JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Chemical Structure Elucidation from ^{13}C NMR Chemical Shifts: Efficient Data Processing Using Bipartite Matching and Maximal Clique Algorithms

Shungo Koichi,*,[†] Masaki Arisaka,[‡] Hiroyuki Koshino,[§] Atsushi Aoki,[∥] Satoru Iwata,[‡] Takeaki Uno,[⊥] and Hiroko Satoh[⊥]

[†]Department of Information Systems and Mathematical Sciences, Nanzan University, Seto 489-0863, Japan

[‡]Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

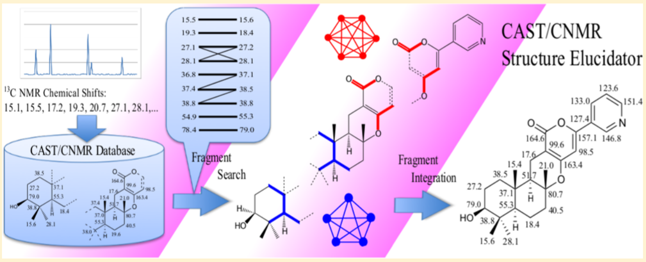
[§]RIKEN, Saitama 351-0198, Japan

[∥]Department of Computer Science, Kyoto Sangyo University, Kyoto 603-8555, Japan

[⊥]National Institute of Informatics, Tokyo 101-8430, Japan

S *Supporting Information*

ABSTRACT: Computer-assisted chemical structure elucidation has been intensively studied since the first use of computers in chemistry in the 1960s. Most of the existing elucidators use a structure−spectrum database to obtain clues about the correct structure. Such a structure−spectrum database is expected to grow on a daily basis. Hence, the necessity to develop an efficient structure elucidation system that can adapt to the growth of a database has been also growing. Therefore, we have developed a new elucidator using practically efficient graph algorithms, including the convex bipartite matching, weighted bipartite matching, and Bron−Kerbosch maximal clique algorithms. The utilization of the two matching algorithms especially is a novel point of our elucidator. Because of these sophisticated algorithms, the elucidator exactly produces a correct structure if all of the fragments are included in the database. Even if not all of the fragments are in the database, the elucidator proposes relevant substructures that can help chemists to identify the actual chemical structures. The elucidator, called the CAST/CNMR Structure Elucidator, plays a complementary role to the CAST/CNMR Chemical Shift Predictor, and together these two functions can be used to analyze the structures of organic compounds.



1. INTRODUCTION

Automatic processing of chemical structures using computers is one of the most enduring subjects in the realms of chemical information and modeling. The idea of using computers to solve chemical problems dates from the 1960s. Since the beginning of computer chemistry, expert systems for structural elucidation have attracted a great deal of attention not only from chemists but also from information scientists. Extensive efforts have been devoted to combining the techniques of chemistry and information science, and these have resulted in many elucidation systems, for example, DENDRAL,[1,2] CHEMICS,[3] SpecInfo,[4−6] MOLGEN,[7] ACD,[8] KnowItAll,[9] Spec2D,[10] X-PERT,[11] StrucEluc,[12,13] and OMG.[14] Moreover, there are commercial systems with well-designed user interfaces (SpecInfo, ACD, KnowItAll). Although some of these systems can generate structures simply from compositional formulas (CHEMICS, MOLGEN, OMG), most take a database-oriented approach to structural elucidation: Candidate structures are built up from fragment/component structures relevant to spectral data, such as from NMR, IR, or mass spectrometers, and those fragment/component structures are obtained from chemical structure−spectrum databases. At present, the total number of identified chemical structures is reported to be about 70 million, and this number is estimated to be increasing by hundreds of thousands per year. Even though the number of chemical structures assigned to spectral data is assumed to be less than this total, the size of chemical structure−spectrum databases is also growing. The topic of computer-assisted structure elucidation has been intensively studied for half a century, and it forms the basis for rigorous chemical data processing. Nonetheless, there still remain challenges as to how to deal with large-scale databases that are being continuously updated and how to obtain candidate/correct structures as fast as possible.

This is the underlying motive that led us to develop a new chemical structure elucidation system. A new feature of our development is that we used a number of graph algorithms, including the convex bipartite matching, weighted bipartite matching, and Bron−Kerbosch maximal clique algorithms, to efficiently obtain a correct structure even from a large-scale

Received: October 18, 2013
Published: March 23, 2014

ACS Publications

database. The utilization of the two matching algorithms especially is a novel point of our elucidator. The new elucidator uses a database of chemical compounds with ^{13}C NMR chemical shifts assigned to their carbons. We use ^{13}C NMR chemical shifts because ^{1}H NMR chemical shifts are very sensitive to through-space interactions of functional groups such as aromatic rings and carbonyl groups. Besides, it is difficult to standardize ^{1}H NMR data by excluding several influences. The database was originally developed for the CAST/CNMR Chemical Shift Predictor (CShift-Predictor),[15−20] which is a ^{13}C NMR chemical shift prediction system. Our new elucidator plays a role complementary to that of CShift-Predictor and is called the *CAST/CNMR Structure Elucidator*, or simply, Structure Elucidator. It uses a set of ^{13}C NMR chemical shifts as a query. Relevant fragment structures are retrieved from the database by mapping of the query chemical shifts to part of each compound in the database. The relevant fragments are assembled to form substructures of an estimated structure. In the final stage, the most probable structures are formed by merging the substructures that have common parts.

Because of the advanced graph algorithms mentioned above, our elucidator efficiently produces a correct structure if all of the fragments are included in the database. Even if not all of the fragments are in the database, it proposes relevant substructures that will help chemists to get ideas about the actual chemical structures. Furthermore, since the entries in our chemical structure−NMR database have information on the stereochemistry of compounds, an estimated structure of our elucidator is also equipped with stereochemistry.

In the present paper, we explain the algorithms of this new elucidation method in accordance with the flow of our elucidator, shown in Figure 1. We also demonstrate how our elucidator is used in real structure analysis of natural organic compounds.

2. FRAGMENT SEARCH

Collection of fragments is a critical step of our elucidation method. If an essential fragment were unable to be found, our elucidator would never obtain the correct chemical structure. Therefore, it is important to have a method to find all essential fragments.

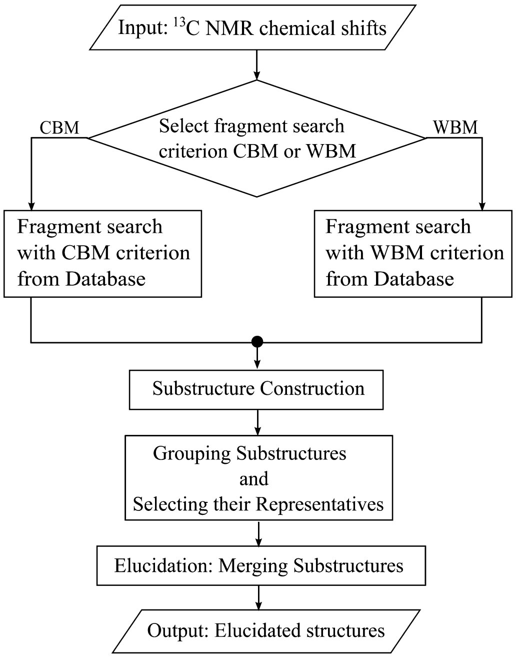

Input: ^{13}C NMR chemical shifts
Select fragment search criterion CBM or WBM
CBM
WBM
Fragment search with CBM criterion from Database
Fragment search with WBM criterion from Database
Substructure Construction
Grouping Substructures and Selecting their Representatives
Elucidation: Merging Substructures
Output: Elucidated structures

Figure 1. Flowchart for our elucidator.

However, if the number of selected fragments is large, the computational cost to elucidate a structure blows up with the number of candidate structures. In the worst case, it is impossible to complete the elucidation within a realistic time frame. Hence, it is necessary to develop an efficient method to make exhaustive and nonredundant searches of relevant fragments. We propose two search criteria for such a method. The two criteria can be evaluated efficiently by the convex bipartite matching (CBM) and weighted bipartite matching (WBM) algorithms. These algorithms are well-known in the context of graph algorithms.

Before we describe the details of the two criteria, we should explain how to use them to make exhaustive and nonredundant fragment searches. The two criteria have the following properties: The search with WBM is laxer than the one with CBM; that is, the WBM search always finds fragments retrieved by the CBM search. Therefore, the WBM search is more exhaustive and more redundant than the CBM search. By combining these two criteria, the user can obtain an exhaustive and nonredundant search result and thereby achieve an efficient and reliable structural elucidation.

2.1. Fragments and Bipartite Graph. Let $G = (U, V; E)$ be the *bipartite graph* with a left vertex set U, right vertex set V, and edge set E. An edge $e \in E$ is a pair $e = (u, v)$ of a vertex $u \in U$ and a vertex $v \in V$. A *matching* M is a subset of E whose edges do not share a vertex. A vertex x of G is called *covered* by a matching M if M has an edge e incident with x. In the following, the left vertex set U corresponds to a query of 1D NMR data that is a set of NMR chemical shifts.

In order for it to be the object of a search, each fragment must be limited to having a certain form; that is, a fragment has a central atom, and the other atoms in the fragment are connected within a user-designated number β of bonds from the central atom. Since the compounds in the database are associated with NMR chemical shifts, one obtains from a fragment a set of NMR chemical shifts that are assigned to the carbons in the fragment. The right vertex set V corresponds to this set of NMR chemical shifts. It should be noted that if the central atom is a carbon, its assigned NMR chemical shift is in V. A basic difference between the two criteria is in how they define the edge set E.

2.2. Convex Bipartite Matching Criterion. When one uses the CBM search, one sets an edge $e = (u, v)$ in E for NMR chemical shifts $u \in U$ and $v \in V$ if the absolute difference $|u - v|$ is less than or equal to a user-designated allowance α. It is known that a bipartite graph defined in this way is a *convex bipartite graph*. Using the CBM algorithm described in section 2.3, one can efficiently compute a maximum matching M_c that has the maximum number of edges among all matchings in G. If the number of edges in M_c is equal to the minimum of $|U|$ and $|V|$, that is, if at least one of U and V is covered by M_c, we say that the central atom of a fragment is *hit*. If a hit central atom a is a carbon and its assigned NMR chemical shift v_a is covered by an edge (u, v_a) in M_c, a is called *hit with respect to* the query NMR chemical shift u. It should be noted that the difference between each pair in M_c is at most the user-designated allowance α.

2.3. CBM Algorithm. This subsection describes the algorithm to obtain the maximum matching M_c of $G = (U, V; E)$ as defined in section 2.2. Let U and V be defined by $U = \{u_i \mid i = 1, 2, ..., m\}$ and $V = \{v_j \mid j = 1, 2, ..., n\}$. We assume that $u_1 \le u_2 \le \cdots \le u_m$. For each vertex $x \in U \cup V$, let $N(x)$ denote the set of all vertices adjacent to x. The algorithm consists of the following steps (where the symbol ":=" denotes the assignment operator, i.e., $x := y$ assigns the value of y to the variable x):

```
Step 1: Let $Q := U$, $D := V$, and $i := 1$.
Step 2: If $N(u_i) \cap D$ is not empty, go to step 2-1; otherwise, go to step 2-2.
    Step 2-1: Let $v_{j*}$ be the element in $N(u_i) \cap D$ that has the minimum of the maximum values of $N(v_j)$ for $v_j \in N(u_i) \cap D$. Put $M_c := M_c \cup \{(u_i, v_{j*})\}$, $Q := Q \backslash \{u_i\}$, $D := D \backslash \{v_{j*}\}$, and $i := i + 1$. Go to step 3.
    Step 2-2: Mark the central atom as *not hit*, and terminate the procedure.
Step 3: If $Q$ or $D$ is empty, mark the central atom as *hit* and terminate the procedure; otherwise, go back to Step 2.
```

Figure 2 is an illustration to help the reader to understand step 2 of the algorithm. Recall that $m = |U|$ and $n = |V|$. In step 2, for each u_i with nonempty $N(u_i) \cap D$, we may do a comparison of two numbers a total of $n - 1$ times to obtain v_{j*} because $N(u_i) \cap D$ may contain n members. Since the number of u_i with nonempty $N(u_i) \cap D$ is at most m, step 2 may be repeated a total of m times. Hence, the CBM algorithm has $(n - 1)m$ operations, and therefore, with the aid of the so-called "big O" notation, its computation cost is represented as $O(nm)$. (In the big O notation, the term m in $nm - m$ is ignored since nm dominates m if n is large.)

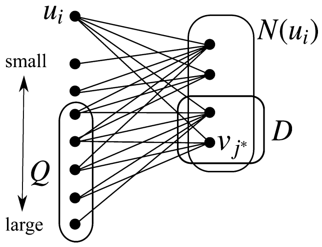


Figure 2. Situation of step 2 in the CBM algorithm.

2.4. Weighted Bipartite Matching Criterion.

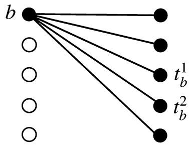
When one uses the WBM search, one defines E to be the set of all pairs (u, v) of NMR chemical shifts $u \in U$ and $v \in V$ and gives a weight $w_e = |u - v|$ to each edge $e = (u, v)$ in E. In this way, one obtains a weighted bipartite graph. Using the WBM algorithm described in section 2.5, one can efficiently compute the minimum weighted maximum matching M_w that has the minimum weight $w(M_w)$ among all matchings covering at least U or V; the weight $w(M_w)$ is given by the sum of the weights w_e on all edges e in M_w: $w(M_w) = \sum_{e \in M_w} w_e$. If $w(M_w)$ is less than or equal to a user-designated allowance α times the minimum of $|U|$ and $|V|$, we say that the central atom of a fragment is *hit*. If a hit central atom a is a carbon and its assigned NMR chemical shift v_a is covered by an edge (u, v_a) in M_w, then similarly to the CBM criterion, a is said to be *hit with respect to* the query NMR chemical shift u. It should be noted that the average of the differences between pairs in M_w is less than or equal to the user-designated allowance α.

2.5. WBM Algorithm.

This subsection describes the algorithm to obtain the minimum weighted maximum matching M_w of $G = (U, V; E)$ as defined in section 2.4. The algorithm is a so-called auction algorithm in the context of graph algorithms. If $|U| \leq |V|$, let $S := U$ and $T := V$; otherwise, let $S := V$ and $T := U$. Then define $\varepsilon = 1/(|S| + 1)$. The algorithm consists of the following steps:

```
Step 1: Let $p_t := 0$ for each $t \in T$ and set $B_1 := S$ and $M_w := \emptyset$.
Step 2: If $B_1$ is not empty, put $B_2 := \emptyset$ and go to step 3; otherwise, output $M_w$ and stop.
Step 3: For each $b \in B_1$, do the following:
    Step 3-1: Let $t_b^1$ be an element in $T$ such that the value $w_{(b,t)} - p_t$ is the minimum for $t \in T$, and similarly, let $t_b^2$ be the element in $V$ such that the value $w_{(b,t)} - p_t$ is the minimum for $t \in T \backslash \{t_b^1\}$ (i.e., the second minimum for $t \in T$).
    Step 3-2: If $(s, t_b^1) \in M_w$ for some $s \in S$, remove $(s, t_b^1)$ from $M_w$ and put $B_2 := B_2 \cup \{s\}$.
    Step 3-3: Add $(b, t_b^1)$ to $M_w$.
    Step 3-4: Update $p_{t_b^1}$ as $p_{t_b^1} := p_{t_b^1} + [w_{(b,t_b^1)} - p_{t_b^1}] - [w_{(b,t_b^2)} - p_{t_b^2}] - \varepsilon$.
Step 4: Set $B_1 := B_2$ and go back to step 2.
```

Figure 3 is an illustration to help the reader to understand step 3-1 of the algorithm. Let $n = |S|$ and $m = |T|$. The maximum of $w_{(u,\,v)}$ for $u \in U$, $v \in V$ is denoted by $w_{\max}$. The computational cost of the WBM algorithm is $O(n^2 m w_{\max})$, which we can consider to be $O(n^2 m)$ since $w_{\max}$ is at most 600 (ppm) as long as a chemical compound includes only ^{13}C, ^{15}N, ^{31}P, and ^{1}H as atoms with spin of 1/2.



$w_{(b,t)}$	p_t	$w_{(b,t)} - p_t$
1.5	0.0	1.5
2.0	0.0	2.0
0.9	−0.4	1.3
1.0	−0.2	1.2
1.4	0.0	1.4

Figure 3. Situation of step 3-1 in the WBM algorithm.

2.6. Relation between the CBM and WBM Criteria.

As stated above, a central atom is judged to be hit by the CBM criterion if the difference between each pair in M_c is at most the user-designated allowance α and by the WBM criterion if the average of the differences between pairs in M_w is at most α. Hence, for the same α, the WBM criterion is *laxer* than the CBM criterion. In other words, the exhaustiveness of the WBM search is *higher* than that of the CBM search. However, since exhaustiveness is inextricably associated with redundancy, it may be necessary to remove nonessential substructures from the search results. In the case of redundancy, the CBM criterion is preferable to the WBM criterion. In addition, the computational cost of the CBM algorithm [$O(nm)$] is less than that of the WBM algorithm [$O(n^2m)$], and hence, in regard to the computational cost, the CBM algorithm has an advantage. In this way, the two criteria play complementary roles. So far, in this procedure the user needs to choose the CBM or WBM criterion and tune the user-designated allowance α to obtain a preferable search result.

2.7. Options for the CBM and WBM Criteria.

Each carbon of the compounds in the database has the information on its type of carbon (primary, secondary, tertiary, or quaternary). This information is helpful for elucidating a structure more reliably. If the query contains information on the type of carbon, our elucidator can utilize that information as an option for refining the CBM and WBM criteria. When one utilizes the CBM criterion, a convex bipartite graph is definable for each type of carbon, and the CBM algorithm is applicable to each of the convex bipartite graphs. Moreover, the union of the maximum matchings for the convex bipartite graphs turns out to be a matching as well, and it can be used instead of M_c to judge whether a central atom is hit or not. With the WBM criterion, when one sets the weight of a pair of different type carbons to be a sufficiently large number, the minimum weighted maximum matching never contains such a pair of carbons.

2.8. Illustrative Example. Consider a query of 1D NMR data consisting of the following set of NMR chemical shifts: 15.1, 15.5, 17.2, 19.3, 20.7, 27.1, 28.1, 36.8, 37.4, 38.8, 40.3, 51.5, 54.9, 78.4, 81.0, 99.4, 100.3, 123.6, 127.6, 132.8, 146.6, 151.0, 155.5, 162.8, and 164.0 ppm. This set of NMR chemical shifts was reported by Itoh et al.[21] An example of a fragment is shown in Figure 4. This fragment consists of atoms three or fewer bonds from the labeled central atom. The number beside each atom is the reported NMR chemical shift for that atom. This fragment is characterized by the NMR chemical shifts 15.6, 18.4, 27.2, 28.1, 37.1, 38.5, 38.8, 55.3, and 79.0 ppm. In order to demonstrate the difference between the CBM and WBM criteria, we consider the two cases of $\alpha = 0.45$ and $\alpha = 1.1$.

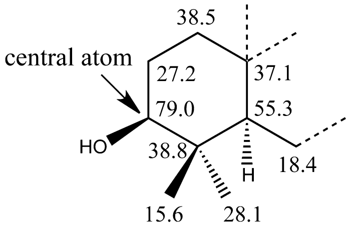


Figure 4. Fragment consisting of atoms connected within three bonds from the central atom (denoted by the arrow). The numbers are ^{13}C NMR chemical shifts assigned to the carbons.

First, we consider the case that the user-designated allowance is $\alpha = 0.45$ and apply the CBM algorithm to the convex bipartite graph $G = (U, V; E)$ shown in Figure 5 (1). Some of the data are not shown in Figure 5 for simplicity. The bold lines denote the maximum matching M_c obtained by the CBM algorithm. In this case, the number of edges in M_c is equal to 6, which is smaller than the number of carbons in the fragment. Therefore, the central atom is judged to be *not hit* according to the CBM criterion. On the other hand, Figure 5 (3) shows the minimum weighted maximum matching M_w obtained by the WBM algorithm. The weight of M_w is $w(M_w) = 3.5$. This is smaller than the user-designated allowance α times the number of carbons in the fragment ($0.45 \times 9 = 4.05$). Therefore, the central atom is judged to be *hit* by the WBM criterion.

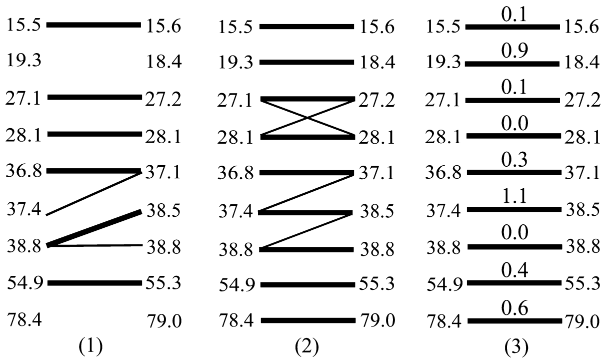


Figure 5. (1) Convex bipartite graph for $\alpha = 0.45$ with no matching that covers at least one of its vertex sets. (2) Convex bipartite graph for $\alpha = 1.1$ with a matching that covers one of its vertex sets (bold lines). (3) Minimum weighted maximal matching that covers one of the vertex sets of a weighted bipartite graph for $\alpha = 0.45$.

Next, we consider the case that the user-designated allowance is $\alpha = 1.1$. In this case, the convex bipartite graph $G = (U, V; E)$ for the CBM criterion is the one shown in Figure 5 (2). The bold lines denote the maximum matching M_c obtained by the CBM algorithm. Since the number of edges in M_c is equal to the number of carbons in the fragment, the central atom is judged to be *hit* by the CBM criterion. Furthermore, according to the relation between the CBM and WBM criteria, the central atom should be *hit* by the WBM criterion. In fact, the minimum weighted maximum matching is the same as the one shown in Figure 5 (3).

3. SUBSTRUCTURE CONSTRUCTION

After we judge whether each atom in a compound is hit or not, we construct a substructure of hit atoms and their neighboring atoms. The constructed substructures are stored as components of a candidate structure of the target compound.

3.1. Definition of Substructures. First, let us rigorously define what a substructure is. Here, an atom that is neither carbon nor hydrogen is called a *heteroatom*. An atom in a substructure is one of the following three types:

(i) a hit atom;
(ii) an atom connected within a user-designated number β of bonds from an atom of type (i) in the substructure;
(iii) a heteroatom connected directly or by a sequence of bonded heteroatoms to an atom of type (i) or (ii) in the substructure.

A substructure includes at least one hit atom, and an atom satisfying (ii) or (iii) must be contained in the substructure. The user-designated number β used in (ii) is assumed to be the same as the one in the definition of fragments in section 2.1. It should be noted that it is possible to obtain more than one substructure from one compound in a database. Figure 6 shows examples of substructures for (1) $\beta = 3$ and (2) $\beta = 1$ (technically, hydrogens are not contained in the substructures; however, hydrogens are included here in order to distinguish them from methyl groups.) The nitrogen designated by the hollow arrow in Figure 6 (2) is included in the substructure because it satisfies part (iii) of the above definition.

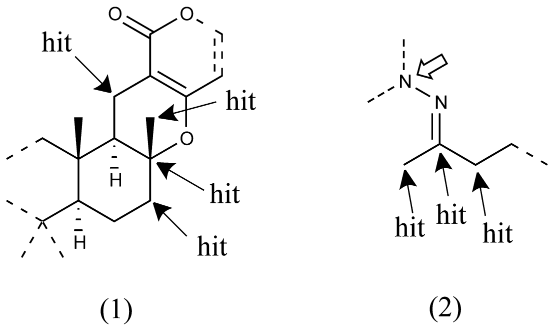


Figure 6. Examples of substructures for (1) $\beta = 3$ and (2) $\beta = 1$, where β denotes the user-designated number of bonds from hit atoms.

Obviously from the definition, if a hit atom *a* is contained in a substructure, then the fragment with the central atom *a* must be contained in the substructure. For example, the fragments for the hit atoms in the substructure shown as Figure 6 (1) are depicted in Figure 7, and they are all included in the substructure.

3.2. Algorithm To Construct a Substructure. To build substructures conforming to the definition in section 3.1, we

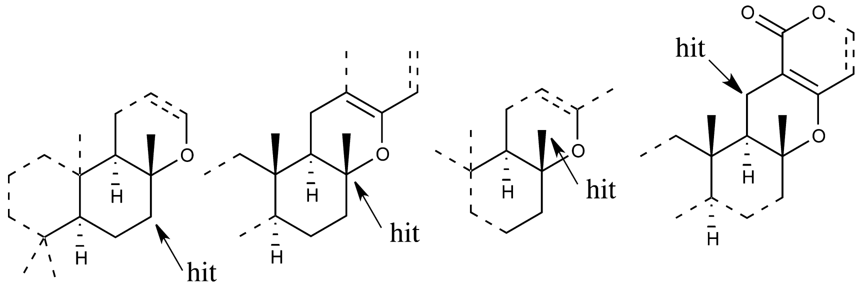


Figure 7. Fragments included in the substructure shown in Figure 6 (1).

devised a construction method based on a depth-first search of a graph, which is a basic algorithm for traversing graphs.

Let a be a hit atom of a compound in a database. Moreover, let d be an argument that indicates the rest of the traversable depth, and initially set $d := \beta$. The following algorithm, named as ConstructSubstructure(a, d), is recursive.

Step 1: If a has not been *visited* yet, then add it to the substructure being constructed and mark a as *visited*.

Step 2: If a is a hit atom, reset $d := \beta$; otherwise, set $d := d - 1$.

Step 3: For each neighboring atom a' of a, do either step 3-1 or step 3-2 if their conditions are fulfilled; otherwise, return.

Step 3-1: If a' has not been *visited* yet and (1) a' is a hit atom, (2) a' is a heteroatom, or (3) a' is a carbon and $d > 0$, then recursively call ConstructSubstructure(a', d).

Step 3-2: If a' has already been *visited*, a' is a carbon or heteroatom, $d > 1$, and a' is not a hit atom, then recursively call ConstructSubstructure(a', d).

If there is a hit atom a'' that is not marked as *visited* after the above method is applied for a, we reset $d := \beta$ and run ConstructSubstructure (a'', d) again. In this case, one obtains more than one substructure from a compound in the database.

4. GROUPING SUBSTRUCTURES AND SELECTING THEIR REPRESENTATIVES

Executing the algorithm to construct substructures creates a list of candidate substructures of the target compound. Before we can proceed to the elucidation step, the duplicate substructures have to be eliminated from the list. However, even after this elimination of duplicates, it would still be computationally intractable to elucidate a structure by using all of the remaining substructures in the list because they would produce too many candidates. Hence, the substructures need to be grouped in order to reduce the number of candidate structures further.

The list of candidate substructures is used to divide the query NMR chemical shifts into groups g in the following way:

(1) For each pair (q_i, q_j) of query NMR chemical shifts q_i and q_j, the number of substructures that include both a hit atom with respect to q_i and a hit atom with respect to q_j, denoted by $c(q_i, q_j)$, is counted. [It should be noted that $c(q, q)$ denotes the number of substructures that have a hit atom with respect to q.]

(2) Let t be a threshold. If $c(q_i, q_j)$ for query NMR chemical shifts q_i and q_j is greater than or equal to the threshold t plus half of $|c(q_i, q_i) - c(q_j, q_j)|$, then q_i and q_j are classified into the same group g.

For example, a list of candidate substructures without duplication is depicted in Figure 8, and part of its corresponding $c(\cdot,\cdot)$ table is shown in Table 1. For each hit atom a, the query NMR chemical shifts with respect to which a is hit are written beside a. If the threshold t is equal to 1, one obtains the following groups of query NMR chemical shifts for the table:

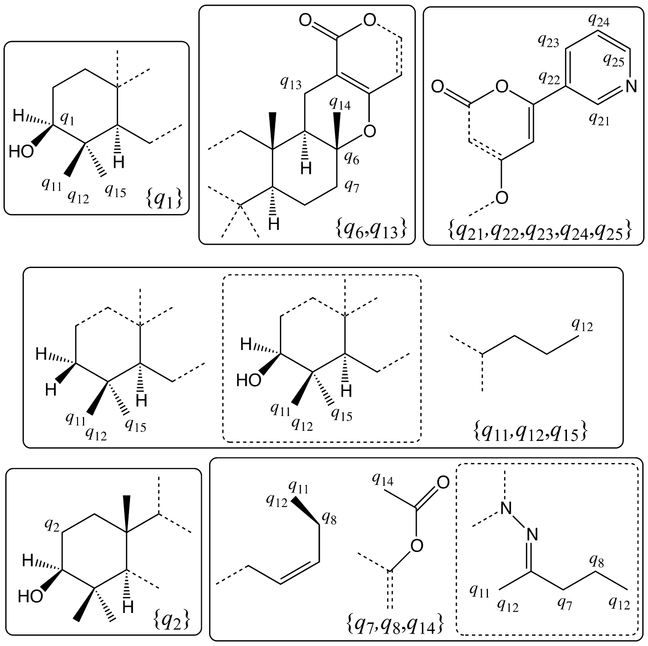


Figure 8. List of candidate substructures without duplication. The substructures are divided into groups as indicated by solid rectangles, and the representatives of the groups with more than one substructure are surrounded by dashed rectangles. The final candidate structures are elucidated by merging the common parts of the representatives.

Table 1. Part of the $c(\cdot,\cdot)$ Table

	q_7	q_8	q_9	q_{10}	q_{11}	q_{12}	q_{13}	q_{14}	q_{15}
q_7	2								
q_8	1	2							
q_9	0	0	0						
q_{10}	0	0	0	0					
q_{11}	1	2	0	0	5				
q_{12}	1	2	0	0	5	6			
q_{13}	1	0	0	0	0	0	1		
q_{14}	1	0	0	0	0	0	1	2	
q_{15}	0	0	0	0	3	3	0	0	3

$$\{q_1\}, \{q_2\}, \{q_3\}, \{q_4\}, \{q_5\}, \{q_6, q_{13}\}, \{q_7, q_8, q_{14}\}, \{q_9\},$$
$$\{q_{10}\}, \{q_{11}, q_{12}, q_{15}\}, \{q_{16}\}, \{q_{17}\}, \{q_{18}\}, \{q_{19}\}, \{q_{20}\},$$
$$\{q_{21}, q_{22}, q_{23}, q_{24}, q_{25}\}.$$

Since $t + |c(q_8, q_8) - c(q_{11}, q_{11})|/2 = 2.5$ is greater than $c(q_8, q_{11}) = 2$, the query NMR chemical shifts q_8 and q_{11} belong to distinct groups.

We classify the substructures S according to the groups g of query NMR chemical shifts as follows:

(1) For each substructure S in the list of candidate substructures, let g_S be the group of query NMR chemical shifts determined in the following way: (i) g_S is one of the groups g such that S has hit atoms with respect to some of the NMR chemical shifts in g; (ii) g_S has the minimum number of members in those groups g; and (iii) if there are several groups $\tilde{g}$ that have the minimum number of members as specified in (ii), g_S is the one that has the minimum $c(q, q)$.

(2) Put S and S′ into a group if g_S is the same as $g_{S'}$.

For each group g of substructures, the substructure in g with the maximum size, denoted as S_g, is designated to be the representative of g. The set $\{S_{g_1}, S_{g_2}, ..., S_{g_k}\}$ of such representatives S_{g_i} $(i = 1, 2, ..., k)$ is used to elucidate the structure of the target compound. The threshold t is adjusted so that the number of representatives k is not too large, as a large number of representatives would yield a huge number of candidate structures and thus make it harder to elucidate the structure of the target compound.

The substructures listed in Figure 8 are grouped as shown by the solid rectangles in the figure, and the representatives of groups with more than one substructure are surrounded by dashed rectangles.

5. ELUCIDATION: MERGING SUBSTRUCTURES

5.1. How To Elucidate a Structure.

Let $\{S_{g_1}, S_{g_2}, ..., S_{g_k}\}$ be the set of representative substructures obtained by grouping the substructures as described in section 4. A permutation $(i_1, i_2, ..., i_k)$ is a particular order of $\{1, 2, ..., k\}$. The following procedure is executed for each permutation $(i_1, i_2, ..., i_k)$.

Step 1: Let $S := S_{g_{i_1}}$ and $j := 2$.

Step 2: Find the (maximal) common part of substructure $S_{g_{i_j}}$ and substructure S.

Step 3: Merge the two substructures S and $S_{g_{i_j}}$ according to their (maximum) common part and update S with the resulting substructure.

Step 4: If $j = k$, terminate; otherwise, set $j := j + 1$ and go to step 2.

As a result, one obtains an elucidated structure S for each permutation $(i_1, i_2, ..., i_k)$. For each atom in S, the judgment as to whether the atom is hit or not is made using the same criterion as in the fragment search. If S is correct or nearly correct, the number of hit atoms is expected to be almost the same as the number of query NMR chemical shifts. The elucidated structures are ranked in terms of the numbers of hit atoms, and highly ranked structures are considered to be reliable elucidated structures. In this way, a highly accurate elucidation is achieved.

5.2. How To Find Common Parts.

We utilize a well-established technique to find the common parts of the substructures.[22−24] Because we use the database for CAST/CNMR Chemical Shift Predictor, each atom in a substructure has a CAST code,[15,20] which is determined by the functional group that includes the atom. The modular product of two substructures S and S′ is a graph $M = (V, E)$ whose vertex set V and edge set E are defined as follows. The vertex set V is composed of all pairs (a, a') of atoms a of S and a' of S′. The edge set E contains an edge $((a, a'), (b, b'))$ if a, a', b, and b' satisfy conditions (1)−(5) below:

(1) Atoms a and a' have the same CAST code.

(2) Atoms b and b' have the same CAST code.

(3) If a and a' are carbons, then a and a' satisfy the following conditions:

  (3-1) If a and a' are both hit, then the absolute difference of the NMR chemical shifts of a and a' is at most the user-designated allowance α;

  (3-2) If a or a' is not hit, then the absolute difference of the NMR chemical shifts of a and a' is at most the user-designated allowance α times a user-designated factor γ.

(4) If b and b' are carbons, then b and b' satisfy the following conditions:

  (4-1) If b and b' are both hit, then the absolute difference of the NMR chemical shifts of b and b' is at most the user-designated allowance α;

  (4-2) If b or b' is not hit, then the absolute difference of the NMR chemical shifts of b and b' is at most the user-designated allowance α times a user-designated factor γ.

(5) Either atoms a' and b' are neighbors of a and b, respectively, or a' and b' are not neighbors of a and b, respectively.

A clique of a graph is a complete subgraph, that is, a subgraph in which each vertex is adjacent to all other vertices. A clique is said to be *maximal* if there is no other clique that contains it, and it is called the *maximum* if the clique has more vertices than any of the other maximal cliques. For a clique C of the modular product M, let S(C) and S′(C) be the parts of S and S′, respectively, that are induced in such a way that atoms a and a' are members of S(C) and S′(C), respectively, if and only if C has a vertex (a, a') and, if there is a bond joining atoms a and b of S(C) in S [or a' and b' of S′(C) in S′], S(C) [or S′(C)] contains this bond. It is known that the clique C of the modular product M represents the common part of the two substructures S and S′; that is, S(C) is identical to S′(C). Therefore, the maximum clique of the modular product M yields the maximum common part of the two substructures S and S′. The problem of finding a maximum clique is NP-hard, but it can be practically solved by enumerating all of the *maximal* cliques if the modular product is not too large.[25] For the enumeration, the Bron−Kerbosch algorithm[26] can be used.

It should be noted that the maximum clique is not necessarily unique. Thus, we define the weight of a clique C as follows. For each vertex (a, a') of the clique C, (a, a') is given a weight $w(a, a')$ as follows:

(i) If both a and a' are carbons, $w(a, a')$ is defined to be the absolute difference between the NMR chemical shifts of a and a'.

(ii) If neither a nor a' is a carbon, $w(a, a')$ is defined to be zero.

The weight of clique C is defined to be the sum of the weights for all of the vertices (a, a') of C. Since the minimum weighted maximum clique is practically unique, we take it to be the maximum common part of two substructures.

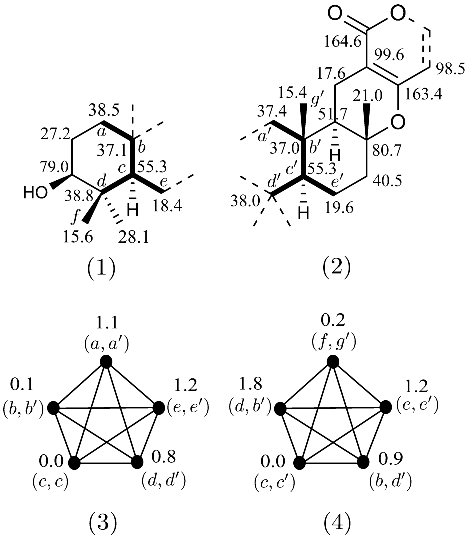


Figure 9. (1) and (2) Substructures and (3) and (4) maximum cliques C_1 and C_2, respectively, of the modular product of substructures (1) and (2). The numbers in (1) and (2) are ^{13}C NMR chemical shifts assigned to the carbons. The weight of clique C_1 shown in (3) is equal to 1.1 + 0.1 + 0.0 + 0.8 + 1.2 = 3.2, and that of clique C_2 in (4) is equal to 0.2 + 1.8 + 0.0 + 0.9 + 1.2 = 4.1. Hence, the parts corresponding to clique C_1 are chosen as the maximum common part of substructures (1) and (2) (denoted by bold bonds). Merging the maximum common parts results in an enlarged substructure that grows into the final elucidated structure.

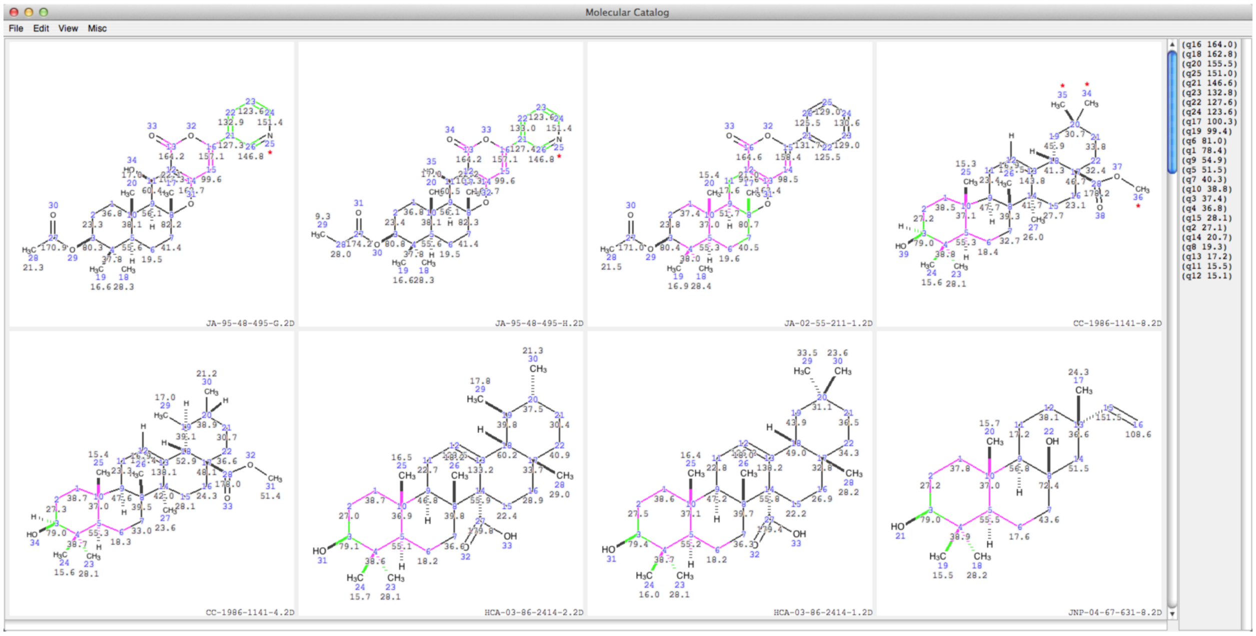
For example, let the substructures shown in Figure 9 (1) and (2) be S and S′ as above. Their modular product for $\alpha = 1.6$ and $\alpha \times \gamma = 4.0$ has two maximum cliques C_1 and C_2, as illustrated in Figure 9 (3) and (4). The number beside each vertex of the cliques represents the weight of the vertex. The weight of clique C_1 is equal to 1.1 + 0.1 + 0.0 + 0.8 + 1.2 = 3.2, and that of C_2 is equal to 0.2 + 1.8 + 0.0 + 0.9 + 1.2 = 4.1. Hence, the weight of C_1 is smaller than that of C_2. Therefore, the parts $S(C_1)$ and $S'(C_1)$ corresponding to clique C_1 are selected as the maximum common part of the two substructures, as shown by the bold bonds in Figure 9 (1) and (2).

5.3. How To Merge Two Substructures. Let C^* be the minimum weighted maximum clique of the modular product M of two substructures S and S′. Moreover, let $S(C^*)$ and $S'(C^*)$ be the (induced) maximum common part of S and S′, respectively, defined as in section 5.2. Merging of the two substructures S and S′ is basically realized by identifying $S(C^*)$ and $S'(C^*)$. What we should take care about is the valence of an atom in the merged (sub)structure. Let (a, a') be a pair of atoms in C^*. The atoms a and a' may have nonidentical neighboring atoms, which should be in $S\backslash S(C^*)$ or $S'\backslash S'(C^*)$. Then, because of the limit of the valence, the merged atom of a and a' cannot have all of the nonidentical neighboring atoms. In that case, a nonidentical neighboring atom of a or a' that has more neighbors is given priority over other nonidentical neighboring atoms of a and a' to be involved in the merged (sub)structure.

6. APPLICATION AND DISCUSSION

We used the chemical structure−NMR database originally developed for CShift-Predictor. The database consists of 2083 organic compounds, including mainly natural products and their synthetic intermediates, and it contains 50 334 carbons.

As query 1D NMR data, we again used the set of NMR chemical shifts in section 2.8: 15.1, 15.5, 17.2, 19.3, 20.7, 27.1, 28.1, 36.8, 37.4, 38.8, 40.3, 51.5, 54.9, 78.4, 81.0, 99.4, 100.3, 123.6, 127.6, 132.8, 146.6, 151.0, 155.5, 162.8, and 164.0 ppm. Each fragment was composed of the atoms connected within three bonds from a central atom (i.e., $\beta = 3$). We used the CBM criterion to search fragments and a user-designated allowance of $\alpha = 1.6$. As a result, our fragment search method output a list of candidate substructures; part of this list is shown in Figure 10,



(q16 164.0)
(q18 162.8)
(q20 155.5)
(q25 151.0)
(q21 146.6)
(q23 132.8)
(q22 127.6)
(q24 123.6)
(q17 100.3)
(q19 99.4)
(q6 81.0)
(q1 78.4)
(q9 54.9)
(q5 51.5)
(q7 40.3)
(q10 38.8)
(q3 37.4)
(q4 36.8)
(q15 28.1)
(q2 27.1)
(q14 20.7)
(q8 19.3)
(q13 17.2)
(q11 15.5)
(q12 15.1)

Figure 10. Part of the list of candidate substructures output by our fragment search method. The blue numbers are serial numbers of the atoms, and the black numbers are ^{13}C NMR chemical shifts assigned to the carbons. In each compound, a hit atom is colored green, and its candidate substructure consists of green atoms and purple atoms. The substructures depicted in Figure 9 (1) and (2) and Figure 11 (ii) are found in the fourth, third, and second molecules of the first row, respectively.

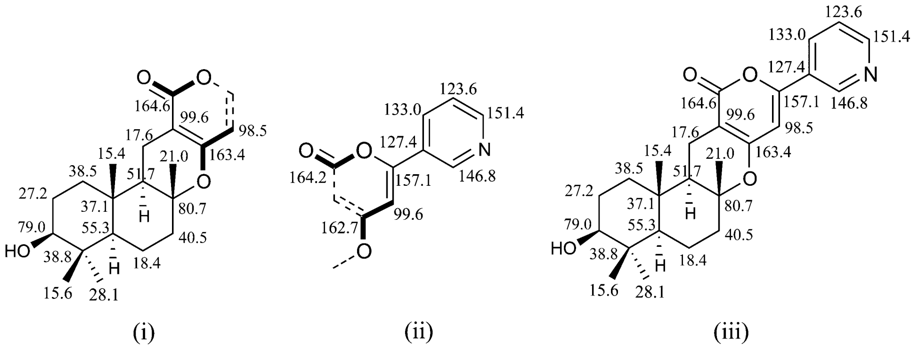


Figure 11. (i) Substructure obtained by merging the substructures in Figure 9 (1) and (2). (ii) Another representative substructure. (iii) (Sub)structure obtained by merging the substructures shown in (i) and (ii); this is the final elucidated structure. The numbers are ^{13}C NMR chemical shifts assigned to the carbons.

and the full list is available in the Supporting Information. In each compound in Figure 10, each hit atom is colored green, and its candidate substructure consists of green atoms and purple atoms. By eliminating duplicate substructures from the list, we obtained the list depicted in Figure 8, and moreover, by grouping the substructures in the list as described in Section 4, we got the groups of candidate substructures illustrated in Figure 8. Two of the representative substructures of the groups are shown in Figure 9 (1) and (2). Since the maximum common part of the two substructures is the part corresponding to the clique in Figure 9 (3), the merged substructure turns out to be the one depicted in Figure 11 (i). Another representative substructure is shown in Figure 11 (ii). The maximum common part of the substructures in Figure 11 (i) and (ii) is emphasized by bold bonds, and by merging the two substructures, we finally obtained the elucidated structure shown in Figure 11 (iii). The NMR chemical shifts of the elucidated structures are assigned preferentially in the order of the substructures Figure 9 (1), Figure 11 (i), and Figure 11 (ii). The set of query chemical shifts was reported by Itoh et al.[21] The elucidated structure in Figure 11 (iii) is the same as the one also reported by Itoh et al.,[21] that is, our elucidator succeeded in elucidating the structure. As is shown in the figure, our elucidator has the capability of elucidating a correct structure.

By the use of more sophisticated strategies, some processes could be modified in order to distinguish subtle differences in the (sub)structures they handle, and this would lead to a more accurate elucidation. In fact, we have already implemented a number of sophisticated strategies in our elucidator. For example, the conditions (3-1) and (4-1) in the definition of the modular product of two substructures in section 5.2 can be replaced by the following:

(3-1)′ If a and a' are both hit, then the fragments with a and a' as the center atoms correspond to each other by the CBM criterion for the user-designated allowance α; that is, the fragments have the same size, and the CBM algorithm returns a matching of that size.

(4-1)′ If b and b' are both hit, then the fragments with b and b' as the center atoms correspond to each other by the CBM criterion for the user-designated allowance α.

These conditions are basically stricter than conditions (3-1) and (4-1) and hence are sensitive to differences in substructures.

7. CONCLUSION

This paper has described a new chemical structure elucidation system, called the CAST/CNMR Structure Elucidator (Structure Elucidator). While Structure Elucidator embraces the conventional protocol used in many elucidation systems, it focuses especially on the efficiency of the methods it employs in order to address the issue of continuously growing databases. For this purpose, Structure Elucidator uses practically efficient graph algorithms, including the convex bipartite matching and weighted bipartite matching algorithms for its fragment search of a chemical structure−NMR database and the Bron−Kerbosch maximal clique algorithm for finding the common parts of the substructures. We applied Structure Elucidator to 1D ^{13}C NMR data and found that it can efficiently evaluate a correct structure, including its stereochemistry, by using substructures containing stereochemical information. We have already applied CShift-Predictor to make structural revisions of several natural products.[27,28] We believe that Structure Elucidator working in combination with CShift-Predictor will be a powerful tool to help users to perform practical structure analyses that yield more efficient and precise structure elucidations.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Full list of candidate substructures output by our fragment search method for the 1D ^{13}C NMR data enumerated in section 6 with $\alpha = 1.6$. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: shungo@nanzan-u.ac.jp.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors were supported by a Joint Research Grant from the National Institute of Informatics. S.K. was supported by Nanzan University Pache Research Subsidy I-A-2 for the 2011 academic year.

■ REFERENCES

(1) Lederberg, J.; Sutherland, G. L.; Buchanan, B. G.; Feigenbaum, E. A.; Robertson, A. V.; Duffield, A. M.; Djerassi, C. Applications of Artificial Intelligence for Chemical Inference I. The Number of Possible Organic Compounds: Acyclic Structures Containing Carbon, Hydrogen, Oxygen, and Nitrogen. *J. Am. Chem. Soc.* **1969**, *91*, 2973−2976.

(2) Buchanan, B. G.; Feigenbaum, E. A. Dendral and Meta-Dendral: Their Applications Dimension. *Artif. Intell.* **1978**, *11*, 5−24.

(3) Funatsu, K.; Sasaki, S. Recent Advances in the Automated Structure Elucidation System, CHEMICS. Utilization of Two-Dimensional NMR Spectral Information and Development of Peripheral Functions for Examination of Candidates. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 190−204.

(4) Bremser, W. Structure Elucidation and Artificial Intelligence. *Angew. Chem., Int. Ed. Engl.* **1988**, *27*, 247−260.

(5) Bremser, W.; Grzonka, M. SpecInfo—A Multidimensional Spectroscopic Interpretation System. *Microchim. Acta* **1991**, *104*, 483−491.

(6) Will, M.; Fachinger, W.; Richert, J. R. Fully Automated Structure Elucidation—A Spectroscopist's Dream Comes True. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 221−227.

(7) MOLGEN. http://www.molgen.de (accessed Feb 28, 2014).

(8) Advanced Chemistry Development, Inc. http://www.acdlabs.com (accessed Feb 28, 2014).

(9) Bio-Rad Laboratories Inc. http://www.bio-rad.com (accessed Feb 28, 2014).

(10) Masui, H.; Hong, H. Spec2D: A Structure Elucidation System Based on ^{1}H NMR and H−H COSY Spectra in Organic Chemistry. *J. Chem. Inf. Model.* **2006**, *46*, 775−787.

(11) Elyashberg, M. E.; Martirosian, E. R.; Karasev, Y. Z.; Thiele, H.; Somberg, H. Expert Systems as a Tool for the Molecular Structure Elucidation by Spectral Methods. Strategies of Solution to the Problems. *Anal. Chim. Acta* **1997**, *348*, 443−463.

(12) Elyashberg, M. E.; Blinov, K. A.; Martirosian, E. R. A New Approach to Computer-Aided Molecular Structure Elucidation: The Expert System Structure Elucidator. *Lab. Autom. Inf. Manage.* **1999**, *34*, 15−30.

(13) Blinov, K. A.; Elyashberg, M. E.; Molodtsov, S. G.; Williams, A. J.; Martirosian, E. R. An Expert System for Automated Structure Elucidation Utilizing ^{1}H−^{1}H, ^{13}C−^{1}H and ^{15}N−^{1}H 2D NMR Correlations. *Fresenius' J. Anal. Chem.* **2001**, *369*, 709−714.

(14) Peironcely, J. E.; Rojas-Chertó, M.; Fichera, D.; Reijmers, T.; Coulier, L.; Faulon, J.-L.; Hankemeier, T. OMG: Open Molecule Generator. *J. Cheminf.* **2012**, *4*, 21.

(15) Satoh, H.; Koshino, H.; Funatsu, K.; Nakata, T. Novel Canonical Coding Method for Representation of Three-Dimensional Structures. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 622−630.

(16) Satoh, H.; Koshino, H.; Funatsu, K.; Nakata, T. Representation of Molecular Configurations by CAST Coding Method. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1106−1112.

(17) Satoh, H.; Koshino, H.; Nakata, T. Extended CAST Coding Method for Exact Search of Stereochemical Structures. *J. Comput. Aided Chem.* **2002**, *3*, 48−55.

(18) Satoh, H.; Koshino, H.; Uzawa, J.; Nakata, T. CAST/CNMR: Highly Accurate ^{13}C NMR Chemical Shift Prediction System Considering Stereochemistry. *Tetrahedron* **2003**, *59*, 4539−4547.

(19) Satoh, H.; Koshino, H.; Uno, T.; Koichi, S.; Iwata, S.; Nakata, T. Effective Consideration of Ring Structures in CAST/CNMR for Highly Accurate ^{13}C NMR Chemical Shift Prediction. *Tetrahedron* **2005**, *61*, 7431−7437.

(20) Koichi, S.; Iwata, S.; Uno, T.; Koshino, H.; Satoh, H. Algorithm for Advanced Canonical Coding of Planar Chemical Structures That Considers Stereochemical and Symmetric Information. *J. Chem. Inf. Model.* **2007**, *47*, 1734−1746.

(21) Itoh, T.; Tokunaga, K.; Matsuda, Y.; Fujii, I.; Abe, I.; Ebizuka, Y.; Kushiro, T. Reconstitution of a Fungal Meroterpenoid Biosynthesis Reveals the Involvement of a Novel Family of Terpene Cyclases. *Nat. Chem.* **2010**, *2*, 858−864.

(22) Raymond, J. W.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.

(23) Raymond, J. W.; Willett, P.; Gardiner, E. J. Heuristics for Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Model.* **2002**, *42*, 305−316.

(24) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631−644.

(25) Kawabata, T. Build-Up Algorithm for Atomic Correspondence between Chemical Structures. *J. Chem. Inf. Model.* **2011**, *51*, 1775−1787.

(26) Bron, C.; Kerbosch, J. Algorithm 457: Finding All Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575−577.

(27) Takahashi, S.; Satoh, H.; Hongo, Y.; Koshino, H. Structural Revision of Terpenoids with a (3*Z*)-2-Methyl-3-penten-2-ol Moiety by the Synthesis of (23*E*)- and (23*Z*)-Cycloart-23-ene-3*β*,25-diols. *J. Org. Chem.* **2007**, *72*, 4578−4581.

(28) Koshino, H.; Satoh, H.; Yamada, T.; Esumi, Y. Structural Revision of Peribysins C and D. *Tetrahedron Lett.* **2006**, *47*, 4623−4626.