A Simple Algorithm to Unique Representation of Chemical Structure - Cyclic/ Acyclic Functionalized Non-Chiral Hydrocarbons

Yenamandra S. Prabhakar*

Medicinal Chemistry Division, Central Drug Research Institute, Lucknow-226 001 (U.P), India

Internet Electronic Conference of Molecular Design **2003**, November 23 – December 6

Abstract

Motivation. Unique representation of adjacency matrix of a labeled chemical structure, a graph, is of great importance in the formation of various characteristic matrices based on the vertex connectivities. It finds application in places where two or more structures or parts thereof are to be compared and evaluated for any given purpose.

Method. An algorithm, based on the vertex connectivity values, atomic weights, subgraph lengths, loops, heteroatom content, etc. has been developed to uniquely sequence and represent connectivity matrix of chemical structure of cyclic/ acyclic fictionalized non-chiral hydrocarbons. In this the terminal vertices are considered separately after completing the sequencing of the core vertices.

Results. The proposed algorithm uniquely sequences a given chemical structure according to the vertex priority. Also, this sequencing procedure generates traditional connection table of the chemical graph. This approach provides a kind of multilayered connectivity graphs which can be put to use in comparing two or more structures or parts thereof for any given purpose.

Conclusions. The present algorithm may serve multipurpose utilities such as examination of the terminal and core vertices of given chemical graphs, in the formation of connectional tables, computation of characteristic matrices/ topological indices and in storing, sorting and retrieving of chemical structures and databases.

Keywords. Chemical structure representation; graph theory; connectivity table; cyclic/ acyclic functionalized non-chiral hydrocarbons.

1 INTRODUCTION

A graph with n vertices can be sequenced in n! (factorial n) ways. If the graph is asymmetric, then each one of these n! representations will be different from the rest. To overcome this problem there are a number of methods for unique sequencing of chemical graphs, e.g., Morgan's algorithm [1,2], Wiswesser Line Notation (WLN) [3], Balaban's Hierarchically Ordered extended Connectivities (HOC) procedure [4], etc. In Morgan's algorithm, the current atom is the one with highest extended connectivity (EC) value, and if there are any attachments to the current atom which have not been assigned sequence numbers, then they are assigned sequence numbers in the decreasing order of EC value of the attachment, which includes terminal attachments (EC value is one) even before other atoms with higher EC values. Here, we propose a simple algorithm, which uniquely sequences a given chemical structure according to the vertex priorities generated in the algorithm, where the terminal vertices are considered separately after completely sequencing the core vertices. Also, this sequencing procedure generates traditional connection table of the chemical graph and finds application in different chemical graph related operations.

* Correspondence author; phone: +91-0522-2212411; fax: +91-0522-2223405; E-mail: yenpra@yahoo.com

2 METHOD

In chemical graph, a hydrogen suppressed chemical structure, the connectivity value of a vertex (an atom) is equal to the number of edges (bonds) with which it is joined to all other immediate neighboring vertices (non-hydrogen atoms). In these graphs, an edge represents either a sigma-bond or 'sigma + pi'-bonds. An algorithm has been designed to sequentially prioritize the vertices of chemical graphs in a hierarchical manner based on the connectivity values and several other associated characteristics, such as atomic weights, sub-graph lengths, loops, heteroatom content. Here, the vertices are sequenced in decreasing order of priority - that is a vertex with higher priority will be addressed and labeled first. In this procedure, the connectivity values, number of sigma bonds as well as 'sigma + pi' bonds, of all vertices will be computed. Based on the vertex connectivity values (sigma bond alone), the vertices of the graph will be divided into two groups – one group corresponds to vertices with connectivity values more than or equal to two and the other group corresponds to vertices with connectivity value equal to one. The vertices with connectivity value more than or equal to two form the core of the graph and will be prioritized successively based on their connectivity values and other associated characteristics. Once the priority of a vertex in the graph is identified and fixed, the sequencing of subsequent vertices will proceed by locating a vertex with next highest priority that is directly connected to the just prioritized vertex. Once the sequencing process encounters an end point in the current fragment propagation 'direction', a successive stepwise backward integration starts to vertex **1** to prioritize any vertices left behind. The sequencing end point on any current fragment propagation 'direction' arises due to the completion of prioritization all vertices with connectivity value more than or equal to two in that 'direction'. After fixing the priority of vertices with connectivity value more than equal to two (core vertices), the priority of vertices with connectivity value one (terminal vertices) will be fixed using the same basic rules of core vertices. The various steps of the sequencing algorithm are described below.

The sequence of various steps of the algorithm in the vertex prioritization of chemical graph

Step	**Computation** **Decision**
1	Compute connectivity (**Cn**) (sigma bond) of all vertices (**Vts**) (non-hydrogen atoms) in a given hydrogen suppressed chemical graph. Segregate all **Vts** into two groups – one with **Cn** values more than or equal to two and the other with **Cn** value equal to one. Consider all **Vts** with **Cn** value as two or more as competing **Vts**.
2	Find number of **Vts** with the highest **Cn** (sigma bond only). If only one **Vt**, then go to step **13**
3	Find number of **Vts** with highest **Cn** (sigma + pi bonds) If only one **Vt**, then go to step **13**
4	Find atomic weights (**Wt**) of **Vts** with highest **Cn** (sigma + pi bonds) If only one **Vt**, then go to step **13**

5 Divide the molecule into fragments (**Frs**) in such a way that each **Fr** contains one and only one competing **Vt** with maximum **Wt** and highest **Cn**.

6 Find the maximum **length** of **Frs**.
If only one **Fr**, then go to step **13**

7 Find the highest number of **loops** in **Frs** with maximum **length**
If only one **Fr**, then go to step **13**

8 Among **Frs** with maximum **length** and highest number of **loops**, find maximum **chain length** with competing **Vt**.
If only one **Fr**, then go to step **13**

9 Among **Frs** with maximum **length**, highest number of **loops** and maximum **chain length** with competing **Vt**, find the maximum number of **heteroatoms**.
If only one **Fr**, then go to step **13**

10 Among **Frs** with maximum **length**, highest number of **loops**, maximum **chain length** with competing **Vt** and also **heteroatoms**, find the maximum **weight** of Frs.
If only one Fr, then go to step **13**

11 Compute distance matrices for **Frs** with maximum **length**, and having the highest number of **loops**, maximum **chain length** with competing **Vt**, highest number of **heteroatoms** and also **weight**. Compare the **distances** between competing **Vts** and **heteroatoms** of each **Fr**. Find **Frs** with compactly connected competing **Vt** and **heteroatoms**.
If only one Fr then go to step **13**

12 Element of symmetry exists. Arbitrarily consider one of the competing **Vts**.

13 Prioritize (label) the **Vt** as 1(one) (subsequently with successive numbers).
If all **Vts** with **Cn** value more than or equal to two are prioritized then go to step **14** else consider the **Vts** connected to the just prioritized vertex as competing **Vts** after excluding already prioritized **Vts**, if any, from the list and go to step **2**.

14 Prioritize the **Vts** with **Cn** value one according to the priority set by competing **Vt** (steps **3** and **4**), **Cn** of immediate neighboring **Vt**.

15 End of graph sequencing.

2.1 Chemical Data

This algorithm has been used to prioritize 1-(6-methyl-8-aza-tricyclo[7.1.1.$0^{1,6}$]undeca-2,4-dien-4-yl)-ethanone (Chart 1; Figure 1) and toluene (Chart 1; Figure 2). The vertex priorities of the present algorithm have been compared with those of Morgan's algorithm.

3 RESULTS AND DISCUSSION

The vertex prioritization, according to Morgan's as well as the present algorithm, of the hydrogen suppressed chemical graphs of 1-(6-methyl-8-aza-tricyclo[7.1.1.$0^{1,6}$]undeca-2,4-dien-4-yl)-ethanone (Figure 1) and toluene (Figure 2) have been shown in Chart 1.

Chemical Formula Representation

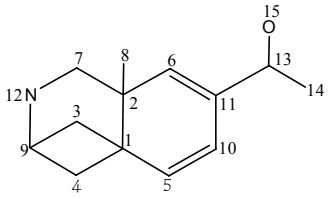

Figure: 1a

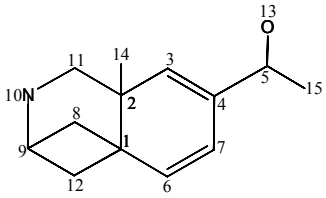

Figure: 1b

Graphical Representation

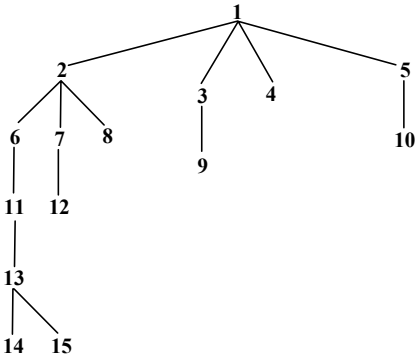

Figure: 1c

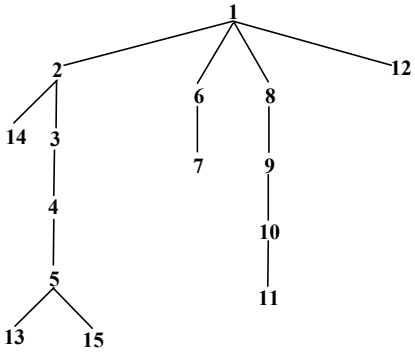

Figure: 1d

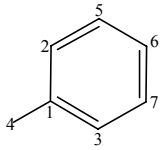

Figure: 2a

Morgan's Algorithm

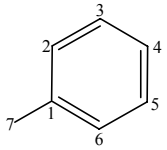

Figure: 2b

Present Algorithm

Chart 1: A comparison of vertex prioritization of chemical graphs in Morgan's algorithm and the algorithm presented here.

In 1-(6-methyl-8-aza-tricyclo[7.1.1.$0^{1,6}$]undeca-2,4-dien-4-yl)-ethanone (Chart 1; Figure 1), in terms of core vertices, a vertex path **1-2-3-4-5** of present algorithm (Figure 1d) is identical with that of Morgan's vertex path **1-2-6-11-13** (Figure 1c)**.** In Morgan's algorithm, the oxygen of the ethanone fragment (Figure 1a) has been sequenced as vertex **15** and its methyl carbon as vertex **14.** As due consideration has been given to the atom types in the present algorithm, the oxygen of the

ethanone fragment (Figure 1b) gets higher priority (vertex **13**) over the methyl carbon (vertex **15**). Also in the present algorithm the methyl carbon (vertex **14**) at the ring junction (vertex **2**) and methyl carbon of ethanone fragment (vertex **15**) (Figure 1b) have been demarcated with distinct priorities. The vertex path **1-8-9-10-11** of 1-(6-methyl-8-aza-tricyclo[7.1.1.$0^{1,6}$]undeca-2,4-dien-4-yl)-ethanone in figure 1d is characteristic to the present algorithm. In Morgan's algorithm vertices corresponding to this path have been distributed in two fragments namely vertex path **1-3-9** and vertex path **1-2-7-12**. Also, an examination of figures 1c and 1d in Chart 1 indicate that the present algorithm leads to less segmented (or less fragmented) and more compact graphs. Similarly in toluene, in Morgan's algorithm the methyl carbon of toluene has been sequenced as vertex **4** much before the other carbons of phenyl ring (Figure 2a). In the present algorithm, this methyl carbon has been sequenced in the end as vertex **7** (Figure 2b). In both the procedures, Morgan's and the algorithm presented here, the vertex with maximum connectivity gets the highest priority. In Morgan's algorithm, irrespective the vertex position – core or terminal (or peripheral) – in the graph, all the connected vertices of the just prioritized vertex will be sequenced simultaneously. However, in the present algorithm the vertices of the graph are at first demarcated in terms of core vertices and terminal vertices. The prioritization of the terminal vertices will be addressed after sequencing the core vertices. This provides an easy handle to study the core and terminal vertices. Moreover, this approach provides a kind of multilayered connectivity graphs which can be put to use in comparing two or more structures or parts thereof for any given purpose. Also, the algorithm allows the progress of vertex sequencing in specific direction till exhausting all the core vertices of the subgraph under consideration. This facilitates the identification of subgraphs with the maximum lengths and allows compact representation of the graphs.

4 CONCLUSIONS

In brief, the present algorithm provides an efficient and easy handle to examine the terminal as well as core vertices of given chemical graphs. The sequential identification of vertices of subgraphs facilitates the formation of connectional tables. This can be used for the computation of characteristic matrices and topological indices. Also, it finds application in storing, sorting and retrieving of chemical structures and databases. As this typically demarcates the core and terminal vertices in graphs, it can be put an easy use in comparing two or more structures or parts thereof for any given purpose.

5 REFERENCES

[1] H.L. Morgan, Generation of a Unique Machine Description for Chemical Structure – A Technique Developed at Chemical Abstract Service, *J. Chem. Doc.* **1965,** *5*, 107-113.

[2] W. T. Wipke and T. M. Dyott, Stereochemically Unique Naming Algorithm, *J. Am. Chem. Soc.* **1974,** *96*, 4834-4842.

[3] W.J. Wiswesser, *A Line Formula Chemical Notation*, T.Y. Crowell Comp., New York, 1954.
[4] A. T. Balaban, O. Mekenyan and D. Bonchev, Unique Description of Chemical Structures Based on Hierarchically Ordered Extended Connectivities (HOC Procedures). I. Algorithms for Finding Graph Orbits and Canonical Numbering of Atoms, *J. Comput. Chem.* **1985**, *6*, 538-551.

Biographies

Dr. Y.S. Prabhakar is assistant director in the Medicinal Chemistry Division, Central Drug Research Institute, Lucknow, India. After obtaining a Ph.D. degree in chemistry and theoretical aspects of drug design from the Birla Institute of Technology and Science, Pilani, India, Dr. Prabhakar undertook postdoctoral research in ethno biology with Professor P. Pushpangadan at the Regional Research Laboratory, Jammu. Dr. Prabhakar's current research interests include development of newer QSAR software and ways of addressing tetrahedral centers in modeling studies.