Gates are accessed in the following cases:

When the parent structure consists of a chain of connected R groups.

For the bubble-up of accumulated information for nodes which involve doubly connected substituents, as described later.

If a substituent has arisen from a doubly connected R group that is connected to further R groups, e.g., R_1 and R_2 of Figure 8, then connections between atoms of the substituent definitions of R_1 and atoms of the substituent definitions of R_2 must be established. This is achieved as the substituents of R_1 are processed by global tree trace, by ascending to the parent and descending the Child Gates of connected R groups. Therefore, an additional local trace is implemented via the Parent Gate of the partial structure. When a connection is established one of three cases arises:

(1) The connecting atom is hydrogen—any nodes which are derived from the non-hydrogen alternatives are made optional neighbors.

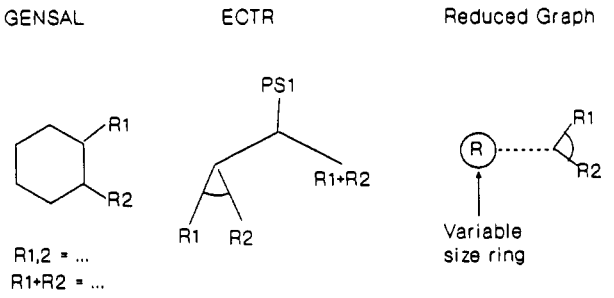


Figure 9. Combined substituents on a ring.

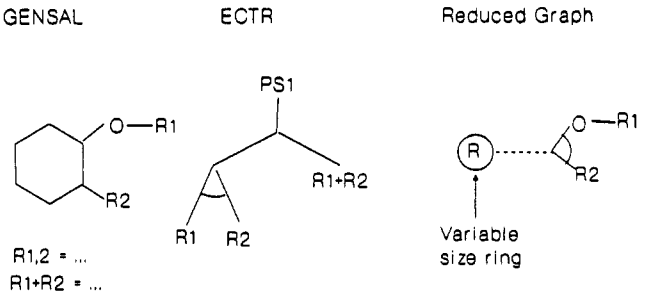


Figure 10. Variable atom environments due to optional combined substitution.

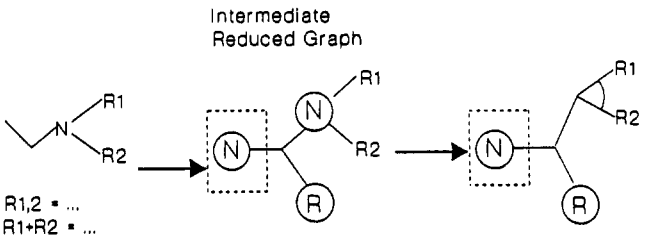


Figure 11. Variable atom environments due to combined optional substitution.

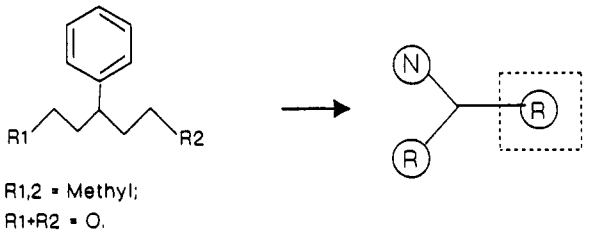


Figure 12. Variable atom environments in the root partial structure.

(2) The connection is to an R group—recursive tracing is required.
(3) The connecting atom forms a node boundary—an external neighbor is recorded.

The operations performed for a doubly connected substituent can be summarized as:

(1) Path tracing within the partial structure to identify all connected atoms of the same type (intra-PS path tracing) and internal node neighbors.
(2) Tracing node neighbors via the parent partial structure.
(3) Local tracing of all Child Gates to extend the node and to construct further external node neighbors.
(4) The accumulation of node descriptors—described in a later section.

7. COMBINED SUBSTITUENTS

Combined substituents are R groups which may have independent existence as singly connected substituents or can combine to form rings. Difficulties arise when atoms in the structure can exist in two environments, i.e., within rings or within chains. In this case, the bonds associated with such atoms are set as variable ring/chain. Bonds are identified as ring, non-ring, or variable during a ring perception phase that is carried out within the GENSAL interpreter. Since the criterion for reduction is the identification of atoms as belonging either to ring or to non-ring nodes, atoms which exist in either topology call for special treatment. In general, the variable atoms become incorporated within more than one node and are no longer uniquely reduced to a single node. When variable atoms occur in the simplest case, the atoms are reduced once as if they are non-ring, and again as if they are connected by ring bonds. When variable atoms are encountered during tracing of ring or non-ring nodes, node boundaries are always constructed and the variable atoms are cited twice as neighbors in order that they are eventually attached to both alternative nodes, e.g., the ring version and the non-ring version.

When all the variable groups that can have independent existence or can combine together are attached directly to a ring, no special treatment is required. Figure 9 illustrates this case. The logical relationship between the various substituents is given directly in the ECTR. The ring node is traced within the parent partial connection table. During Child Gate tracing, the independent definitions of the R groups are accessed first and the connecting atoms in the children noted. The combined definition is traced next, and the ring node extended, with the variable composition of the ring system noted. If the extended ring is bounded by non-ring bonds or by further children, these are in OR relationship with the R_1 and R_2 singly connected children, otherwise R_1 and R_2 substituents are given as optional.

Figure 10 illustrates a structure that contains variable atoms. In this example, one of the R groups is separated from the ring by an O atom, and the O atom is connected both to the ring and to the R group by variable bonds. The consequence of this for graph reduction is that the O atom can exist in one of two environments; either as part of a ring system when the combined definition of $R_1 + R_2$ is considered or as initiating a non-ring node when the substituents are taken independently. The O atom is no longer uniquely reduced to a single node. The ring is identified first and is traced internally with the O atom included, but with the variable composition of the ring system noted. Only the combined definitions of the R groups are traced. After completion of the ring node, the O atom is processed a second time as a non-ring node; this time only the singly connected definitions of the R groups are processed.

There are cases in which all of the atoms between the R groups that may be combined together are input as non-ring; an example is given in Figure 11. The atoms have already been identified as variable by the ring perception routines and are initially processed as if they are non-ring. A non-ring node is created, and the internal trace collapses connected variable atoms together; any definite non-ring atoms form node boundaries. The combined definition of the R groups is ignored during Child Gate tracing, and only the independent definitions are included. The node is not extended into these children, as above, but they are considered as forming node boundaries. When processing of the non-ring node is complete, the atoms are reprocessed as if they are ring atoms. This time, when the Child Gates are traced, only the combined definition of the R groups is traced and becomes collapsed within the node. If variable atoms occur in the root partial structure, then in some cases the structure is reduced to disconnected graphs, e.g., Figure 12.

Unfortunately, the variability is not always restricted to two alternative bonding patterns, as demonstrated by Figure 13. The highlighted atom in this example belongs to four alternative environments, i.e., it forms part of a non-ring node, it is incorporated in one of two different rings bearing different

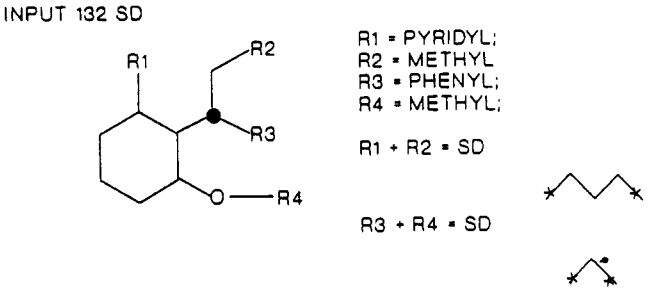


Figure 13. Four alternative atom environments.

substituents, or it is a ring fusion atom. This type of structure is not handled at present.

8. R/NC/NZ REDUCED GRAPHS

A further level of discrimination may be introduced, if desired, in order to enhance the resolution to be gained in the search. The criteria for the generation of reduced graphs from specific partial structures can be extended so that the non-ring components are further differentiated into aggregates of non-ring carbon atoms and non-ring heteroatoms, respectively.[5] The effectiveness of this measure will be reported in a later paper.

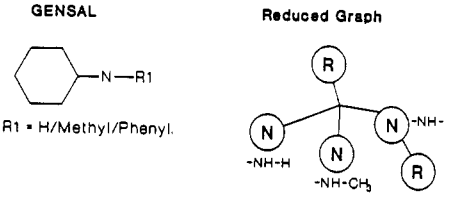
9. REDUCED GRAPHS WITH NON-A-VARIANT NODES

The role of non-a-variant reduced graph nodes in the refined search has already been described,[5] (where a structure is non-a-variant if it is non-p- *and* non-s- *and* non-f-variant). The procedures described above are readily modified to produce such nodes. In the simplest case, when a node boundary coincides with a partial structure boundary, the reduced graph with non-a-variant nodes is produced and further processed to collapse alternative nodes of the same type. The production of reduced graphs with non-a-variant nodes requires only the omission of this further step. When nodes extend from a partial structure into child partial structures, the final node encompasses branches of a subtree of the ECTR. Rather than collapse this subtree into a single reduced graph node, it is now necessary to create a new node for each leaf of the subtree. Instead of the local trace accumulating atoms from partial structures at lower levels of the tree, nodes are created in the terminal partial structures of the local trace. This can be summarized by the pseudocode of Figure 14.

10. NODE DESCRIPTORS

The discrimination of reduced graph nodes is increased by the accumulation of node descriptors, which can be derived from both specific partial structures and generic partial structures. The applicability of Enumerative Parameter Values (EPVs) and Reduced Parameter Values (RPVs) to reduced graph nodes has already been discussed.[5] Node descriptors have been implemented to include EPVs and RPVs. Molecular formula ranges, characterizing reduced graph nodes in a similar manner to molecular formulas characterizing specific full structures, are used in the form of EPVs and include counts of the individual occurrences of carbon, nitrogen, oxygen, sulfur, phosphorus, halogen, and a count of *other heteroatoms*. The parameters given below are evaluated as RPVs. The RPV of

t indicates the presence, absence, or possible presence within the node of 3-connected atoms, i.e., atoms which are connected to three non-hydrogen atoms. This RPV is distinct from that of the parameter T, which indicates the presence, absence, or possible presence of carbon atoms attached to three other carbon atoms within an acyclic (partial) structure.




```
PROCEDURE Trace_Local_ChildGates(ChildGate : );

BEGIN
FOR Each_Child_Partial_Structure DO
  IF Connected_To_Atoms_In_Node(ChildGate)
    THEN
      CASE Connecting_Atom OF
        Hydrogen    : Assign_Neighbours_As_Optional;
        R-group     : Trace_Local_ChildGates(ChildGate^.ChildPs^.ChildGate);
        RingAtom    : Add_To_External_Neighbours;
        NonRingAtom : BEGIN
                        Create_New_Node;
                        Trace_Intra_Ps_Path(ChildPs);
                        Trace_Local_ChildGates(ChildPs^.ChildGate);
                        Combine_Internal_And_External_Neighbours;
                      END
      END
END;

PROCEDURE Generate_NonRing_Node(Ps : );

BEGIN
Trace_Intra_Ps_Path(Ps);
Trace_Local_ChildGates(Ps^.ChildGate);
END;
```

Figure 14. Pseudocode description for non-a-variant node generation (including the GENSAL description and the reduced graph representation of the generic structure).

q indicates the presence, absence, or possible presence within the node of 4-connected atoms, i.e., atoms which are connected to four non-hydrogen atoms. This value, again, is distinct from that of the parameter Q.

b, a general branching parameter, indicates the presence, absence, or possible presence of 3- or 4-connected atoms, i.e., atoms that are connected to three or four non-hydrogen atoms within the node. b is present if either t or q is present.

E indicates the presence, absence, or possible presence of double bonds within the node.

Y indicates the presence, absence, or possible presence of triple bonds within the node.

U, a general unsaturation parameter, indicates the presence, absence, or possible presence of either double or triple bonds within the node. U is present if either E or Y are present.

As indicated, RPVs can take one of three values, specifically, Obligatory Presence ($\mathbf{P}_o$), Obligatory Absence ($\mathbf{A}_o$), and Possible Presence ($\mathbf{P}_p$) of the respective structural feature. EPVs, containing integer ranges (which may be genuine ranges or distinct integers), can be reduced to RPVs in accordance with the following:

$\mathbf{P}_o$, if the integer range does not include 0

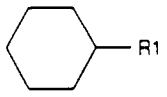
$\mathbf{A}_o$, if the integer range is 0

$\mathbf{P}_p$, if the integer range includes 0 but is not identical with 0

Each of these values can be represented by 2 bits for rapid searching:

Presence ($\mathbf{P}_o$)	represented by	1 0
Absence ($\mathbf{A}_o$)	represented by	0 1
Possible presence ($\mathbf{P}_p$)	represented by	1 1

Nodes in the reduced graph may span many partial structures, and this superposition of partial structures is equivalent to the superposition of branches in the ECTR. The bubble-up process for accumulating information for a reduced graph node is almost identical therefore to bubble-up over the whole ECTR. The only difference is that the accumulation of information is carried out during a local tracing of each node



R1 = Methyl SB (Cl & Br)

Figure 15. A three-connected group.

and not during a global trace through the whole ECTR.

The process of accumulating RPV values at AND and OR nodes of the ECTR within each reduced graph node is essentially the same for RPVs and EPVs. For EPVs, the operation at AND nodes is an ADDing of ranges; the operation at OR nodes is an ORing of ranges. For RPVs, the operation at AND nodes is an ORing of the first bit for each value and an ANDing of the second bit for each value; the operation at OR nodes is an ORing of each bit for each respective value.

10.1. BRANCHING RPVS

The branching node descriptors record the presence, absence, or possible presence of 3-connected and 4-connected atoms within a node. Connections across node boundaries are also included. As an example, a node containing one atom with three non-hydrogen connections, and a second atom with four non-hydrogen connections, has node descriptors representing branching as: $b\langle P_o\rangle t\langle P_o\rangle q\langle P_o\rangle$.

The connectivity of an atom is obviously dependent on its environment both within the partial structure in which it occurs and external to the partial structure, i.e., possible connections to other partial structures (parent or child connections). Connections are not always given explicitly in the connection table, e.g., substituents introduced via the "SB" and "OSB" operators. Even where connections to child partial structures are given explicitly within the connection table (by the inclusion of an R group), this does not guarantee an external connection since the R group may be defined as a series of alternative substituents that includes hydrogen. Thus, individual atoms within a partial structure may have variable connectivity, and this can be determined only by examining the connecting atoms in child partial structures.

Parent Gates also have to be examined; consider the structure in Figure 15 where the "methyl" group is substituted by both "Cl" and "Br". Here the carbon atom of the methyl group has no internal (or intra-PS) neighbors, but has one connection through the Parent Gate and two connected atoms accessed via Child Gates; it is therefore 3-connected.

The routine CountBranchPts examines each atom within a node and is summarized by the following pseudocode:

```
for Each_Atom_In_Node do
  begin
    Count_Intra_Ps_Congeners;
    Count_Parental_Congeners;
    Count_Child_Congeners;
    Add_Into_Node_Descriptors;
  end;
```

Count_Intra_Ps_Congeners counts the congeners of an atom directly from the connection table. The Parent Gate of a partial structure contains information about the atom (or atoms if the position of attachment in the child is variable) within the connection table that forms the connection between the parent and child partial structures, and it is not actually necessary to access the parent partial structure itself. As each atom is processed, the Parent Gate is examined by Count_Parental_Congeners to see, firstly, if the atom can be connected to the parent, and secondly to see if this is the only atom which can be connected or if the position in the child is variable, in which case the atom has variable connectivity. For example, consider the expression "phenyl substituted by 1- or 2-pyridyl" (given in GENSAL as "phenyl SB pyridyl-[1,2]"); here, atoms 1 and 2 of pyridyl are variably 2- or 3-connected. A doubly connected substituent may have both connections to the same atom, e.g. "*–CH_2–*", in which case the number of congeners is incremented by two.

An atom can also have variable connectivity through its child connections, e.g., if the substituents are variably positioned, e.g., "*n*-propyl substituted in positions 2 or 3 by Cl" (given in GENSAL as "*n*-propyl SB[2,3] Cl") or contain "hydrogen" as an alternative child, e.g., "isopropyl substituted in position 2 by H, Cl, or Br" [given in GENSAL as "isopropyl SB [2] (H/Cl/Br)"]. In the first example, atom 2 of *n*-propyl is variably 2- or 3-connected, and the node generated from these partial structures has branching parameter values of: $b\langle P_p\rangle t\langle P_p\rangle q\langle A_o\rangle$. In the second example, atom 2 of isopropyl is variably 3- or 4-connected, and the node has branching parameter values of: $b\langle P_o\rangle t\langle P_p\rangle q\langle P_p\rangle$.

A partial structure may have a number of Child Gates, arising from different R groups within the partial connection table, or substituents introduced by the "SB" and "OSB" operators. The points of attachment of the substituents in the parent structure are indicated within the Child Gates, and each Child Gate is examined for each atom to see if a connection is possible. This procedure can be summarized by the partial code below:

```
for Each_Atom_Within_Node do
  for Each_Child_Gate do
    if Connection_To_Atom
      then
        .
        . {Child Gate is relevant}
        .
      else {Child Gate is not relevant}
```

Only the immediate environment of the atom is important, and the trace terminates when all atoms one bond away have been identified. This may require tracing down more than one level in the ECTR to find the connecting atom if the immediate connection is to an R group. Finally, when the connectivity of an atom is determined, its value (or range of values) is incorporated within the node descriptors. CountBranchPts is called whenever new atoms are added into a node, i.e., for each partial structure which contributes atoms to the node.

11. RING SCREENS

Information on ring nodes is further augmented with ring screens as node descriptors. The actual information that is contained within the ring screen has been described,[2] as has the bubble-up process for the accumulation of ring screens to represent an entire generic structure by two bit strings: the MUST screen and the POSS screen. The same algorithm is applied during reduced graph generation, only the method of accumulation differs. In this case, bubble-up is performed for each node over selected branches of the ECTR, i.e., those branches which lead to partial structures that contribute to the given node.

12. CONCLUSIONS

Reduced graphs provide a powerful and flexible means of representation of generic structures. The distinction between partial structures that are described specifically and generically is removed, allowing for transparency of search. Many different kinds of reduction can be envisioned, and some are described here. There is a trade-off between the degree of reduction and search time and storage requirements, and it may be efficient to implement different levels of reduced

graphs in a sequence of search operations, e.g., R/N reduced graphs, followed by R/NC/NZ reduced graphs, followed by reduced graphs with non-a-variant nodes. In addition, it is possible to vary the amount of information within nodes, while for the refined search all the parameters must be evaluated by EPVs. The effectiveness of the reduced graph as a screening method will be discussed in the next paper in the series.

ACKNOWLEDGMENT

We gratefully acknowledge financial support by International Documentation in Chemistry GmbH (IDC), and an early contribution to the redesign of the ECTR by Dr. G. U. Schwarz of BASF.

REFERENCES AND NOTES

(1) Schoch-Grubler, U. (Sub)structure Searches in Databases Containing Generic Chemical Structure Representations. *Online Rev.* **1990**, *14*, 95-108.
(2) 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.
(3) 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.
(4) Fisanick, W. The Chemical Abstracts Service Generic Chemical (Markush) Structure Storage and Retrieval Capability. 1. Basic Concepts. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 145-155.
(5) 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.
(6) 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.
(7) Welford, S. M.; Barnard, J. M.; Lynch, M. F. 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.
(8) Nilsson, N. *Principles of Artificial Intelligence.* Springer-Verlag: Berlin, 1982.
(9) 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.
(10) Downs, G. M.; Gillet, V. J.; Holliday, J. D.; Lynch, M. F. Computer Storage and Retrieval of Generic Chemical Structures in Patents. 9. An Algorithm To Find the Extended Set of Smallest Rings (ESSR) in Structurally Explicit Generics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 207-214.

A LISP Program for the Generation of IUPAC Names from Chemical Structures

K. W. RAYMOND

Department of Chemistry and Biochemistry, Eastern Washington University, Cheney, Washington 99004

Received February 21, 1990

The nomenclature software described in this paper is designed to assist students in becoming proficient at using the nomenclature rules of organic chemistry. Features of the software include a graphics routine that converts structural formulas into drawings, the ability to determine correct IUPAC names for depicted molecules, and the capacity to identify errors in names that are input by the user. This paper describes the current stage of development of the LISP-based software and discusses its effectiveness in dealing with alkanes, alkenes, alkynes, and related halides.

INTRODUCTION

The work described in this paper results from the ongoing development of software[1] designed to help beginning students of organic chemistry learn to use the IUPAC rules.[2] The goal of the project is to give these students the opportunity to practice assigning IUPAC names to a great number of organic compounds and to have the computer supply them with the same type of assistance that an instructor might provide. The initial version of the software has been used in two upper division and two health sciences organic chemistry courses at Eastern Washington University. Although no formal study to judge the effectiveness of the software has been undertaken, student response has been exceptionally favorable.

Recent reports in this journal have documented studies related to the application of the IUPAC rules by computer analysis. Davidson[3] has described an IUPAC-based method for naming complex alkanes, and Cooke-Fox and co-workers[4-6] have reported on their work with development of a grammar system to interpret IUPAC names. Starting from structural formulas, the nomenclature software described in this paper draws molecules and determines their correct IUPAC names. Additionally, errors in IUPAC names supplied by the user are identified and assistance in making corrections is provided.

The computer programs were developed using the LISP language.[7-9] LISP is an acronym for LISt Processing, which refers to the form of both LISP programs and data. Lists are defined as sequences of objects (or other lists) embedded within a pair of parentheses. The great flexibility of the language makes LISP especially suitable for applications related to organic nomenclature. Built-in and programmer-defined functions allow lists to be manipulated in terms of any number of parameters including length, alphabetical position, and numerical value. Selection of an appropriate representation for a structural formula or name allows direct application of the IUPAC rules.

GRAPHICS OUTPUT

The first stage of program use involves the drawing of a molecule of choice. For acyclic compounds, the molecule to be drawn is specified by providing its structural formula. The formula may be provided by the user or taken from a supplied library of molecules. Once entered, the formula is parsed into a list in which each atom other than hydrogen is given an individual identity and is described in terms of its position relative to other atoms, the number of attached hydrogen atoms it carries, and the number of bonds required to complete its valency. Analysis of this list allows correct bond orders to be determined and permits any errors in the structural formula to be identified. Bond orders are not specified in the original formula, since they are determined as part of the graphics routine. Using built-in LISP graphics commands, the molecule is then drawn. Entry of the formula

0095-2338/91/1631-0270$02.50/0