Graphical Approach for Defining Natural Internal Coordinates

CORY C. PYE, RAYMOND A. POIRIER

Department of Chemistry, Memorial University of Newfoundland, St. John's, Newfoundland, Canada

Received 21 February 1997; accepted 6 October 1997

ABSTRACT: A simple, customizable connectivity scheme is rigorously defined in which pairs of atoms are classified into three categories. The tools of graph theory are used to analyze the molecular graph and to efficiently find rings and ring assemblies through a combination of pruning and homeomorphic reduction. The definition of natural internal coordinates is extended in a nonredundant fashion for the various cases of weakly interacting components and for fused ring systems. The ring system coordinates were tested and found to be superior to *Z*-matrix coordinates. © 1998 John Wiley & Sons, Inc. J Comput Chem 19: 504–511, 1998

Keywords: geometry optimization; natural internal coordinates; connectivity; graph theory; ring systems; intermolecular interactions

Introduction

G eometry optimization is arguably the most important first step in any theoretical investigation of molecular properties. The cost of such a calculation depends primarily on the number of function evaluations (energy and gradient) required to achieve proper convergence. This, in turn, depends on the efficiency of the optimization algorithm, the accuracy of the initial geometry and

Correspondence to: C. C. Pye at present address: Chemistry Department, University of Calgary, Calgary, Alberta T2N 1N4, Canada

Contract/grant sponsor: Natural Sciences and Engineering Research Council of Canada

Hessian, the choice of coordinates, and to some extent, the accuracy of the function evaluation. For *ab initio* wave functions, the function evaluation is usually the most expensive part of the calculation.

The simplest choice of coordinates is Cartesian. These have the advantage that they are easily generated by model-building programs such as SPARTAN.¹ In addition, because energy derivatives are usually evaluated directly in Cartesian coordinates,^{2,3} no transformation to another coordinate system is required. Some disadvantages are that there is no intuitive way to generate an initial Hessian (second derivative) estimate, and the imposition of internal coordinate constraints is not straightforward.⁴⁻⁶ In addition, there are 6 degrees of freedom corresponding to overall translation and rotation, which have a zero force constant in

the absence of an external field, and are not relevant to a discussion of molecular structure. However, these can be projected out.⁷

Another popular coordinate system consists of internal Z-matrix coordinates. In these, the position of the atoms is determined by bond lengths, angles, and torsions, which parallels the way chemists tend to think about molecular structure. Z-matrices for small systems are usually easily constructed, and the imposition of symmetry constraints (sometimes by using dummy atoms) is generally straightforward. Because Z-matrix coordinates correspond closely to valence coordinates, initial diagonal Hessian approximations can easily be generated. A big disadvantage of Z-matrix coordinates is that there is at least one "missing" bond whenever rings are part of a structure, which introduces strong coupling between the parameters. One scheme for correcting these inherent difficulties in Z-matrix coordinates uses a combination of Cartesian and Z-matrix coordinates, where Cartesian coordinates are used for cyclic portions of the molecule and internal coordinates are used for the noncyclic portions, 8,9 whereas another scheme is used for constrained optimization.¹⁰ In addition, the assignment of initial force constants for coordinates involving dummy atoms has no physical basis.

An attractive choice of coordinates is the natural internal coordinates. 11-13 The features of these coordinates are the correspondence to molecular topology (all bond stretching motions are included) and the elimination of redundancy by using the concept of local symmetry about rings and nonring atoms. Because the quadratic force constant matrix in this coordinate system is diagonally dominant, the common use of a diagonal initial Hessian lends itself well with this technique. In addition, cubic and higher-order force constants are also likely to be well approximated by a diagonal tensor (or even ignored), and thus, for optimizations starting far from equilibrium, natural internal coordinates are the method of choice for minimization of organic¹⁴ and inorganic¹⁶ structures. However, for transition states, currently available natural internal coordinate generators have some drawbacks, because the assigning algorithm may leave out some very long bonds. 1

Another more recent coordinate choice is delocalized internal coordinates,¹⁷ in which a numerical scheme is used to decide which primitive internal coordinates mix to form the new coordinates. Redundant coordinates have been applied to minimum structures, ¹⁸ to transition states, ¹⁹ and to atomic clusters. ²⁰

Molecular Graph

As mentioned previously, one of the drawbacks of the automatic generation of natural internal coordinates is the possibility that for especially long bonds, the algorithm may fail to find a connection, and thus the generated coordinates may not span all of the molecular degrees of freedom. We chose to address this problem²¹ in a very fundamental fashion, using an efficient graph theoretical approach.

A molecule may be regarded as a graph, G, consisting of a nonempty set of vertices, $\mathscr{V} \neq \emptyset$ (atoms), and a distinct set, \mathcal{E} , of edges ("bonds"), where each element of & consists of an unordered pair, vw, of distinct elements, v, w of \mathcal{V} , as shown in Figure 1. Multiple occurences of the same vw (multiple edge) or edges incident with only one vertex vv (loops) are not permitted, but these are allowed in the more general object known as a pseudograph. Two vertices (or edges) sharing a common edge (or vertex) are said to be adjacent. An edge, vw, is *incident* with the vertices v and w. The number of edges incident with v is called the degree of v. In our scheme, the degree is less than or equal to the valence of an atom; for example, an sp³ carbon has degree 4, and an sp² carbon has degree 3. Multiple bonds are considered as one edge. A subgraph, G', of G is a graph whose vertex and edge sets are subsets of those of \mathcal{G} . To determine the edge set & from a given molecular

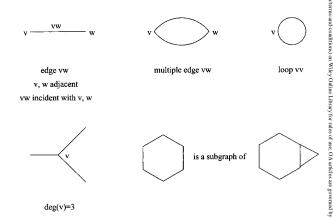


FIGURE 1. Illustration of some basic graph theoretical definitions.

structure, we use the formula:

$$C_{ij} = \left| \frac{\operatorname{sign} \left(d_{ij} - \lambda_{BS} \left(R_i^{BS} + R_j^{BS} \right) \right) - 1}{2} \right| \quad (1)$$

where C_{ij} is a binary element (0 or 1) of the adjacency matrix, or connection matrix, of the graph; d_{ij} is the distance between atoms i and j; $R_i^{\bar{B}S}$ and R_j^{BS} are the Bragg–Slater radii²² of atoms *i* and *j*; and λ_{BS} is a scale factor (default value = 1.2). Both the scale factor and the individual atomic radii can be customized by the user on input. With this definition one can find most bonds of a chemical structure, from which we can define angles and torsions. The default specification will not find extremely long "bonds" corresponding to transition vectors, hydrogen bonds, or van der Waals contacts. In general, the graph it generates will consist of several disjoint components corresponding to "fragments" of the given structure. This may consist of, for example, just one molecule (i.e., the graph is connected), a bimolecular transition state consisting of two (e.g., Diels-Alder) or three components (e.g., $S_N 2$), or a collection of molecules (e.g., lithium ion surrounded by eight water molecules). We use a very efficient graph theoretical algorithm for determining the number of these components.²³

For describing other connections such as hydrogen bonds and bonds corresponding to transition vectors, a more relaxed distance criteria is required. We use the formula:

$$C_{ij}^* = \left| \frac{\operatorname{sign} \left(d_{ij} - \lambda_{vdW} \left(R_i^{vdW} + R_j^{vdW} \right) \right) - 1}{2} \right| - C_{ij} - \max_{k} C_{ik} C_{kj}$$
 (2)

The matrix C^* is referred to as the *close contact matrix*. In this case, we use the van der Waals radii,²⁴ which are typically larger than the Bragg–Slater radii, and a different scale factor $\lambda_{\rm vdW}$ (default value = 0.8). Both the scale factor and the individual van der Waals radii can be customized by the user on input. Subtracting C_{ij} ensures that we do not count normal bonds twice. Atoms bonded to a mutual third atom usually possess overlapping van der Waals radii. Subtracting $\max_k C_{ik} C_{kj}$ ensures that these pairs are not counted.

For the case of extremely weakly bound species, it is desirable to have a fail-safe mechanism for connecting widely separated fragments. Suppose that the extended graph, consisting of the union of

the two edge sets as determined by eqs. (1) and (2), is not connected. We then find the closest two atoms belonging to different components (where the van der Waals radii have already been subtracted):

$$\min_{i,j; \operatorname{comp}(i) \neq \operatorname{comp}(j)} d_{ij} - \left(R_i^{\operatorname{vdW}} + R_j^{\operatorname{vdW}} \right) \quad (3)$$

and force a connection between *i* and *j*. This edge is added to the total edge set, the component mapping is updated, and the procedure is iterated until the graph is connected. This results in what we call the *intercomponent connection matrix*.

An example of the application of this procedure to the pentahydrate of lithium ion is given in Figure 2. The edge set of the adjacency matrix is represented by solid lines, the edge set of the close contact matrix by short-dashed lines, and the edge set of the intercomponent connection matrix by long-dashed lines.

Topological Analysis

Once a molecular graph, \mathscr{G} , has been constructed, several related graphs may be derived, as illustrated in Figure 3. A *pruned* subgraph of \mathscr{G} is

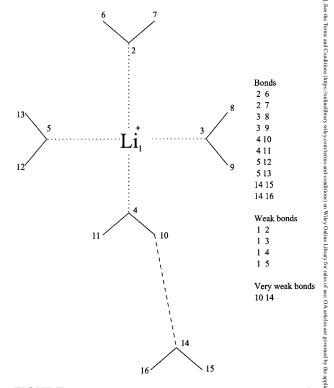


FIGURE 2. Application of connection analysis to LiAq₅⁺.

506 VOL. 19, NO. 5

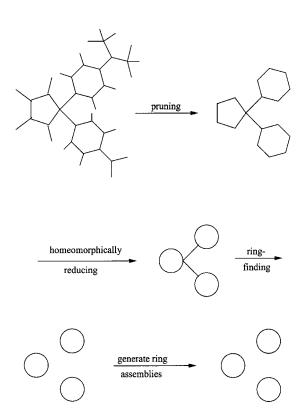


FIGURE 3. Application of topological analysis to Cp(PhNO₂)(PhNMe₂).

the graph remaining when vertices of degree 0 or 1 are recursively removed until no more remain. A homeomorphically reduced graph associated with $\mathscr G$ is a graph produced by recursively replacing vertices of degree two (and thus their incident edges) with an edge. It is clear that the two adjacent vertices must be distinct, otherwise a pseudograph with multiple edges may be produced (cf. Fig. 4). If this procedure is continued, the resulting object we define as the homeomorphically reduced pseudograph which contains multiple edges but no loops. Continuing once more, some vertices of degree two would be converted to a loop, which we call a minimal homeomorphically reduced pseudograph. If a loop is considered as adding two to the degree of a vertex, then for isolated loops the next iteration would annihilate the loop, resulting in a (possibly empty) object that we call the fully homeomorphically reduced pseudograph.

Other graph theoretical concepts that are important in characterizing the molecular topology are defined as follows:

 Walk—an alternating sequence of vertices and edges, starting and ending with vertices in which each object is incident to those

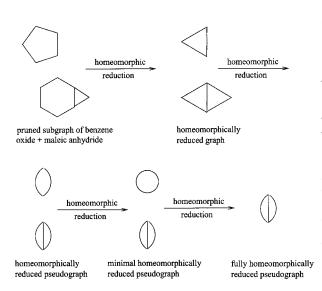


FIGURE 4. Various degrees of homeomorphic reduction.

immediately preceding and following it in the sequence.

- *Trail*—a walk with distinct edges.
- *Path*—a walk with distinct vertices.
- Closed walk—a walk in which the first and last vertices are equal.
- *Open walk*—a walk in which the first and last vertices are not equal.
- Circuit—a closed trail.
- Cycle—a circuit with distinct vertices (except the first and last). This is also known as a ring. A numerical prefix indicates the number of vertices.
- Ring assembly—a set of rings which cannot be subdivided into two or more disjoint edge sets, at least one of which is a ring.
- Separable graph—a connected graph for which the removal of one vertex (the cutvertex) disconnects the graph. The resulting fragments are called blocks.
- vertex connectivity of a connected graph—the minimum number of vertex removals required to disconnect the resulting subgraph.
- 1-isomorphic—two graphs \mathcal{G}_1 and \mathcal{G}_2 are 1-isomorphic if graphs \mathcal{G}'_1 \mathcal{G}'_2 , derived by splitting the cut-vertices between blocks, are isomorphic.
- 2-isomorphic—two graphs \mathcal{G}_1 and \mathcal{G}_2 are 2-isomorphic if there exists a cut-set of two or fewer cut-vertices for each graph, for which splitting the cut-vertices generates graphs \mathcal{G}'_1 and \mathcal{G}'_2 that are isomorphic.

These terms are useful in classifying chemical structures. Molecules with simple spiro centers are separable, with the cut-vertex being the spiro center. If there are two cut-vertices of degree three or more in a graph of vertex connectivity 2, and if they are adjacent, then we have a fused-ring system such as propellane or decalin. If these two vertices are not adjacent, then the atoms are called bridgeheads. It is clear that the minimal homeomorphically reduced pseudograph preserves the ring structure of the full molecule.

To find the ring systems, we use an algorithm similar to that of Matyska.²⁵ First, the molecular graph is pruned to remove all paths not terminated at both ends by a cycle. Second, from the pruned graph, the minimal homeomorphically reduced pseudograph is constructed. Once this graph is determined, a fundamental set of rings is constructed and the ring assemblies determined. We do not use the binary representation of Matyska because difficulties arose with the spanning tree construction if loops were present. Once the fundamental rings and ring assemblies are found, a mapping array from each edge to a ring assembly is constructed, with a map to zero indicating that the bond is not a member of any ring assembly. A pictorial representation of the application of the pruning, homeomorphic reduction, and ring finding procedure to a 5,5-disubstituted 1,3-cyclopentadiene [Cp(PhNO₂)(PhNMe₂)] is given in Figure 3.

Coordinate Generation

We have implemented the graph theoretical tools described above in our *ab initio* program, MUNGAUSS 1.0,²⁶ which was written according to the Open Structured Interfaceable Programming Environment (OSIPE).²⁷ A natural internal coordinate generation program for MUNGAUSS 1.0 was then written which uses these connectivity definitions and the suggestions of Pulay¹³ for acyclic structures and simple ring systems.

INTERNAL COORDINATES OF ACYCLIC STRUCTURES

All stretches (from the connectivity matrix) are defined as individual coordinates. Nonring bending modes are defined as symmetrized linear combinations about a given atom. Any valence bend that is totally enclosed in one ring assembly is not considered. However, the relative motion of spiro

rings are included in this category (rock, wag, twist) because the two bonds comprising each bending component belong to two different ring assemblies. Out-of-plane bending modes are also handled. Nonring torsional modes (with a central bond determined from the connectivity matrix) are defined as the sum of all torsions about this bond.

SIMPLE RING SYSTEMS

The deformation modes of simple rings are formed as linear combinations of the valence bends of the ring according to:

$$S_m^a = \sum_{k=1}^n \cos\left\{\frac{(k-1)m2\pi}{n}\right\} q_k$$

and:

$$S_m^b = \sum_{k=1}^n \sin\left\{\frac{(k-1)m2\pi}{n}\right\} q_k$$

where q_k is a ring bend or torsion, and $m = 2, ..., \lfloor n/2 \rfloor$. The first n - 3 bends and torsions are chosen from the above set.

WEAK CONNECTIONS

As discussed previously, weak connections such as hydrogen bonds and breaking/forming bonds in transition states may need special treatment. The components of the given molecular structure are connected by close contacts (weak bonds). If three or more components are connected by weak bonds to form a ring, as in Figure 5, then one of the weak bonds is excluded from further consideration. Weak bonds between atoms of the same component are likewise not considered. The stretching coordinates of the remaining weak bonds are then added.

We now consider the bending and torsion coordinates corresponding to the relative orientation of component pairs. If for each pair of components there is exactly one weak bond under consideration, then the five coordinates relating to their relative position/orientation are generated. If there are greater than three weak bonds between a pair of components, then the first three encountered are chosen and the rest dropped from further consideration. One could also select these three using another criteria such as those of shortest distance. For two weak bonds with noncoincident atoms, the coordinates recommended by Pulay¹³ are suggested, namely, the "fourring" deformation, and

508 VOL. 19, NO. 5

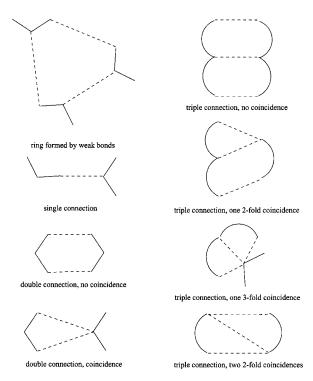


FIGURE 5. Weak connection arrangements.

the two "butterfly" coordinates. For coincident atoms on component A, we suggest the use of the butterfly motion of component B, and three deformation coordinates of A. For the case of three connections, we may have a threefold (3-fold) coincidence on A or B, a twofold (2-fold) coincidence on both A and B, a 2-fold coincidence on either A or B, or no coincidences. In these cases we would suggest:

- 3-fold A: B-twisting, B-deformation (2).
- **2** × **2-fold**: four-ring pucker, A, B-butterfly.
- 2-fold A: shearing along A, rocking A, twisting.
- **None**: twisting, shearing (2).

Only the single-connection deformations are currently automatically generated. The intercomponent connections (very weak bonds) are treated in the same manner as the single close contacts.

COMPLICATED RING ASSEMBLIES

More complicated ring assemblies pose special problems. Ring assemblies not composed of a single ring must have bridgehead atoms. If a pair of bridgehead atoms are adjacent, then it may be possible to decompose the ring assemblies further into subassemblies sharing a common edge (2-isomorphic to the original). When this can be done, the relative motion of the subassemblies can be described by valence parameters involving this edge or the incident atoms. Pulay treats the bicyclic and propellane cases separately, but these are simply subcases of two and three subassemblies, respectively. This also allows us to treat the subassemblies separately, combining them with the appropriate coordinate. Pulay only explicitly mentions the cases where the subassemblies themselves are simple rings, but this can clearly be extended to more general cases. The use of the torsion components for this mode (as recommended by Pulay) is possible, but one could also use symmetrized bending components as well, although more coupling to the subassembly deformation modes would be expected.

For more complicated cases than these, Pulay recommends the use of redundant coordinates. It would be convenient to construct topologically derived coordinates for any system, regardless of the molecular topology. Numerical approaches to generating coordinates (the "delocalized" internal coordinates¹⁷) look promising, although it is not clear how transferable the coordinates and, more importantly, the force constants would be from one system to the next. For bicyclic [n.m.k] systems the ring coordinates of the two rings with (n + k)+ 2) and (m + k + 2) atoms are used, which introduces redundancy. It is implied that one should use the two smallest rings, but there is no reason *a* priori to choose these two of the (redundant) set of three possible rings [the third containing (m + n)+ 2) atoms]. In addition, if m = k, then the choice becomes arbitrary and would destroy symmetry. Inspired by the aforementioned propellane coordinates, one could suppose that a bond exists between the two bridgeheads, and write the appropriate propellane distortions, or perhaps, as just suggested, write the propellane distortions in terms of bendings rather than torsions.

We consider bicyclo[1.1.1]pentane (C_5H_8) as an example (see Fig. 6). The 13 atoms require 33 internal coordinates, of which 14 are simple stretches. Each of the three CH_2 bridges requires a scissor, rock, wag, and twist (four coordinates) to describe the hydrogen deformation motions (total of 12), and each of the two bridgehead hydrogens requires a degenerate rocking mode (2) for a total of 4 and a total of 30 coordinates. The propellane-like distortions (which we call ring wag and scissor) are defined as appropriate bending combinations about the bridgeheads, and the final

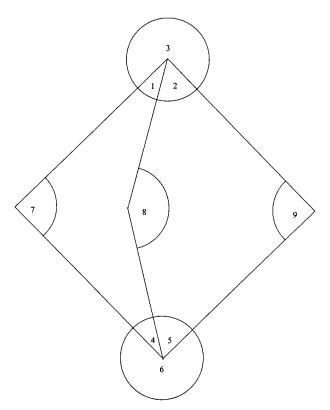


FIGURE 6. Bicyclo[1.1.1] pentane skeletal valence internal coordinates.

coordinate (ring "squish") as either the sum of the bridge bends or as an appropriate combination of all nine skeletal angles, as shown in Table I. The angles in Table I are defined in Figure 6. We note the remarkable coincidence between the coordinates used and the symmetrized linear combinations of the ring deformations. This suggests that some supersymmetric principle may be used to generate suitable coordinates whereby the symmetry about the bridgeheads is used on coordinates that are themselves symmetric combinations of ring deformations. Use of these coordinates were tested with Davidon's optimally conditioned (OC) optimization method²⁸ starting from a point 0.868 hartrees above the final geometry. The results were impressive. The *Z*-matrix optimization converged in 29 function evaluations, whereas optimization with natural internal coordinates converged in 11 function evaluations.

Conclusions

A simple, customizable connectivity scheme is rigorously defined in which pairs of atoms are classified into three categories that correlate roughly with bonds, weak bonds such as hydrogen bonds and transition bonds, and very weakly interacting atoms. The tools of graph theory can then be used to analyze the molecular graph and efficiently find rings and ring assemblies through a combination of pruning and homeomorphic reduction. The definition of natural internal coordinates can be extended for the various cases of weakly interacting components and for fused ring systems in a nonredundant fashion. For bicyclo[1.1.1]pentane, these new natural internal coordinates are shown to be superior to *Z*-matrix coordinates.

Description	Angle Coefficients								
	1	2	3	4	5	6	7	8	9
Ring A def.	1	0	0	1	0	0	-1	– 1	0
Ring B def.	0	1	0	0	1	0	0	-1	-1
Ring C def.	0	0	1	0	0	1	-1	0	-1
Ass. sciss	2	-1	-1	2	-1	-1	0	0	0
Ass. rock	0	1	-1	0	1	-1	0	0	0
Ass. sq. (1)	0	0	0	0	0	0	1	1	1
Ass. sq. (2)	1	1	1	1	1	1	-2	-2	-2
A + B + C	1	1	1	1	1	1	-2	-2	-2
2A - B - C	2	– 1	-1	2	– 1	– 1	-1	– 1	2
B-C	0	1	-1	0	1	-1	-1	1	0

510 VOL. 19, NO. 5

Acknowledgment

C. C. P. thanks Dr. P. Pulay for providing source code for comparison purposes.

References

- 1. Spartan V 4.1, Wavefunction, Inc., Irvine, CA (1996).
- 2. P. Pulay, Mol. Phys., 17, 197 (1969).
- 3. P. Pulay, In *Modern Theoretical Chemistry, Vol 4, Applications of Electronic Structure Theory*, H. F. Schaefer III, Ed., Plenum, New York, 1977, p. 153.
- D. Lu, M. Zhao, and D. G. Truhlar, J. Comput. Chem., 12, 376 (1991).
- 5. J. Baker, J. Comput. Chem., 13, 240 (1992).
- 6. J. Baker and D. Bergeron, J. Comput. Chem., 14, 1339 (1993).
- 7. J. Baker and W. J. Hehre, J. Comput. Chem., 12, 606 (1991).
- 8. H. Nakano, T. Nakajima, and S. Obara, *Chem. Phys. Lett.*, **177**, 458 (1991).
- H. B. Schlegel, Int. J. Quant. Chem. Quant. Chem. Symp., 26, 243 (1992).
- 10. J. D. Head, J. Comput. Chem., 11, 67 (1990).
- 11. P. Pulay, G. Fogarasi, F. Pang, and J. E. Boggs, J. Am. Chem. Soc., 101, 2550 (1979).
- G. Fogarasi and P. Pulay, In NATO ASI Series, Series C, Vol. 410, Structures and Conformations of Non-Rigid Molecules, J. Laane et al., Eds., Kluwer, Dordrecht, 1993, p. 375.
- G. Fogarasi, X. Zhou, P. W. Taylor, and P. Pulay, J. Am. Chem. Soc., 114, 8191 (1992).

- 14. J. Baker, J. Comput. Chem., 14, 1085 (1993).
- 15. J. Baker and F. Chan, J. Comput. Chem., 17, 888 (1996).
- 16. A. Bérces, J. Comput. Chem., 18, 45 (1997).
- J. Baker, A. Kessi, and B. Delley, J. Chem. Phys., 105, 192 (1996).
- 18. P. Pulay and G. Fogarasi, J. Chem. Phys., 96, 2856 (1992).
- C. Peng, P. Y. Ayala, H. B. Schlegel, and M. J. Frisch, J. Comput. Chem., 17, 49 (1996).
- 20. J. Baker and P. Pulay, J. Chem. Phys., 105, 11100 (1996).
- 21. C. C. Pye, Ph.D. Thesis, Memorial University of Newfoundland, St. John's, NF Canada, 1997.
- 22. The Bragg-Slater radii, which are defined for most atoms, are taken from J. E. Huheey, *Inorganic Chemistry—Principles of Structure and Reactivity, 2nd Ed.*, Harper & Row, New York, 1978. For the noble gases the estimated covalent radii are used.
- N. Deo, Graph Theory with Applications to Engineering and Computer Science, Prentice-Hall, Englewood Cliffs, NJ, 1974.
- 24. The van der Waals radii are taken from either J. E. Huheey, Inorganic Chemistry—Principles of Structure and Reactivity, 2nd Ed., Harper & Row, New York, 1978; L. Pauling, The Nature of the Chemical Bond, 3rd Ed., Cornell University Press, Ithaca, NY, 1960; or the "xpt" X-Windows program (J. P. Lord, 1993).
- 25. L. Matyska, J. Comput. Chem., 9, 455 (1988).
- R. A. Poirier, Y. Wang, and C. C. Pye, MUNGAUSS 1.0 (OSIPE Version), Chemistry Department, Memorial University of Newfoundland, St. John's, NF, Canada (1996).
- F. Colonna, L.-H. Jolly, R. A. Poirier, J. G. Ángyán, and G. Jansen, *Comput. Phys. Comm.*, 81, 293 (1994).
- 28. W. C. Davidon, Math. Prog., 9, 1 (1975).