Extended CAST Coding Method for Exact Search of Stereochemical Structures

Hiroko Satoh^{* a), b), c)}, Hiroyuki Koshino^{* a)}, Tadashi Nakata^{a)}

a) RIKEN (The Institute of Physical and Chemical Research),
2-1 Hirosawa, Wako, Saitama 351-0198, Japan

b) PRESTO, Japan Science and Technology Corporation,
4-1-8 Honcho, Kawaguchi, Saitama 332-0012, Japan

c) National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430, Japan

(Received January 18, 2002; Accepted March 7, 2002)

Accurate recognition of differences and similarities in stereochemical structures is achieved by extended CAST (CAnonical-representation of STereochemistry) coding method. Using the CAST notations, complete search of partial structures as well as whole structures by matching with the query in any level from a specific atom considering planar, configurational, and conformational information has been achieved. Differences and similarities of stereochemistry in four aldopentoses of D-xylose, D-ribose, D-arabinose, and D-lyxose, which have three chiral centers, are clearly represented by the extended CAST. Applications for some organic compounds containing more complicated stereochemical structures are also demonstrated.

Key Words: Stereochemistry, Configuration, Conformation, Canonical Coding Method, CAST, Three-dimensional Structure Database

Introduction

Stereochemistry is important information for treating molecular data. Exact structure search with configurational and conformational information is essential for a database system including three-dimensional structural information. Some coding

* *hsatoh@nii.ac.jp*

* *koshino@postman.riken.go.jp*

methods for three-dimensional chemical structures have been developed and used for structural representation in some chemical software, however, an essential problem of exact search of structures considering configurational and conformational structural environments has been remained.¹⁻⁶

We reported CAST (CAnonical-representation of STereochemistry) coding method in the previous paper^{7, 8}. The basic CAST gives canonical codes based on the dihedral angles. Two types of CAST notations for configurational and conformational structures can be flexibly used depending on the users together with CANOST notation⁹ for a planar structure. However, there are many ways to possibilities in selection of the CAST notations for comparison between structures, canonical rule for the selection is needed. This paper describes extended CAST coding method, which makes it possible to represent stereochemical environments for a specific site, and solves the problem of exact search of structures in a three-dimensional database system.

Representation by the Basic CAST Method^{7, 8}

The basic CAST coding method represents a molecular structure with conformational information as well as relative/absolute configurational information based on the dihedral angles. Twelve types of clock-like CAST codes are assigned for representation of stereochemical structures (Figure 1). For example, staggered and eclipsed conformations are represented by the set of CAST codes of (tw, si, te) and (ze, fo, ei), respectively. The dihedral angle is uniquely defined by using a molecular tree. The molecular tree is constructed from every atom in a molecule. The CAST notation is finally converted to a linear representation, which consists of the CAST codes with some separators that are a semi-colon ';', a comma ',', and an asterisk '*'. The semi-colon, the comma, and the asterisk are code-, group- and level- separators, respectively. The group consists of atoms that connect to the same atom at one higher level in a molecular tree. The CAST notations for representation of conformation and configuration are given, which are called conformational CAST and configurational CAST, respectively. The conformational CAST represents an input conformation. The configurational CAST represents relative and absolute configuration, which is constructed by canonical rotation of the dihedral angles of the input structure.

CAST can generally represent whole of a molecule by the two linear notations started from two atoms that are apart from each other through more than three bonds, because the dihedral angles are not defined for a bond belonging to the starting atom in a molecular tree. For

structural comparison between whole structures by CAST, two notations are needed, and for comparison between partial structures around a specific atom, notations started from another atoms are needed. There are many ways to select notations for the comparison, and some canonical rule is needed. This is the problem to use the basic CAST for structural search.

Figure 1 CAST codes defined for dihedral angles. Twelve types of the CAST code are defined to twelve areas of the dihedral angle. The CAST codes are represented as the first two letters of clockwise numbering from zero to eleven.

Figure 2 Model structures. These four acyclic aldopentoses have the same planar structure and different stereochemistry. The planar structure is shown with the numbering.

Figure 3 The basic configurational CAST notations for the aldopentoses **1**, **2**, **3**, and **4**. The notations are started from C1 and C4, respectively.

For example, the basic CAST notations for four acyclic aldopentoses having the same planar structure and different stereochemistry, D-xylose (**1**), D-ribose (**2**), D-lyxose (**3**), and D-arabinose (**4**) (Figure 2) are shown in Figure 3. The absolute configurations of the model structures are (2R, 3S, 4R), (2R, 3R, 4R), (2S, 3S, 4R), and (2S, 3R, 4R) for **1**, **2**, **3**, and **4**, respectively. In Figure 3 the basic configurational CAST notations started from C1 and C4, which are apart from each other through three bonds, are shown. The CAST notations are labeled by structural no. with subscript of its starting point like as 1_{C1}, namely, which indicates the CAST notations started from C1 for **1**. The linear notations correspond to the molecular tree with CAST codes. The relationships between codes in the molecular tree are translated to linear notations using some separator codes, a semi-colon, a comma, and an asterisk as mentioned above, according to the CAST coding rule.^{7,8} In Figure 3, the notations that related to R and S-configurations are bold- and double-underlined, respectively. In the notations from C1, the absolute configurational difference for C3 and C4 are represented, and in the notations from C4, the differences for C2 are represented. Both notations are needed to distinguish the four pentoses. In general, stereochemistry for the starting atom in the first level, which is called A-atom, is represented in the CAST notations from C-atoms that are

apart from A-atom through two bonds, and stereochemistry for an adjacent atom to A-atom, which is called B-atom, is represented in the CAST notations from the other adjacent atoms.

For solving the problem for the structural search, extra CAST notations were added. Procedures for making the additional CAST notations and constructing the extended CAST notations are described, below.

The Extended CAST Notations

For descriptions in general, the starting atom in the basic CAST notation is denoted as A, atoms connecting A is denoted as B₁, B₂, ..., B_n, where n is the number of atoms connecting the atom A, and they are ordered according to the priority defined in CANOST coding method.⁹ Atoms connecting to the atom B_n are denoted as C_{n-1}, C_{n-2}, ..., C_{n-m}, where m is the number of atoms connecting B_n, but excepting A, and they are also ordered according to the CANOST priority. Atoms connecting to the atom C_{n-m} are denoted as D_{n-m-1}, D_{n-m-2}, ..., D_{n-m-k}, where k is the number of atoms connecting C_{n-m}, but excepting B_n, and they are also ordered according to the CANOST priority. Namely, an A-atom, B-atoms, C-atoms, and D-atoms are in the first,

Figure 4 Additional configurational CAST notations. (a) A model structure. **A** is the starting atom. **B**, **C**, and **D** are shown with the number according to CANOST priority. Lower number is higher priority. (b) **A**-CAST: Additional CAST for stereochemical information on **A**. (1) The contents of the **A**-CAST notation described by the Newman projection with the corresponding CAST notations. The CAST codes are assigned to **B**-atoms from the circled **C**-atom. Grayed and squared **B**-atoms are the most and the second CANOST prior atoms, respectively. (2) Final form of the **A**-CAST notations. (c) **B**-CAST: Additional CAST for stereochemical information on **B**. (1) The contents of the **B**-CAST notation described by the same way as (b)-1. (2) Final form of the **B**-CAST notations.

second, third, and fourth level. Two types of the additional CAST notations defined for representation of stereochemical information on **A**-atom and **B**-atoms, which are called as **A**-CAST and **B**-CAST notations, respectively. Figure 4 shows the additional configurational CAST notations for structure (a), which is drawn according to the way described above. Figure 4-(b) shows contents of **A**-CAST, which is a set of CAST codes defined by the dihedral angle of **C-B-A-B**, where CAST codes are assigned to the fourth **B**-atoms. In (b)-1 of Figure 4, they are represented by the Newman projection, where CAST codes are assigned to the three **B**-atoms from a circled **C**-atom. The most and squared, respectively. The corresponding additional CAST notations are shown under the Newman projection. The six additional CAST notations ((b)-1) are finally joined by an asterisk "*" and inputted between two percent "%" shown in (b)-2. Figure 4-(c) shows contents of **B**-CAST, which is a set of CAST codes defined by the dihedral angle of **B-A-B-C**, where CAST codes are assigned to

the **C**-atoms. In (c)-1 of Figure 4, they are also represented by the Newman projection with the corresponding CAST notations. The final representation is shown in (c)-2. For the **B**-CAST notations, a sharp "#" is used instead of "%" in the **A**-CAST notations.

If any rotational operation is not performed, the dihedral angle defined by **B-A-B-C** is the same as that defined by **C-B-A-B**. However, the dihedral angle is rotated to make the basic configurational CAST notations according to the canonical rearrangement rules as described in a previous paper,⁸ thus, both notations started from **C** and **B** are needed.

The conformational **A**- and **B**-CAST notations are defined as the same way as the configurational CAST. Figure 5 shows the additional conformational CAST for structure (a) by the same style as Figure 4. An extended linear conformation is used for the coding, namely, **A/D**₁₋₁₋₁, **B**₂/**C**₁₋₁, **A/D**₂₋₁₋₁, and **B**₁/**C**₂₋₁ are anti positions. Figure 5-(b) and -(c) show the contents of **A**- and **B**-CAST, respectively.¹⁰

Figure 5 Additional conformational CAST notations. (a) A model structure. **A** is the starting atom. **B**, **C**, and **D** are shown with the number according to CANOST priority. Lower number is higher priority. (b) **A-CAST**: Additional CAST for stereochemical information on **A**. (1) The contents of the **A-CAST** notation described by the Newman projection with the corresponding CAST notations. The CAST codes are assigned to **B**-atoms from the circled **C**-atom. Grayed and squared **B**-atoms are the most and the second CANOST prior atoms, respectively. (2) Final form of the **A-CAST** notations. (c) **B-CAST**: Additional CAST for stereochemical information on **B**. (1) The contents of the **B-CAST** notation described by the same way as (b)-1. (2) Final form of the **B-CAST** notations.

Representation of the Model Structures by The Extended CAST

The extended configurational and conformational CAST notations started from C1 for the four model structures of Figure 2 are shown in Figure 6.

Figure 6-(a) and -(b) are configurational and conformational CAST, respectively. The numbers 1-4 correspond with the numbers of the model structures, D-xylose (1), D-ribose (2), D-lyxose (3), and D-arabinose (4). As shown in (a), the absolute configurations of **1** (2*R*, 3*S*, 4*R*), **2** (2*R*, 3*R*, 4*R*), **3** (2*S*, 3*R*, 4*R*), and **4** (2*S*, 3*S*, 4*R*) were successfully distinguished by using the

extended notations, where notations corresponding with *R* and *S* configurations are bold and double underlined, respectively. In the extended conformational notations (b), both conformational and configurational similarities and differences are represented. In this execution, three-dimensional structures were constructed by a CSC Chem3D software and energy optimization was performed by MM2 calculations.

Thus, the extended CAST notations can successfully represents similarities and differences in both configuration and conformation for a molecule by a single linear notation.

a)

```

1) %si;ze*si;ze*si;ze%#,si;te;tw,si;te;tw,#A*B;B*B*C;C;C,,si;tw;te,OH,si;te;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2R                        3S                        4R
2) %si;ze*si;ze*si;ze%#,si;te;tw,si;te;tw,#A*B;B*B*C;C;C,,si;te;tw,OH,si;te;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2R                        3R                        4R
3) %si;ze*si;ze*si;ze%#,si;tw;te,si;tw;te,#A*B;B*B*C;C;C,,si;tw;te,OH,si;te;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2S                        3S                        4R
4) %si;ze*si;ze*si;ze%#,si;tw;te,si;tw;te,#A*B;B*B*C;C;C,,si;te;tw,OH,si;te;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2S                        3R                        4R

```

b)

```

1) %ei;tw*ze;si*th;te%#,ei;ze;th,tw;si;te,#A*B;B*B*C;C;C,,si;tw;te,OH,si;te;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2R                        3S                        4R
2) %ei;tw*ze;si*fo;te%#,ei;ze;fo,tw;si;te,#A*B;B*B*C;C;C,,si;te;tw,OH,si;te;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2R                        3R                        4R
3) %fo;te*ze;si*ei;tw%#,fo;ze;ei,te;si;tw,#A*B;B*B*C;C;C,,si;tw;te,OH,si;el;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2S                        3S                        4R
4) %fo;te*ze;si*ei;tw%#,fo;ze;ei,te;si;tw,#A*B;B*B*C;C;C,,se;el;th,OH,se;el;tw,OH,,*si;H;H,OH,,*OH,,,*
                        2S                        3R                        4R

```

Figure 6 The extended CAST notations for the model aldopentoses **1**, **2**, **3**, and **4**. They are started from C1. (a) The extended configurational CAST notations. (1)-(4) are the notations for **1-4**. The notations corresponding with *R* and *S* configurations are bold and double underlined, respectively. (b) The extended conformational CAST notations.

Application to More Complicated Organic Molecules

The extended CAST notation was applied to representation of more complicated organic molecules of pironetin (**5**)¹⁰ and 9-*epi*-pironetin (**6**). The structures with the atom nos. are shown in Figure 7, where the gray areas are the different sites. Figure 8 shows the extended configurational CAST notations from C9 for **5** and **6**. As shown by gray areas in Figure 8, the stereochemical differences for C9 are represented in the A-CAST notations. The rest codes for **5** and **6** are the same.

The application demonstrated that the extended CAST successfully represents the stereochemistry even for a molecule with sequential stereo centers and makes it possible to exactly search a structural database with stereochemical information.

Figure 7 Structures of pironetin (**5**) and 9-*epi*-pironetin (**6**) with the numbering. The stereochemical difference is shown by the gray circle in the each structure. **6** is not a natural product.

```

5C9  %si;te;tw*si;te;tw*si;te;tw*si;twite*si;twite*si;twite*si;te;tw%
      #si;twite,si,*si;twite,si,*si;twite,si;twite,*si;twite,si;twite,si#
      A*B;B;B*B*C;C;C,C;C;C,C,*si;te;tw,H;H;H,,si;H;H,H;H;H,,H;H;H
      *si;H;H,OH,,,,,size,,,,,,*si;twite,,,size,*si;te;tw,si,,H;H;H,
      *size,si;H;H,,size,,,*&l;si,,H;H;H,,,,*

6C9  %si;twite*si;twite*si;twite*si;te;tw*si;te;tw*si;te;tw*si;twite%
      #si;twite,si,*si;twite,si,*si;twite,si;twite,*si;twite,si;twite,si#
      A*B;B;B*B*C;C;C,C;C;C,C,*si;te;tw,H;H;H,,si;H;H,H;H;H,,H;H;H
      *si;H;H,OH,,,,,size,,,,,,*si;twite,,,size,*si;te;tw,si,,H;H;H,
      *size,si;H;H,,size,,,*&l;si,,H;H;H,,,,*

```

Figure 8 The extended configurational CAST notations for pironetins **5** and **6**. The stereochemical difference for C9 is represented by the extended notations shown as grayed areas.

Stereochemical Database and Structural Search

We are constructing a structural database with stereochemistry using the extended CAST notations with the CANOST notations as the methods for molecular descriptions. In the database, a structure is described by the configurational and conformational CAST and CANOST linear notations started from every atom, and they are related each other. Original 3D-molecular structure is also related to the notations. Structure search is performed by simple comparison of the linear notations from the starting code to a specified matching level. Details are described in elsewhere.

Conclusion

Representations of chemical structures with configurational and conformational information for exact search of three-dimensional database have been achieved by the extended CAST coding method. The extended method can canonically represent both partial and whole stereochemical environments. An application of the three-dimensional structural database to accurate prediction of ^{13}C -NMR chemical shifts considering stereochemistry are now under going, and they will be reported, elsewhere, in near future.

Acknowledgements

We thank Dr. Jun Uzawa in RIKEN (The Institute of Physical and Chemical Research) and Prof. Kimito Funatsu in Toyohashi University of Technology for his

helpful advises and encouragement. This research was supported in part by President's Special Research Grant from RIKEN, Shiseido Fund for Science and Technology, and a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

References and Notes

- [1] W. T. Wipke, T. M. Dyott, *J. Am. Chem. Soc.*, 96, 4834–4842(1974).
- [2] F. Choplin, R. Dorschner; G. Kaufmann, W. T. Wipke, *J. Organometallic Chem.*, 152, 101–109(1978).
- [3] W. C. Herndon, J. E. Leonard, *Inorg. Chem.*, 22, 554–557(1983).
- [4] T. Akutsu, *J. Chem. Inf. Comput. Sci.*, 31, 414–417(1991).
- [5] A. T. Balaban, C. Rucker, *J. Chem. Inf. Comput. Sci.*, 41, 1145–1149(2001).
- [6] S. Fujita, *Bull. Chem. Soc. Jpn.*, 74, 1585–1603(2001).
- [7] H. Satoh, H. Koshino, K. Funatsu, T. Nakata, *J. Chem. Inf. Comput. Sci.*, 40, 622–630(2000).
- [8] H. Satoh, H. Koshino, K. Funatsu, T. Nakata, *J. Chem. Inf. Comput. Sci.*, 41, 1106–1112(2001).
- [9] H. Abe, Y. Kudo, T. Yamasaki, K. Tanaka, M. Sasaki, S. Sasaki, *J. Chem. Inf. Comput. Sci.*, 24, 212–216(1984).
- [10] For conformational CAST notations, the dihedral angles defined by **B-A-B-C** and **C-B-A-B** are the same, however, both are used for compatibility with the additional configurational CAST notations.
- [11] S. Kobayashi, K. Tsuchiya, T. Kurokawa, T. Nakagawa, N. Shimada, Y. Iitaka, *J. Antibiotics*, 47, 703–707(1994).

立体化学構造の正確な検索のための拡張CASTコード化法

佐藤 寛子^{* a), b), c)}, 越野 広雪^{* a)}, 中田 忠^{a)}

a) 理化学研究所, 〒351-0198 埼玉県和光市広沢2-1

b) 科学技術振興事業団さがけ研究 2 1, 〒332-0012 埼玉県川口市本町4-1-8

c) 国立情報学研究所, 〒101-8430 東京都千代田区一ツ橋2-1-2

立体化学の類似と相違の正確な認識のための拡張CAST (CAnonical-representation of STereochemistry)コードについて述べる。拡張CASTコードによって、特定部位周りの平面 / 立体配置 / 立体配座構造の構造環境を正確に考慮した検索が可能となった。D-xylose, D-ribose, D-arabinose, and D-lyxoseの4種のaldopentoseについて拡張CASTコード化法により立体化学の類似と相違を表現した。より複雑な分子構造についての応用例も併せて報告する。

キーワード: 立体化学, 立体配置, 立体配座, 規範的コード化法, CAST, 三次元構造データベース

* hsatoh@nii.ac.jp

* koshino@postman.riken.go.jp