Correct Bond Order Assignment in a Molecular Framework Using Integer Linear Programming with Application to Molecules Where Only Non-Hydrogen Atom Coordinates Are Available

Matheus Froeyen* and Piet Herdewijn

Laboratory for Medicinal Chemistry, Rega Institute, Catholic University of Leuven, Minderbroedersstraat 10, B-3000 Leuven, Belgium

Received November 26, 2004

We describe a method based on linear programming, for deducing the correct bond orders in small molecules, which only needs the sigma bonds connectivity and atom symbols as input. The procedure checks whether the current structure can be written as a valid Lewis structure by assigning double and triple bonds by translating the octet equations into an integer linear program that is solved by an external solver. The procedure was intensively tested on some ligands from the protein data bank as well as some other exotic molecules, for which first the hydrogen topology is generated by a standard procedure from literature. The most stable Lewis structure is retained for which hydrogen coordinates are generated.

INTRODUCTION

The Octet Rule and Lewis Drawings Revisited. The Lewis structure (electron dot structure) of a molecule or polyatomic ion shows how the valence electrons are arranged among the atoms in the molecule or the ion. In the first part of the previous century, scientists observed that chemical bonding can be explained by the valence electrons, i.e., the electrons of the atoms (except the hydrogens) in molecules have a tendency to attract eight valence electrons to create a filled s and p orbital.[1] Indeed, this is based on the chemist's experience: the most important requirement for the formation of a stable compound is that atoms achieve the noble gas electronic configuration by sharing electrons with other atoms, i.e., satisfying the octet rule. This rule can be applied to main group elements (except the elements in IA, IIA, and IIIA) in the second period. More formally, the octet rule says the following: main group elements (except elements in IA, IIA, and IIIA) in the second period lose, gain, or share electrons in such a way as to achieve an outer shell with eight electrons, ns^2np^6 that matches the nearest rare gas (noble gas) electronic configuration.[2]

For example carbon has an electron configuration $1s^2\ 2s^2\ 2p^2$ of which 4 electrons are in the valence shell (n=2). If bonded in a molecule, carbon will try to get 8 electrons, like neon, in the periodic system following noble gas. For instance the chlorine atom has 7 valence electrons, but we know that Cl^- is a stable ion having 8 electrons by making 4 covalent bonds. For hydrogen, H, in the first period the objective is getting 2 valence electrons such as helium.[1,2]

Lewis diagrams are a crude description of the electronic structure of a molecule. The best way to derive the 'correct' Lewis structure of a molecule would be to use the actual electronic density distribution calculated by a quantum chemical program and then use the Natural Bond Orbital method (NBO) for getting the bond orders which can be used to construct the Lewis diagram.[3] However this procedure would take a lot of time to calculate. The attentive reader may recognize that the problems discussed in this paper are related to a set of problems variously termed "kekulization"; that is, the change of an aromatic system depicted with a circle into a Kekulé form with an alternating path of single and double bonds.[4,5] The distance geometry program DGEOM95 is also able to assign automatically double bonds and aromatic bonds; however, this program needs as input an accurate set of starting coordinates.[6] Other methods have been published to assign correct bond orders based on the existing 3D coordinates.[7−11]

New Procedure To Generate Lewis Drawings. In the following paragraphs we propose a new procedure, based on linear programming, which only needs the sigma bonds connectivity and atom symbols as input. The procedure outputs a 'best' Lewis structure. Best does not always mean that this drawing represents the real electronic distribution of the molecule. In many cases the electronic configuration in a molecule needs to be represented by not one but several Lewis drawings. Also in the case of the elements phosphorus and sulfur we propose a way to generate drawings as found in recent peer reviewed scientific literature.

THE OCTET RULE METHOD IN EQUATIONS

In this paragraph the octet rule equations are presented (eqs 1−5) together with an additional restraint on the formal charge (eq 6).

Consider that we know all the atoms belonging to a molecule (or ion) and know their charges and also the bonding matrix, i.e., we know which atoms are bonded to which other atoms. The purpose of the procedure is to derive the bond order of all bonds (single, double, triple) and calculate the number of free electron pairs for each atom.

Count the total number of electrons V available in the system summing up the valence electrons of N atoms (V_i =

* Corresponding author phone: +32 16 33 73 79; fax: +32 16 33 73 40; e-mail: matheus.froeyen@rega.kuleuven.ac.be.

10.1021/ci049645z CCC: $30.25
Published on Web 06/17/2005

number of valence electrons in atom i and charge = the total charge of the molecule or ion):

$$V = \sum_{i}^{N} V_i - \text{charge} \tag{1}$$

Count the number of 'octet' electrons OCT: add an 8 for every non-H atom, a 2 for every hydrogen, $k_i \in \{2,8\}$.

$$\text{OCT} = \sum_{i}^{N} k_i \tag{2}$$

The total number of bond electrons B in the molecule is

$$B = \text{OCT} - V \tag{3}$$

Subtract the number of bond electrons B from the valence electrons V

$$F = V - B \tag{4}$$

resulting in the total number of electrons not participating in bonds. We limit the number F to even numbers, i.e., the electrons present in the free electron pairs, because no unpaired electrons are allowed in our program.

The octet rule for individual atoms restricts the number of electrons around every heavy atom to 8 (2 for a hydrogen)- with B_{ij} being the number of bond electrons to bound

$$\sum_{j} B_{ij} + F_i = k_i \tag{5}$$

neighbors and F_i being the number of electrons in the free electron pairs of atom i, and $k_i \in \{2,8\}$.

The formal charge FC_i for every atom i should be minimized in absolute value:

$$\text{FC}_i = V_i - (F_i + 0.5 \sum_{j} B_{ij}) \tag{6}$$

V_i is the number of valence electrons of atom i, and j runs over all bound neighbors of atom i.

TRANSLATION OF THE OCTET RULE INTO AN INTEGER LINEAR PROGRAM

General Rule. A linear programming problem can be formulated as the minimization (or maximization) of a linear function in N variables X_k

$$Z = \sum_{k}^{N} a_k X_k \tag{7}$$

subject to M constraints (i goes from 1 to M)

$$\sum_{j}^{N} d_{ij} X_j \{\leq, \geq, =\} b \tag{8}$$

The additional constraint

$$\text{integer } X_k \tag{9}$$

limits the variables to the set of integers and defines the problem as an integer problem. Implicitly in the used linear programming program, the integers are limited to positive values.

The octet rule procedure can be translated into an integer linear program. Let us define the integer variables X_{ij}, the number of electrons in every bond (single, double, triple) from atom i to atom j and X_{ii}, and the free electron pair electrons of every atom i. If atom i is not bonded to j, the $X_{ij} = 0$ is well defined and is not taken as a variable in the integer linear problem.

Three constraints are about the total number of electrons in the system. From eq 1 we know that the total number of electrons is limited to the number of valence electrons, minus the charge:

$$\sum_{i}^{N} \sum_{j}^{N} X_{ij} - V = 0 \tag{10}$$

The total number of bond electrons is limited by eq 3

$$\sum_{i}^{N} \sum_{j}^{N} X_{ij} - B = 0,\ i \neq j \tag{11}$$

and the total number of free electron pair electrons is constrained as well, as follows from eq 4

$$\sum_{i}^{N} X_{ii} - F = 0 \tag{12}$$

The octet rule for every atom i as given in eq 5 may be formulated

$$\sum_{j}^{N} X_{ij} + X_{ii} - k_i = 0 \tag{13}$$

with $k_i = 8$ for heavy atom, 2 for hydrogen. In this equation j runs over the atoms directly bonded to atom i. This gives us N additional equations for N atoms in the molecule. Additional bounds may be defined:

$$X_{ii} = 0 \text{ (hydrogen, carbon, ...)}$$

$$X_{ij} > 0 \text{ (for bonded atoms)}$$

$X_{ii} \leq 4, 6$

(a maximum of 2 free electron pairs on nitrogen atoms, 3 free electron pairs on oxygen)

$$X_{ij} = 2 \text{ (halogen or hydrogen bonded)} \tag{14}$$

The objective function to be minimized can be set pro forma to the sum of all variables

$$\sum_{i}^{N} \sum_{j}^{N} X_{ij} \tag{15}$$

Additional Restraints Needed: The Formal Charge. In the resonance form of nitromethane (Figure 1) the number of electrons around the oxygen with the single bond can be counted as having 3 free electron pairs and 1 electron from the ON bond (the other electron belongs to the N atom),

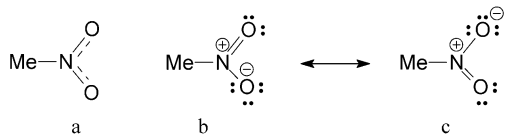

Figure 1. Nitromethane resonance forms.

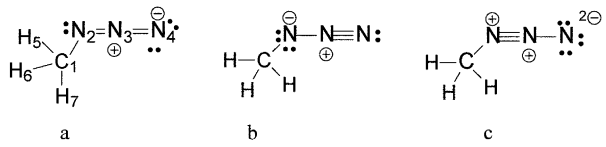

Figure 2. Lewis structures for methyl azide.

which adds up to 7. A neutral oxygen atom has 6 valence electrons, so the formal charge of this atom O is $6 - 7 = -1$. If several Lewis diagrams are available for one molecule, the one(s) where the atoms has(have) the lowest formal charge must be selected. An example run using lp_solve 4.0 [18] to solve the integer linear system for methyl azide may result in structure a, b, or c in Figure 2. All three structures obey the octet rule. Which structure is returned by the solver, implementing eq 7 to eq 15 depends on the sequence of equations in the lp_solve input file, i.e., is nondeterministic. However if we calculate the formal charges in these Lewis structures, structure 2c has nonzero formal charges +1, +1, and −2 at azide base nitrogen atoms N_2, N_3, and N_4, respectively. In fact, we did not apply yet the "minimal formal charge on each atom" rule. This rule changes our integer linear problem into a quadratic integer problem. We have to add the following terms to the objective function for every atom i

$$FC_i^2 = \sum_j (V_i - (0.5X_{ij} + X_{ii}))^2 \qquad (16)$$

in which j is a sum over all bound neighbors. In this way we get an optimization problem belonging to the class of integer quadratic problems. At the moment of this writing there exists no open source software to handle this kind of problem. The Lancelot package can handle nonlinear objective functions[19] but works with real variables, not with integers. When including the equations above (eqs 10−15) and minimizing the square of the formal charges (eq 16), Lancelot proposes nonphysical solutions. Indeed, a minimal solution exists for the set of equations with a noninteger bond and free electron pair charges. For example, we have solved the equations for methyl azide using the Lancelot package. The proposed optimal solution is $Z = +1.5$, $X_{0101} = 0$, $X_{0102} = 2$, $X_{0202} = 3$, $X_{0203} = 3$, $X_{0303} = 0$, $X_{0304} = 5$, $X_{0404} = 3$ (for the atom numbering see Figure 2, structure a; in this calculation eq 15 was omitted from the objective function), giving unpaired electrons in the bonds and free electron pairs, which is physically nonsense.

Therefore, to add the formal charge constraint, we propose the following procedure. Consider the square of the formal charge in eq 16 for the nitrogen atom N_3 in the adenine base. The solution of this equation is $FC_{N3}^2 = 1$, 0, and 0 for configurations a, b, and c in Figure 2, respectively. These squared formal charge equations for all heavy atoms in the objective function can be linearized using a similar procedure as in Fortet's linearization method for 0/1 variables. [20] Let us add an extra variable C_i to the objective function for every heavy atom i and add the following constraints

$$(Y_i - V_i) + C_i \geq 0 \qquad (17)$$

$$-(Y_i - V_i) + C_i \geq 0 \qquad (18)$$

with

$$Y_i = X_{ii} + \sum_j X_{ij} \qquad (19)$$

Table 1. Application of the Fortet Method for Converting the Quadratic Formal Charge Eq 16 To Be Optimized, into 2 Linear Inequalities[a]

	5	4	3	2	1	0
2	7/3	6/2	5/1	4/0	**3/−1**	**2/−2**
1	6/4	5/3	4/2	3/1	2/0	**1/−1**
0	5/5	4/4	3/3	2/2	1/1	0/0
−1	4/6	3/5	2/4	1/3	0/2	*−1/1*
−2	3/7	2/6	1/5	0/4	*−1/3*	*−2/2*

[a] Example given for deviations from the number of valence electrons dY= Y−V = 2, 1, 0, −1, −2. Equation 17 dY+C ≥ 0 is the left number in every field, eq 18 C-dY ≥ 0 is the number at the right. Code: boldface: allowed eq 17 only, italic: allowed eq 18 only, lightface roman: allowed both eqs 17 and 18. C is the column variable and dY is the row variable.

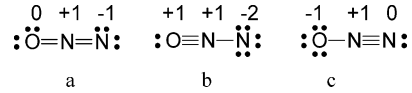

Figure 3. Lewis structures for N_2O.

Table 1 summarizes all the possible (allowed and forbidden) values that can be taken by C_i for the deviation values $Y_i - V_i$. Minimizing C_i will bring us to an optimal electron configuration. In the case of MeN_3 this leads us thus to configuration b or c ($Y_i - V_i = 0$, $C_i = 0$). This procedure will be applied only to all heavy atoms except from carbon (because of its straightforward electron configuration).

Expansion of the Octet Rule for S and P Elements. Often, Lewis drawings in published scientific articles containing third period elements such as phosphorus (P) or sulfur (S) do not follow the octet rule. Initially this was justified by the theory that P and S have empty d-orbitals.[12] For instance, the element phosphorus in phosphate was thought to have empty 3d orbitals available into which it can shuffle electrons resulting in one double bond and 3 single bonds, i.e., 10 valence electrons for P. However recent quantum chemical computational studies have convincingly shown that the octet rule is obeyed by all second-row main group elements in hypervalent molecules and ions by not neglecting the ionicity of the bonds.[13−16] To get drawings containing third period elements and violating the octet rule, often found in publications (Figure 4a),[21] we have to remove some restraints from our equalities/inequalities. Consider the phosphate group in a hexitol nucleotide[22] as shown in Figure 4. Structure 4b would be obtained when applying the strict octet rules on the PO_4 group. Experimentally and by quantum chemical computations, it has been observed that the terminal PO bonds have a 42−48% π bonding contribution.[23] In this view, structure 4c exists in 3 different electron configurations, one for each terminal oxygen. To get those drawings being made by the program, the octet rule must be made less strict, eqs 11 and 12 must be dropped and the octet equation for

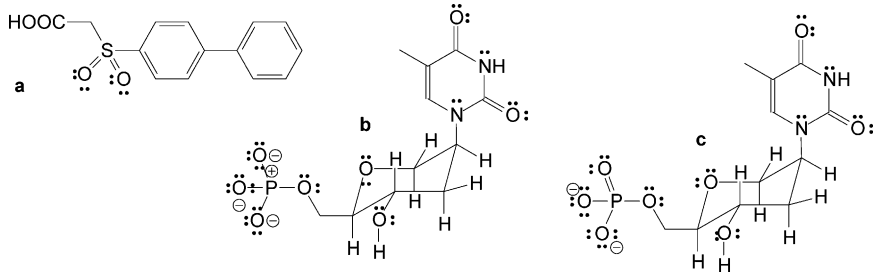

Figure 4. Lewis structures of (a) (biphenyl-4-sulfonyl)acetic acid, an inhibitor of the ATPase activity of human papillomavirus E1 helicase and (b,c) in the phosphate group of a hexitol nucleotide monophosphate.

Table 2. lp_solve Input and Results for Methyl Azide

```
min:
+x0101+x0102 +x0202+x0203 +x0303+x0304 +x0404
+ c0202+ c0303+ c0404;
/*all valence electrons equalities */
+x0101+x0102 +x0202+x0203 +x0303+x0304 +x0404 = 16;
/*all bonds electrons equalities */
+x0102 +x0203 +x0304 = 10;
/*all free electron pairs electrons equalities */
+x0101 +x0202 +x0303 +x0404 = 6;
/*all octets equalities */
+x0101+x0102 = 2;
+x0102+x0202+x0203 = 8;
+x0203+x0303+x0304 = 8;
+x0304+x0404 = 8;
/*all inequalities (bounds) */
x0102 >= 2;
x0203 >= 2;
x0304 >= 2;
/*additional (in)equalities C lone pairs */
x0101 <= 0;
x0202 <= 4;
x0303 <= 4;
x0404 <= 4;
/*fortet linearization */
/*first inequality */
- 0.50 x0102−1.00 x0202−0.50 x0203 + 5 +c0202>= 0;
- 0.50 x0203−1.00 x0303−0.50 x0304 + 5 +c0303>= 0;
- 0.50 x0304−1.00 x0404 + 5 +c0404>= 0;
/*second inequality */
+ 0.50 x0102+ 1.00 x0202 + 0.50 x0203−    5+c0202>= 0;
+ 0.50 x0203+ 1.00 x0303+ 0.50 x0304−    5+c0303>= 0;
+ 0.50 x0304+ 1.00 x0404−5+c0404>= 0;
/*integer definitions */
INT
x0101,x0102,x0202,c0202,x0203,x0303,c0303,x0304,x0404,c0404;

value of objective function: 18
actual values of the variables:
c0202   1   c0303   1   c0404   0   x0101   0   x0102   2
x0202   4   x0203   2   x0303   0   x0304   6   x0404   2
```

the atom P must be changed into two inequalities (in eq 22 replace 10 by 12 for the S atom.)

$$8 \le X_{ii} + \sum_j X_{ij} \tag{20}$$

$$X_{ii} + \sum_j X_{ij} \le 10 \tag{21}$$

Resonance Forms. For several molecules one can write more than one valid Lewis structure. Consider nitromethane in Figure 1a. All atoms together have a total of 24 unshared electron pair and shared bond electrons around them. By writing down its structure, we take care that every atom obeys the octet rule. For instance N has $4 \times 2 = 8$ bond electrons. But it has been seen experimentally (electron diffraction) that each NO bond has the same length (1.21 Å).[17] This molecule cannot be represented by only one Lewis structure but by two hybrid electronic structures b and c in Figure 1. The linear programming implementation of the octet rule returns only one but correct Lewis structure when more than one electron configurations are possible.

Adjacent Charge Rule. Consider the azido group in methyl azide (Figure 2). We can write three Lewis structures. The experimentally determined NN bond distances are 1.24 and 1.10 Å, respectively, which can only be explained by a combination of resonance forms 2a and 2b. The c form is excluded. This rule was called by Linus Pauling the adjacent charge rule.[17] Note however that the exclusion of structure c is taken care of by the 'minimal formal charge rule'.

Electronegativity Rule. This rule says that we should place the electrons or the lone pairs preferentially at those atoms exhibiting a large electronegativity. Let us consider the resonance forms of the N_2O molecule in Figure 3. Though both forms a and c have the same formal charge, Lewis structure c is more stable than structure a, because the negative formal charge is on the most electronegative atom. (The Pauling electronegativity of oxygen is 3.5 and of nitrogen is 3.0). If in the objective function the variables X_{ii} are divided by the electronegativity value of atom i, the correct solution is found.

Additional Restraints. It is very easy to add known additional chemical restrictions as restraints to the linear problem:

The restraint that electrons should be paired is added to the linear program (closed shell at the electron pair level).

If one is sure that a bond is double or single (by dihedral angle check) and not part of an aromatic system, its bond order is fixed as an additional restraint.

If the valence angle of an atom N is different from 180, i.e., not linear (for instance N is part of a ring), a double bond =N= is not possible. This translates in the restraint $X_{12} + X_{23} \le 6$ (N has atom number 2 and is bound to two other atoms 1 and 3, X represent the number of electrons in a bond).

Additional bounds can be applied to individual atoms: for example atoms H and C have no free electron pairs (Fi = 0); the maximum bond order of a carbon atom is three, etc.

APPLICATION

The algorithm described in the previous paragraph can be used to deduce the correct bond orders in small molecules, which only needs the sigma bonds connectivity and atom symbols as input. The molecular configuration (described by the connectivity and atom types) is tested for 'closed shell'

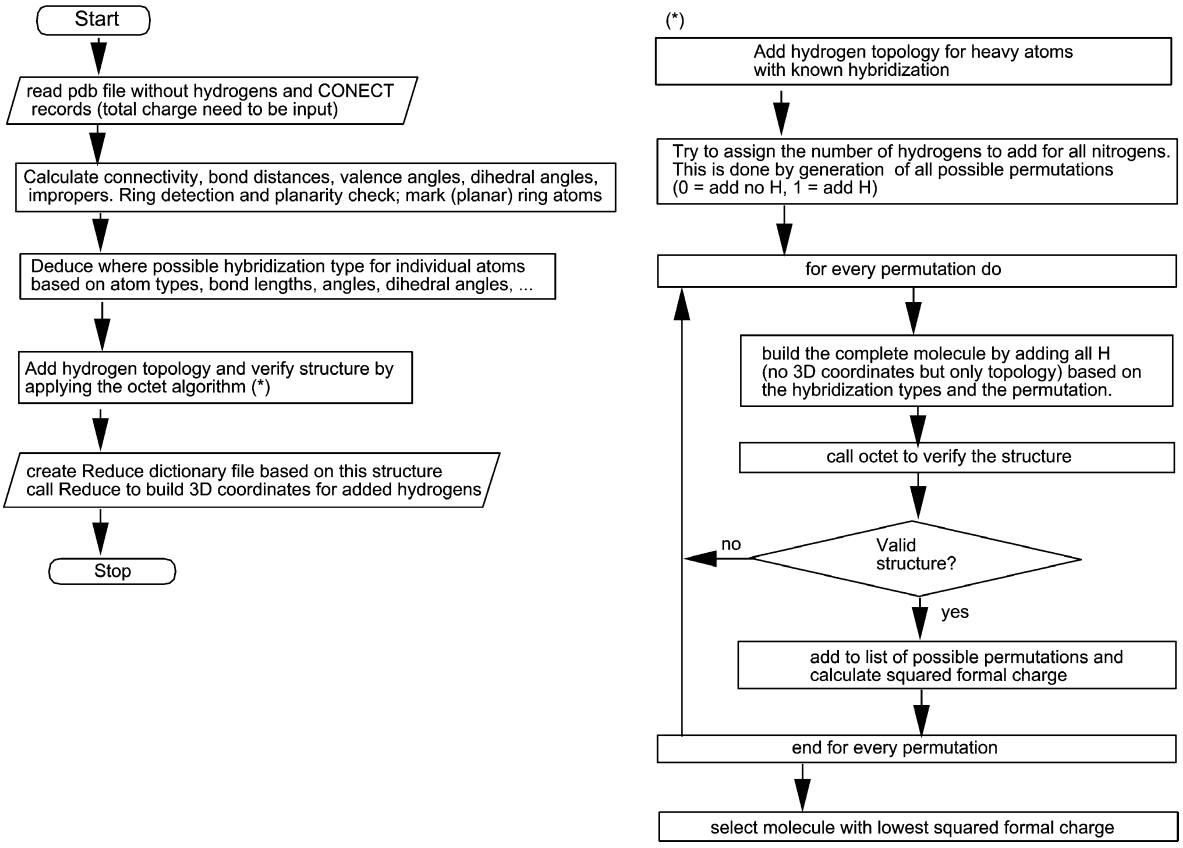

Figure 5. Flowchart of the application of the octet procedure in a program for deducing the correct structure if just the heavy atom positions are given.

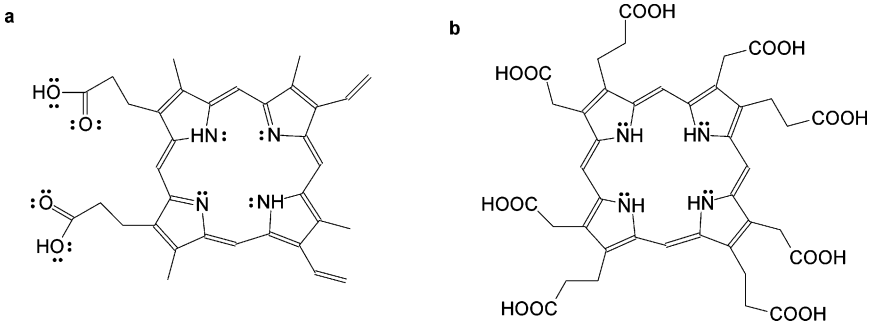

Figure 6. Example porphine ring systems: (a) protoporphyrin IX and (b) uroporphyrinogen III.

property, and the octet procedure, described higher, is applied on it, to verify its electron distribution. A set of equations based on the rules given in the previous section are built from the atom types and bonding information and translated in the command language of the lp_solve 4.0 program. Hydrogen atom variables and equations are removed because of their straightforward electronic configuration. The resulting linear programming problem is then solved using the lp_solve 4.0 program.[18] Table 2 shows the example run of the azide CH_3N_3 from Figure 2. The electron distribution proposed by the program (Table 2, bottom) is resonance form b.

We wanted to test the capabilities of the algorithm intensively on the vast amount of ligand structures available through the Protein Data Bank (PDB). For this purpose, however, we needed first a procedure to add hydrogen atoms to the molecular structure files containing only non-hydrogen atoms. In a second step, the octet algorithm is run on the hydrogen-containing ligand files. It is used as a tool to verify the electron distribution in single, double, and triple bonds and free electron pairs and decide also about the guessed protonation state of some nitrogens (see below).

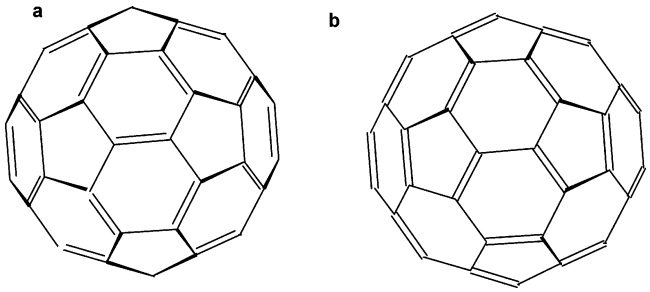

Figure 7. Correct (left) and wrong (right) resonance forms of C60.

In short we will describe this program here. This program was realized in fortran-77 (Figure 5).[24] A protein databank file with only heavy (non-hydrogen) atoms and without CONECT records is needed as input.[25] Molecules are considered as 'closed shell', i.e., they have an even number

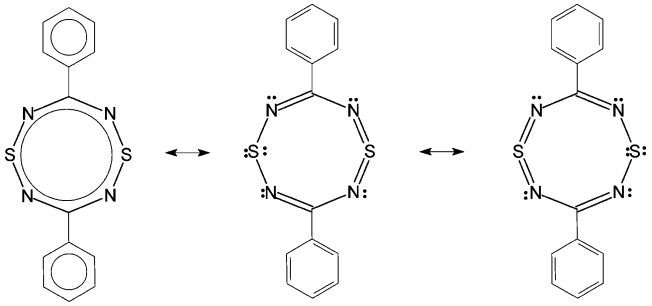

Figure 8. 3,7-Diphenyl-1,5-dithia-2,4,6,8-tetrazocine.

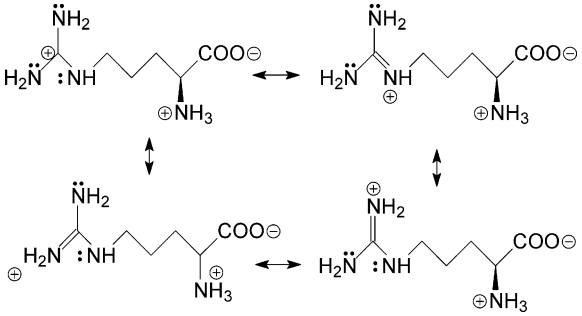

Figure 9. L-Arginine.

of electrons. If the molecule is charged, the charge can be entered. The deduction of the hybridization of each heavy atom is not new but based on the molecular geometry using similar methods as described in refs 7−11 and 26. The number of heavy atoms bound to bond distances, valence angles, dihedral angles, improper angles, detection of functional groups, and planar ring detection is used in assigning the correct hybridization type of individual atoms. For instance if an atom has 3 connections to 3 other heavy atoms, an improper dihedral angle can be calculated. A value close to 0 degrees indicates sp^2 hybridization and 36 degrees sp^3 hybridization. For atoms with two connections to heavy atoms, a valence angle can be calculated deciding about the hybridization of the central atom: 109 degrees (sp^3) or 120 degrees (sp^2). For atoms connected to one other heavy atom the bond length is determining. Reference bond lengths were taken from ref 27. The functional group detection routine starts from atoms connected to one other heavy atom and recognizes the most common functional groups found in organic molecules. The cycle detection routines are based on routines found in ref 28 as implemented by ref 29. Planarity for every ring atom is calculated as the deviation from a least squares plane routine using Eispack.[30,31] In a first test of the hydrogen addition part, the planarity check was implemented using the Cremer/Pople method.[32] However tests on the HETATM ligand PU in 1ffz show that the algorithm did not assign the correct planarity. The Cremer/Pople algorithm finds a mean plane based on the assumption that successive points in the ring have the same angle between them with respect to the center of geometry, so for kinky rings this may not work. The mean plane algorithm was then replaced by the classic least squares plane routine.

The assignment of atom types and hybridization depends heavily on the accuracy of the coordinates. But also for nitrogens in a planar ring system, it is not obvious to tell ad hoc that they are protonated or not. If there is doubt about the number of hydrogens to add to an atom, mostly being nitrogen, this atom gets a tag. Examples are for instance N1 and N3 in a guanine base. A simple binary counter is used: every heavy atom with an uncertain hydrogen count gets one bit assigned where 0 means no addition of H and 1 means add a hydrogen. All binary values are processed giving a number of chemical structures. To get rid of the uncertain protonation states, the octet procedure is called for every chemical structure proposed. If an acceptable solution is returned, this configuration with added hydrogen topology assigned is saved. For all the configurations that pass the test, the sum of squared formal charge is calculated and the one with the lowest sum is retained. However, if more configurations have the same lowest sum, additional rules are considered: for instance, configurations where hydrogens are added to nonplanar nitrogens are preferred. Finally if still no exclusive solution can be selected, the structure with the lowest number of H added is taken.

The three-dimensional (3D) coordinates of the new hydrogen atoms are generated by Reduce for which first a dictionary file of the molecule is generated.[33,34] The CONECT records are added explicitly to the Reduce output PDB file. They are also used to define the bond order of a bond by entering them twice for a double and trice for triple bond. For instance, 'CONECT 1 2 2' defines a double bond from atom 1 to 2. The electron pairs are also written in the CONECT records as for example 'CONECT 1 1 1', which means two free electron pairs. This format is recognized by rasmol 2.6 and will allow visualization of the multiple bonds (but unfortunately not the free electron pairs).[35]

RESULTS AND DISCUSSION

Examples from the Protein Data Bank. The octet procedure embedded in a hydrogen addition program described higher in the application section was tested on different HETATM molecules from the PDB.[25] While developing and testing the program, problems in the first step, the atomic hybridization assignment process, were encountered, resulting in wrongly perceived molecular structures. Looking at the individual test cases, it was found that the perception of the correct structure fails in those cases due to incorrect bond lengths, angles, or dihedral angles or close atom contacts in the ligands in some PDB files (see the Supporting Information). These are possibly a consequence of incorrect parametrization during X-ray refinement. Examples are found in PDB entries 2dhf, 1fx1, 2trm, 1opb, 1bib, 2ttd, 1pmp, 4fbp, 1aqb, 1dih, 2r05, 2sns, 3dfr, 4gr1, 1ffz, 1erb, and 1ouk. As already indicated by Labute, these cases will present difficulties to all chemical perception methods.[11] The second step, however, the interplay between the octet verification procedure and hydrogen addition for atoms with an uncertain protonation state, performed well on the tested PDB HETATM entries. In the cases 4fbp, 1aqb, 3dfr, 4gr1, and 1erb, no valid electron distribution was found by the octet procedure, proving its value as a verification tool in identifying chemical impossible configurations, generated in the first step. In the other cases 2dhf, 1fx1, 2trm, 1opb, 1bib, 2ttd, 1pmp, 1dih, 2r05, 2sns, 1ffz, and 1ouk, chemical valid structures were generated and perceived by the octet procedure as correct. However, these structures are different from the structures found in the original publications.

Other Examples. Tests on the heme molecule of PDB entry 2hhb showed that applying the procedure to porphyrin

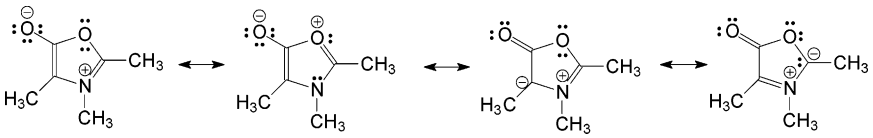

Figure 10. A mesoionic compound.

molecules can lead to more than one valid solution. Extracting the porphyrin molecule from 2hhb (without central iron ion) and running the program resulted in 2 possible solutions when not considering the side chains: protoporphyrin IX and uroporphyrinogen III (Figure 6). [36] It is impossible to predict the correct protonation state of the porphyrin nitrogens based on the 3D coordinates of the non-hydrogen atoms alone. Both solutions (one with 2 central protons and one with 4 central protons) were encountered as correct structures in the octet verification step.

The crystal structure of C60 (entry NOBLEV[37] from the Cambridge Structural Database)[38] initially failed in the detection of sp^2/sp^3 hybridization based on an improper dihedral angle of less/higher than 18 degrees (halfway between sp^2, 0 degrees, and sp^3, 36 degrees) because some improper angles in this X-ray structure of C60 are 25 degrees or more. When changing this cutoff to 30 degrees, the octet procedure performed well in predicting the distribution of double and single bonds depicted in Figure 7 at the right. However, the average C−C bond lengths are 1.39 Å for 6-6 ring fusions and 1.43 Å for 6-5 ring fusions.[39] Hence, the bonding in C is not completely delocalized, which means structure 7 at the right, which has aromatic or double bonds at long bonds and short bonds that have a bond order of one, is not correct. Rather, the dominant resonance structure (see Figure 7, left) is one in which the double bonds are located exocyclic to the five-membered rings and between the six-membered rings. When defining the 6-5 fusion bonds as single in the linear programming problem, based on the bond lengths, the octet procedure returns the electron distribution of Figure 7, left.

The bond orders and electron pairs of the molecule 3,7-diphenyl-1,5-dithia-2,4,6,8-tetrazocine (structures in Figure 8, CSD entry PTZCNA)[40] and a charged L-arginine (structures in Figure 9, CSD entry ARGIND11)[41] were reproduced without any problem. For the dithiatetrazocine, one of the two resonance forms was reproduced while for L-arginine one of the two bottom resonance forms was found.

Another exotic molecule from the class of mesoionic compounds was subjected to the test.[42] This compound is not charged but has a negative and positive charge delocalized around the ring system. The program validated the heavy atom structure and returned the resonance form at the left in Figure 10.

CONCLUSION

A new procedure to assign double and triple bonds in small molecules from simple PDB files proves to be very simple. The heart of the program is the definition of the problem as an integer linear problem, which can be solved by any external solver (if the correct syntax is used). The procedure proves to be very useful for verifying chemical structures when just the heavy atom positions are given. While adding the hydrogens, the electron distribution is verified by the procedure and approved or rejected resulting in mostly one correct structure. Finally, the three-dimensional coordinates of the new hydrogen atoms are generated by Reduce for which first a dictionary file of the molecule is generated. The bond orders found can be visualized by Rasmol.

Supporting Information Available: Chemical drawings of the ligand molecules in the PDB entries discussed in the Results and Discussion section and possible coordinate errors. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Jensen, W. B. Abegg, Lewis, Langmuir, and the octet rule. *J. Chem. Educ.* **1984**, *61*, 191−200.

(2) Ahmad, W. Y.; Zakaria, M. B. Drawing Lewis structures from Lewis symbols: a direct electron pairing approach. *J. Chem. Educ.* **2000**, *77*, 329−331.

(3) Reed, A. E.; Curtiss, L. A.; Weinhold, F. Intermolecular interactions from a natural bond orbital, donor−acceptor viewpoint. *Chem. Rev.* **1988**, *88*, 899−926.

(4) Hansen, P.; Jaumard, B.; Sachs, H.; Zheng, M. Finding a Kekule Structure in a benzenoid System in Linear Time. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 561−567.

(5) Balaban, A. T.; Randic, M. Partitioning of pi-Electrons in Rings of Polycyclic Benzenoid Hydrocarbons. 2. Catacondensed Coronoids. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 50−59.

(6) Blaney, J. M.; Dixon, J. S. Distance Geometry in Molecular Modeling. *Rev. Comput. Chem.* **1994**, *5*, 299−335.

(7) Leach, A. R.; Dolata, D. P.; Prout, K. Automated Conformational Analysis and Structure Generation: Algorithms for Molecular Perception. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 316−324.

(8) Meng, E. C.; Lewis, R. A. Determination of Molecular Topology and Atomic Hybridization Stats from Heavy Atom Coordinates. *J. Comput. Chem.* **1991**, *12*, 891−898.

(9) Baber, J. C.; Hodgkin, E. E. Automatic Assignment of Chemical Connectivity to Organic Molecules in the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 401−406.

(10) Hendlich, M.; Rippmann, F.; Barnickel, G. BALI: Automatic Assignment of Bond and Atom Types for Protein Ligands in the Brookhaven Protein Databank. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 774−778.

(11) Labute, P. On the Perception of Molecules from 3D Atomic Coordinates. *J. Chem Inf. Model.* **2005**, *45*, 215−221.

(12) Gilheany, D. G. No d orbitals but Walsh diagrams and maybe banana bonds: chemical bonding in phosphines, phosphine oxides, and phosphonium ylides. *Chem. Rev.* **1994**, *94*, 1339−1374.

(13) Reed, A. E.; von Rage Schleyer, P. Chemical Bonding in Hypervalent Molecules. The Dominance of Ionic Bonding and Negative Hyperconjugation over d-Orbital Participation. *J. Am. Chem. Soc.* **1990**, *112*, 1434−1445.

(14) Magnusson, E. Hypercoordinate molecules of second-row elements: d functions or d orbitals? *J. Am. Chem. Soc.* **1990**, *112*, 7940−7951.

(15) Cioslowski, J.; Mixon, S. T. Rigorous interpretation of electronic wave functions. 2. Electronic structures of selected phosphorus, sulfur, and chlorine fluorides and oxides. *Inorg. Chem.* **1993**, *32*, 3209−3216.

(16) Dobado, J. A.; Martinez-Garcia, H.; Molina, J. M.; Sundberg, M. R. Chemical bonding in hypervalent molecules revised. Application of the atoms in molecules theory to Y_3X and Y_3XZ (Y = H or CH_3; X = N, P or As; Z= O or S) compounds. *J. Am. Chem. Soc.* **1998**, *120*, 8461−8471.

(17) Pauling, L. *The Nature of the Chemical Bond*; Cornell University Press: New York, 1945; pp 199−200.

(18) Berkelaar, M. lp_solve() 4.0, Eindhoven University of Technology. http://groups.yahoo.com/group/lp_solve/

(19) Conn, A. R.; Gould, N. I. M.; Toint, P. L. *LANCELOT: a Fortran package for large-scale nonlinear optimization (Release A)*; Springer-Verlag: Berlin, 1992.

(20) Fortet, R. Applications de l'algebre de Boole en recherché operationnelle. *Rev. Fr. Rech. Operationelle* **1960**, *4*, 17−26.

(21) Faucher, A. M.; White, P. W.; Brochu, C.; Grand-Maitre, C.; Rancourt, J.; Fazal, G. Discovery of small-molecule inhibitors of the ATPase activity of human papillomavirus E1 helicase. *J. Med. Chem.* **2004**, *47*, 18−21.
(22) Vanheusden, V.; Van Rompaey, P.; Munier-Lehmann, H.; Pochet, S.; Herdewijn, P.; Van Calenbergh, S. Thymidine and thymidine-5′-O-monophosphate analogues as inhibitors of mycobacterium tuberculosis thymidylate kinase. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 3045−3048.
(23) Saenger, W. *Principles of Nucleic Acid Structure;* Springer-Verlag: New York, 1984; p 85.
(24) The program source code is available from the author M. Froeyen.
(25) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(26) van Aalten, D. M. F.; Bywater, R.; Findlay, J. B. C.; Hendlich, M.; Hooft, R. W. W.; Vriend, G. PRODRG, a program for generating molecular topologies and unique molecular descriptors from coordinates of small molecules. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 255−262.
(27) *CRC Handbook of Chemistry and Physics*, 59th ed.; 1978−1979.
(28) Nijenhuis, A.; Wilf, H. *Combinatorial Algorithms*, 2nd ed.; Academic Press: 1978.
(29) Burkardt, J. graphpack, http://orion.math.iastate.edu/burkardt/.
(30) Schomaker, V.; Waser, J.; Marsh, R. E.; Bergman, G. To Fit a Plane or a Line to a Set of Points by Least Squares. *Acta Crystallogr.* **1959**, *12*, 600−604.
(31) B. T. Smith, B. T.; Boyle, J. M.; Dongarra, J. J.; Garbow, B. S.; Ikebe, Y.; Klema, V.; Moler, C. B. *Matrix Eigensystems Routines − EISPACK Guide. Lecture Notes in Computer Science*; Springer-Verlag: 1976.
(32) Cremer, D.; Pople, J. A. A General Definition of Ring Puckering Coordinates. *J. Am. Chem. Soc.* **1975**, *97*, 1354−1358.
(33) Word, J. M.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. Asparagine and Glutamine: Using Hydrogen Atom Contacts in the Choice of Side-chain Amide Orientation. *J. Mol. Biol.* **1999**, *285*, 1735−1747
(34) Word, J. M.; Lovell, S. C.; LaBean, T. H.; Taylor, H. C.; Zalis, M. E.; Presley, B. K.; Richardson, J. S.; Richardson, D. C. Visualizing and Quantifying Molecular Goodness-of-Fit: Small-probe Contact Dots with Explicit Hydrogen Atoms. *J. Mol. Biol.* **1999**, *285*, 1711−1733.
(35) Sayle, R. A.; Milner-White, E. J. RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.* **1995**, *20*, 374.
(36) Battersby, A. R. Tetrapyrroles: the pigments of life. *Nat. Prod. Rep.* **2000**, *17*, 507−526.
(37) Atwood, J. L.; Barbour, L. J.; Raston, C. L.; Sudria, I. B. N. C60 and C70 Compounds in the Pincerlike Jaws of Calix[6]arene. *Angew. Chem. Int. Ed.* **1998**, *37*, 981−983.
(38) Allen, F. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr.* **2002**, *B58*, 380−388.
(39) Hawkins, J. M.; Meyer, A.; Lewis, T. A.; Loren, S.; Hollander, F. J. Crystal Structure of Osmylated C_{60}: Confirmation of the Soccer Ball Framework. *Science* **1991**, *252*, 312−313.
(40) Ernest, I.; Holick, W.; Rihs, G.; Schomburg, D.; Shoham, G.; Wenkert, D.; Woodward, R. B. 1,5-Dithia-2,4,6,8-tetrazocine: A Novel Heterocycle of Unusual Properties. *J. Am. Chem. Soc.* **1981**, *103*, 1540−1544.
(41) Lehmann, M. S.; Verbist, J. J.; Hamilton, W. C.; Koetzle, T. F. Precision Neutron-Diffraction Structure Determination of Protein and Nucleic-Acid Components. 5. Crystal and Molecular-Structure of Amino-Acid L-Arginine Dihydrate. *J. Chem. Soc., Perkin Trans. 2* **1973**, 133−137.
(42) Huisgen, R.; Gotthardt, H.; Bayer, H. O.; Schaefer, F. C. 1,3-Dipolar cycloadditions. LVI. Synthesis of N-substituted pyrroles from mesoionic oxazolones and alkynes. *Chem. Ber.* **1970**, *103*, 2611−2624.

CI049645Z