ARTICLES

On the Perception of Molecules from 3D Atomic Coordinates

Paul Labute*

Chemical Computing Group Inc., 1010 Sherbrooke Street, Suite 910, Montreal, Quebec, Canada H3A 2R7

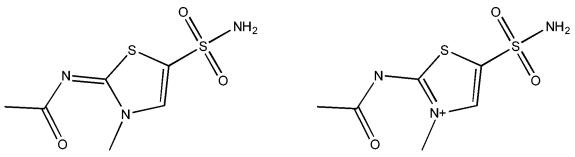
Received March 4, 2004

A method is presented for perceiving chemical types of atoms in molecules given 3D atomic coordinates and element identities. The method assigns hybridizations, bond orders, and formal charges for structures whether hydrogen atoms are present. The Maximum Weighted Matching algorithm for nonbipartite graphs is used to assign bond orders with weights derived from statistics of a large collection of organic molecules. Results form tests on a collection of functional groups, heterocycles, entries from the Protein Data Bank, and Cambridge Structural Database as well as a comparison to other methods, are presented and discussed.

INTRODUCTION

The rapid growth and large size of the Cambridge Structural Database[1] (CSD) and the Protein Data Bank[2] (PDB) have made them important resources in both the life and materials sciences, especially for statistical or knowledge-based analysis of molecular structure and molecular interactions. Such databases contain information about element identities, atomic coordinates, and, usually, molecular bond connectivity. Atom types and bond orders are not provided which makes automated processing of these databases a nontrivial proposition. Making matters worse, PDB structures are of poorer quality than the CSD, with lower resolution, missing hydrogen atoms, larger coordinate errors, more instances of incorrect connectivity, and even erroneous element labels. Hand curation efforts such as Relibase[3] and the Macromolecular Structure Database[4] are attempts to correct these errors, at least for PDB ligands about which errors are often found. Unfortunately, questions often remain surrounding PDB ligands; for example, a hand-curated database may represent ligands in neutral form prior to reaction with an enzyme while the crystallographic structure contains the ligand in an intermediate geometry or product or nonneutral form.

These considerations and the need to automatically interpret the contents of the CSD and PDB have inspired the development of methods to automatically perceive chemical atom types when given experimental atomic coordinates and element identities. Over the past decade, methods such as those of Meng,[5] Baber,[6] Hendlich,[7] and Sayle[8] have been developed, each attempting to solve the problem, and many computer programs contain some sort of perception algorithm used when reading CSD or PDB files. For the most part, these methods comprise bond length and bond angle analysis to determine hybridization in combination with functional group pattern matching or other rules (such as maximizing the number of aromatic rings) and possibly followed by a "fix-up" pass do deal with inconsistent hybridization and bond order assignments made by the less reliable geometric tests. The ad hoc nature of these methods often limits their utility because they have been tuned for use with the PDB. One example of this is Sayle's perception of benzamidine in 2trm of the PDB despite its grossly distorted structure (which we describe later on in this paper). Careless application of patterns for peptide recognition and heuristics such as aromaticity maximization may lead to incorrect structures such as the ligand of PDB entry 1bzm:



1bzm (correct) *1bzm (incorrect)*

A perception method that insists on making the ring aromatic or making an a priori peptide assignment to nitrogen in the =N−C=O group will not perceive this structure correctly.

In an attempt to overcome some of the shortcomings of previous methods, we have developed a new algorithm to perceive molecules from 3D atomic coordinates and element identities. The method makes use of the well-known Maximum Weighted Matching graph algorithm from the field of combinatorial optimization. We present details of the method in the Methods section and present results of its application in the Results and Discussion section. We draw some conclusions in the final section.

METHODS

Let $x_1,...,x_n$ denote the 3D coordinates of n atoms with atomic numbers $Z_1,...,Z_n$. and let r_{ij} denote $|x_i - x_j|$. Chemical bonds are perceived by first producing a list of candidate bonds based upon proximity and reference covalent radii and then by refining the list after consideration of the geometry

* Corresponding author e-mail: paul@chemcomp.com.

10.1021/ci049915d CCC: $30.25
Published on Web 01/19/2005

of the proposed bonded neighbors. The method of Meng[5] is used to produce the candidate list. A candidate bond between two atoms i and j is generated if

$$0.1 < r_{ij} < R_i + R_j + 0.4$$

where R_i is the "covalent radius" of atom i and the tolerances are measured in angstroms. A radius of 0.23 is used for hydrogen, 0.68 for carbon, 0.68 for nitrogen, 0.68 for oxygen, 0.64 for fluorine, 1.05 for phosphorus, 1.02 for sulfur, and 0.99 for chlorine (a complete list is given by Meng which is based on the results of Allen et al.[9]).

It often happens that spurious bonds arise from coordinate errors (especially in the PDB) or strained geometries; such bonds require removal. For example, Sayle relies on bond count limits based upon the atomic number Z_i. In our experience a more detailed but general analysis is required. For each atom i, a "dimension" d_i is calculated as follows. A given atom with coordinates q_0 has k candidate bonds to atoms with coordinates $q_1,...,q_k$. A covariance matrix, A, is calculated using

$$A = \sum_{j=0}^{k} (q_j - \bar{q})(q_j - \bar{q})^T, \bar{q} = \frac{1}{k}\sum_{j=0}^{k} q_j$$

The value of d_i is set to k if $k < 2$ otherwise, d_i is the number of positive eigenvalues of A with square root greater than 0.2. The value d_i will be 0 for isolated atoms, 1 for terminal and linear atoms with at least two bonds, 2 for planar atoms (e.g., sp^2 or square-planar), and 3 otherwise (e.g., tetrahedral or dsp^3 hybridized). The value 0.2 was determined by inspection of numerous structures and is required to correctly perceive pi systems in strained environments (such as C_{60}). The d_i and Z_i are used to remove bonds from the candidate list. Each pair (d_i,Z_i) determine an upper bound, B_i, on the number of bonds of each atom according to the following:

1. $B_i = 0$ if $d_i = 0$ (disconnected atoms);
2. $B_i = 1$ if $Z_i < 3$ (hydrogen and helium);
3. $B_i = 2$ if $d_i = 1$ and $Z_i > 2$ (sp hybridizations and linear geometries);
4. $B_i = 3$ if $d_i = 2$ and $Z_i < 11$ (sp^2 hybridizations for second row elements);
5. $B_i = 4$ if $d_i = 2$ and $Z_i > 10$, or $d_i = 3$ and $Z_i < 11$ (square planar or sp^3 hybridizations);
6. $B_i = 7$ otherwise.

For each atom, only the shortest B_i bonds are retained from the candidate list. This removal may (rarely) result in an asymmetric connection table; i.e., atom i is in atom j's list but not vice versa. Such asymmetric bonds are also removed resulting in a final, symmetric connection table.

Bond orders and hybridizations are perceived after the connectivity has been determined. The most complex part of this procedure is dealing with resonance structures in pi systems. For the present purposes, designation as sp, sp^2, sp^3, etc. is symbolic and only sp and sp^2 atoms can participate in higher order bonds. For example, sp^3 is used for tetrahedral atoms and amide nitrogen since neither has a formal double bond. Additionally, dative bonds are perceived in charge-separated notation (e.g., a sulfoxide is perceived as (S+)-(O-) with both S and O designated as sp^3). In drawings of chemical structures, dative bonds are postprocessed from charge-separated notation to double bond representation. Each atom is initially assigned a hybridization code of "?" (for "unknown"), and each bond order b_{ij} in the connection table is initially assigned a bond order of zero (for "unknown").

Let Q_i denote the number of connections of atom i. The following algorithm is used to assign obvious initial hybridizations based on d, Z, and Q. Each step applies to atoms for which the hybridization is "?" after application of the previous steps.

1. $Z_i = \{1,2\}$ atoms are set to sp^3 (to indicate non-pi participation).
2. ($Q_i > 4$, $Z_i = \{$Group 5$\}$) and ($Q_i = 5$, $Z_i = \{$Group 4,5,6,7,8$\}$) atoms are set to dsp^3.
3. ($Q_i > 4$, $Z_i = \{$Group 6$\}$) and ($Q_i = 6$, $Z_i = \{$Group 4,5,6,7,8$\}$) atoms are set to d^2sp^3.
4. ($Q_i > 4$, $Z_i = \{$Group 7$\}$) and ($Q_i = 7$, $Z_i = \{$Group 4,5,6,7,8$\}$) atoms are set to d^3sp^3.
5. ($Q_i = 4$, $Z_i > 10$, $d_i = 2$) atoms are set to d^2sp^3 (square planar).
6. $Z_i = \{$transition metal$\}$ atoms are set to d^2sp^3.
7. $Z_i > 10$ and not {Si,P,S,Se} atoms are set to d^2sp^3 if $Q_i > 4$ and sp^3 otherwise.
8. ($Q_i = 4$) and ($Q_i = 3$, $d_i = 3$) atoms are set to sp^3 (tetrahedral).
9. ($Q_i > 2$, $Z_i = \{$Group 6,7,8$\}$) atoms are set to sp^3.
10. Z_i not {C,N,O,Si,P,S,Se} atoms are set to sp^3.
11. All atoms such that none of their bonded neighbors have a hybridization of "?" are set to sp^3. Repeat this step until no new atoms are assigned.

After application of the foregoing steps the only atoms with unassigned hybridizations have $d < 3$, $Z = \{$C,N,O,-Si,P,S,Se$\}$, $Q < 4$ and at least one bonded neighbor with an unassigned hybridization. All bond orders b_{ij} in which either atom i or j has non-"?" hybridization are set to 1 (since they are involve one atom that is not sp or sp^2 and cannot be of higher order). The preceding analysis obviates the need for further (and riskier) geometric tests on terminal atoms bonded to tetrahedral groups (e.g., sulfone oxygens) or $-CH_2-$ groups between clearly tetrahedral groups.

A conservative dihedral angle test is then applied to identify bonds of order 1. For each bond (i,j) not involving linear atoms the smallest out-of-plane dihedral is computed according to

$$\min_{a,b} \{|p_{aijb}|,|\pi - p_{aijb}|,|-\pi - p_{aijb}|\}$$

$$p_{aijb} = \cos^{-1}\frac{(x_a - x_i) \times (x_j - x_i)\cdot(x_i - x_j) \times (x_b - x_j)}{|(x_a - x_i) \times (x_j - x_i)||(x_i - x_j) \times (x_b - x_j)|}$$

where a ranges over the bonded neighbors of i (other than j) and b ranges over the bonded neighbors of j (other than i). If the computed smallest dihedral is greater than 15 degrees, then b_{ij} is set to 1. The value of 15 degrees was derived by manual experimentation. This test constrains double bonded atoms to have at least one near-planar dihedral. In most cases double bonded atoms are such that all of their dihedral angles are close to planar, but in strained environments (such as C_{60}) this is not the case. A further advantage of the dihedral test is that many $-CH_2-CH_2-$ atoms can be identified in hydrogen-suppressed situations

Table 1. Reference Single Bond Lengths Used To Eliminate Cases of Higher Order Bonds

bond	length	bond	length
C−C	1.54	C−N	1.47
C−O	1.43	C−Si	1.86
C−P	1.85	C−S	1.75
C−Se	1.97	N−N	1.45
N−O	1.43	N−Si	1.75
N−P	1.68	N−S	1.76
N−Se	1.85	O−O	1.47
O−Si	1.63	O−P	1.57
O−S	1.57	O−Se	1.97
Si−Si	2.36	Si−P	2.26
Si−S	2.15	Si−Se	2.42
P−P	2.26	P−S	2.07
P−Se	2.27	S−S	2.05
S−Se	2.19	Se−Se	2.34

without the need for riskier bond length or bond angle analyses.

A conservative bond length test is then applied to identify single bonds; note that the test is not used to determine double or triple bonds but only to identify bonds that cannot be of higher order. The bond order b_{ij} is set to 1 if $|x_i - x_j| > L_{ij} - 0.05$ where L_{ij} is the reference bond length between atoms i and j taken from Table 1 which was derived from MMFF94 parameters.[10]

After the bond length and dihedral angle tests have been applied, the hybridizations of all uncharacterized atoms not involved in a bond of unknown order are set to sp^3. The remaining bonds shall be subjected to pi system analysis although not all such bonds are necessarily part of the pi system (e.g., methyl groups with short bonds attached to aromatic rings in hydrogen suppressed situations). The bonds with thus far unassigned bond order induce a subgraph of the original molecular system. The induced subgraph contains a collection of connected components all of whose bonds have unassigned bond order. Each such component is analyzed independently and bond orders assigned.

Thus far, our method largely resembles that of previous efforts except for the fact that our geometric tests are conservative rather than definitive. We come now to the novel technique of our method: that of making a reasonable and consistent assignment bond orders. We apply the Maximum Weighted Matching algorithm[11] for nonbipartite graphs to the problem of assigning a consistent set of bond orders to each candidate pi system. A *matching* in a graph is a subset of edges such that no two edges in the subset have a vertex in common. A *maximal matching* is a matching such that no additional edge can be added without that edge having a vertex in common with an edge already in the matching. A *maximum matching* is a matching such that no other matching contains more edges. Assigning double bonds to the edges in a maximum matching corresponds to maximizing the number of double bonds in a pi system. The Maximum Weighted Matching Problem is related to the problem of determining a maximum matching except that each edge in the graph has an associated weight, and the problem is to determine a matching that has the largest sum of edge weights. Our algorithm to assign higher order bonds is to (a) assign a weight w_{ij} to each unassigned bond order b_{ij}; (b) apply the Maximum Weighted Matching algorithm to determine a matching with largest weight; and (c) assign double or triple bonds to the edges of the match.

Table 2. Log Likelihood Ratios Scoring the Preference of Atoms for Higher Order Bonds

atom[a]	Q=1[b]	Q=2	Q=3
C−O[c]	1.3[d]	4.0	4.0
C−N	−6.9	4.0	4.0
C	0.0	4.0	4.0
N−C−O	−2.4	−0.8	−7.0
N−C−N	−1.4	1.3	−3.0
N	1.2	1.2	0.0
O−C−O	4.2	−8.1	−20.0
O−C−N	4.2	−8.1	−20.0
O	0.2	−6.5	−20.0

[a] An atom denoted in the manner of N−C−O indicates nitrogen bonded to a carbon bonded to an oxygen. [b] Q denotes the coordination number of the atom. [c] Top-to-bottom precedence is used: if an atom matches two patterns, then the topmost pattern is used. [d] The given weight applies only to the first atom of the pattern.

In our method, the edge weights are set to a sum of atom weights and bond length properties according to the following formula

$$w_{ij} = u_i + u_j + 2\delta(r_{ij} < L_{ij} - 0.11) + \delta(r_{ij} < L_{ij} - 0.25)$$

where u_i is the weight scoring atom i's participation in a higher order bond; the delta terms add additional preference for higher order bonding for short bond lengths. The atom weights are derived statistically according to the formula

$$u = \log\left[\frac{\Pr(\text{atom has a double bond})}{\Pr(\text{atom has no double bond})}\right]$$

where the probabilities are estimated by counting frequencies of single and double bonding in a collection of 200 000+ organic compounds taken from vendor catalogs. In the frequency counting, dative bonds were treated as single bonds and octet rules were strictly followed; i.e., nitro and carboxylate groups have one double-bonded oxygen and one single-bonded oxygen. Positive atom weights indicate preference for double bonding and negative weights indicate preference for single bonds only. The atom weights are listed in Table 2. In Table 2, connectivity context is denoted with single bonds and lower numbered rows take precedence over higher numbered rows. Elements with atomic number above 10 are mapped to second row elements in the same column of the periodic table and 0.1 is subtracted from the weight in the table. A weight of −20.0 is used for elements not in the table. For example, N with two bonded neighbors one of which is C bonded to O has an atom weight of −0.8. Hendlich's method[7] is somewhat similar in that a weighting scheme is used to assign multiple bonds once hybridizations are assigned; his weights are derived solely from geometric information and ad hoc rules provide hard a priori preferences.

Once the maximum weighted matching is calculated, a bond length test is used to determine which of the match bonds should marked as triple bonds ($|x_i - x_j| < L_{ij} - 0.25$). This assignment is made only for linear atoms, and two double bonds are assigned to a linear atom only if it and its neighbors are not matched. Finally, linear atoms with triple or two double bonds are set to sp, and other atoms with double bonds are set to sp^2.

Ionization states and formal charges are perceived after the connectivity and bond orders have been determined. If

the system includes hydrogen atoms and is in the gas phase, then the ionization state of all atoms is set to the formal charge: $f_i = c_i - o_i + b_i$ where c_i is the atom's column number in the periodic table, o_i is the nominal "octet" (e.g., 2 for hydrogen, 6 for boron, 8 for carbon, and all other sp^3 atoms in Groups V, VI, VII, and VIII), and b_i is the sum of its bond orders.

If the system does not include hydrogen atoms, then the following steps are applied to set the ionization state:

1. $Z_i = 1$ atoms are set to 0 (hydrogen).

2. ($Z_i = \{$transition metal$\}$, $Q_i > 0$) atoms are set to f_i, isolated transition metals are set to 0.

3. ($Q_i = 4$, sp^3), ($Q_i = 3$, sp^2), and ($Q_i = 2$, sp) atoms are set to f (buried atoms).

4. All atoms with $f_i > 0$ are set to f_i (no possibility of implicit hydrogens).

5. All atoms with $f_i < 0$ not bonded to an atom with $f_i > 0$ are set to 0 (implicit hydrogens).

6. The first k most electronegative neighbors with $f_i < 0$ of each atom with $k = f_i > 0$ are set to -1 (dative bonds).

7. Remaining atoms are set to 0.

Last, if the system is to be considered in solution, then a conservative protonation/deprotonation/ionization is performed as follows: ($Z_i = \{$Group 1$\}$, $Q_i = 0$) atoms are set to $+1$; ($Z_i = \{$Group 2$\}$, $Q_i = 0$) atoms are set to $+2$; ($Z_i = \{$Group 7$\}$, $Q_i = 0$) atoms are set to -1; and alphiatic amine nitrogens (sp^3 nitrogens not bonded to an atom with a higher order bond) are set to $+1$, sulfonic, phosphonic, and carboxylic acid oxygens are set to -1, one nitrogen from amidine or guanidine is set to $+1$.

RESULTS AND DISCUSSION

Accurate assignment of bond orders and hybridization states is generally not possible in the presence of large coordinate errors and suppressed hydrogen atoms. This is due to the fact that there are, in general, many molecules that share the same hydrogen-suppressed sigma skeleton. Ultimately, the perception of molecules from 3D atomic coordinates is a guessing game, and the utility of any particular algorithm is a function of how often it guesses the right answer. With the foregoing in mind, we present some validation results. We proceed from situations with reasonably accurate geometries (to test the graph theoretical aspects of the present method) to PDB structures (to test the geometric aspects of our method).

A collection of common functional groups, depicted in Figure 1, was assembled and subjected to energy minimization with the MMFF94 force field using the Molecular Operating Environment (MOE) software.[12] Hydrogen atoms were then removed, and the resulting geometry was submitted to our perception method. In each case the formal charges and bonding pattern was correctly perceived (remembering that dative bonds are perceived in charge-separated form and postprocessed into double bond form). Degenerate cases such as nitro and carboxylate groups are automatically handled by the Maximum Weighted Matching algorithm which will always produce a consistent double bonding pattern (although arbitrary selections will be made in the case of equal weights). Similarly, a small set of ring structures, depicted in Figure 2, was assembled (with emphasis on biologically relevant rings) and subjected to energy minimization. Our perception method was applied to the hydrogen-suppressed structure. In each case the formal charges and bonding pattern was correctly perceived. Unlike previous work, these groups

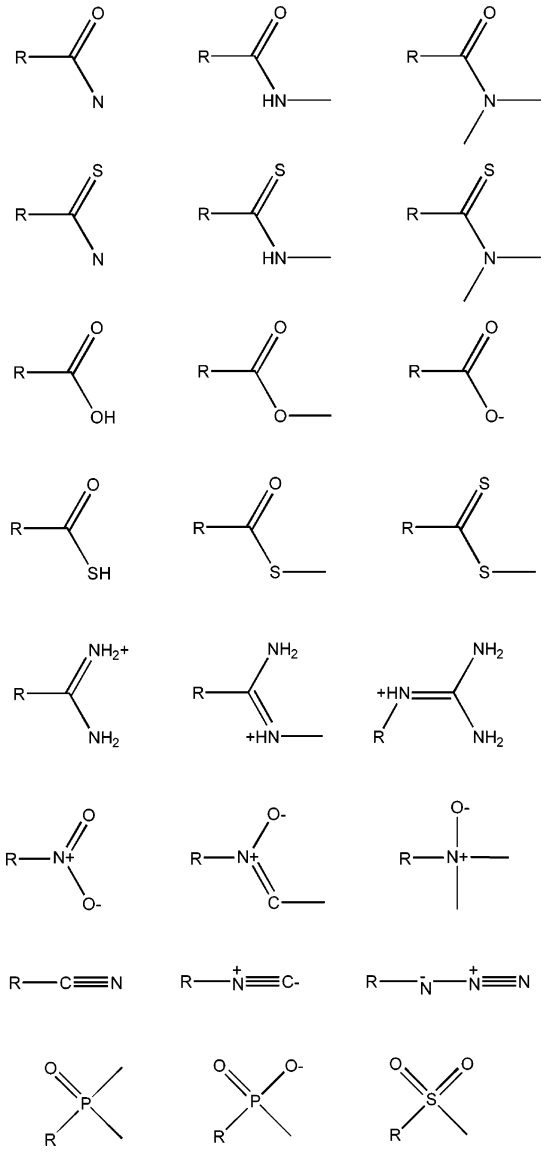


Figure 1. A collection of common functional groups whose formal charges and bonding pattern were correctly perceived by the method described herein (with hydrogens suppressed).

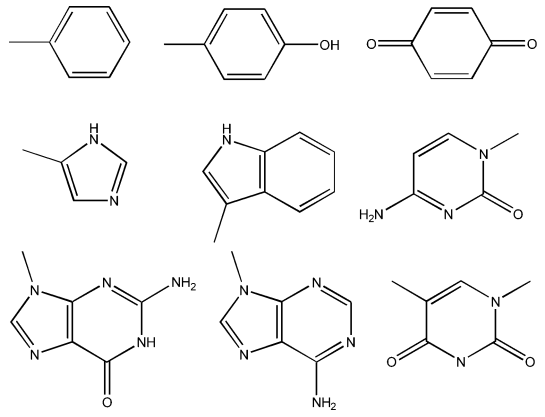


Figure 2. A small set of ring structures whose formal charges and bonding pattern were correctly perceived by the method described herein (with hydrogens suppressed).

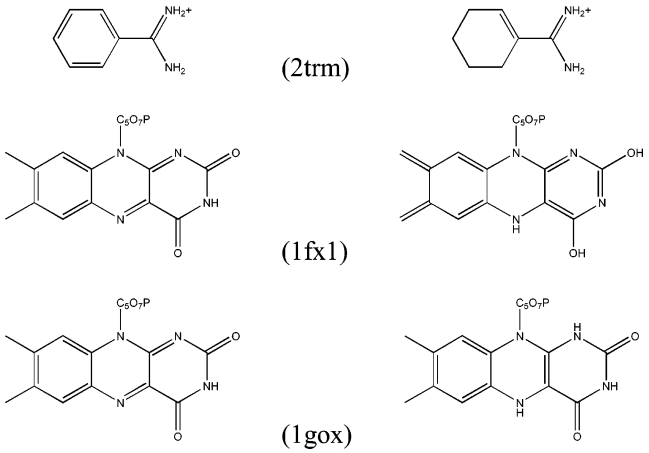


Figure 3. Three ligand structures from the PDB that our method failed to perceive (left). The perceived structures are to the right.

and rings were recognized without special-purpose ring analysis or heuristic bias toward aromatic rings.

The crystal structures of caffeine (CAFINE01), C_{60} (NOBLIZ), ibuprofen (IBPRAC02), and acrylic acid (ARCLAC) from the Cambridge Structural Database were subjected to the perception method described here. In each case the connectivity, bond order, hybridization, and formal charges were correctly assigned. The most notable of these is C_{60} in which the sp^2 carbons are highly nonplanar and participate in dihedral angles of 30-degree magnitude. It is a nontrivial matter to assign bond orders to C_{60}, yet our method correctly perceives the bonding pattern without special rules or patterns.

Ricketts et al.[13] presented the results of several molecular perception methods on 17 PDB structures with codes 1cla, 1fcb, 1fx1, 1gox, 2aat, 2dhf, 2gbp, 2trm, 3cpp, 3ptb, 4dfr, 4xia, 5xia, 7dfr, 8atc, 8ldh, and 8rsa. Sayle presented a comparison summary including his own method. Of the 17 structures, the method of assigning all single bonds to the molecule correctly perceived 3 structures, Baber's method 7, Meng's method 5, and Sayle's method 17. Our method correctly perceived 14 structures (with hydrogen atoms suppressed) and failed on 3 structures (1fx1, 1gox, and 2trm) depicted in Figure 3. The ligand of 1fx1 is FAD, but all of the bond lengths in the isoalloxazine ring and its immediate substituents are in the range (1.39,1.41) with most equal to 1.40 Å, suggesting poor ligand parametrization during refinement. The ligand of 1gox is also FAD (oxidized form), but our method perceives $FADH_2$ (reduced form). The geometric differences of FAD and $FADH_2$ are slight (optimized structures superpose to 0.067 Å RMSD), and our perception is reasonable given the hydrogen-suppressed 2.0 Å resolution PDB entry. The claimed ligand of 2trm is benzamidine, and our method detects a cyclohexene ring instead of a benzene ring because the ring carbons are highly nonplanar participating in dihedral angles of −35.6, 41.8, −33.9, 21.0, −15.8, and 23.6 degrees, which are in excellent agreement with cyclohexene when optimized with MMFF94. In our opinion a perception method should not perceive the 2trm ligand as benzamidine—perceiving benzene in this situation implies that cyclohexene could not be perceived correctly. Although our method nominally failed on three structures, we believe that the failures were understandable, and, indeed, these cases ought to present difficulty to all but biased methods. Sayle notes that these 17 structures were used in the development of his method, and only Sayle perceived these three structures correctly.

The data set of 120 entries from the PDB reported by Hendlich was assembled. In each case, our method was applied to the hydrogen-suppressed structure. Of the 120 entries, 112 (93%) were correctly perceived; Hendlich reported a 90% success rate (109 correctly perceived). Our method correctly perceived 5 of Hendlich's 11 failures (8cat, 1opb, 1fbp, 1bib, 2tdd) and failed on 6 of Hendlich's failures (1pcr, 1erb, 1aia, 1h7a, 1pmp, 4fbp). Only two of Hendlich's successful heme identifications (2mm1, 1yst) were incorrectly perceived by our method. With the exception of 1yst, all of our failures were due to bond lengths that were incompatible with the correct assignment (e.g., C═C bonds with 1.54 Å bond lengths) and 1yst failed due to a 16 degree C═C smallest dihedral angle.

As a further validation test, a collection of 177 entries from the PDB commonly used for the training of docking scoring functions was assembled. The codes for this collection are presented in Table 3. In each case our perception method was applied, and the results were manually compared to the intended ligand. Of the 177 ligands 166 (94%) were correctly perceived by our method and 11 were incorrectly perceived. The problematic cases were as follows:

1. 1aaq − the method failed to assign the carbonyl double bond in the terminal −C(═O)-O-Me in the ligand ALA-ALA-PHE-PSI(CHOH−CH2)-ALA-VAL-VAL-OME. This failure was due to carbonyl C═O bond length of 1.47 Å. The carbonyl carbon is quite nonplanar having bond angles of 112, 113, and 123 degrees. We note that our method correctly perceived the same similarly strained groups in 1apt and 1apu which had more reasonable C═O bond lengths.

2. 1aqb − the method failed to assign one vinylic double bond in the terminal −C(C)═C−C−OH group of the bound

Table 3. List of PDB Structure Codes Used To Validate the Molecular Perception Method

1AAQ	1ABE	1ABF	1ADB	1ADD	1ADF	1APB	1APT	1APU	1APV	1APW	1AQB	1BAP	1BRA
1BZM	1CBX	1CLA	1CPS	1CSC	1CTT	1DBB	1DBJ	1DBK	1DBM	1DHF	1DIH	1DR1	1DRF
1DWB	1DWC	1DWD	1EBG	1ELA	1ELC	1ETR	1ETS	1ETT	1FBC	1FBF	1FBP	1FKB	1FKF
1G6N	1HBV	1HPV	1HSL	1HTF	1HTG	1HVI	1HVJ	1HVK	1HVR	1HVS	1L83	1LDM	1LGR
1LYB	1MBI	1MCB	1MCF	1MCH	1MCJ	1MCS	1MDQ	1MFE	1MNC	1NNB	1PGP	1PHE	1PHF
1PHG	1PHH	1PPC	1PPH	1PPK	1PPL	1PPM	1PSO	1RBP	1RNE	1RNT	1RUS	1SNC	1SRE
1THA	1TLP	1TMN	1TMT	1TNG	1TNH	1TNI	1TNJ	1TNK	1TNL	1ULB	1XLI	2AK3	2CGR
2CSC	2CTC	2DBL	2DRI	2ER6	2GBP	2IFB	2LDB	2MCP	2PHH	2PK4	2R04	2RNT	2RTD
2SNS	2TMN	2XIM	2XIS	2YPI	3CLA	3CPA	3CSC	3DFR	3FX2	3PGM	3PTB	3TMN	3TPI
4CLA	4DFR	4FAB	4GR1	4HVP	4MDH	4PHV	4SGA	4TIM	4TLN	TMN	4TS1	4XIA	5ABP
5ACN	5CNA	5CPP	5ENL	5HVP	5ICD	5LDH	5P21	5SGA	5TIM	5TLN	5TMN	5XIA	6ABP
6APR	6CPA	6ENL	6GST	6RNT	6TIM	6TMN	7ABP	7ACN	7CAT	7CPA	7EST	7HVP	7TIM
7TLN	8ABP	8ATC	8CPA	8HVP	8ICD	8XIA	9AAT	9ABP	9HVP	9RUB			

retinol due to the fact that the C═C bond participated in dihedral angles of magnitude 18 and 23 degrees and is quite nonplanar. We note that our method correctly perceived the vinylic groups of the bound FK506 ligand in 1fkb and the retinol in 1rbp.

3. 1dih − the method failed to assign the amide substitutent on the heterocycle of the NAD+ ligand due to a C═O bond length of 1.44 Å; interestingly, the carbonyl carbon's C−N bond had a bond length of 1.32 Å. The resolution of the entry was 2.2 Å, and the NAD+ reactive group was highly strained. One possible improvement to our method would be to postprocess terminal N═C−O groups and convert them to N−C═O.

4. 1mnc − the method incorrectly assigned a double bond to the methylamino methyl in the methylamino-phenylalanyl-leucyl-hydroxamate ligand due to a short bond length of 1.44 Å.

5. 2cgr − the method failed to predict one of the benzene rings in the N-(p-cyanophenyl)-N′-(diphenylmethyl)guanidineacetic acid ligand due to near identical bond lengths of 1.4 Å for the ring carbons and the substitutent to which our method assigned an exocyclic double bond. One possible improvement to our method would be to slightly decrease the preference for higher order bonding according to the nonplanarity of an atom, which would have corrected this situation.

6. 2r04 − the method failed to perceive the (OCCNC) heterocycles in the antivirual agent WIN IV ligand due to a long C−C bond length of 1.52 Å in one of the rings and a short beta exocyclic C−C bond length of 1.44 Å in the other ring. The resolution of the entry was 3.0 Å; however, the ligand only exhibited unusual strain for these heterocycles. It is possible that they were incorrectly parametrized during refinement.

7. 2sns − the method failed to perceive the heterocyle of the thymidine ligand due to a long exocyclic carbonyl bond length of 1.39 Å. That same carbonyl carbon had a C−N bond length of 1.6 Å suggesting a poorly resolved ligand even though the resolution of the PDB entry was 1.5 Å.

8. 3dfr − the method incorrectly assigned *sp* hybridization to the amide nitrogen in the methotrexate ligand due to the nitrogen's bond angle of 143 degrees. In addition the benzene ring was not perceived due to two bond lengths of 1.49 and 1.55 Å. The resolution of the PDB entry was 1.7 Å; however, we suspect inadequate refinement of the ligand geometry. We note that our method correctly perceived the reduced form of NAD also present.

9. 4gr1 − the method incorrectly assigned *sp* hybridization to the amide nitrogen in retro-CSSG ligand due to the nitogren's near linearity (158 degrees) and short N−C bond length of 1.16 Å. In addition, our method failed to perceive an amide group and a carboxylate group due to long bond lengths. The PDB entry's resolution of 2.4 Å together with the strained geometry suggests a poorly resolved retro-CSSG structure. We note that our method correctly perceived the FAD ligand in this entry.

10. 5tln − the method incorrectly assigned a double bond to the beta carbon of the HONH-benzylalonyl-L-alanylglycine-*p*-nitroanilide's alanyl group due to the alpha carbon's apparent near-planarity and a short C−C bond length of 1.35 Å. The coordinates of the *p*-nitroanilide group were absent from the PDB entry due to the absence of electron density.

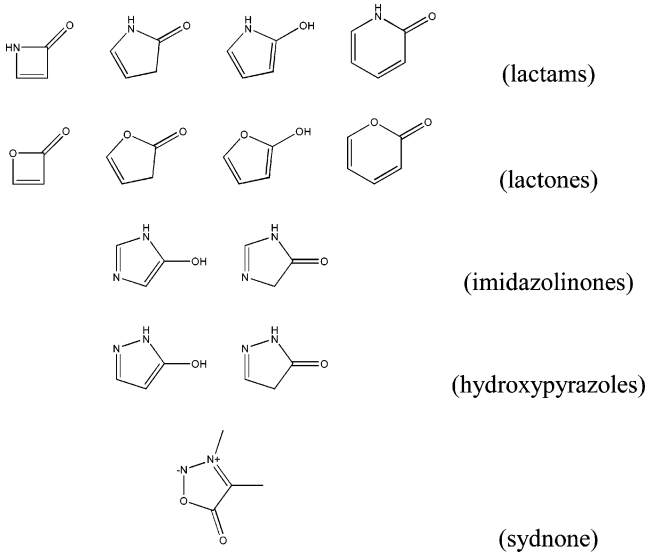


Figure 4. A small collection of possible problematic heterocycles used to investigate bias in the weights given to the Maximum Weighted Matching algorithm.

This together with the resolution of 2.3 Å and generally strained ligand geometry suggests a poorly resolved structure.

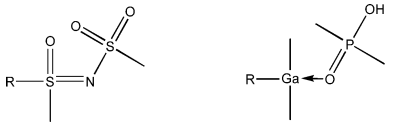
11. 7tln − the method incorrectly assigned a double bond to the N−OH group in the apparently covalently bound CH_2-CO(N−OH)LEU−OCH_3 due to a short bond length of 1.35 Å. Our method correctly perceives this ligand after it was subjected to coordinate optimization with MMFF94. Perhaps an improved single bond/double bond cutoff in the edge weighting scheme would lead to a correct perception of this group.

Out of the 11 failures, five (1aaq, 3df4, 4gr1, and 5tln) were due to questionable geometry; the remaining six failed due to strained or ambiguous geometry, especially unusual bond lengths. Our method uses hard cutoffs to detect single bonds and weight possible double bonds; one possible improvement to our method would be to use a more continuous weighting scheme to judge the more questionable geometries found in the failure cases. However, strained cases will always present problems; for example, ambiguous bonds between C and N will always arise since coordinate or refinement errors may just as easily shorten a C−N bond or lengthen a C═N bond, and if these atoms are both 2-coordinated there may not be sufficient information to make a bond order assignment reliably.

Egregious bond length error is a reasonable justification for failures; however, certain heterocylces are problematic with suppressed hydrogens. In such cases, bond lengths will likely provide little information, and the weights provided to the Maximum Weighted Match algorithm will determine the bond order assignment. To investigate possible unreasonable bias in the weights, a small collection of potentially problematic heterocycles was assembled; these structures are depicted in Figure 4. Each structure was optimized with AM1 as implemented in MOPAC 7,[14] and the hydrogen-suppressed structure was submitted for automatic typing. The lactams were typed correctly; however, it should be noted that, with the exception of the unsaturated 5-ring case, when the ═N−C−OH tautomer geometry was submitted the −N−C═O tautomer resulted. Thus, there is a bias toward the amide tautomer when hydrogens are suppressed. The lactones,

imidazolinones, and hydroxypyrazoles were all perceived correctly. The sydnone was incorrectly perceived: an exocylic C═C double bond was perceived due to the 1.46 Å bond length. Perhaps a softer criterion on bond lengths (instead of the hard cutoffs) would solve this problem. With hydrogens included, the sydnone is correctly perceived. These results suggest that with good geometry, different tautomers can be correctly perceived; however, with hydrogens suppressed difficulties may be encountered with certain heterocycles, which is likely true for the previously published methods. Choosing the "correct" tautomer in the presence of coordinate errors or strained geometries would likely require some sort of dictionary or analysis of the nonbonded environment.

Finally, it should be noted that structures with complex dative bonds present difficulties in the absence of hydrogen atoms. For example, our method applied to the following two structures (with hydrogen atoms suppressed)



results in a correct perception of the molecule on the left and an incorrect perception of the molecule on the right. This is due to the fact that our method detects dative bonding only for atoms with less than full valence coordination. The terminal −OH in the structure on the right was perceived as the coordination bond and the H was lost; such examples require more sophisticated analysis and are beyond the capabilities of the present method (and probably the other published methods).

CONCLUSIONS

We have presented a method for perceiving chemical types of atoms in molecules given 3D atomic coordinates and element identities. The method assigns hybridizations, bond orders, and formal charges for structures with or without hydrogen atoms (although greater accuracy is achieved with hydrogen atoms present). Unlike previous work, no ad hoc chemical group patterns nor ring analyses are used. Instead, the Maximum Weighted Matching algorithm for nonbipartite graphs is used to assign bond orders with weights derived from statistics of a large collection of organic molecules.

The method performed well on a collection of common functional groups and ring structures as well as on examples from the CSD and the PDB. However, the method has some limitations: (a) delocalized dative bond complexes are not handled well; (b) specific tautomeric forms cannot be controlled in hydrogen suppressed structures; (c) perception of the ionization state of transition metals is problematic; and (d) large coordinate errors and strained geometries will lead to perception errors (the most likely cause of failures). Our results suggest that a more continuous weighting scheme (other than the hard cutoffs used) may lead to better performance; this will be the subject of future work.

The method was implemented in the SVL programming language of the Molecular Operating Environment and is available for MOE version 2004.03.

REFERENCES AND NOTES

(1) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; et al. The Development of Version 3 and Version 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−201.

(2) Berstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: A Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.

(3) Hendlich, M. Databases for Protein−Ligand Complexes. *Acta Crystallogr. D. Biol. Crystallogr.* **1998**, *54*, 1178−1182.

(4) Macromolecular Structure Database Project; EMBL − European Bioinformatics Institute; 2003, http://www.ebi.ac.uk/msd.

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

(6) 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.

(7) 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.

(8) Sayle, R. PDB: Cruft to Content (Perception of Molecular Connectivity from 3D Coordinates); *Daylight Chemical Information Systems Inc. MUG'01 Presentation*, 2001, http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html.

(9) Allen, F. H.; Kennard, O.; Watson, D. G. Tables of Bond Lengths Determined by X-ray and Neutron Diffraction. Part 1: Bond Lengths in Organic Compounds. *J. Chem. Soc., Perkin Trans. 2* **1987**, S1−S19.

(10) Halgren, T. A. The Merck Molecular Force Field. I. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(11) Papadimitriou, C. H.; Steiglitz, K. *Combinatorial Optimization*; Prentice-Hall: New Jersey, 1982; pp 247−270.

(12) MOE software available from Chemical Computing Group Inc., Montreal, Canada. Consult http: www.chemcomp.com for further information.

(13) Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N.; Ricketts, D. M. Comparison of Conformations of Small Molecule Structures from the Protein Data Bank with those Generated by Concord, Cobra, ChemDBS-3D and Convertor and those Extracted from the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 905−925.

(14) Stewart, J. J. P. *MOPAC Manual*, 7th ed.; 1993.

CI049915D