Automatic Perception of Organic Molecules Based on Essential Structural Information

Yuan Zhao, Tiejun Cheng, and Renxiao Wang*

State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, 354 Fenglin Road, Shanghai 200032, People's Republic of China

Received January 25, 2007

Format conversion is very common in structure preparation in molecular modeling studies. Unfortunately, format conversion cannot always be executed precisely. We have developed an automatic method, called I-interpret (available on-line at http://www.sioc-ccbg.ac.cn/software/I-interpret/), for interpreting the chemical structure of a given organic molecule merely from its essential structural information, including element identities and three-dimensional coordinates of its component atoms. I-interpret uses standard geometrical parameters of organic molecules in atom/bond-type assignment. A series of elaborate considerations are arranged in a logical sequence for this purpose. I-interpret was tested on a set of 179 small organic molecules from the Protein Data Bank and a set of 1990 organic molecules from the NCI diversity set. On both sets, it achieved a success rate of over 95% in interpreting the correct chemical structures, outperforming other programs under our evaluation. I-interpret also provides users some optional functions, which makes it more flexible and powerful in practice. It may serve as a valuable tool for processing chemical structures in molecular modeling.

INTRODUCTION

Chemical structures need to be presented in certain formats to be processed by computer programs. With respect to organic molecules, essential information in such a format may include identities of atoms and bonds, connection table, atomic coordinates, etc. For obvious reasons, it is difficult to formulate a standard format which is efficient for all molecular modeling purposes. Thus, various formats are currently in existence, each of which has its unique syntax rules and is typically optimized for particular applications. This circumstance unfortunately hampers direct data flow among different programs. To overcome this problem, a decent molecular modeling program typically supports multiple file formats as valid inputs and outputs. Some special programs, such as OpenBabel,[1,2] have also been developed for format conversion.

Conversion a structural file from one format to another, surprisingly, is not a trivial task. It cannot always be executed precisely. One possibility is what we call "hard error", that is, the input format does not include certain information required by the output format. In such a scenario, a direct format conversion will result in an incomplete description of the given chemical structure. Another possibility is what we call "soft error", that is, the input structural file has some defects in its contents. For example, it may have missing sessions or contain minor errors. This occurs very often in reality because one has to deal with structural files from various resources, some of which are simply not compiled carefully enough. In such a scenario, a direct format conversion cannot identify and fix such problems in the input file and therefore may not be able to produce the correct output file. Because of these difficulties, we find that even specially developed format-converting programs often fail to produce satisfactory results. Structural preparation is typically the very first step in any molecular modeling study. Incorrectness in structural preparation may lead to uncertainties in consequent analyses. This type of uncertainty is difficult to find especially in high-throughput studies, such as database processing, which can be quite disturbing.

We believe that an ideal format-converting program should be able to correctly interpret a given chemical structure only with the very essential information from the input. In this way, format conversion will be least affected by the incompatibility between different formats or the defects in inputs. As a matter of fact, some researchers have also realized this. Since the 1990s, a number of approaches[3−8] have been reported for interpreting chemical structures merely from atomic identities and coordinates. Meng's[3] and Baber's[4] methods perceived atom types and bond types according to standard bond angles and lengths. Hendlich[5] and Sayle[6] introduced additional procedures for recognition of functional groups and hybridization states to improve accuracy. While all of the above methods were based on analysis of bond lengths and angles, Labute[7] and Froeyen[8] applied the maximum weighted matching algorithm and the octet rule, respectively, to assign bond types. These approaches collectively represent an interesting trend in chemical structure processing.

Aiming at the same problem, we have developed a new program which is able to interpret a given chemical structure with only atomic identities and three-dimensional coordinates. This program is a component of our in-house software package, namely, the Integrated Toolkits for Drug Design (ITDD), and therefore will be referred to as I-interpret throughout this paper (available on-line at http://www.sioc-ccbg.ac.cn/software/I-interpret/). As validated on two data sets, it achieved higher overall success rates than the other programs

* To whom correspondence should be addressed. E-mail: wangrx@mail.sioc.ac.cn. Phone: 86-21-54925128.

10.1021/ci700028w CCC: $37.00
Published on Web 05/27/2007

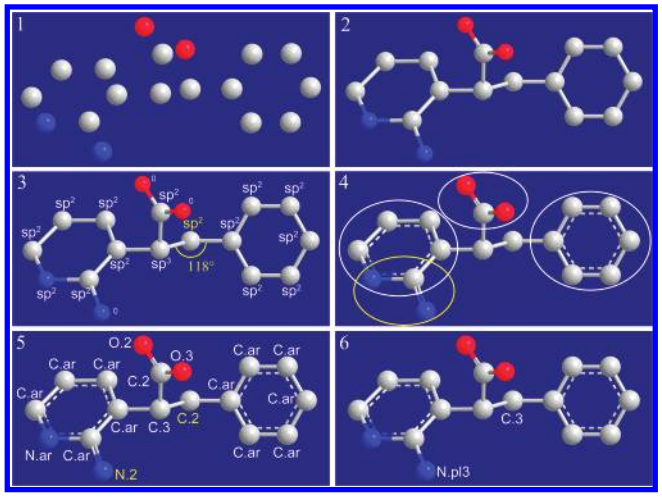


Figure 1. Interpreting the chemical structure of an organic molecule from nude model. (1) Only identities and three-dimensional coordinates of all non-hydrogen atoms are needed as inputs. (2) Covalent connections in the given molecule are established. (3) Hybridization states are assigned to atoms. Terminal atoms remain unresolved until later steps. (4) Functional groups and aromatic rings are identified. (5) All bonds and atoms are resolved. Two conflicts remain in the structure, which are colored in yellow. (6) Conflicts are resolved by resetting certain atom types and bond types.

under our evaluation. I-interpret also provides a number of optional functions to users, which makes it more powerful and flexible in practice. We expect that I-interpret will be applied as a valuable tool for processing chemical structures in various molecular modeling studies.

METHODS

On the Input. The very essential information of a three-dimensional molecular structure includes the element identity and Cartesian coordinates of each component atom, which is basically what one can observe directly from structural determination, such as X-ray diffraction. We call a molecular structure with only such information a "nude model" in this paper. Our program, that is, I-interpret, is designed to interpret the correct chemical structure of a given organic molecule from its nude model. The PDB format[9] was chosen in our study for recording nude models of organic molecules. This format was originally designed for characterization of biological macromolecules, such as proteins and nucleic acids. When used for normal organic molecules, it only provides information of atomic identities and coordinates, which makes it ideal for our purpose. Our program may also accept other formats as valid inputs because apparently all file formats for presenting three-dimensional molecular structures must contain the essential information required by our program.

Basic Algorithm. Our method consists of multiple steps. Each major step will be explained below. An example is given in Figure 1 to illustrate the entire procedure.

(1) Detect Connectivity. Given the coordinates of all component atoms, the first step of our method is to detect the covalent connections within the given molecule. A covalent bond between atoms i and j is recorded if

$$d_{ij} < R_i + R_j + 0.4$$

where d_{ij} is the distance between the two given atoms and R_i and R_j are the covalent radii of atom i and j, respectively.

Table 1. Covalent Radii (Å) Used in I-interpret[a]

atom	radius	atom	radius	atom	radius	atom	radius
Ag	1.59	F	0.64	Mo	1.47	Se	1.22
Al	1.35	Fe	1.34	N	0.68	Si	1.20
Ar	1.51	Ga	1.22	Na	0.97	Sm	1.80
As	1.21	Gd	1.79	Nb	1.48	Sn	1.46
Au	1.50	Ge	1.17	Ni	1.50	Sr	1.12
B	0.83	H	0.23	O	0.68	Ta	1.43
Ba	1.34	Hf	1.57	Os	1.37	Tb	1.76
Be	0.35	Hg	1.70	P	1.05	Te	1.47
Br	1.21	Ho	1.74	Pb	1.54	Ti	1.47
C	0.68	I	1.40	Pd	1.50	Tl	1.55
Ca	0.99	In	1.63	Pr	1.82	U	1.58
Cd	1.69	Ir	1.32	Pt	1.50	V	1.33
Cl	0.99	K	1.33	Rb	1.47	W	1.37
Co	1.33	Kr	1.50	Re	1.35	Xe	1.50
Cr	1.35	La	1.87	Rh	1.45	Y	1.78
Cs	1.67	Li	0.68	Ru	1.40	Yb	1.94
Cu	1.52	Lu	1.72	S	1.02	Zn	1.45
D	0.23	Mg	1.10	Sb	1.46	Zr	1.56
Eu	1.99	Mn	1.35	Sc	1.44		

[a] Cited from the publications by the Cambridge Crystallographic Data Center (ref 10).

Table 2. Maximal Valences Used in I-interpret

atom	max valences	atom	max valences	atom	max valences
H	1	As	4	F	1
B	3	O	2	Cl	4
C	4	S	4	Br	4
Si	4	Se	4	I	4
N	4	Te	4		
P	4	At	4		

All data are given in angstroms. The covalent radii of common elements used in our program are summarized in Table 1, which are cited from the publication from the Cambridge Crystallographic Data Center.[10] Covalent connections identified by the above judgment, however, may contain false ones because of distortion from ideal geometry in the given molecule. Our method then examines the valence, that is, total number of covalently connected neighbors, of each atom to determine if it violates the theoretical limit and consequently breaks the false covalent connections. The maximal valences of nonmetal elements used in our method are summarized in Table 2.

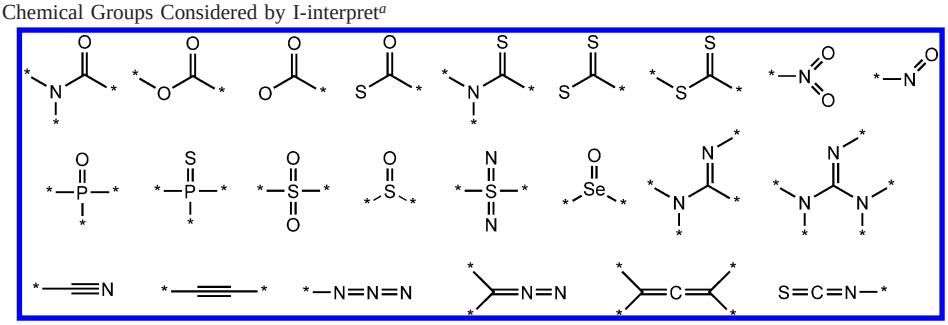
After the connection table is established, our method determines if there are any rings in the given structure according to the Cauchy formula[11]

$$N_{\text{ring}} = N_{\text{bond}} - N_{\text{atom}} + N_{\text{segment}}$$

where N_{ring}, N_{bond}, N_{atom}, and N_{segment} are the numbers of smallest rings, bonds, atoms, and segments in the given structure, respectively. Note that N_{segment} equals to 1 for an integrated molecule. It could be greater than 1 if the given structure contains multiple isolated segments. If N_{ring} is not zero, our method detects all of the rings in the given molecule using an algorithm similar to the one used in Corey's approach.[12] In brief, terminal and branched structures are "collapsed" first in an iterative manner. Rings are then detected among remaining structures through a depth-first searching algorithm. All of the rings that have been detected are recorded by their sizes for later steps.

(2) Determine the Hybridization States of Atoms. An initial hybridization state is assigned to each atom by consideration

Scheme 1. Common Chemical Groups Considered by I-interpret[a]



[a] Asterisk denotes any group.

of the geometries of the covalent bonds connecting this atom. An atom will be marked as sp-hybridized if it is the center of a bond angle greater than 155°. Otherwise, it will be marked as sp^2-hybridized if it is the center of a bond angle greater than 115°. Other atoms will be marked as sp^3-hybridized. All terminal atoms remain unclassified until later steps because they cannot be the centers of any bond angles. Hydrogen atoms, if given in the input, are marked as sp^3-hybridized at this step.

The above algorithm may fail to assign correct hybridization states to certain atoms in rings. For example, all of the bond angles in pyrrole are smaller than 110 degrees. Consequently, atoms in pyrrole will be mistakenly marked as sp^3-hybridized. To tackle this problem, we have implemented special considerations for cyclic structures. For a five-member or smaller ring, it will be identified as planar if the average in-ring torsion angle of this ring is smaller than 7.5°. All component atoms in this ring will be marked as sp^2-hybridized. For a six-member or larger ring, the cutoff is 15°.

(3) Recognize Functional Groups. The advantage of defining functional groups is that multiple atoms/bonds can be resolved simultaneously. The recognition results are also more robust because such a process relies less on the geometry of a given molecule. Our program carries out this task based on the hybridization states of all component atoms assigned at the previous step. A functional group will be recognized if a group of atoms match a certain pattern defined in our program. If so, the appropriate hybridization states, atom types, and bond types of this group of atoms will be assigned accordingly. A complete list of the functional groups defined in our program is summarized in Scheme 1. As demonstrated later in this paper, our method has an overall promising performance with the use of these functional groups. Nevertheless, this list of functional groups in our method can be easily expanded to produce even more accurate results in structure processing.

Aromatic rings can be considered as a special class of functional groups. Our method only considers planar cyclic structures as valid candidates for aromatic rings. If a planar ring is found to contain two consecutive single bonds, it will not qualify for an aromatic ring either. Then, the classical Huckel's rule of $4n + 2$ π-electrons is adopted to determine if a given ring is aromatic. Once a ring is identified as aromatic, all bonds in this ring will be set in either the Kekule mode, that is, in alternating single and double bonds, or the delocalized mode upon user's choice.

(4) Determine Bond Types. All of the covalent bonds that remain unresolved so far are analyzed at this step. Their types are set according to their lengths and the dihedral angles in which they are involved. A given bond will be interpreted as a single bond if it is the center of a dihedral angle greater than 30°. A given bond will be interpreted as a double bond if it connects at least one sp^2-hybrized atom and

$$d_{ij} < L_{ij} - 0.10$$

Here, d_{ij} is the length of the given bond between atoms i and j, and L_{ij} is the standard length of a single bond between atoms i and j. Both d_{ij} and L_{ij} are given in angstroms. Similarly, a given bond will be interpreted as a triple bond if it connects at least one *sp*-hybridized atom and

$$d_{ij} < L_{ij} - 0.25$$

Table 3. Standard Lengths (Å) of Single Bonds Used in I-interpret[a]

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

[a] Cited from the MMFF94 force field implemented in the SYBYL software.

The standard single bond lengths used in our method (Table 3) are cited from the MMFF94[13] force field as implemented in the SYBYL software. After all of the above analyses, the remaining bonds are set as single bonds. Once all of the bond types are set, the hybridization states of the atoms that remain unresolved so far, such as the terminal atoms, are set by consideration of the types of their neighboring bonds.

(5) Resolve Conflicts. As one may have noticed, atomic hybridization states and bond types are derived largely in an independent manner through the above steps. In some cases, this leads to conflicts in the resulting chemical structure. Such conflicts often occur in delocalized chemical moieties. To illustrate this issue, a molecule containing the 2-amino-pyridine moiety is shown in Figure 1. The exocyclic amino group is partially conjugated with the aromatic ring, and thus the bond connecting this amino group is often considerably shorter than a typical single bond. Thus, this bond will be interpreted as a double bond at step 4, which makes the valences on a relevant carbon atom exceed theoretical limit (Figure 1).

Therefore, it is necessary to conduct an additional examination on the resulting chemical structure after all previous

steps to identify and resolve conflicts. Our method applies the following priority rule: functional groups > bond types > atom types. Features with lower priorities will be considered first to be reset to resolve conflicts. For example, the incorrect exocyclic double bond shown in Figure 1 will be changed to a single bond to keep the aromatic ring intact. The other conflict shown in Figure 1 will be resolved by changing the hybridization state of a troublesome carbon atom from sp^2 to sp^3 to keep related bonds intact.

Optional Functions. A variety of optional functions are also provided by our program, I-interpret, for processing chemical structures. These optional functions are designed to offer the users greater convenience in practice.

(1) Ability to Utilize Additional Structural Information. In addtion to atomic identities and coordinates, sometimes a connection table is also available from input, providing explicit information of each bond in a given molecule. One such example is the MDL SD format. If user decides to trust such information from the input, I-interpret can use it as well in processing a given molecule. Since the identity of each bond is already provided in such a case, the remaining main task is to set the appropriate type of each atom required by the output format, which is relatively straightforward.

(2) Adjustable Geometrical Parameters. I-interpret uses a number of cutoffs of bond lengths, bond angles, and dihedral angles in its algorithm. These default parameters were derived from a statistical analysis of over 400 000 organic molecules, which were structurally optimized by the MMFF94 force field, in the MDL Available Chemical Directory database. All of these cutoffs can be adjusted by users through an external parameter file.

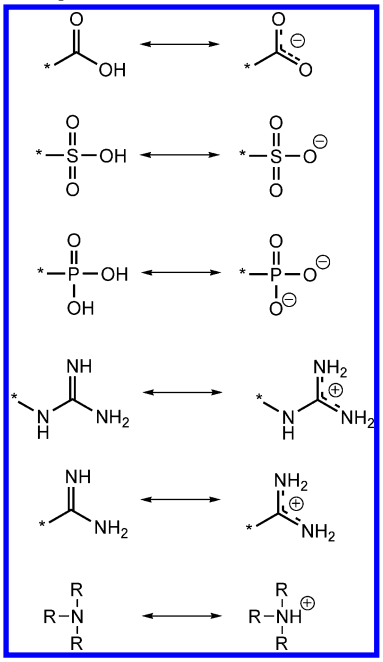
(3) Ability to Fill Up Hydrogen Atoms. I-interpret is designed to interpret the chemical structure of a given molecule correctly even if hydrogen atoms are not provided implicitly. It certainly works when hydrogen atoms are provided. Upon user's choice, I-interpret can fill up the hydrogen atoms on an organic molecule using standard bond lengths and angles.

(4) Ability to Process Metal-Containing Organic Molecules. Few format-converting programs are able to handle such molecules. For the sake of convenience, I-interpret labels all coordinate bonds formed between metal atoms and organic molecules as single bonds in its outputs.

(5) Adjustable Protonation States. User is allowed to specify the protonation states of certain chemical groups in the final outputs. Some chemical groups are not in their neutral forms under a physiological pH condition, such as carboxylic acids, sulfuric acids, aliphatic amino groups, guanidines, and others (Scheme 2). This is of course a very crude method for setting the protonation states of a given molecule. In the future, we plan to implement in I-interpret a more rigorous method for prediction of pK_a values for this purpose.

(6) Multiple Supported Formats. I-interpret can output results in either the MDL SD format or the tripos Mol2 format. These two particular formats are chosen because they are accepted by many molecular modeling programs. These two formats provide concise yet fairly complete descriptions of the chemical structures of organic molecules. If necessary, they can be converted into other formats, with the aid of another program, without losing critical information. Future versions of I-interpret will support more formats directly.

Scheme 2. Chemical Groups with Neutral and Charged Forms Considered by I-interpret[a]



[a] Asterisk denotes any group; R denotes an aliphatic group.

RESULTS AND DISCUSSION

To assess the performance of I-interpret, we tested it on the same data set used in Labute's study,[7] an assembly of 179 entries from the Protein Data Bank (Table 4). These entries are all protein−ligand complexes which are "commonly used for the training of docking and scoring functions" as described by Labute. The ligand in each complex was retrieved from the original structural file from PDB and then saved in a PDB-format file. During this process, connectivity information in the original structural file, if any, was ignored so that the resulting PDB-format file of each ligand only contained information of atoms. Then, I-interpret was applied to these specially compiled PDB-format files. The outcome in each case was saved in a Mol2-format file. It was visually inspected and compared to the correct chemical structure of the given molecule. The correct chemical structure of each molecule can be found in the primary reference cited in the corresponding PDB file. Among all 179 molecules, our program correctly interpreted the chemical structures of 170 molecules, achieving a success rate of 95.0% (Table 5). Our program failed in 9 cases, while Labute's method was reported to fail in 11 cases in a similar test.

The 9 failed cases for our method are summarized in Table 6, which can be roughly classified into two groups. In the first group, including entries 1CPS, 1ETR, 1NNB, 2R04, and 4GR1, hybridization states of certain atoms were misjudged. Consequently, some unsaturated bonds were mistakenly interpreted as saturated bonds or vice versa. We noticed that the three-dimensional structures of these molecules, especially in the cyclized parts, exhibited notable distortions from ideal geometries, which were probably attributed to the uncertainties in structure determination. This accounts for the failure of I-interpret in these cases. Note that a set of relatively tight criteria are implemented in I-interpret for determination of the hybridization states of atoms by examination of relevant bond lengths, bond angles, and dihedral angles. We found that if a set of less rigorous criteria

Table 4. PDB Entries Used in Evaluation[a]

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	4TMN	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	

[a] Cited from Labute's study (ref 7).

Table 5. Success Rates of Three Methods on the 179 Molecules from PDB

	success rates	
	original structures from PDB as inputs	optimized structures as inputs[a]
I-interpret	95.0%	100%
Labute's method	93.9%[b]	94.4%[b]
OpenBabel	79.3%	86.6%

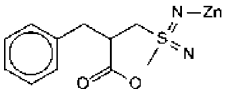
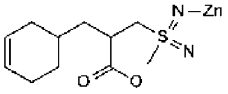
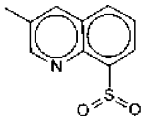
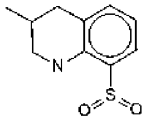
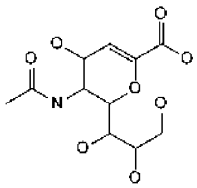
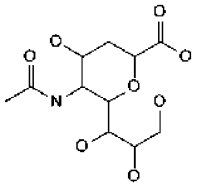
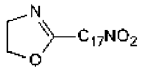
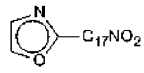
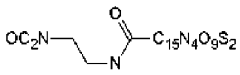
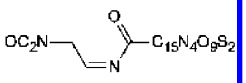
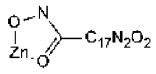
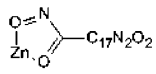
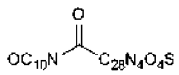
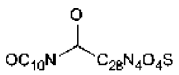
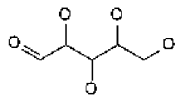
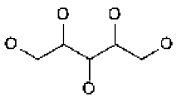
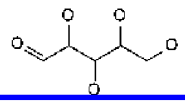
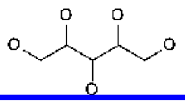
[a] All molecules were subjected to structural optimization by using the MMFF94 force field in SYBYL. [b] Results cited from Labute's study (ref 7).

were applied for instead, some of these failed cases could be corrected without jeopardizing the results of any other molecules in this test set. However, it may lead to more failures on other data sets. Therefore, our program still adopts those relatively tight criteria by default. The second group includes entries 1MNC, 1RNE, 2XIM, and 8XIA. In each of these molecules, a double bond connecting a terminal oxygen atom was misjudged as a single bond or vice versa. Given the fact that hydrogen atoms are unavailable from input, I-interpret may make mistakes when interpreting a terminal bond especially when the input is not a high-resolution structure. Sometimes, the nature of such a terminal bond is ambiguous by itself. For example, the ligand from PDB entry 1MNC has a hydroxamate moiety which forms coordinate bonds with a zinc ion in the binding pocket of neutrophil collagenase. The O−N bond in this moiety is considerably shorter than a standard single bond. It is not surprising that our program interpreted it as a double bond.

We then took the correct chemical structures of the nine problematic cases and subjected them to structural optimization using the MMFF94 force field implemented in SYBYL.[14] The resulting models were saved in the PDB format and were processed by I-interpret once more. With these optimized structures, all nine molecules were interpreted correctly, implying that the errors made by I-interpret in the previous test were caused primarily by the faulty geometries of input structures. Labute also did the same experiment on the 11 failed cases using his method. Only one molecule was correctly processed by his method after structural optimization.

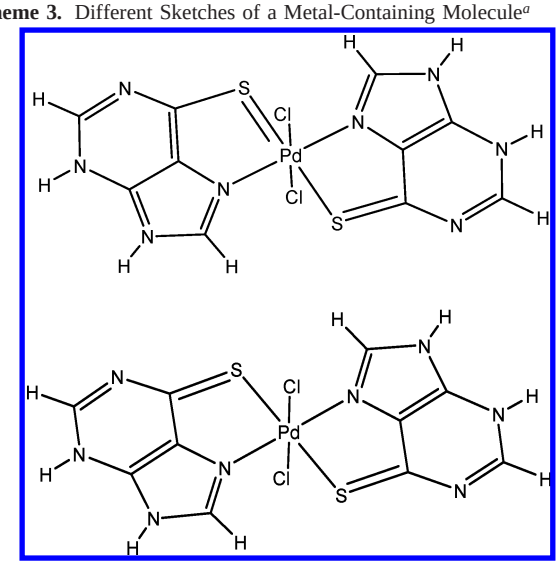
We also included OpenBabel (version 2.0.0) [1,2] in our evaluation. OpenBabel is a popular publicly available program, which is developed for format conversion and some other purposes. We found that OpenBabel did not function properly when converting a PDB-format file into a Mol2-format file. Thus, we applied OpenBabel to convert the PDB-format files of the 179 molecules in Labute's test set into SD-format files first and then into Mol2-format files. The final outcomes were considered in our evaluation. In this

Table 6. Chemical Structures Perceived Incorrectly by I-interpret

PDB Code	Correct structure [a]	Perceived structure
1CPS		
1ETR		
1NNB		
2R04		
4GR1		
1MNC		
1RNE		
2XIM		
8XIA		

[a] Chemical structure of the ligand in the given protein−ligand complex.

Scheme 3. Different Sketches of a Metal-Containing Molecule[a]



[a] Above: The one in the original SD file from NCI. Below: The one produced by I-interpret.

test, OpenBabel failed to interpret the correct chemical structures in 37 cases, achieving a success rate of 79.3% (Table 5). If the three-dimensional structures of these 37 molecules were optimized prior to processing, OpenBabel still failed to process 24 of them correctly. I-interpret apparently outperformed OpenBabel in this test.

The NCI structural diversity set (http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html) was used as another test set in our evaluation. This data set was derived from ~140 000 selected molecules from the NCI database. It consists of 1990 druglike molecules, each of which has a unique structural scaffold. This data set obviously represents a much greater structural diversity than Labute's set. This data set was downloaded from the NCI web site as SD-format files. They were converted to PDB-format files first so that only essential structural information still remained. Our program was then applied to process these PDB-format files. The outcomes were saved in SD-format files. The resulting SD file of each molecule was first compared to the original one from NCI with a computer program. If the two SD files were found not identical, we then inspected them visually to judge if the SD file produced by our program was correct or not. In some cases, discrepancy between the two SD files does not necessarily signal an error made by our program. For example, a nitro group sometimes is saved as the dative bond form in the original SD file, while our program saves it as the neutral form, both of which are valid. For metal-containing molecules, coordinate bonds between metal atoms and other atoms are sometimes labeled as double bonds in the original SD files from NCI, while our program always labels such bonds as single bonds for convenience (Scheme 3). These cases were not counted as errors made by our program.

After careful examination, we concluded that the chemical structures of 1954 molecules among all 1990 molecules were correctly perceived by I-interpret, achieving a high success rate of 98.2%. The success rate was 98.3% when the Mol2-format was chosen as output for instead (Table 7). We examined the 36 problematic cases and found many of them were in fact caused by the faulty geometries of given molecules. This is understandable because the three-dimensional structures of the molecules in the NCI database were not experimentally determined but computer-generated models. If such cases are ignored, our program has achieved a very high success rate in this test.

Table 7. Success Rates of Two Methods on the 1990 Molecules from the NCI Diversity Set

format conversion conducted	success rates	
	I-interpret	OpenBabel
PDB to SD	98.2%	91.6%
PDB to Mol2	98.3%	82.3%
SD to Mol2	100%	91.0%

We also tested I-interpret by using the original SD-format files of those 1990 molecules as inputs, that is, interpreting chemical structures with additional structural information from inputs. In this test, no error was found in the resulting Mol2-format files produced by our program. Interestingly, I-interpret prompted that the original SD files of nine molecules contained certain errors in atom/bond types, that is, soft errors. Such errors in inputs, however, did not prevent our program from perceiving the correct chemical structures of these molecules. This finding supports our statement that public data sets often contain errors. Thus, an ideal format-converting program should be robust enough to tolerate such errors rather than just execute a straightforward translation.

We also tested OpenBabel on the NCI diversity set. The first test was to apply OpenBabel to convert the PDB-format files of this data set into SD-format files. In this test, a total of 168 molecules were found to contain at least one error in atom/bond type assignment in final outputs, yielding a success rate of 91.6%. The second test was to convert the PDB-format files of this data set into Mol2-format files. As mentioned earlier, for some reasons OpenBabel does not function properly in such a conversion. Therefore, technically this task was completed by converting PDB-format files into SD-format files first, which had been completed in the first test, and then into Mol2-format files. In this way, a total of 352 molecules were found to contain at least one error in atom/bond type assignment in final outputs, yielding a success rate of 82.3%. When OpenBabel was applied to convert the original SD-format files from NCI into Mol2-format files, a total of 179 molecules were found to contain at least one error in final outputs, yielding a success rate of 91.0% (Table 7). Apparently, I-interpret conducted more precise conversions either way.

In summary, I-interpret has been tested on two sets of diverse organic molecules. On Labute's test set, its performance was at least comparable to that of Labute's method; on both sets, it outperformed OpenBabel, one of the most popular format-converting programs. Compared to the methods developed by other researchers, ours is relatively straightforward to understand. Our method uses standard geometrical parameters of organic molecules in decision making. A series of steps are arranged in a logical sequence, which seems to be the key factor accounting for its promising performance.

I-interpret still produces errors on some chemical moieties showing tautomerism. Our future efforts will focus on this issue. If using nude models as inputs, I-interpret may not correctly perceive chemical moieties with apparent charges, such as ylides and dative bonds. Awareness of this limitation

will help users apply I-interpret properly. In addition, the current version of I-interpret only considers the PDB format, the Tripos Mol2 format, and the MDL SD format as valid inputs and outputs. If necessary, it can be expanded to support other formats as well.

CONCLUSIONS

Our study has demonstrated that it is possible to interpret the chemical structure of a given molecule with high accuracy based on the essential structural information of it. As demonstrated on two sets of diverse organic molecules, the performance of I-interpret apparently surpassed OpenBabel, one of the most popular structure-processing programs. Since structure preparation is normally the starting point of a molecular modeling study, we expect that I-interpret will find its application in various molecular modeling studies. Its high accuracy in format conversion makes it particularly suitable for high-throughput projects, such as processing large databases of organic molecules. As we have emphasized, format conversion should be executed in an intelligent and robust manner rather than a straightforward translation. We believe that this is the direction for a new generation of format-converting programs.

ACKNOWLEDGMENT

The authors are grateful for financial support from the Chinese National Natural Science Foundation (Grant 20502031) and the Chinese Ministry of Sciences and Technology (the 863 program, Grant 2006AA02Z337). The authors are also grateful for the technical aid provided by Chunni Lu and Weiqi Zhang at the Shanghai Institute of Organic Chemistry.

REFERENCES AND NOTES

(1) Guha, R.; Howard, M. T.; Hutchison, G. R.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Wegner, J.; Willighagen, E. L. The Blue Obelisk—Interoperability in Chemical Informatics. *J. Chem. Inf. Model.* **2006**, *46*, 991−998.

(2) Open Babel: The Open Source Chemistry Toolbox. http://openbabel.sourceforge.net (accessed March 28, 2006).

(3) 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.

(4) 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.

(5) 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.

(6) Sayle, R. PDB: Cruft to Content (Perception of Molecular Connectivity from 3D Coordinates); Presented at MUG'01, Santa Fe, NM, March 6−9, 2001; http://www.daylight.com/meetings/mug01/Sayle/m4xbondage.html (accessed Apr 10, 2006).

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

(8) Froeyen, M.; Herdewijn, P. Correct Bond Order Assignment in a Molecular Framework Using Integer Linear Programming with Application to Molecules Where Only Non-Hydrogen Atom Coordinates Are Available. *J. Chem. Inf. Model.* **2005**, *45*, 1267−1274.

(9) 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.

(10) Allen, F. H. The Cambridge Structural Database: A Quarter of a Million Crystal Structures and Rising. *Acta Crystallogr.* **2002**, *B58*, 380−388.

(11) Petitjean, M.; Fan, B. T.; Panaye, A.; Doucet, J. P. Ring Perception: Proof of a Formula Calculating the Number of the Smallest Rings in Connected Graphs. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1015−1017.

(12) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules. *J. Am. Chem. Soc.* **1972**, *94*, 431−439.

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

(14) *SYBYL*, version 7.1; Tripos Inc.: St. Louis, MO, 2005; http://www.tripos.com.

CI700028W