A Study of Chemical Compound of Graph with help of Computer Coding

Sudhir Prakash, Rakesh Kumar, Piyush Rai, Shiksha Jain, Manish Singh, Rajnish Pandey, Saanidhya Dubey, Shobhit Srivatava, Anoop K Srivastava

IET, Dr.Ram Manohar Lohia Avadh University, Ayodhya, 224001, India
Corresponding Author: sudhir_i.etfzd@yahoo.com

Abstract

In this research work we study about the outline of graph structures algorithms required for the recognition of chemical graphs by a computer automatic encoding of organic chemical structures into the line formula notation. During the study we will first classify chemical graph then find a planar projection and fundamental cycles of a chemical graph with its characterization. Then we draw that for automatic construction code from the connection table for this we need some algorithms. Thus a graph structure language like GRAAL should prove useful for chemical coding systems. Further development along these lines should be helpful in coding knotted chemical structures.

1-INTRODUCTION

In this study we outline the graph structures algorithms required for the recognition of chemical graphs by a computer automatic encoding of organic chemical structures into the line formula notation such as ALWIN (Algorithmic Wiswesser Notation) [1]. ALWIN is a new chemical coding system which retains all the salient features of the well-known WIN [2]. The logical and information structure of ALWIN has been designed to algorithmically efficiently the encoding and decoding procedures. Infect the procedures for encoding into ALWIN is very strongly dependent on the availability of many new graph algorithms and parsing algorithms developed during recent years. Since the chemical graphs vary widely in type and complexity, a programming language for graph algorithms, such as GRAAL [3, 4, 5] might turn out to be efficient for encoding into ALWIN.

The purpose here is to indicate the type of graph structure algorithms needed for automatic encoding of chemical graphs and also provide a good bibliography of the available algorithms. It is also expected that this report will stimulate interest in related areas such as Graph Grammars [6 – 9], and also the study of automata [10] for the solution of chemical graph problems [10 – 12].

In view of the enormous complexities involved in the coding of chemical graphs we will deal the graph structure algorithms qualitatively, omitting the additional precedence rules derived from chemistry [1] to obtain a unique notation. The actual algorithms and their practical implementation for chemical coding are available elsewhere [39, 40].

For convenience, important graph theoretic definitions are given in the Appendix. Other important definitions used in chemical coding are, however, included in the text.

2-CLASSIFICATION OF CHEMICAL GRAPHS

It is well known that [1, 2] chemical graphs can be classified as.....

A. Acyclic graphs
B. Cyclic graphs

Beside these two major classification chemists have subdivided these into further categories.

The acyclic graphs are subdivided thus:

A. Unbalanced structure (trees without branches).
B. Branched structure (trees).

The cycle graphs are subdivided thus:

a. Monocycles: A system of rings which consist of single cycles which are connected through a cycle structures (as substituents).
b. Fused rings systems: A system of rings in which two or more rings share a common node or edge is called a fused rings system; the common edge is called the fusion edge and common node is called the fusion edge the fusion lucent or node.
A multicycle point is defined as a node which is common to at least rings.

Example

Multicyclic point

The fused rings systems are further subdivided as follows:

I. The bicycle fused system consisting of only two rings sharing one common edge.
II. The polycyclic fused system consisting of more than two rings, every pair of rings sharing only one common edge, but having no multi-cyclic point.
III. The perfuse ring system consisting of a minimum of three rings with at least one multicyclic point.
IV. The Spiro ring system in which a fused system has a Spiro atom – an atom connected to four other ring atoms. Accordingly, three different kinds of Spiro atoms can occur.

Kind 1: Spiro atom shared by two rings:

Kind 2 : Spiro atom shared by three rings:

Kind 3 : Spiro atom shared by four rings:

V. The poly-edge shared ring system in which at least a pair of rings shared two or more edges. There are three kinds of this system:

Kind 1 : Bridged ring system in which two or more rings share more than a single edge. In such a case we define as the bridge (not in the graph theoretic sense) the shortest acyclic segment of the ring sharing more than one edge.

Example

In this case the ring 1 – 5 – 4 – 8 shares more than one edge with the ring 1 – 2 -3 – 4 – 5. The segment 1 – 8 is treated as the bridge.

Kind 2: Connected ring of ring systems:

A ring system which is connected back to itself through one or several other substituent’s ring systems.
Kind 3: Fused ring of ring systems:

A ring system where many rings are fused together in such a way as to enclose completely another ring which shares two or more edges with at least one of the outer rings.

3-AUTOMATIC PROCEDURES FOR CLASSIFICATION

For classification and coding we need to

i) Partition the graph into blocks.

ii) Obtain a planar projection of each block to choose a set of chemically significant rings.

iii) Differentiate the various categories of ring systems.

3.1. Partitioning into blocks

The problem of partitioning a graph into blocks in relatively easier than the other problems.

We know that a graph is singly connected if the removal of any single node disconnects the graph; such a point is called a cut point [13-16]. A graph is doubly connected if it can be disconnected only by removing a minimum of two nodes in it. A doubly connected graph is known as a book.

The chemical ring systems such as monocycles, bicycle, polycyclic and perfuse ring systems as well as Spiro fused systems of kind 2 and 3 are all blocks; however, the Spiro fused ring system of kind 1 is not a block, since it contains a cut-point.

Among the available algorithms for detecting blocks [14-16], Paton’s algorithm [16] seems efficient. This algorithm uses the theorem that a graph G (nodes ≥ 3) is a block if every two nodes of G lie on a common cycle of G. Accordingly this procedure consists in growing a rooted tree [17] and a set of chords. Then for each chord the corresponding fundamental circuit is formed. The edges of this circuit are labelled. Then a new circuit is found which shares at least one edge with an already labelled circuit; the edges of the new circuit are now given the same label. This procedure is iterated until no other ring is found to share an edge with an already labelled ring.

If there are no more rings, then the given system is a single block. However, if there are still many rings with a different label, we continue as before until all the rings are labelled. Then there are as many blocks as the number of distinct labels.

Detection and partitioning into blocks is a first step in chemical coding.

3.2- Finding a planar projection and fundamental cycles

This problem is easily explained with an example:

Consider the chemical graph

By the problem of finding a planner projection and fundamental cycles we mean that above graph is redrawn in a plane thus:
Note here that the graph is a redrawn to find the fundamental rings $1 \rightarrow 2 \rightarrow 3 \rightarrow 4 \rightarrow 5$, $1 \rightarrow 5 \rightarrow 6 \rightarrow 7 \rightarrow 9$, $7 \rightarrow 8 \rightarrow 9$, $1 \rightarrow 9 \rightarrow 8 \rightarrow 4 \rightarrow 3 \rightarrow 2$ and no two edges cross anywhere.

The problem of finding a planar projection is one of the celebrated problems in graph theory [13 – 15, 18 – 29]. There is a well known criterion due to Kuratowski [22] which says that a graph is planar if and only if it does not contain as a sub-graph a homographic of one of the two Kuratowski graph, viz. the complete graph $K_5$ and the complete bipartite graph $K_{3,3}$. A homeomorph of a graph is one that can be obtained from it by inserting nodes of degree or valence 2 into its edges. This criterion does not provide a practical algorithm to determine planarity.

Another interesting feature is that there always exists a realization of a planar graph in the plane such that each edge is a straight line segment [Fary [25], Wagner [26]]. An interesting algorithm for obtaining a straight-line representation of a planar, imbedded graph has been given by Davis [38]; however, no good algorithm exists still to project or imbed a planar graph.

Yet another criterion for planarity is Maclane’s criterion which says that a planner graph is a graph whose edges are shared only by two circuits. This criterion can be used for developing a trial and error procedure for projecting a graph.

There are other algorithms developed by Tuttle [27], Weinberg [15], Fisher and Wing [28], Nicholson [29], Dunn and Cahn [30]. It is not clear whether these algorithms can give a maximally planar projection of a non-planar graph.

The problem of finding chemically significant rings has been extensively studied by Corey and Peterson [31], Gotlieb and Corneil [32], Paton [33], Weinblatt [34], Fugman, et al. [35]. These algorithms can tell what the chemically significant rings are; however, these are also still complex and demand a large computer memory and time.

In any case, there are no automatic procedures which can help us redraw a graph on the plane so that no two edges intersect or to get a maximally planar projection.

### 3.3- Characterization

For the purpose of characterization we introduce the following concepts. Consider the fundamental rings in a planar projection of the chemical graph. From this graph $G$ we construct another graph, called the weighted reduced graph $R(G)$ as follows:

a. Represent each fundamental ring by a node.
b. Join each of these nodes by an edge if their corresponding rings; assign a weight equal to the number of edges shared; however, if only nodes are shared then assign a zero weight.

**Examples**

1. ![Example 1](attachment:image1)

2. ![Example 2](attachment:image2)
3.

From the nature of the reduced graph it is possible to identify the ring system using the characterization chart (Table I).

<table>
<thead>
<tr>
<th>No.</th>
<th>Nature of the Reduced graph</th>
<th>Further comments</th>
<th>the system to which Which it belongs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A single node</td>
<td>-</td>
<td>Monocycle</td>
</tr>
<tr>
<td>2</td>
<td>A null graph</td>
<td>-</td>
<td>A system of k monocycles connected to each other</td>
</tr>
<tr>
<td></td>
<td>With k nodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>An edge (unit weight)</td>
<td>-</td>
<td>Bicyclic ring system</td>
</tr>
<tr>
<td>4</td>
<td>A tree (all edges of unit weight)</td>
<td>-</td>
<td>Polyfused ring system.</td>
</tr>
<tr>
<td>5</td>
<td>A cyclic graph with All edges of unit weight</td>
<td>No ring atom of degree 4</td>
<td>Perifused ring system</td>
</tr>
<tr>
<td></td>
<td>weight</td>
<td>A ring atom of degree 4 shared by 3 fundamental rings</td>
<td>Spiro rings system II kind.</td>
</tr>
<tr>
<td>6</td>
<td>A single ledge of Spiro ring system I kind.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>or more with zero weight and the rest of unit weight</td>
<td>-</td>
<td>Spiro ring system I kind.</td>
</tr>
<tr>
<td>7</td>
<td>At least one edge with weight two or Polyedge shared type II.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>The WRG is a tree</td>
<td>Polyedge shared type I</td>
<td></td>
</tr>
</tbody>
</table>

![Diagram](image_url)
For differentiating ring of ring and bridged ring systems, we cannot arrive at a definite algorithm unless the chemical documentation lists agree upon the basic definitions of these types of systems. Therefore, we use the categorization as in Case 7, Table I.

It is possible to categorize the ring system further as multi-angular, branched or un-branched respectively, according as the WRG is cyclic, branched or un-branched tree. These algorithms are straightforward and simple [39].

We call an un-branched WRG as a linear tree when we consider in addition to topology, the geometric property that the WRG is a straight line passing through all the nodes. This implies that in a totally fused system all the fusion edges are parallel; in a totally Spiro system all the Spiro-nodes lie in a straight line; in a mixed Spiro-fused system the middle points of fusion edges and Spiro nodes lie along a straight line.

The above definition implies that for the WRG to be linear it is necessary that

a. A totally fused or Spiro linear system consists only of even node rings (excepting the end-rings).

b. A mixed-Spiro-fused linear system (excepting the end-rings) consists of odd node rings between Spiro nodes and fusion edges/ or fusion edges and Spiro nodes/ and node rings between fusion edges or Spiro-nodes.

c. If condition a is satisfied then if we count the nodes starting from one of the Spiro nodes (or the fusion nodes) of any particular ring, the next Spiro node (or the closest fusion node) is exactly the \( n+2/2 \) node where \( n \) is the number of nodes in that ring.

d. If condition b is satisfied, then if we count the nodes starting from one of the Spiro (fusion) nodes of any particular ring, the next fusion (Spiro) node is exactly the \( n+1/2 \) th node where \( n \) is the number of nodes in that ring. If we count the nodes starting from one of the Spiro (fusion) node of any particular ring to the next Spiro (fusion) node in the same ring, the same condition c holds.

4-CODING OF THE GRAPHS

The format of the ALWIN Code [1] depends on the chemical nature of the molecule. However, for convenience we will mention the graph theoretical algorithm used for coding acyclic and cyclic structures.

4.1-Acyclic structures

The coding for acyclic structures involves finding the longest path in the tree [1] such that the path contains a maximum number of branching nodes and uses a polish notation for representing the various side a branches. This in turn involves finding the centre or bi-centre of the tree for selecting the longest path. Although one could think of other coding scheme such as [36] for acyclic structures, in ALWIN we follow the logic of WLN [2] and polish notation of obtain a code [1]. In a more recent approach we use the Morgan labelling scheme to reduce the complexity [39].

Example

Consider

Table I. CHARACTERIZATION CHART

more.

8 A graph with many Components - A set of independent ring systems (blocks) connected With each other.
The structural formula is first translated into the symbolic form [1]

and the main chain (one longest path) is $G \times X \times Y \times 2 \times Y \times A$ (using precedence given in [1]).

The ALWIN code is $G \times E \times A \times G \times 1 \times A \times Y \times A \times Y \times A \times A$ including the branches in the polish notation.

4.2-Cyclic structures

For obtaining ALWIN code to planar projection of the given chemical structure consisting of rings is treated as a tessellation or tile-filling problem. The order sequence of such tile-filling operations completely describes the fused cyclic structures. This coding system uses many of the concepts used by the well-known international notations – in particular it agrees with WLN except for the description of the topology. As already mentioned in [1] WLN [2] makes use of a Hamiltonian path passing through the chemical graph, whenever it exists, or a minimal spanning tree. For large chemical graphs such a description is not desirable, since the existence and the choice of a Hamiltonian path in these graphs appear to be very difficult problems in graph theory.

The construction of ALWIN code is best explained by means of an example.

Consider the following chemical graph (Fig. 1a)
Choose any particular ring and label its nodes (locants) clockwise (Fig. 1a). The choice of the first ring and the sense of labelling are governed by certain rules, if one wants to obtain an unique code; this aspect will be treated later. By convention we name each edge by its cyclically preceding node label. Then, we choose a ring which shares a common edge with the first ring and label its nodes (continuing, from the previous ring), clockwise, starting from that node which has not been labelled and which would permit us to give a consecutive clock-wise assignment of the locants (the sense of labelling is totally clockwise or anticlockwise so as to obtain an unique code). We reiterate this process until all the rings in the given structure are exhausted.

The topology is now described by first specifying the ring numerals in the order in which the Locants have been assign to them, and then the common edges (when a ring shares only one common edge with any other ring) and connectivity – called knit (when two or more edges are shared by other rings); this can be implemented easily by drawing the reduced graph.

The reduced graph is an edge or node relational graph in which the rings in the original ring system are thought of as the nodes and the sharing or common edges and nodes between the rings as connections between the corresponding set of node. In obtaining this reduced graph, priority is given to the edge relationship over the node relationship, when both these relationships coexist; this eliminates redundancy in the description.

The reduced graph for Fig. 1a is shown in Fig. 1b. Here the nodes 1, 2 and 3 correspond to the three rings in the order in which they were assigned locants; the arcs 1 – 2, 2 – 3 and 3 – 1 in the reduced graph correspond to the common edge AB, JB and CB respectively. The code, in general, consists of the following parts.

The tessellation part. Here each common edge is represented by a single locant; this corresponds to the description of the edges of elementary trees – maximal sub-trees derived from the spanning tree such that the nodes specified in the sub-trees do not constitute cycles in the reduced graph. This gives the tessellation code.

The knit part. After describing the tessellation corresponding to the elementary tree, the remaining rings in the structure can be described by citing ring segments which complete the new rings. It will be observed that in some cases a single connected segment suffices to complete a ring, while in other cases two or more independent segments are required to complete a new ring. The completion of a ring by a single, connected segment is denoted by citing the beginning and ending locants of the particular segment; these new locants which lie in between this pair of locants are easily labelled by continuing from the shared edges of the previous rings, using the new ring numeral. The completion of a ring by two or more segments is a little more involved. In this case, there is no simple algorithm to obtain the number of locants that are included in each one of these segments; hence, we cite the number of included locants that lie between the beginning and the ending locants of each segment.

The process of completing a ring by a single segment is called a “knit” and that which involves more than a single segment is called a “multi-knit”. The “knit” is separated from the tessellation by a colon (:) and the multi-knit is separated from the knit by a semi-colon (;).

Both the knit and multi-knit correspond to those edges of a spanning tree (of the reduced graph) which are not realized in tessellation; using this the given structure is encoded.
In fig 1b one spanning tree is 1–2–3 and the corresponding elementary tree is 1–2, since 3 is connected to 1. Thus, the edge 1–2 of the reduced graph corresponds to the tessellation in which the rings 1 and 2 share the common edge AB (this is denoted in the notation by A) while the edge 2–3 corresponds to the knit; this corresponds to the segment JKLMC of rig 3 (Fig. 1a) and hence is indicated by the pair JC. The code for this graph is 6666A:JC

In fig. 2b one spanning tree of the reduced graph of Fig. 2a is 1–2–3–4–5–6–7–8–9–10. The elementary tree is 1–2–3–4–5–6–7; this corresponds to a tessellation code A, I, M, Q, T, AB. The remaining edges of the spanning tree 7–8–9–10 correspond to knit or multi-knit as explained below. The edges 7–8 and 9–9 correspond to the two ring segments (AF-AH-AI-E) and (D-AG) which close the ring 8 in Fig. 2a. This gives the multi-knit part of the code AF2E, D, AG. The edge 9–10 however, stands for the single connected segment (C-V) which closes rings 9 and 10 simultaneously in the original graph. This gives the knit part (C-V). Thus, the code is 666666666A, I, M, Q, AB; AF2E, D, AG; C,V

Since the algorithm we use is sequential in nature, it is possible that tessellation, knit and multi-knit occur many times in any order. This will be seen from the repeated occurrence of “;” and “;” in the appropriate places which indicate that the succeeding set of symbols correspond to the knit or multiunit respectively.

Also, note that although corresponding to every tessellation there is an elementary tree, the converse need not be true, since it is possible that some of the other edges in the reduced graph might have been realized by knit or multi-knit during the intermediate steps in the sequential algorithm and would only permit us to knit or multi-knit rather than tessellate. Thus, during the intermediate steps it is necessary to check which of the edges in the reduced graph have already been realized.

During the reverse process of decoding the entire code is fully read and tessellation, knit or multi-knit are performed in the required order.

It is clear that for automatic construction of the code from the connection table we need the following algorithms:

i. Finding blocks.
ii. Finding planar projection and fundamental rings.
iii. Formation of reduced graph.
iv. Finding spanning tree.
v. Finding longest path in a tree.
vi. Formation of elementary tree.

vii. Formation of tessellation—knit code.

Thus a graph structure language like GRAAL should prove useful for chemical coding systems. Further development along these lines should be helpful in coding knotted chemical structures such as globular proteins [37].

For those interested, references [39] to [42] will be very useful.

APPENDIX

The following are some of the important terms used from graph theory:

Nodes or vertices: Points of connections (atoms)

Edges or aces: Line segment connecting two nodes—(bonds).

Graph: A collection of nodes connected through arcs.

Chemical graph: A structural diagram or a molecule represented as an abstract graph.

Degree of a node: The total number of arcs connected to a node.

Tree: A connected graph having at least two nodes, that has no circuits; there is a unique arc connecting any pair of nodes.

Spanning tree: A tree which spans all the nodes of a given graph.

Elementary path: A path which visits each of the nodes contained in it only once—a tree in which there are no branches.

Minimal spanning tree: Any tree can be thought of as the union of a set of spanning tree is a spanning tree composed of the least number of elementary paths.

Hamiltonian path: An elementary path which passes through all the nodes of the given graph.

Reduced graph: The reduced graph is an edge or node relational graph in which the rings in the original chemical graph (consisting of only rings) are represented as nodes and the sharing edges or nodes between them as edges in the reduced graph.

Elementary tree: A maximal sub-tree derived from a spanning tree of the reduced graph, such that the nodes specified in this sub-tree, do not constitute a cycle in the reduced graph.

Cut point: A node, the removal of which increases the number of connected components in the original graph.

Singly connected graph: A graph in which the removal of any single node disconnects the graph.

Doubly connected: A graph in which the removal of any two nodes disconnects the graph.

Block: A doubly connected graph or a graph with no cut point.

Planar graph: A graph which can be embedded in a two-dimensional plane with no two edges intersecting except at nodes.

Plane graph: A planar graph embedded in the plane.

Non-planar graph: A graph which cannot be embedded in the plane.

Maximal planar graph: A group in which the addition of an edge between any two non-adjacent vertices makes it non-planar.

References


[7]. Pfaltz, J. and Rosenfeld, A., Web Grammars, TR-69-84, University of Maryland, Computer Science Center, College Park, Maryland, 1969.


[18]. Harary, F., Graph Theory, Addison-Wesley, 1969.


[40]. Krishnamurthy, E.V., Data-base Interface between ALWIN and fragmentary codes, Information Processing and Management, in Press.
