An Explicit Representation of Molecular Geometry and Topology for Small Molecules

Jarosław Tomczak

Institute of Chemistry, University of Wrocław, F. Joliot-Curie 14, 50-383 Wrocław, Poland

Received August 17, 1994[⊗]

A novel method for the representation of the geometry and constitution of chemical structures is presented. It enables both molecular topology manipulation and geometry calculation. The concepts of weighted bond−valence (WBV) and weighted oriented bond−valence (WOBV) graphs are also introduced. An application to the representation of the elements of the inverse kinetic energy matrix is described as an example.

INTRODUCTION

A number of methods for implementing the representation of chemical structures has been proposed in the past; most of them concern the molecular topology[1−7] and conformations.[8] Unfortunately, usually they cannot store information about molecular geometry encoded in a compact way, which would be very useful for numerous computational problems, e.g., vibrational calculations or quantum chemistry. In the present work a novel specification of molecular structure is proposed. It allows storing and manipulating the molecular topology, on one hand, and access to the information about the internal coordinates, on the other hand. The method was primarily created for the purpose of the normal coordinate analysis, but it seems to have more general applications.

SOME GRAPH-THEORETICAL CONCEPTS

The molecular information, which must be stored, is classified into several groups:

(1) nature of atoms
(2) bond lengths
(3) values of valence and torsional angles
(4) constitution (topology)
(5) optionally, Cartesian coordinates of all atoms

Usually a molecule is regarded as a graph whose nodes and edges correspond to atoms and bonds, respectively. Since such an approach to the problem of chemical structure does not allow explicit coding of molecular geometry (particularly the information abot angles within a molecule), it is necessary to introduce some new ideas.

The concept will be presented in two steps, to make it more clear. First let a molecule be represented by an undirected graph, where each node describes one bond, and each weighted edge corresponds to one equilibrium angle and its value. Throughout this paper such a graph will be called the WBV (weighted bond−valence) representation. Using this concept one can easily describe molecular constitution together with two main groups of geometrical parameters (bond lengths and angles). An analogous representation can be obtained by dividing a given molecule into two-atom blocks and treating these blocks as nodes and cutpoints as edges of a certain graph. Unfortunately, the projection from a "normal" molecular graph (where atoms correspond to nodes and bonds to edges) to a WBV graph is not unequivocal, as shown, e.g., for the two graphs presented in Figure 1.

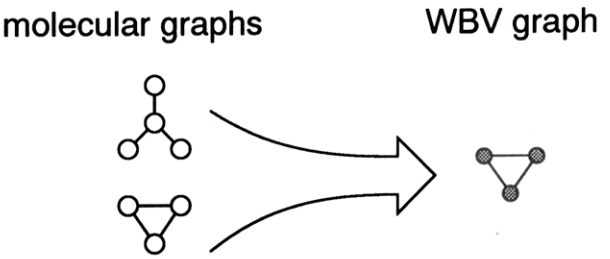

Figure 1. The ambiguous projection to the WBV graph.

To overcome this problem the concept of the WBV graph was extended, i.e., all the nodes (called below *external*) were expanded and oriented, meaning that they are no longer points, but a kind of internal two-node graphs made up by two *internal* nodes (atoms) and an *internal* edge (the real bond between two atoms). Furthermore, only the same *internal* nodes can be connected by the *external* edges. Thus each atom of a molecule is represented by as many *internal* nodes as many bonds it creates, and all these *internal* nodes have to be connected with one another by *external* edges. This graph is named further on the WOBV (weighted oriented bond−valence) graph. From the other point of view one can treat a WOBV graph as constructed from a classical molecular graph by expanding all nodes to code the information about angles related to them. The relationship between the molecular structure, the WBV and WOBV graphs for a sample molecule of methylamine, is shown in Figure 2.

It is possible to include also information about torsional angles into this scheme, but it complicates so much the calculations and algorithms that it is better kept in a separate list.

MACHINE REPRESENTATION

Due to the complexity of the WOBV graphs, the normal ways of representing molecular graphs (using different kinds of matrices) cannot be applied. In order to get a compact data structure describing a WOBV graph the idea of adjacency list was used as a starting point. It is constructed from a vector of records representing nodes, and each record contains a pointer to a one-direction list of the boundaries of the given node. Figure 3 presents the adjacency list for the WBV graph for methylamine. Each record of the vector **bonds** stores information about one bond and one pointer (represented by an arrow) to the list of its neighbors ("nil" for the last one). Such a record is called a header of the list. Remaining elements of the list are constructed from records, whose one field contains a bond number that is a

[⊗] Abstract published in *Advance ACS Abstracts,* February 15, 1995.

0095-2338/95/1635-0199$09.00/0

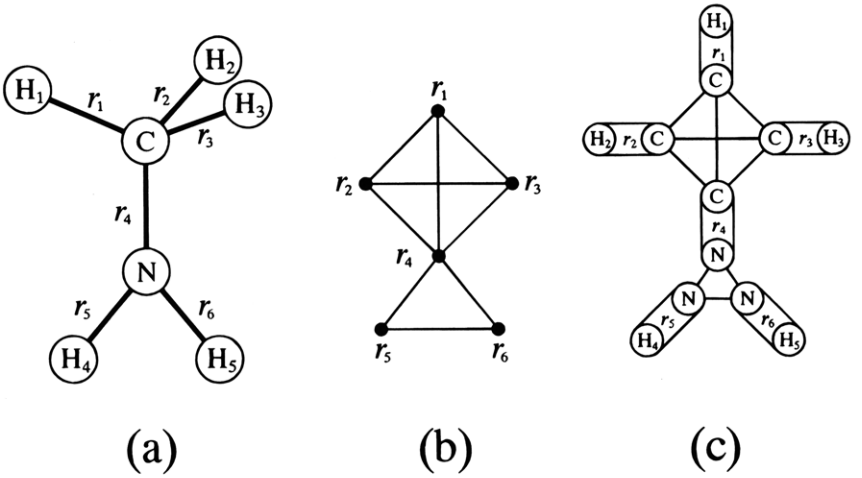

Figure 2. The methylamine molecule: (a) structure, (b) WBV graph, and (c) WOBV graph.

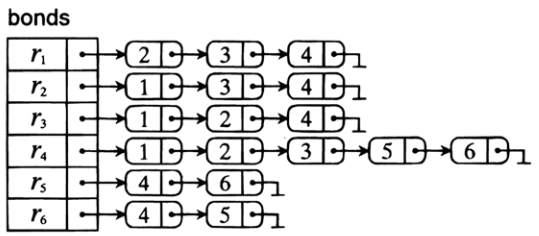

Figure 3. The adjacency list representing the WBV graph for methylamine.

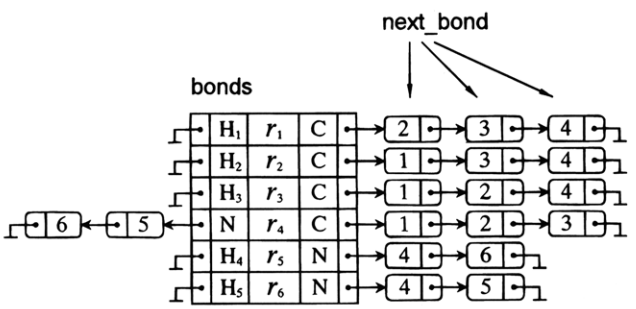

Figure 4. The doubled adjacency list for methylamine.

neighbor of the header node, and another field contains a pointer to the next element on the list.

In comparison with the representation based on matrices, the adjacency list represents a considerable advantage because it is more efficient when we look at time complexity for typical searching tasks. For example finding all the boundaries of a given node in the case of the edge adjacency matrix requires verification of all n elements of the appropriate row (column), where n is the total number of bonds in the molecule. For the adjacency list the required time is proportional to the number of neighbors of the given node, which is less or equal to n. Another advantage of the adjacency list is the fact that it requires less memory than the matrix representation for a huge set of molecules, whose connection tables are sparse.

This representation suffices in the case of WBV graphs, but it must be extended for the WOBV graphs.

For this purpose the adjacency list is doubled, i.e., each element of the vector is a record **one_bond** containing the bond's data:

(1) indexes of atoms that are bonded

(2) bond length

(3) two pointers to the lists of bonds that are neighbors and possess the same *internal* node (The lists are constructed from records called **next_bond**.)

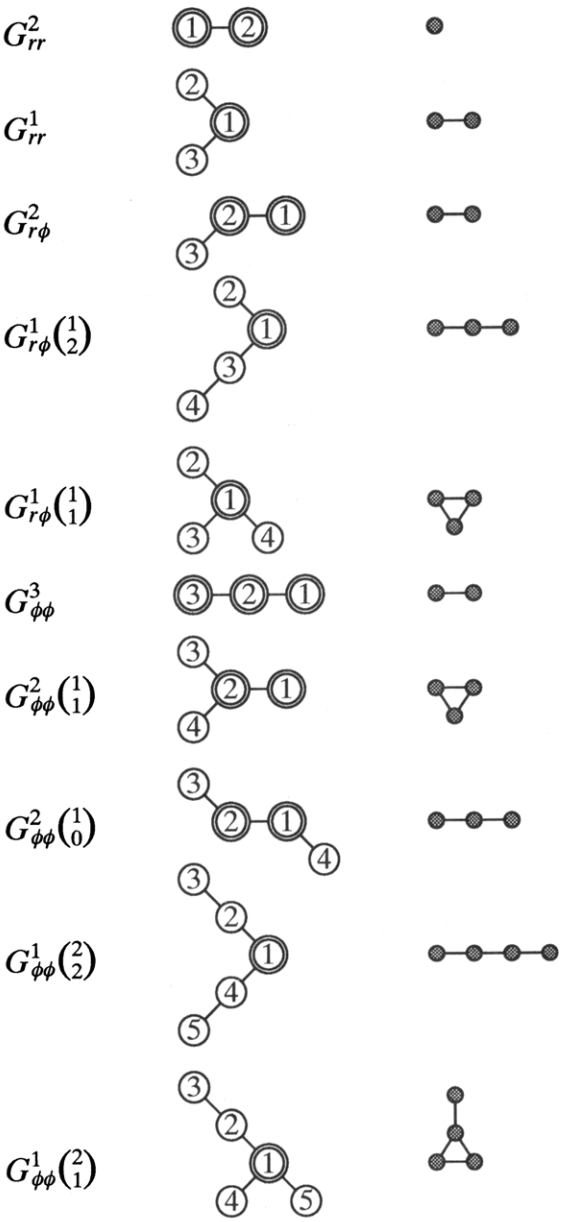

Figure 5. The WBV graphs for the **G** matrix elements.

Full information about atoms (their atomic weights, names, and optionally Cartesian coordinates) is stored in a separate list. The **next_bond** structure is similar to that of the WBV graph and contains only the index of a bond that is a neighbor

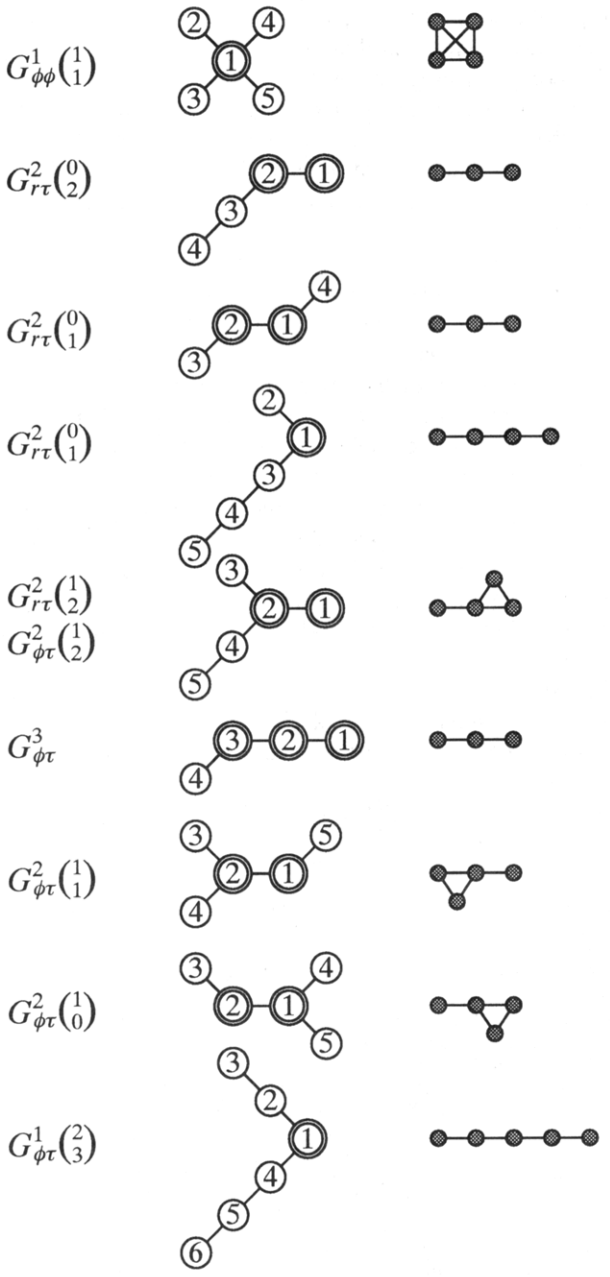

Figure 6.

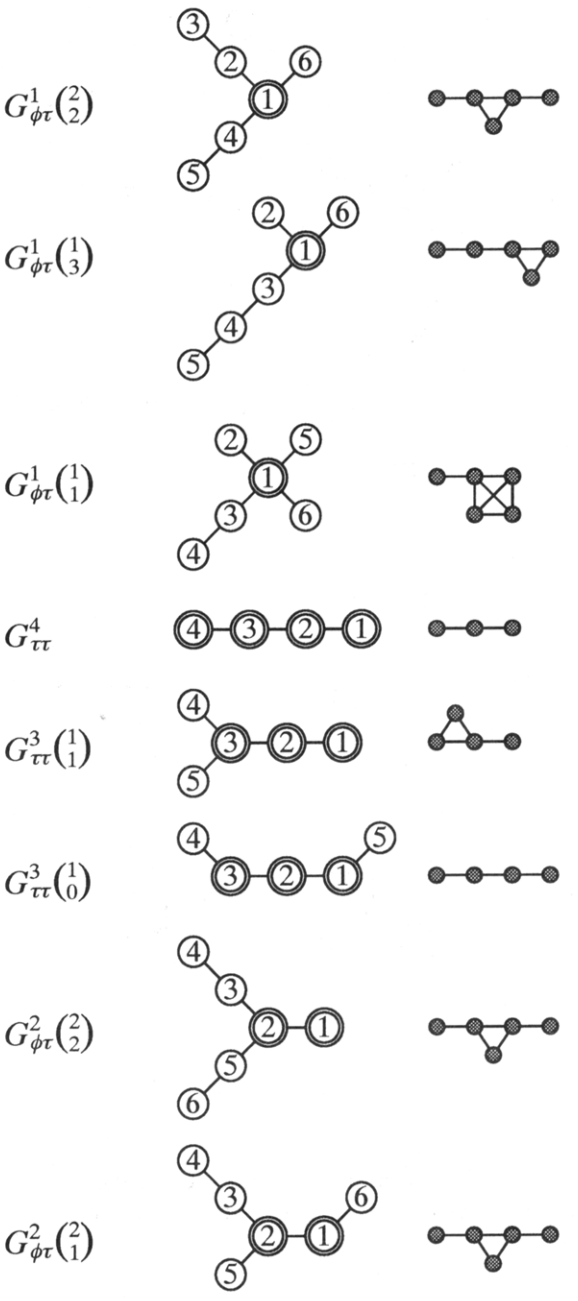

Figure 7.

of the bond encoded in the **one_bond** record and the value of angle between these two bonds. Additionally the **torsions** list supports information about the torsional angles in a given molecule in terms of bonds that are bonded. A simplified data structure for the example of methylamine is shown in Figure 4.

Both the WBV and WOBV graphs are undirected, a fact that leads to an inefficient usage of memory, since each angle is stored twice. However, on the other hand, the algorithms using doubled adjacency list data representation are faster in the case of undirected graphs. Actually the WBV and WOBV graphs can be also directed, but this requires small differences in the approach to algorithm construction.

REPRESENTATION OF G MATRIX ELEMENTS

The WOBV graph constructed as described above may be used to perform the calculation of **G**, the inverse kinetic energy matrix in the internal coordinate space. The knowledge of **G** is the basis for normal coordinates analysis, whose principles were given by Wilson[9,10] and Decius.[11] The complete theory was published by Wilson, Decius, and Cross.[12]

The vibrational kinetic energy can be easily expressed in the Cartesian or mass-weighted Cartesian coordinates, but internal coordinates provide the most physically significant set for use in describing the potential energy of the molecule. They can be expressed as a function of $3N$ Cartesian coordinates

$$S_n = S_n(X_1, X_2, ..., X_{3N}) \tag{1}$$

Their relationship can be obtained through a Taylor expansion about some initial configuration

$$S_n = S_n^0 = \sum_p \left(\frac{\partial S_n}{\partial X_p}\right)_0 \Delta X_p + \frac{1}{2}\sum_{p,t}\left(\frac{\partial^2 S_n}{\partial X_p \partial X_t}\right)_0 \Delta X_p \Delta X_t + \ldots \tag{2}$$

Using the first-order approximation one can write

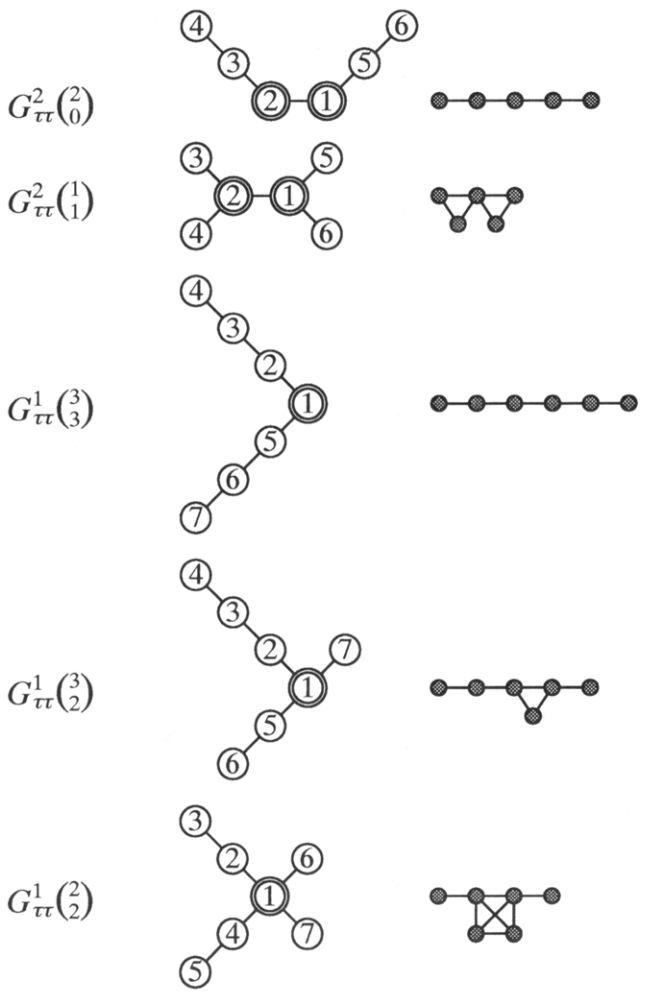

Figure 8.

$$S_n = \sum_t B_{tn}\Delta X_t \quad \text{where} \quad B_{tn} = \left(\frac{\partial S_n}{\partial X_t}\right)_0 \tag{3}$$

and ΔX_t are Cartesian displacements coordinates. The coefficients B_{tn} were used to define the **G** matrix, as shown in eq 4

$$G_{tt'} = \sum_{i=1}^{3N} \frac{1}{m_i} B_{ti}B_{t'i} \qquad t,t' = 1,2,...,3N-6 \tag{4}$$

The coefficients B_{ti} can be grouped into sets of three represented by a vector $s_{t\alpha}$ and then (3) and (4) take on respectively the simpler forms

$$S_t = \sum_{\alpha=1}^{N} s_{t\alpha} \cdot \varrho_\alpha \tag{5}$$

$$G_{tt'} = \sum_{\alpha=1}^{N} \frac{1}{m_\alpha} s_{t\alpha}s_{t'\alpha} \tag{6}$$

in which ϱ_α is a vector of Cartesian displacement coordinates for an atom α.

Based on the fundamental formula 6, Decius[11] has derived analytical expressions for $G_{tt'}$ for 33 possible acyclic configurations involving such internal coordinates as bond stretching, bending, and the torsional type.

The computer program for the calculation of **G**, written primarily by Schachtschneider,[13] evaluates the **B** matrix by using the direction unit vector method and needs both Cartesian coordinates of all atoms of a given molecule and information about the internal coordinates encoded in a special way.

The implementation of the graph-theoretical concepts outlined above enables a simplification of the representation of **G** matrix elements and of their evaluation. For this purpose the relations given by Decius (without intermediate calculation of the **B** matrix) were used. Each kind of the **G** matrix element can be easily represented as a WBV (or WOBV) graph. The complete set of **G** matrix elements with the corresponding graphs is given in Figures 5−8.

The calculations are performed separately for each kind of **G** matrix elements. First a WOBV graph, which describes a given molecule, is searched for all its subgraphs that are isomorphic with the graph representing the desired **G** element. If the subgraph has been found the calculations are performed. The program is written in the C++ language, and it uses extensively its object-oriented features. Several numerical examples and a detailed discussion of the efficiency of the procedure are presented in a separate paper.[14]

REFERENCES AND NOTES

(1) Rush, J. E. Status of Notation and Topological System and Potential Future Trends. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 202−210.
(2) Rouvray, D. H. In *Chemical Applications of Graph Theory*; Balaban, A. T., Ed.; Academic Press: London, 1976; Chapter 7, p 175.
(3) Nakayama, T.; Fujiwara, Y. Computer Representation of Generic Chemical Structures by an Extended Block-Cutpoint Tree. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 80−87.
(4) Tokizane, S.; Monjoh, T.; Chihara, H. Computer Storage and Retrieval of Generic Chemical Structures Using Structure Attributes. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 177−187.
(5) Barnard, J. M.; Lynch, M. F.; Welford, S. M. Computer Storage and Retrieval of Generic Structures in Chemical Patents. 4. An Extended Connection Table Representation for Generic Structures. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 160−164.
(6) Gillet, V. J.; Downs, G. M.; Ling, A.; Lynch, M. F.; Venkataram, P.; Wood, J. V.; Dethlefsen, W. Computer Storage and Retrieval of Generic Chemical Structures in Patents. Reduced Chemical Graphs and Their Applications in Generic Chemical Structure Retrieval. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 126−137.
(7) Ryhänen, T.; Bermejo, F. J.; Santoro, J.; Rico, M. Moltw: A Program For Conformational Studies Using Potential Functions−II. Algorithms For Molecular Coordinates And Topology Manipulation. *Comput. Chem.* **1987**, *11*, 13−18.
(8) Fella, A. L.; Nourse, J. G.; Smith, D. H. Conformation Specification of Chemical Structures in Computer Programs. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 43−47.
(9) Wilson, E. B. A Method of Obtaining the Expanded Secular Equation for the Vibrational Frequencies of a Molecule. *J. Chem. Phys.* **1939**, *7*, 1047−1052.
(10) Wilson, E. B. Some Mathematical Methods for the Study of Molecular Vibrations. *J. Chem. Phys.* **1941**, *9*, 76−84.
(11) Decius, J. C. A Tabulation of General Formulas for Inverse Kinetic Energy Matrix Elements in Acyclic Molecules. *J. Chem. Phys.* **1948**, *16*, 1025−1034.
(12) Wilson, E. B.; Decius, J. C.; Cross, P. *Molecular Vibrations*; Dover: New York, 1980. Republication of the original work of 1955 published by McGraw-Hill, New York.
(13) Schachtschneider, J. H. Technical Report no. 57−65, Shell Development Co., 1966.
(14) Tomczak, J.; Hawranek, J. P. An Object-Oriented Approach to the Calculation of the Inverse Kinetic Energy Matrix. *Computers Chem.*, in press.

CI940345Z