A Simple Algorithm for Unique Representation of Chemical Structures—Cyclic/Acyclic Functionalized Achiral Molecules

Yenamandra S. Prabhakar[†] and Krishnan Balasubramanian*[,‡,§]

Medicinal and Process Chemistry Division, Central Drug Research Institute, Lucknow-226 001 (U.P.), India, Chemistry and Material Science Directorate, University of California, Lawrence Livermore National Laboratory, Livermore, California 94550, and Department of Mathematics and Computer Science, California State University, East Bay, Hayward, California 94542-3092

Received March 27, 2005

An algorithm, based on "vertex priority values" has been proposed to uniquely sequence and represent connectivity matrices of chemical structures of cyclic/acyclic functionalized achiral hydrocarbons and their derivatives. In this method, "vertex priority values" have been assigned in terms of atomic weights, subgraph lengths, loops, and heteroatom contents. Subsequently, the terminal vertices have been considered upon completing the sequencing of the core vertices. This approach provides a multilayered connectivity graph, which can be put to use in comparing two or more structures or parts thereof for any given purpose. Furthermore, the basic vertex connection tables generated here are useful in the computation of characteristic matrices/topological indices and automorphism groups and in storing, sorting, and retrieving chemical structures from databases.

1. INTRODUCTION

Graph isomorphism and vertex-based graph automorphism-partitioning problems have a long history in both the mathematical and chemical literature.[1−15] In graph theoretical connotation, a chemical structure is a collection of vertices (atoms) connected by means of edges (bonds) in a predefined fashion. This has a diverse role in the elucidation and comprehension of various physical, chemical, and biological behavioral aspects of chemical entities. Quantum chemistry, spectroscopy, molecular symmetry, and physical/biological property predictions are a few aspects, among many others, that revolve around graph theoretical concepts of chemical structure.[1−7] In all of the graph theoretical approaches, partitioning of vertices is crucial, and this gives rise to characteristic applications to the defined vertex-partitioning procedure. Among many graph theoretical applications, Balasubramanian and co-workers[4,5] have demonstrated the usefulness of automorphism partitioning of vertices in spectroscopy and quantum chemistry. Moreover, vertex-partitioning procedures have a significant role in chemical database operations. Most of the chemical database operations involve sequenced chemical structures. In the absence of any priority (for a vertex), a chemical graph with **n** vertices can be sequenced in **n!** (factorial **n**) ways. If the graph has no symmetry, then each one of these **n!** representations will be different from the rest. In this environment, sequencing of the chemical graph on the basis of a preset vertex priority offers a unique advantage to many complex problems such as similarity searches, and so forth. There are a number of methods for unique sequencing of chemical graphs, for example, Morgan's algorithm,[8] Wipke and Dyott's algorithm,[1] Wiswesser line notation (WLN),[9] Balaban's hierarchically ordered extended connectivities (HOC) procedure,[10] Randić's canonical labeling method,[11] and so forth. Many of these techniques, while they are very useful, have difficulties with highly transitive graphs. In one of the oldest algorithms, known as the Morgan algorithm, the current atom is the one with the highest extended connectivity (EC) value, and if there are any attachments to the current atom that have not been assigned sequence numbers, then they are assigned sequence numbers in a decreasing order of EC values of the attachment, which includes terminal attachments (EC value is 1) even before other atoms with higher EC values. Here, we present a rule-based algorithm to partition the vertices of chemical graphs with an approach that we call "inside-in and outside-out". This results in a unique sequence for 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. In this method, the progress of sequencing of the core vertices take place along the uninterrupted vertex paths. Also, this sequencing procedure generates a traditional connection table of the chemical graph and finds application in different chemical-graph-related operations. The method can also be extended to generate the automorphisms of a given structure or a graph.

2. METHOD

In a chemical graph, (a hydrogen-suppressed chemical structure), it is well-known that 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, edges represent either sigma bonds or "sigma + pi" bonds. An algorithm has been designed to sequentially prioritize the vertices of chemical graphs in a hierarchical manner on the

* Corresponding author e-mail: balu@mcs.csuhayward.edu.
† Central Drug Research Institute, Lucknow, India.
‡ University of California, Lawrence Livermore National Laboratory.
§ California State University.

10.1021/ci050096m CCC: $33.50
Published on Web 07/12/2005

basis of the connectivity values and several other associated characteristics, such as atomic weights, subgraph lengths, loops, and heteroatom content. A loop in this algorithm represents a cyclic system identified (or encountered) in the direction of vertex sequencing or propagation. Here, the vertices are sequenced in a decreasing order of priority; that is, a vertex with a higher priority will be addressed and labeled first. In this procedure, the connectivity values (numbers of sigma bonds as well as "sigma + pi" bonds) of all vertices will be computed and used for prioritization. On the basis of 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 2, and the other group corresponds to vertices with connectivity values equal to unity. The vertices with connectivity values more than or equal to 2 form the core of the graph and will be prioritized successively on the basis of 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 because of the completion of prioritization of all vertices with connectivity values more than or equal to 2 in that "direction". After fixing the priorities of vertices with connectivity values more than or equal to 2 (core vertices), the priorities of vertices with unit connectivity values (terminal vertices) will be fixed using the same priority rules of core vertices. The successive steps of the sequencing algorithm are as follows:

Step 1. Compute connectivity (Cn; sigma bond) of all vertices (Vt's; non-hydrogen atoms) in a given hydrogen-suppressed chemical graph. Segregate all Vt's into two groups, one with Cn values more than or equal to 2 and the other with Cn values equal to one. Consider all Vt's with Cn values of 2 or more as competing Vt's.

Step 2. Find the number of Vt's with the highest Cn (sigma bonds only). If there is only one Vt, then go to step 13.

Step 3. Find the number of Vt's with the highest Cn (sigma + pi bonds). If there is only one Vt, then go to step 13.

Step 4. Find the atomic weights (Wt's) of Vt's with the highest Cn (sigma + pi bonds). If there is only one Vt, then go to step 13.

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

Step 6. Find the maximum lengths of the Fr's. If there is only one Fr, then go to step 13.

Step 7. Find the highest number of loops in the Fr's with the maximum length. If there is only one Fr, then go to step 13.

Step 8. Among the Fr's with the maximum length and highest number of loops, find the maximum chain length with the competing Vt. If there is only one Fr, then go to step 13.

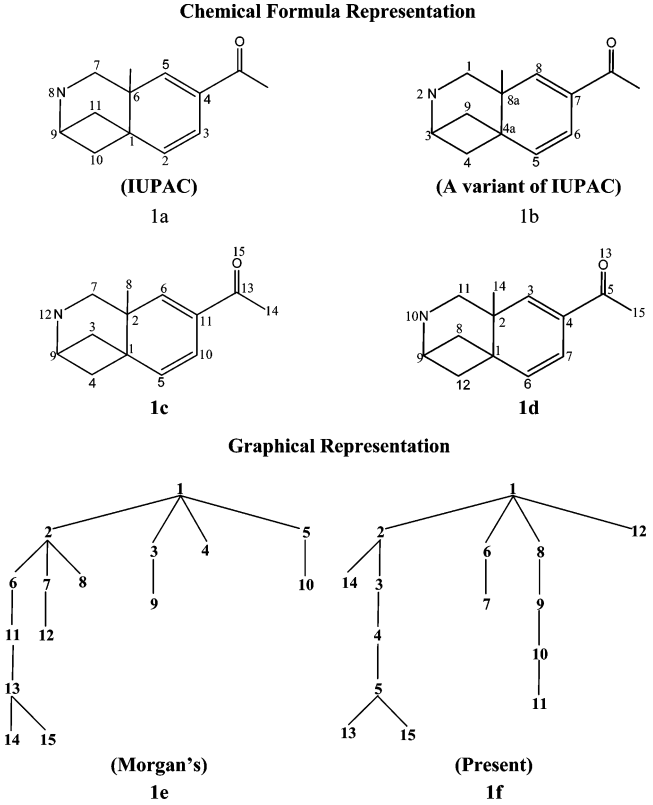

Figure 1. Vertex prioritization of 1-(6-methyl-8-aza-tricyclo-[7.1.1.0^{1,6}]undeca-2,4-dien-4-yl)ethanone in the IUPAC system, Morgan's, and the present algorithms.

Step 9. Among the Fr's with the maximum length, highest number of loops, and maximum chain length with the competing Vt, find the maximum number of heteroatoms. If there is only one Fr, then go to step 13.

Step 10. Among the Fr's with the maximum length, highest number of loops, maximum chain length with the competing Vt, and highest number of heteroatoms, find the maximum weight of the Fr's. If there is only one Fr, then go to step 13.

Step 11. Compute the distance matrices for the Fr's with the maximum length and having the highest number of loops, maximum chain length with the competing Vt, highest number of heteroatoms, and highest Wt. Compare the distances between the competing Vt's and heteroatoms of each Fr. Find the Fr's with compactly connected competing Vt and heteroatoms. If there is only one Fr, then go to step 13.

Step 12. An element of symmetry exists. Arbitrarily consider one of the competing Vt's.

Step 13. Prioritize (label) the Vt as 1 (one) (subsequently with successive numbers). If all Vt's with Cn values more than or equal to 2 are prioritized, then go to step 14; otherwise, consider the Vt's connected to the just-prioritized vertex as competing Vt's after excluding the already-prioritized Vt's, if any, from the list, and go to step 2.

Step 14. Prioritize the Vt's with a Cn value of unity according to the priority set by the competing Vt (steps 3 and 4) and Cn of the immediate neighboring Vt.

Step 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 (Figure 1) and toluene (Figure 2). We

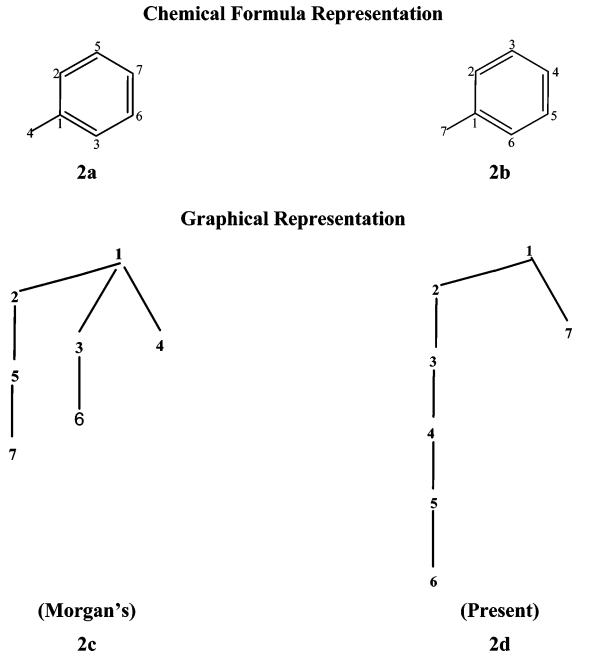

Figure 2. Vertex prioritization of toluene in Morgan's and the present algorithms

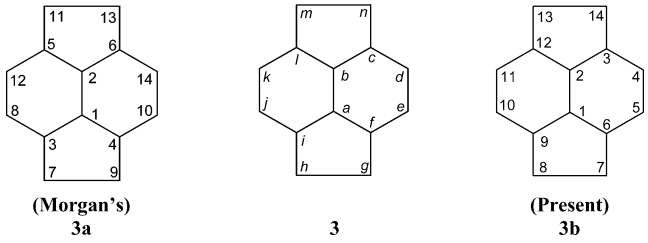

Figure 3. Vertex prioritization of tetradecahydrocyclopenta-[*fg*]acenaphthylene in Morgan's and the present algorithms.

have also considered a few more examples representing chemical graphs with a high degree of symmetry and loops such as tetradecahydrocyclopenta[*fg*]acenaphthylene (Figure 3), octadecahydrochrysene (Figure 4), octadecahydrobenzo-[*c*]phenanthrene (Figure 5), and cubane (Figure 6). For all the chemical graphs, the vertex priorities assigned by the present algorithm have been compared with those of Morgan's algorithm.

3. RESULTS AND DISCUSSION

In the IUPAC system of nomenclature for the compound 1-(6-methyl-8-aza-tricyclo[7.1.1.0^{1,6}]undeca-2,4-dien-4-yl)-ethanone (Figure 1), the priorities of molecular subunits will result in its expression as the ethanone system (the main frame) carrying the (6-methyl-8-aza-tricyclo[7.1.1.0^{1,6}]undeca-2,4-dien-4-yl) fragment on it. Accordingly, Figure 1a shows the numbering (priorities) of the various vertices of the 8-aza-tricyclo[7.1.1.0^{1,6}]undeca-2,4-dien-4-yl fragment. The same molecule, in a variant of IUPAC nomenclature procedure, may also be expressed as 1-(1,3,4,8a-tetrahydro-8a-methyl-2*H*-3,4a-methanoisoquinolin-7-yl)ethanone. Figure 1b shows the numbering of vertices of the 1,3,4,8a-tetrahydro-2*H*-3,4a-methanoisoquinolin-7-yl fragment of the same molecule. The IUPAC and other similar nomenclature procedures are based on a very high level of human visualization perceptions and comprehension. One motivation for the investigations into alternative procedures of vertex prioritization approaches is to embed the highest possible information into the graph system at the earliest possible level of vertex sequencing. With this in view, the present algorithm has been explained in comparison with Morgan's algorithm in sequencing the vertices of chemical graphs.

The vertex prioritizations, according to Morgan's as well as the present algorithm, of the hydrogen-suppressed chemical graph of 1-(6-methyl-8-aza-tricyclo[7.1.1.0^{1,6}]undeca-2,4-dien-4-yl)ethanone have been shown in the chemical formula representation form (Figure 1c,d) as well as in graphical representation form (Figure 1e,f). In this example, in terms of core vertices, a vertex path 1−2−3−4−5 of the present algorithm (Figure 1f) is identical to that of Morgan's vertex path 1−2−6−11−13 (Figure 1e). In Morgan's algorithm, the oxygen of the ethanone fragment (Figure 1c) 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 1d) 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 the methyl carbon of the ethanone fragment (vertex 15) (Figure 1d) 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 1f is characteristic

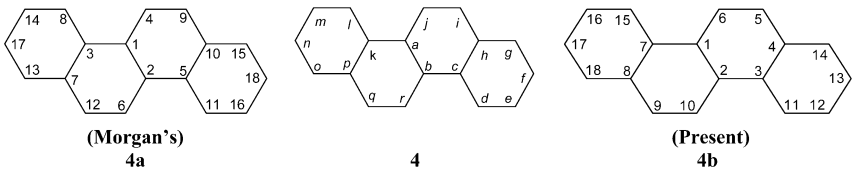

Figure 4. Vertex prioritization of octadecahydrochrysene in Morgan's and the present algorithms.

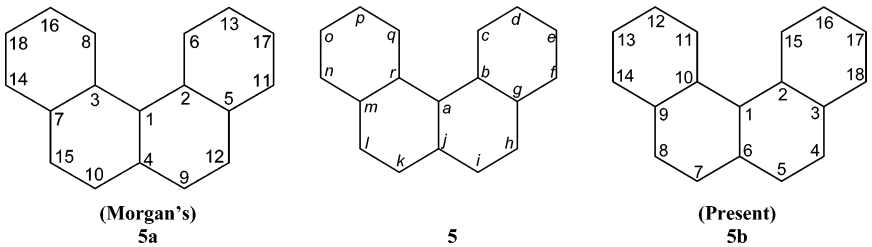

Figure 5. Vertex prioritization of octadecahydrobenzo[*c*]phenanthrene in Morgan's and the present algorithms.

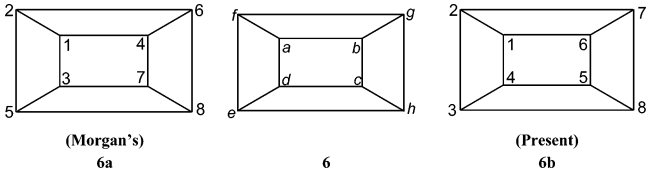

Figure 6. Vertex prioritization of cubane in Morgan's and the present algorithms.

of the present algorithm. In Morgan's algorithm, vertices corresponding to this path have been distributed in two fragments, namely, the vertex path 1−3−9 and the vertex path 1−2−7−12. Moreover, further scrutiny of Figure 1e,f indicates 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 the phenyl ring (Figure 2a). In the present algorithm, this methyl carbon has been sequenced as the end as vertex 7 (Figure 2b). Also, the graph of toluene is less segmented in the present algorithm (Figure 2d) when compared to that of Morgan's algorithm (Figure 2c). For the purpose of brevity, the remaining illustrations of polycyclic hydrocarbons have been limited to the structures of the chemical formulas. In these cases, the figures labeled **a** show the priorities of the vertices according to Morgan's algorithm and those labeled **b** correspond to the present algorithm. These examples have a varying degree of symmetry embedded in the structures. Thus, the vertex prioritization, in these cases, has been explained with the help of alphabet-labeled (in italics) chemical graphs.

In tetradecahydrocyclopenta[*fg*]acenaphthylene (Figure 3), because of the symmetry in the graph, the vertices *a* and *b* have equal priority. Arbitrarily, vertex *a* has been assigned as 1 and *b* has automatically become vertex 2. Again, for the same reason of symmetry in the system, the next competing vertices *c* and *l* have equal priority to take the position 3. In the arbitration, vertex *c* has been assigned as 3. The vertices competing for prioritization in the fourth position are *d* and *n*. Here, *d* gets priority over *n* because of the fact that the smallest loop involving *d* is a larger ring (six-membered ring) than that of its competing vertex *n* (five-membered ring). In this molecule, the priority of *d* over *n* is a clear case of the participation of loop size in the decision making process. In the vertex propagation direction of *d*, after two vertices (*e* and *f*), the propagation encounters a vertex *a* that is already a prioritized vertex, and thereby, the existence of a loop (six-membered ring) becomes relevant. Likewise, prioritization of vertices *m* and *l* is also determined by the encounter of prioritized vertex *b* and, thus, the existence of a loop (in this direction, that is a five-membered ring). As a six-membered loop is larger than one of five, the vertex that is part of the former one will get a higher priority. Hence, *d* has been labeled as 4. The prioritization of the remaining vertices proceeds in the same direction until the final vertex is reached.

As a result of symmetry in octadecahydrochrysene, shown in Figure 4, both *a* and *b* have equal priorities. Hence, arbitrarily, *a* has been assigned as 1 and *b* has been assigned as 2. After this, *c* and *r* become the competing vertices; as *c* is more heavily substituted (when compared to *r*), it has been assigned as 3. The next competing vertices are *d* and *h*; because of the heavier substitution, *h* becomes vertex 4. Now, the competing vertices are *i* and *g*. The vertex propagation in the *i* direction encounters an early branching compared to that of the *g* chain direction. Accordingly, *i* has been assigned the label 5. The succeeding end vertex *j* is given as 6. At this stage, the propagation traverses back to the vertex with next highest priority, which is, in this case, *k*, and gets assigned as 7. For the same reasons as discussed above, vertices *p*, *q*, and *r* become 8, 9, and 10, respectively. This leads to the second end point. Now, the vertices to be prioritized stem from *c* (3), *h* (4), *k* (7), and *p* (8). Being equal on all counts, the new vertex prioritization starts adjacent to the vertex with highest priority (lowest vertex number), that is, *c* (3), and proceeds through *d* (11) to *g* (14). This leads to the third end point. The remaining vertices *l*−*o* are prioritized as 15−18, respectively, in the same manner. Interestingly, in this case, the two inner rings remain inside, flanked by the peripheral rings.

Next, we consider octadecahydrobenzo[*c*]phenanthrene, shown in Figure 5, where vertex *a* is the vertex with highest priority and automatically has been assigned as 1. The vertices *b* and *r* attached to *a* have the same priority. Arbitrarily, vertex *b* has been assigned as 2. After this, between the competing vertices *c* and *g*, *g* gets priority over *c* because of the former's "heaviness"; hence, it is labeled as 3. The next available vertices are *f* and *h*. The vertex of choice is *h* as, in this direction, an early heavier substitution is encountered, and thus, *h* is assigned as 4. The vertex propagation proceeds in the same direction, as shown in Figure 5b, until reaching the first end point after prioritizing the vertex *n* as 14. In the backward integration process, the vertices *c*−*f* will be prioritized as 15−18, respectively. In this case, only one end point has been encountered before complete prioritization of all the vertices. Analogous to the previous case, the inside rings have been prioritized much before the outside rings.

The algorithm considered here offers the scope to prioritize even highly symmetric graphs such as cubane (Figure 6). In the first iteration, vertex *a* has been assigned as 1 because all vertices are equivalent. After fixing vertex 1, three vertices, *f*, *b*, and *d*, compete for the second position. In the second iteration, *f* has been assigned as vertex 2. In the third step, the vertex to be prioritized will be either *e* or *g*. Being equal in all respects, either of *e* or *g* can be assigned as vertex 3. In the present case, *e* has been assigned as 3. After fixing the third vertex, the remaining five vertices are amenable for prioritization without any prior arbitration. The vertices *d* and *h* await prioritization after vertex *e* (3). Now, the proximity of vertex *d* to the already-prioritized vertex *a* (1) (formation of a loop also) gives it a competitive edge over vertex *h*; accordingly, *d* has been assigned as vertex 4. Now, vertex *c* becomes the fifth one in the list. Similarly, the proximity of vertex *b* to the already-prioritized vertex *a* (1) gives it a competitive edge over vertex *h* (this is also near to vertex *e* (3) but of lowered priority); accordingly, *b* becomes vertex 6. The remaining two vertices, *g* and *h*, will take the positions 7 and 8, respectively, to complete the cubane graph.

In all procedures, Morgan's and the present algorithms, the vertex with maximum connectivity gets the highest priority. In Morgan's algorithm, irrespective of the vertex position, that is, the core or terminal (or peripheral) vertex

in the graph, all the connected vertices of the just-prioritized vertices will be sequenced together. However, in the algorithm described here, the vertices of the graph are, at first, demarcated in terms of core 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 multilayered connectivity graph, 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 a specific direction until exhausting all the core vertices of the subgraph under consideration. This facilitates the identification of subgraphs with maximum lengths and allows a compact representation.

4. CONCLUSIONS

The present algorithm provides an efficient and more convenient technique to examine the terminal as well as the core vertices of given chemical graphs. If the graph under consideration exhibits automorphic symmetry, this procedure arbitrarily considers one of the equivalent vertices as the vertex of choice for the propagation of the sequence of the graph. Consequently, equivalent vertices would have the same prioritization, paving the way for automatic generation of the automorphism group of the graph. The sequential identification of vertices of subgraphs facilitates the formation of connectivity tables. This can be used for the computation of characteristic matrices, polynomials, and several topological indices. Moreover, the method would find application in storing, sorting, and retrieving chemical structures and databases, as this typically demarcates the core and terminal vertices in graphs and it can be easily used in comparing two or more structures or parts thereof for any given purpose. Another important feature of our current algorithm is that it considers heteroatoms, loops in graphs, and atomids, thus, expanding its applicability to a wider spectrum of chemical graphs. Thus, we believe that the approach developed here has several advantages over other algorithms, including Morgan's algorithm. The applicability of the current algorithm to directed graphs might also be explored.

ACKNOWLEDGMENT

The research at California State, East Bay, was supported by the National Science Foundation under Grant No. CHE-0236434. The work at Lawrence Livermore National Laboratory was performed, in part, under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory, under Contract No. W-7405-Eng-48.

REFERENCES AND NOTES

(1) Wipke, W. T.; Dyott, T. M. Stereochemically unique naming algorithm. *J. Am. Chem. Soc.* **1974**, *96*, 4834−4842.

(2) Kier, L. B.; Hall, L. H. *Molecular connectivity in chemistry and drug research*; Academic Press, Inc.: New York, 1976.

(3) Randić, M.; Brissey, G. M.; Wilkins, C. W. Computer perception of topological symmetry via canonical numbering of atoms. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 52−59.

(4) Balasubramanian, K. Application of combinatorics and graph theory to spectroscopy and quantum chemistry. *Chem. Rev.* **1985**, *85*, 599−618.

(5) Liu, X.; Balasubramanian, K.; Munk, M. E. Computer-assisted graph-theoretical construction of ^{13}CNMR signal and intensity patterns. *J. Magn. Reson.* **1990**, *87*, 457−474.

(6) Liu, X.; Balasubramanian, K.; Munk, M. E. Computational techniques for vertex partitioning of graphs. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 263−269.

(7) Razinger, M.; Balasubramanian, K.; Munk, M. E. Graph Automorphism perception algorithms in computer-enhanced structure elucidation. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 197−201.

(8) Morgan, H. L. The generation of a unique machine description for chemical structure−A technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107−113.

(9) Wiswesser, W. J. *A line formula chemical notation*; T. Y. Crowell Co.: New York, 1954.

(10) Balaban, A. T.; Mekenyan, O.; Bonchev, D. 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.

(11) Randić, M. On canonical numbering of atoms in a molecule and graph isomorphism. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171−180.

(12) Herndon, W. C. In *Chemical applications of graph theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; Vol. 28, pp 231−242.

(13) Balasubramanian, K. Symmetry groups of chemical graphs. *Int. J. Quantum Chem.* **1982**, *21*, 411−418.

(14) Read, R. C.; Corneil, D. G. Graph isomorphism disease. *J. Graph. Theory* **1977**, *1*, 339−363.

(15) King, R. B. In *Chemical applications of graph Theory*; King, R. B., Ed.; Elsevier: Amsterdam, 1983; Vol. 28, pp 108−122.

CI050096M