Three-Dimensional Chemical Structure Search Using the Conformational Code for Organic Molecules (CCOM) Program

Hiroshi Izumi,*[a] Laurence A. Nafie,[b][c] and Rina K. Dukor[c]

Abstract: Searching of three-dimensional structural fragments of organic molecules is challenging because of structural differences between X-ray and theoretically calculated geometries and the conformational flexibility of substituents. The codification program called Conformational Code for Organic Molecules (CCOM) can be used to unambiguously convert three-dimensional conformational data for various molecules to one-dimensional data. Two deviations from Rule E-5.6 of the IUPAC Rules for Nomenclature of Organic Chemistry were introduced to the CCOM program for three-dimensional fragment searching. First, the search for the highest priority atom was limited to a distance of two bonds from the center bond for dihedral angle determination. Second, for indistinguishable atoms in experimentally observed solution structures, the smallest number of atom index in the molecular model was chosen as the priority atom for dihedral angle determination. A search of the three-dimensional conformational fragment ***mb***_3a6c4c of mevastatin (**1**) in combination with the SMARTS description suggested that a change in the conformation of this fragment may be the driving force for dissociation of mevastatin from its target protein.

Keywords: Mevastatin, Pravastatin, 3D Fragment Search, SMARTS, SMILES, Nomenclature, Python, Pybel, SDF Format, Vibrational Circular Dichroism

[a] *Dr. H. Izumi*
National Institute of Advanced Industrial Science and Technology (AIST)
AIST Tsukuba West, 16-1 Onogawa, Tsukuba, Ibaraki 305-8569, Japan
Fax: (+)81-29-861-8252
E-mail: izumi.h@aist.go.jp

[b] *Prof. L. A. Nafie*
Department of Chemistry
Syracuse University
Syracuse, New York 13244-4100

[c] *Prof. L. A. Nafie and Dr. R. K. Dukor*
BioTools Inc.
17546 SR 710 (Bee Line Hwy) Jupiter, Florida 33458

Introduction

Many software programs for visualizing organic molecules, such as GaussView,[1] can be used to represent chirality by means of a one-button operation. However, no program can automatically describe molecular conformations such as *trans* and +*gauche*. Further, in contrast to searching of two-dimensional chemical structures using Simplified Molecular Input Line Entry Specification (SMILES) and SMiles ARbitrary Target Specification (SMARTS) notations,[2] searching of three-dimensional (3D) conformational fragments of organic molecules remains unexplored.

The function of a molecule is strongly related to its conformation; in particular, the binding of pharmaceutical ligands to their targets is regulated by conformational changes.[3–6] Despite the importance of conformation, only the superimposed method is used for conformational comparisons. However, this method cannot be effectively used to study quantitative structure–activity relationships, and therefore the codification of conformations may be useful for optimizing drug development. Further, codification can be expected to improve our understanding of the complex effects of conformational changes.

For this purpose, we have proposed a conformational code with a one-to-one correspondence between conformation and code for the description of conformations of many kinds of chemical compounds; the code is based on structural analysis of chiral bioactive compounds by means of vibrational circular dichroism (VCD).[7–11] For the extraction of the four atomic coordinates necessary for determination of a dihedral angle, Rule E-5.6 of the IUPAC Rules for Nomenclature of Organic Chemistry is used.[10,12] However, this rule was subsequently found to be unsuitable for the selection of four atoms for searching of 3D conformational fragments of organic molecules. Unlike the chiral descriptors D and L, the *R* and *S* descriptors are based on the relative location of the atoms connected to a chiral center and are unsuitable for comparisons between molecules. We have now optimized the conformational code for the purpose of 3D fragment searching.

In this paper, we describe the codification program referred to as Conformational Code for Organic Molecules (CCOM), which uses the Python programming language, and we present the first visualization of conformational codes on molecular models. The difference between the conformational code rules and IUPAC rules for conformations is also described. Furthermore, we discuss the use of the code for a search of a 3D conformational fragment of hypolipidemic agent mevastatin (**1**) and the involvement of the fragment in the dissociation of the molecule from its target protein.

The new and still unpublished results in this paper are the followings: (1) the extraction

of characteristic 3D conformational fragments from hundreds of structural files of a wide range of pharmaceutical molecules, (2) a detailed description of the relationship between conformational fragments and function with CCOM as shown in Figure 6A, and (3) new deviation rules of CCOM from Rule E-5.6 of the IUPAC Rules for Nomenclature of Organic Chemistry[10,12] as shown in Figure 3 and Figure 4.

Materials and Methods

Programming

The CCOM program and the Molecular3DView visualization program were constructed by using various functions of Pybel and Open Babel,[13] wxPython,[14] VPython,[15] and Avogadro.[16] Automatic conversion of conformational information to codes was carried out by means of two procedures: (1) extraction of four atomic coordinates for determination of dihedral angles and (2) conversion from dihedral angles to conformational codes using dot products and cross products of vectors and plane equations. Rule E-5.6 of the IUPAC Rules for Nomenclature of Organic Chemistry[10,12] was the main rule used for the first process, but the following two differences were introduced for searching 3D fragments, as described in Results and Discussion: (1) the search for the highest priority atom was limited to a distance of two bonds from the center bond for dihedral angle determination, and (2) for indistinguishable atoms in experimentally observed solution structures, the atom with the smallest number of atom index[13] (a technical term of Open Babel) was selected as the priority atom for dihedral angle determination. Finally, the converted conformational codes and Gibbs free energies for the conformations were arranged in structure-data file (SDF) format.

Calculations

All geometry optimizations, conformer searches, and calculations of vibrational frequencies, absorption intensities, and VCD intensities were carried out using the Gaussian 03 program[17] on a Pentium 4 (3.2 GHz) PC. Density functional theory with the B3LYP functional and 6-31G(d) basis set was used for the calculations.

Results and Discussion

CCOM Program

The conformational code is composed of a combination of codes for regional angle locations and codes for conformational elements (Figure 1), and the conformational elements representing the classification of dihedral angles are substituted for the symbols indicating bond locations, such as b1–b9 for mevastatin (**1**).[9] The generic notation for **1** from Figure 1A is (*mb*_b1b2b3)*npMe*_b4b5(*hope*_b6b7b8b9) where for a

particular conformation the individual bonds b1, b2, etc., are replaced by conformational elements given in Figure 1B that specify the range of conformational angle of that bond, for example, b1 becomes 3a, and so on for the other eight conformational defining bonds in this molecule. The conformational elements 1, 2, 3, 4, 5, and 6 correspond to the following conformational terms: *ap* (*antiperiplanar*), *+sc* (+*synclinal*), *−sc* (−*synclinal*), *sp* (*synperiplanar*), +*ac* (+*anticlinal*), and −*ac* (−*anticlinal*), respectively (Figure 1B). The letters c and a in the conformational elements designate clockwise and anticlockwise, respectively, and the local structural differences between the optimized conformations determined by means of the density functional theory calculations can be discriminated conclusively by classification of the dihedral angles into 12 segments similar to those on the face of a clock (Figure 1B). The 12 segments were introduced for discrimination between conformers in the occasional cases in which two conformations with different Gibbs free energies can be optimized in the same region, as is the case for *ap* (1) and *sp* (4) by means of theoretical calculations.[9]

For the CCOM program, Rule E-5.6 of the IUPAC Rules for Nomenclature of Organic Chemistry[10,12] was the main rule used, and the following criteria were used to select an atom or group to define the torsion angle: (1) If all the groups or atoms were different, Rule E-4.9 for the definition of chirality as *R* or *S* was applied, and the group or atom with the highest priority was selected. (2) If all but one of the groups or atoms were the same, the unique group or atom was selected. (3) If all the groups or atoms were the same and thus could not be distinguished, the group or atom with the smallest torsion angle was selected. Rule E-4.9 defines that chiral compounds in which the absolute configuration is known are differentiated by the stereodescriptors *R* and *S* assigned by the sequence-rule procedure: (1) A substituent with a higher atomic number takes precedence over a substituent with a lower atomic number. Hydrogen is the lowest possible priority substituent, because it has the lowest atomic number. (2) If there are two substituents with equal rank, proceed along the two substituent chains until there is a point of difference. (3) If a chain is connected to the same kind of atom twice or three times, check to see if the atom it is connected to has a greater atomic number than any of the atoms that the competing chain is connected to. If none of the atoms connected to the competing chain(s) at the same point has a greater atomic number: the chain bonded to the same atom multiple times has the greater priority. If however, one of the atoms connected to the competing chain has a higher atomic number: that chain has the higher priority. Being double or triple bonded to an atom means that the atom is connected to the same atom twice.

The CCOM program satisfies the sequence-rule procedure. For the carboxylic acid

moiety, the C=O oxygen, not the O-H oxygen, was selected according to Rule E-4.9(2), whereas for the carboxylic acid ester moiety, the C–O–C oxygen, not the C=O oxygen, was selected according to Rule E-4.9(3), as shown in Figure 2. In the figure, the codes are indicated on stick models generated by using a VPython function.[15] However, Rule E-5.6 was subsequently found to be unsuitable for selecting four atoms for searching of 3D conformational fragments of organic molecules as stated in the introduction.

Two deviations from Rule E-5.6 were introduced for the purpose of searching 3D fragments. First, the search for the highest priority atom did not extend farther than a distance of two bonds from the center bond for dihedral angle determination. For example, in the structure shown in Figure 3, the methylene carbon atom in the dotted-line box was selected from Rule E-5.6(1), however, the different atom selection occurred in occasional cases for searching of 3D conformational fragments of organic molecules. Therefore, the methyl carbon was selected as the highest priority atom because all the other groups within a distance of two bonds were judged to be the same. Second, for indistinguishable atoms in experimentally observed solution structures, the smallest number of atom index[13] (a technical term of Open Babel) in the molecular model was selected as the priority atom for dihedral angle determination. Comparing theoretically calculated geometries with geometries determined by means of X-ray crystallography is difficult because of the flexibility of conformations and the variety of substituents. In such comparisons, homology was judged to be high if the conformational elements were within 90° of each other.[7] There were 12 sets of conformation elements that met this criterion: {4c, 2a, 2c}, {2a, 2c, 5a}, {2c, 5a, 5c}, {5a, 5c, 1a}, {5c, 1a, 1c}, {1a, 1c, 6a}, {1c, 6a, 6c}, {6a, 6c, 3a}, {6c, 3a, 3c}, {3a, 3c, 4a}, {3c, 4a, 4c}, and {4a, 4c, 2a}. It was supposed that a set of the 12 sets above was X = {p, q, r}, and a conformational element, which was compared at a specific angle location, was y. If all compared conformational elements satisfied the expression $y \in X$, the structural code homology of the angle location was judged to be high.[7] This fuzzy comparison technique can also apply to the case that similar conformations, which can reach to the identical structure by optimizations, are given by different notations and can compare the ligand conformations located at the inside and outside of the binding site of the target protein. For a benzyl group, the conformation in which the phenyl ring is perpendicular to R is stable, and the phenyl ring can easily rotate (Figure 4). PhCH(1) and PhCH(2), in which the numbers 1 and 2 in the parenthesis indicate the atom indexes, in Figure 4 are indistinguishable moieties in experimentally observed solution structures by spectroscopies such as NMR. For the comparison method of conformational elements mentioned above, the conformational elements must be within 90° of each other. In the case of the slight

rotation of PhCH(1) from Figure 4A to Figure 4B, the conformational element 2c changes to 3a, which is not within 90°, according to Rule E-5.6(3). In Figure 4B, the conformational element becomes 5a, which is within 90°, according to the second deviation rule. The second deviation was convenient for this comparison method of conformational elements.

Finally, the converted conformational codes and Gibbs free energies of all the conformations were arranged in SDF format (Figure 5).[13] The optimized output files were easily selected in combination with this SDF format.

SMARTS 3D Search of Mevastatin

The CCOM program was applied to the X-ray structure of mevastatin (**1**) (PDB: 4oqr).[18] The addition of hydrogens and the file conversion from the ent-format to the mol-format were carried out with GaussView.[1] The conformational code used for **1** was **(*mb*_3a6c4c)*npMe*_6a1c(*hope*_1c2c2a2c)** (Figure 6A). To deal with structural flexibility, we used the fuzzy comparison technique described above, and SMARTS notation was used to deal with the variety of substituents.[2,13]

We constructed a database consisting of 635 SDF files for organic molecules, including fragment molecules, from the output files of density functional theory calculations (B3LYP/6-31G*)[17] for VCD conformational analysis; the database included (*R*)-malathion,[8] (*S*)-ibuprofen,[9] paclitaxel,[10] chiral alkyl alcohols,[11] thalidomide,[19] cholesterol derivatives, (–)-*cis*-permethrin, pravastatin, FK506 (Figure 6B),[20] donepezil, quinine, nopol, salvileucalin B,[21] norepinephrine, glutathione, ophthalmic acid, peptides, amino acids, levofloxacin, biotin, aldosterone, corticosterone, cortisol, pregnenolone, longifolene, thujone, warfarin, and many other pharmaceutical candidates. The chemical structure search of the 3D conformational fragment ***mb*_3a6c4c** of mevastatin (**1**) using the CCOM program in combination with the SMARTS notation COC(=O)[C@H](CC)C (Figure 7) extracted the structural homology files for pravastatin sodium (**2**) and two pravastatin fragments (**3** and **4**; Figure 8). Owing to steric repulsion, the −*gauche* form (**b1**: **3a**) of the optimized conformation **(*mb*_3a3a4c)*npOH*_1c2a1c(*hpNa*_1c2c2a3a2a2a3c4c)** for pravastatin sodium (**2**) was about 1.75 kcal/mol less stable than the +*gauche* form (**b1**: **2c**), which was the most stable form [**(*mb*_2c6a4c)*npOH*_1c2a1c(*hpNa*_1c2c2a3a2a2a3c4c)**] (Table 1). For fragment molecule **3** [**(*mb*_b1b2b3)*npOH*_b4b5b6**], the −*gauche* form (**b1**: **3a**) of optimized conformation (***mb*_3a3a4c)*npOH*_1c2a1c** was about 1.50 kcal/mol less stable than the +*gauche* form (**b1**: **2c**), which was the most stable form [**(*mb*_2c6a4c)*npOH*_6a2a1c**]. For fragment molecule **4** [**(*mb*_b1b2b3)*ipr*_b4**], the −*gauche* form (**b1**: **3a**) of optimized conformation **(*mb*_3a3a4a)*ipr*_3c** was about 1.06

kcal/mol less stable than the +*gauche* form (**b1**: **2c**), which was the most stable form [(***mb***_2c6a4c)***ipr***_3c]. These results suggest that the ***mb***_3a6c4c fragment is twisted in its target protein owing to steric repulsion, and a conformational change of the ***mb***_3a6c4c fragment present in the X-ray structure may be the driving force for dissociation for mevastatin (**1**) from its target protein. Accumulation of more density functional theory calculation data for various fragment molecules with conformational codes and Gibbs free energies can be expected to raise the accuracy of the population in experimentally observed solution structures.

The CCOM program will be freely available for academic purposes in the near future as soon as ready. It is also possible to send it via E-mail now.

FIGURE 1 Definition of angle locations and conformational elements for conformational codes of mevastatin (**1**) and pravastatin sodium (**2**). (A) Angle locations consist of prefixes for chemical blocks and symbols indicating bond locations. (B) Conformational elements represent classifications of dihedral angles.

FIGURE 2 Four atoms of (A) carboxylic acid and (B) carboxylic acid ester selected for CCOM. For the carboxylic acid moiety, the C=O oxygen, not the O-H oxygen according to Rule E-4.9(2), was selected, whereas for the carboxylic acid ester moiety, the C–O–C oxygen, not the C=O oxygen, was selected according to Rule E-4.9(3).

FIGURE 3 Search for highest priority atom limited to distance of two bonds from center bond for dihedral angle determination. The methylene carbon atom in the dotted-line box was selected from Rule E-5.6(1), however, the different atom selection occurred in occasional cases for searching of 3D conformational fragments of organic molecules. Therefore, the methyl carbon was selected as the highest priority atom because all the other groups within a distance of two bonds were judged to be the same.

FIGURE 4 Selection of the smallest number of atom index in the molecular model as priority atom for dihedral angle determination for indistinguishable atoms in experimentally observed solution structures. PhCH(1) and PhCH(2) mean CH moieties in the phenyl ring. Numbers 1 and 2 in the parenthesis indicate the atom indexes. In the case of the slight rotation of PhCH(1) from A to B, the conformational element 2c changes to 3a, which is not within 90°, according to Rule E-5.6(3). In B, the conformational element becomes 5a, which is within 90°, according to the second deviation rule.

FIGURE 5 SDF format for CCOM.

FIGURE 6 Conformational codes of (A) mevastatin (**1**) (PDB: 4oqr) on a stick model and (B) FK506 (PDB: 1fkj) on a wireframe-ball model. Conformational element **2a** of bond **b8** on the ring of **1** is omitted.

FIGURE 7 Search of 3D conformational fragment ***mb*_3a6c4c** of mevastatin (**1**) using CCOM in combination with SMARTS notation COC(=O)[C@H](CC)C (PDB: 4oqr).

FIGURE 8 Chemical structures of fragments **3** and **4**.

TABLE 1 Relative Gibbs free energies[a] of pravastatin sodium (**2**) conformations [**(*mb*_b1b2b3)*npOH*_b4b5b6(*hpNa*_b7b8b9b10b11b12b13b14)**]

Conclusion

The CCOM program can be used to unambiguously convert 3D conformational data for various organic molecules to one-dimensional data. The advantage of conformational representation with the CCOM program over the superimposed method is that the former can be used to easily search for conformational changes of characteristic 3D fragments of organic molecules, including theoretically calculated geometries and geometries determined by X-ray crystallography. In the future, a database of DFT calculation data for various fragment molecules with conformational codes and Gibbs free energies will be available for use in drug discovery. We are currently preparing the program to be used for fragment conformational search[8] for precise analysis of VCD structural data.

REFERENCES AND NOTES

1. Dennington R, Keith T, Millam J. GaussView. Shawnee Mission KS: Semichem Inc; **2009**.

2. Weininger D, Weininger A, Weininger JL. SMILES. 2. Algorithm for generation of unique SMILES notation. *J Chem Inf Comput Sci* **1989**;29:97-101

3. Taylor RE, Chen Y, Beatty A, Myles DC, Zhou YQ. Conformation-activity relationships in polyketide natural products: A new perspective on the rational design of epothilone analogues. *J Am Chem Soc* **2003**;125:26-27.

4. Carotenuto A, D'Ursi AM, Mulinacci B, Paolini I, Lolli F, Papini AM, Novellino E, Rovero P. Conformation-activity relationship of designed glycopeptides as synthetic probes for the detection of autoantibodies, biomarkers of multiple sclerosis. *J Med Chem* **2006**;49:5072-5079.

5. Askari JA, Buckley PA, Mould AP, Humphries MJ. Linking integrin conformation to function. *J Cell Sci* **2009**a;122:165-70.

6. Testa C, Scrima M, Grimaldi M, D'Ursi AM, Dirain ML, Lubin-Germain N, Singh A, Haskell-Luevano C, Chorev M, Rovero P, Papini AM. 1,4-Disubstituted- 1,2,3 triazolyl-Containing Analogues of MT-II: Design, Synthesis, Conformational Analysis, and Biological Activity. *J Med Chem* **2014**;57:9424-9434.

7. Izumi H, Wakisaka A, Nafie LA, Dukor RK. Data Mining of Supersecondary Structure Homology between Light Chains of Immunoglobulins and MHC Molecules: Absence of the Common Conformational Fragment in the Human IgM Rheumatoid Factor. *J Chem Inf Model* **2013**;53:584-91.

8. Izumi H, Ogata A, Nafie LA, Dukor RK. Structural determination of molecular stereochemistry using VCD spectroscopy and a conformational code: Absolute configuration and solution conformation of a chiral liquid pesticide, (*R*)-(+)-malathion. *Chirality* **2009**b;21:E172-E180.

9. Izumi H, Ogata A, Nafie LA, Dukor RK. A revised conformational code for the exhaustive analysis of conformers with one-to-one correspondence between conformation and code: Application to the VCD analysis of (*S*)-ibuprofen. *J Org Chem* **2009**c;74:1231-1236.

10. Izumi H, Ogata A, Nafie LA, Dukor RK. Vibrational circular dichroism analysis reveals a conformational change of the baccatin III ring of paclitaxel: Visualization of conformations using a new code for structure-activity relationships. *J Org Chem* **2008**a;73:2367-2372.

11. Izumi H, Yamagami S, Futamura S, Nafie LA, Dukor RK. Direct observation of odd-even effect for chiral alkyl alcohols in solution using vibrational circular dichroism spectroscopy. *J Am Chem Soc* **2004**;126:194-198.

12. Cross LC, Klyne W. REPORT FROM IUPAC COMMISSION ON NOMENCLATURE OF ORGANIC-CHEMISTRY - RULES FOR NOMENCLATURE OF ORGANIC-CHEMISTRY .E. STEREOCHEMISTRY (RECOMMENDATIONS 1974). *Pure Appl Chem* **1976**;45:13-30.

13. O'Boyle NM, Morley C, Hutchison GR. Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem Cent J* **2008**b;2:5.

14. wxPython. http://www.wxpython.org/ (accessed June 17, 2015).

15. VPython. 3D Programming for Ordinary Mortals. http://vpython.org/index.html (accessed June 17, 2015).

16. Avogadro. http://avogadro.cc/wiki/Main_Page (accessed June 17, 2015).

17. Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery JA Jr, Vreven T, Kudin KN, Burant JC, Millam JM, Iyengar SS, Tomasi J, Barone V, Mennucci B, Cossi M, Scalmani G, Rega N, Petersson GA, Nakatsuji H, Hada M, Ehara M, Toyota K, Fukuda R, Hasegawa J, Ishida M, Nakajima T, Honda Y, Kitao O, Nakai H, Klene M, Li X, Knox JE, Hratchian HP, Cross JB, Bakken V, Adamo C, Jaramillo J, Gomperts R, Stratmann RE, Yazyev O, Austin AJ, Cammi R, Pomelli C, Ochterski JW, Ayala PY, Morokuma K, Voth GA, Salvador P, Dannenberg JJ, Zakrzewski VG, Dapprich S, Daniels AD, Strain MC, Farkas O, Malick DK, Rabuck AD, Raghavachari K, Foresman JB, Ortiz JV, Cui Q, Baboul AG, Clifford S, Cioslowski J, Stefanov BB, Liu G, Liashenko A, Piskorz P, Komaromi I, Martin RL, Fox DJ, Keith T, Al-Laham MA, Peng CY, Nanayakkara A, Challacombe M, Gill PMW, Johnson B, Chen W, Wong MW, Gonzalez C, Pople JA. Gaussian 03. Wallingford CT: Gaussian Inc; **2004**.

18. McLean KJ, Hans M, Meijrink B, van Scheppingen WB, Vollebregt A, Tee KL, van der Laan JM, Leys D, Munro AW, van den Berg MA. Single-step fermentative production of the cholesterol-lowering drug pravastatin via reprogramming of Penicillium chrysogenum. *Proc Natl Acad Sci USA* **2015**;112: 2847-52.

19. Izumi H, Futamura S, Tokita N, Hamada Y. Fliplike motion in the thalidomide dimer: Conformational analysis of (*R*)-thalidomide using vibrational circular dichroism spectroscopy. *J Org Chem* **2007**;72:277-279.

20. Wilson KP, Yamashita MM, Sintchak MD, Rotstein SH, Murcko, M. A.; Boger, J.; Thomson, J. A.; Fitzgibbon, M. J.; Black JR, Navia MA. COMPARATIVE X-RAY STRUCTURES OF THE MAJOR BINDING-PROTEIN FOR THE IMMUNOSUPPRESSANT FK506 (TACROLIMUS) IN UNLIGANDED FORM AND IN COMPLEX WITH FK506 AND RAPAMYCIN. *Acta Cryst* **1995**;D51:511-521.

21. Aoyagi Y, Yamazaki A, Nakatsugawa C, Fukaya H, Takeya K, Kawauchi S, Izumi H. Salvileucalin B, A Novel Diterpenoid with an Unprecedented Rearranged Neoclerodane Skeleton from Salvia leucantha Cav. *Org Lett* **2008**;10:4429-4432.