JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

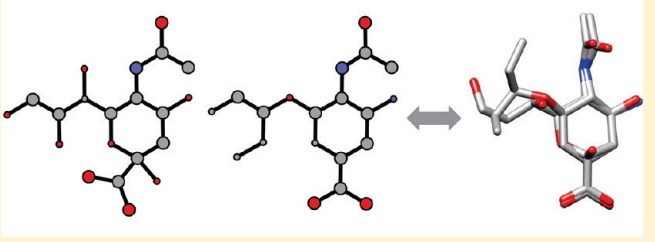
Build-Up Algorithm for Atomic Correspondence between Chemical Structures

Takeshi Kawabata[†,‡]

[†]Institute of Protein Science, Osaka University, Osaka 565-0871 Japan

[‡]Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma, NARA 630-0192 Japan

ABSTRACT: Determining a one-to-one atom correspondence between two chemical compounds is important to measure molecular similarities and to find compounds with similar biological activities. This calculation can be formalized as the maximum common substructure (MCS) problem, which is well-studied and has been shown to be NP-complete. Although many rigorous and heuristic algorithms have been developed, none of these algorithms is sufficiently fast and accurate. We developed a new program, called "kcombu" using a build-up algorithm, which is a type of the greedy heuristic algorithms. The program can search connected and disconnected MCSs as well as topologically constrained disconnected MCS (TD-MCS), which is introduced in this study. To evaluate the performance of our program, we prepared two correct standards: the exact correspondences generated by the maximum clique algorithms and the 3D correspondences obtained from superimposed 3D structure of the molecules in a complex 3D structure with the same protein. For the five sets of molecules taken from the protein structure database, the agreement value between the build-up and the exact correspondences for the connected MCS is sufficiently high, but the computation time of the build-up algorithm is much smaller than that of the exact algorithm. The comparison between the build-up and the 3D correspondences shows that the TD-MCS has the best agreement value among the other types of MCS. Additionally, we observed a strong correlation between the molecular similarity and the agreement with the correct and 3D correspondences; more similar molecule pairs are more correctly matched. Molecular pairs with more than 40% Tanimoto similarities can be correctly matched for more than half of the atoms with the 3D correspondences.



INTRODUCTION

Finding structurally similar molecules is important for solving biochemical and medicinal chemical problems: predicting metabolic pathways of various compounds in a cell and determining new drug candidates from known lead compounds.[1,2] Comparison methods for small chemical compounds can be roughly classified into two approaches: a molecular descriptor approach and an atom correspondence approach. The molecular descriptor approach uses various descriptors, such as molecular weight, log P value, number of hydrogen-bond donors, and presence or absence of predefined substructures. The values of these descriptors are calculated for all the molecules in a library and stored as a set of values, called a "fingerprint." This approach generally requires a lower computational cost, because the molecular similarity is evaluated by their fingerprints and not directly by their molecular structures. In contrast, the atom correspondence approach generates a one-to-one atom correspondence between two molecules often using their atomic properties and covalent-bond connections.[3,4] If the 3D structures of two molecules are available, then its correspondence can be made by superimposing the two 3D structures. Because the atom correspondence approach often requires more computational cost, it is mainly used for rather small sets of molecules. However, the similarities detected by the atom correspondence are more intuitively understood because similar atoms in the molecules are explicitly shown.

The atom correspondence approach between 2D chemical structures often employs the concept of the maximum common subgraph/substructure (MCS).[3] Covalently bonded atoms are regarded as a mathematical "graph" where an atom and a bond correspond to a vertex and an edge, respectively. A common substructure is defined as a substructure present in two molecules with the same atom types and bond connections. The MCS are defined as the common substructures with the largest number of atoms or bonds. Various types of MCS have been proposed so far. A subgraph induced by selecting the largest number of atom is called the maximum induced common subgraph (MCIS). In contrast, a subgraph induced by selecting the largest number of edges is called the maximum edge common subgraph (MCES). MCIS and MCES can differ, particularly for dissimilar molecule pairs. MCS can be classified further into the connected and disconnected MCS. The number of connected components of a connected MCS is one, whereas that of the disconnected MCS can be greater than one, as shown in Figure 1. The substructure search can be regarded as the special case of the connected MCS. This search tests whether all the atoms of one molecule can be mapped to a subset of the atoms of another molecule.[4] Although the

Received: March 1, 2011
Published: July 07, 2011

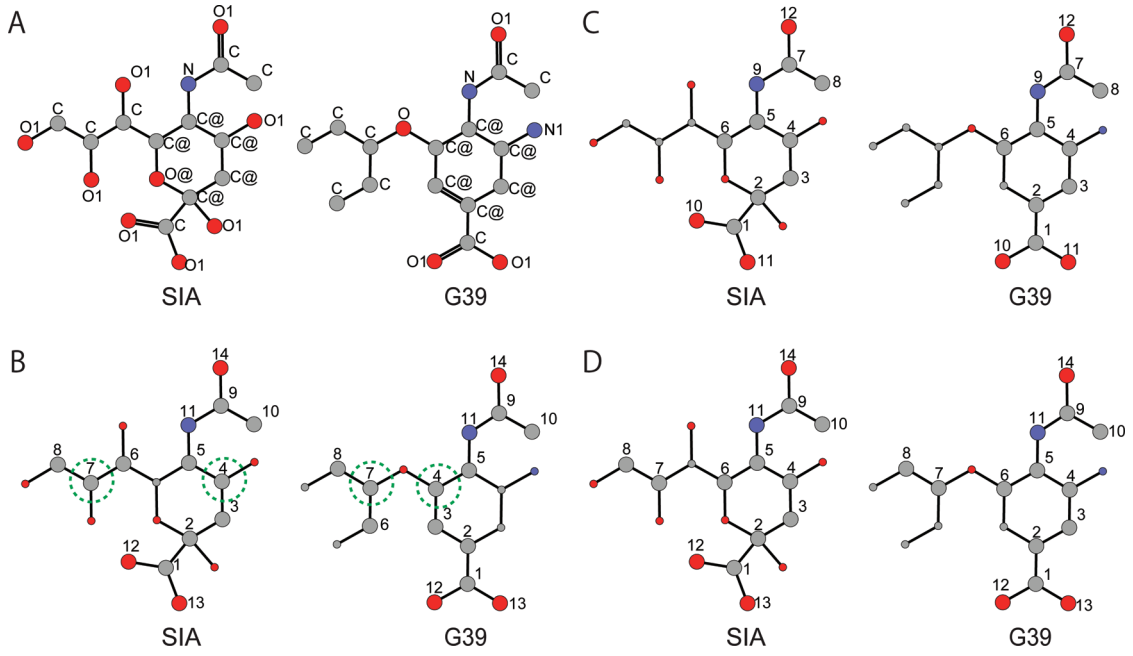


Figure 1. Example of atom types and MCSs for two neuraminidase-binding molecules, sialic acid (SIA) and oseltamivir (G39). Gray, blue, and red circles correspond to carbon, nitrogen, and oxygen atoms, respectively. In B−D, atoms with the same numbers indicate a corresponding atom pair, and noncorresponding atoms are denoted by smaller circles. (A) Atom types for two molecules. (B) D-MCS. This D-MCS does not satisfy the topological condition with $\theta < 2$. The topological distance between the atoms 4 and 7 for SIA is 4, whereas that for G39 is 2. (C) C-MCS. (D) TD-MCS using $\theta = 0$.

sensitivity of the substructure search is limited, it is popularly used because of its fast computation.

Various algorithms to calculate MCS have been proposed since the 1960s.[3,4] They are classified into exact and inexact algorithms. Exact algorithms often employ backtracking (depth-first) search algorithms[5] or maximum clique algorithms.[6−8] The computation time exponentially increases in both of these algorithms as the number of atoms compared increases. Unfortunately, calculating MCS has been proved to be NP-complete, suggesting that an exact polynomial algorithm cannot be invented for MCSs.[9] Therefore, fast approximated heuristic algorithms have also been proposed. Most heuristic algorithms are based on heuristically pruning a search tree of the exact algorithm.[10−16] Standard heuristic techniques, such as the greedy algorithm[17,18] and genetic algorithms,[19,20] have also been introduced. Some researchers have been employed reduced representations of chemical graphs by combining several atoms into one group.[21,22] Although many exact and heuristic algorithms have been developed, fast and accurate algorithms have yet to be established, and the performances of these algorithms are not fairly evaluated.

In this study, we develop a new program kcombu using a heuristic algorithm called the "build-up" algorithm to calculate MCS (MCIS). The build-up algorithm is a type of standard greedy algorithm, which is a popular heuristics in the field of combinatorial optimization. Hadagone proposed a greedy algorithm for MCS almost 20 years ago.[17] However, our build-up algorithm is more powerful than the simple greedy algorithm. The program can calculate various types of MCS (connected and disconnected MCS). To obtain more realistic disconnected substructures, we introduce a topologically constrained disconnected MCS (TD-MCS) . Similar concepts to the TD-MCS have been proposed so far.[14,21,22] TD-MCS can be calculated by the build-up algorithm and the maximum clique algorithms.

Another novel point of our study is the evaluation of calculated MCSs. We initially compare our MCSs to those calculated by the exact maximum clique algorithms.[6−8] However, agreement with exact algorithm does not guarantee biological correctness. Therefore, we also compare our results to atom correspondences obtained by superimposed 3D structures of chemical compounds bound to the same protein. These superimposed compound structures are generated by superimposing two 3D protein structures. The concept behind this test is that similar molecules are defined as molecules bound to the same target protein and that corresponding atoms are defined as atoms located in the same position on the target protein in their binding conformations. These definitions are proper at least for pharmaceutical screening of chemical compounds bound to a target protein. As far as we know, employing 3D structures as the correct standard of MCS has yet to be examined, although the accuracies of 3D molecular alignments are frequently evaluated by their 3D structures.[23,24]

MATERIALS AND METHODS

Atom Classification. Classification of the atom types affects both the sensitivities and the computational speed when calculating molecular similarities. Although a classification solely using an element type remains popular, many researchers employ more complicated classifications. For example, Berglund and Head used Tripos MOL2 atom-type definitions (25 types),[18] and Hattori et al. proposed environment-based 68 atom types.[13] The computational cost decreases as the number of atom types increases, but its recognition power becomes less sensitive to weak molecular similarities.

In this study, we introduced a simple atom classification described as a combination of element name, ring property, and number of bonded heavy atoms (degree). Figure 1A shows an

dx.doi.org/10.1021/ci2001023 |J. Chem. Inf. Model. 2011, 51, 1775–1787

example of our classification. We employ a heavy atom-based classification because we use the 3D complex structures as the correct atom correspondences, which often lack hydrogen atoms. If an atom is included in a ring structure, the character '@' is added after its element name. For example, a carbon and a nitrogen atom in a ring are called "C@" and "N@", respectively. The rings are limited to eight heavy atoms in this study, and we do not distinguish between bond types (single or double) included in rings; carbons in benzene and cycrohexane are both classified into the same class "C@". "O1" and "N1" indicate oxygen and nitrogen atoms bonded to one heavy atom, respectively. This notation distinguishes an oxygen atom in an ester bond from that in a hydroxyl or carboxyl group. The two oxygen atoms (═O and —OH) in a carboxyl group are classified into the same class "O1" because they cannot be distinguished without knowing the position of the hydrogen atom.

MCS. Next, we introduce several definitions for the MCS. This study only considers the style of MCIS.[3] Four types of MCSs are introduced: disconnected MCS (D-MCS), connected MCS (C-MCS), topologically constrained disconnected MCS (TD-MCS), and topologically constrained connected MCS (TC-MCS). The last two are new concepts.

First, we introduce the basic notations. If two molecules A and B are compared, then the set of atom pairs P is defined as follows

$$P = \{(x, y) : x \in V_A, y \in V_B, \text{type}(x) = \text{type}(y)\} \quad (1)$$

where the set V_A and V_B are the set of atoms in molecule A and molecule B, respectively, and type(x) is the atom type of atom x, defined in the previous subsection and Figure 1A. By combining several atom pairs, we introduce a one-to-one atom correspondence m, which is defined as

$$m = \{(a_1, b_1), (a_2, b_2), \ldots, (a_{|m|}, b_{|m|})\} \subseteq P \quad (2)$$

where (a_i,b_i) is the i-th atom pair of molecules A and B, and $|m|$ is the number of atom pairs contained in correspondence m. Because correspondence m is a one-to-one correspondence, it should satisfy a following condition:

$$a_i \neq a_j, b_i \neq b_j \text{ for } 1 \leq i < j \leq |m| \quad (3)$$

To construct a disconnected common substructure (D-CS), an atom correspondence m should satisfy the following connection condition:

$$\text{bond}(a_i, a_j) = \text{bond}(b_i, b_j) \text{ for } 1 \leq i < j \leq |m| \quad (4)$$

The function bond (a_i,a_j) has a value of 1, if atom a_i and a_j are connected. Otherwise the value is 0. The D-CS is defined as the common substructure induced by the atom correspondence m defined by eqs 1−4. The D-MCS is defined as the D-CS with the largest number of corresponding atom pairs $|m|$. Figure 1B and D shows two examples of disconnected MCSs.

The connected common substructure (C-CS) should satisfy the following additional condition:

$$\text{For any}(a_i, b_i) \text{ in } m, \text{ at least one } (a_j, b_j) \text{ in } m \text{ satisfies}$$
$$\text{bond}(a_i, a_j) = 1 \text{ and bond}(b_i, b_j) = 1 \quad (5)$$

The C-CS is defined as the common substructure induced by the atom correspondence m defined by eqs 1−5. The C-MCS is defined as the C-CS with the largest number of corresponding atom pairs $|m|$. Figure 1C shows an example of connected MCS. Because C-CS is a subset of D-CS, the number of corresponding atom pairs of C-MCS is not more than that of D-MCS, as shown in Figure 1B and C.

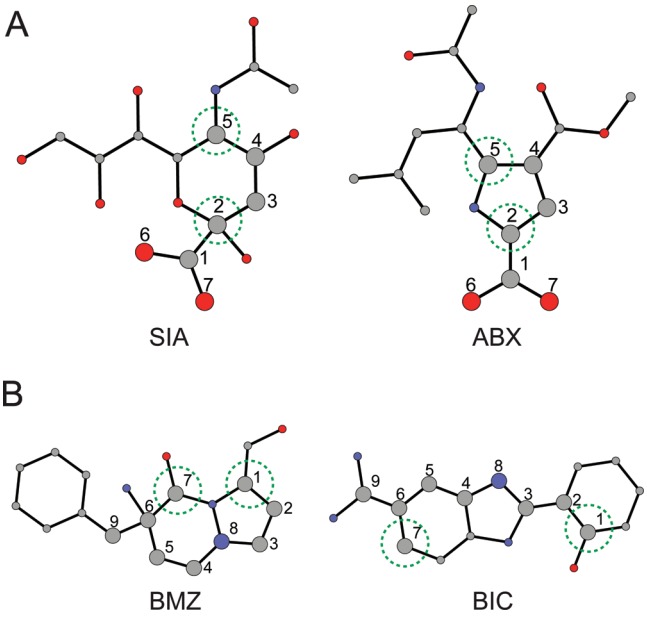


Figure 2. Example of C-MCSs not satisfying the topological constraint using $\theta = 0$. (A) C-MCS for the molecules SIA and the ABX. The topological distance between the atoms 2 and 5 for SIA is 3, whereas that for ABX is 2. (B) C-MCS for the molecules BMZ and BIC. The topological distance between the atoms 1 and 7 for BMZ is 2, whereas that for BIC is 6.

Upon observing calculated disconnected MCSs, we noticed that some of the disconnected MCSs are unrealistic, because they cannot be realized as superimposed correspondences in 3D space. Figure 1B shows an example of a nonsuperimposable correspondence. To avoid these correspondences, we introduced a topological constraint into the disconnected MCS, using topological distance. The topological distance $T(a,b)$ for two atoms a and b, is the number of bonds on the shortest path between atoms a and b. It is also referred to as the minimum path or graph distance. The topological constraint requires the difference in the distance between corresponding atom pairs must be θ or less:

$$|T_A(a_i, a_j) - T_B(b_i, b_j)| \leq \theta \text{ for } 1 \leq i < j \leq |m| \quad (6)$$

The topologically constrained disconnected common substructure (TD-CS) is defined as the common substructure induced by the atom correspondence m defined by eqs 1−4 and 6. The TD-MCS is defined as the TD-CS with the largest number of corresponding atom pairs $|m|$. Figure 1D shows an example of TD-MCS for $\theta = 0$. Figure 1B does not satisfy the topological constraint 6 for $\theta = 0$, although it satisfies that for $\theta = 2$.

Most C-CS satisfy the topological constraint with $\theta = 0$ as shown in Figure 1C, however, we did observe some counter examples (shown in Figure 2). Thus, we also introduce the topological constraint into the connected MCS and define topologically constrained connected common substructure (TC-CS) and topologically constrained connected maximum common substructure (TC-MCS). The TC-CS is defined as the common substructure induced by the atom correspondence m defined by eqs 1−6. The TC-MCS is defined as the TC-CS with the largest number of corresponding atom pairs $|m|$.

dx.doi.org/10.1021/ci2001023 |J. Chem. Inf. Model. 2011, 51, 1775–1787

Figure 3 shows a Venn diagram of the sets D-CS, C-CS, TD-CS, and TC-CS. The set of TD-CS is the subset of D-CS, and the set of TD-CS with smaller θ is the subset of TD-CS with larger θ. Similarly, the set of TC-CS is the subset of C-CS, and the set of TC-CS with smaller θ is a subset of TC-CS with larger θ. In addition, the set of TC-CS with θ is the subset of TD-CS with the same θ value.

Build-Up Algorithm. To calculate the MCS, we employ a build-up algorithm, which is widely used to build low-energy polymer structures[25−27] and to compare the 3D secondary structure elements of proteins.[28,29] The build-up algorithm is an extension of the greedy algorithm. Figure 4 summarizes the outline of the algorithm. The algorithm begins with many atom correspondences with only one atom pair, and these correspondences are sorted by their selection scores. Then the first K correspondences are taken. In the next step, another atom pair is added, and the best K correspondences are again selected. In this study, we employ $K = 40$. The build-up algorithm with $K = 1$ is equivalent to the simple greedy algorithm. The heuristic selection score $d(m)$ is introduced arbitrarily to choose good candidates of atom correspondences.

Figure 5 describes a pseudo code of the build-up algorithm. To keep candidates of atom correspondences, we introduce a set C_n, which contains several atom correspondences:

$$C_n = \{m_1, m_2, \ldots, m_{|C_n|}\} \quad (7)$$

Each correspondence m_i in the set C_n has n atom pairs. This algorithm does not employ any recursive calls, suggesting that its computational cost does not exponentially increase as the number of atoms increases. The program can search four types of MCS: D-MCS, C-MCS, TD-MCS, and TC-MCS, depending on the filtering functions and heuristic selection scores. The algorithm uses four filtering functions: Bond_Equivalence(), Bond_Connection(), Difference_of_Topological_Distance(), and Nonredundancy_by_Selection_Score(). Figure 6 shows the pseudo codes of these four filtering functions. The Bond_Equivalence() function is used for all the four types of MCSs. The Bond_Connection() function is used only in C-MCS, which verifies the condition defined in eq 4. The function Difference_of_Topological_Distance() is used in the topologically constrained MCSs (TD-MCS and TC-MCS), which verifies the condition defined in eq 6. The function Nonredundancy_by_Selection_Score() checks a uniqueness of a given atomic correspondence using a value of each heuristic selection score and a number of atoms with each atom type. The heuristic selection score $d(m)$ is described in the following section.

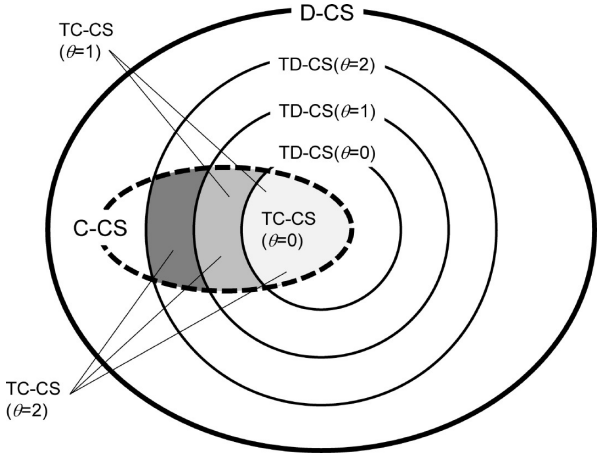


Figure 3. Venn diagram for the sets D-CS, C-CS, TD-CS, and TC-CS. The light-gray region corresponds to TC-CS with $\theta = 0$, the union of the light-gray and gray regions corresponds to TC-CS with $\theta = 1$, the union of the light-gray, gray, and dark-gray regions corresponds to TC-CS with $\theta = 2$.

Sort (x,y) **in** P **by** $d(\{(x,y)\})$
C_1 = first K correspondences in P.
$d_{worst} = d(m_{|C1|})$ for the correspondence $m_{|C1|}$ in C_1
n=1

while (n<=max[$|V_A|$,$|V_B|$]):
  $C_{n+1}=\phi$
  for m **in** C_n:
    for (x,y) **in** P:
      $m' = m + \{(x,y)\}$
      if (x is not in m) **and** (y is not in m) **and**
        Bond_Equivalence(m, x, y) **and** Bond_Connection(m, x, y) **and**
        Difference_of_Topological_Distance(m, x, y,θ) **and**
        (m' is not in C_{n+1}) **and** Nonredandunncy_By_Selection_Score(m', C_{n+1}) **and**
        ((|C_{n+1}|<K) **or** ($d(m')$<= d_{worst})):
          Add m' **to** C_{n+1}
          Sort m_i **in** C_{n+1} **by** $d(m_i)$
          Delete m_i **in** C_{n+1} **for** K+1 <=i<= |C_{n+1}|
          $d_{worst} = d(m_{|Cn+1|})$ for the last correspondence $m_{|Cn+1|}$ in C_{n+1}.
    if ($C_{n+1}=\phi$):
      Output C_n as optimal correspondences. Escape from while loop.
    else:
      $n = n + 1$

Figure 5. Pseudo code of the build-up algorithm.

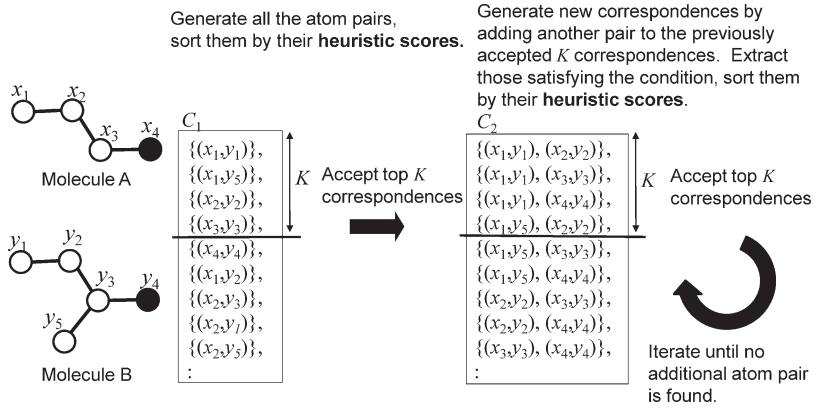


Figure 4. Outline of the build-up algorithm.

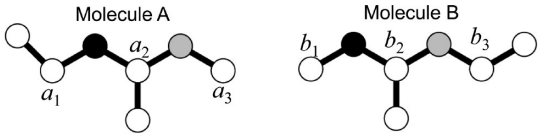
```
Bond_Equivalence(m, x, y):
  for (a,b) in (m):
    if (bond(a, x)!=(bond(b, y)):
      return(false)
  return(true)

Bond_Connection(m, x, y):
  for (a,b) in (m):
    if (bond(a,y)==1) and (bond(b,y)==1):
      return(true)
  return(false)

Difference_of_Topological_Distance(m, x, y, θ):
  for (a,b) in (m):
    if (|T_A(a, x)-T_B(b,y)|>θ):
      return(false)
  return(true)

Nonredandancy_By_Selection_Score(m', C):
  for m in (C):
    if (d_neighbor(m')==d_neighbor(m)) and (d_ec(m')==d_ec(m)) and
      (d_topology(m')==d_topology(m)) :
        for e in (AllTheAtomTypes):
          if (N(e, m')!=N(e,m)):
            return(false)
  return(true)
```

Figure 6. Pseudo codes for the four filtering functions included in the build-up algorithm. $N(e,m)$ is the number of e-type atoms contained in the atom correspondence m.



		b_1	b_2	b_3
	$N_A(e,a)$ / $N_B(e,a)$	C:0,O:1,N:0	C:1,O:1,N:1	C:1,O:0,N:1
a_1	C:1,O:1,N:0	1	1	2
a_2	C:1,O:1,N:1	2	0	1
a_3	C:0,O:0,N:1	2	2	1

Figure 7. Example of the d_{neighbor} score for atom types of neighboring atoms. White, gray, and black circles correspond to carbon, nitrogen, and oxygen atoms, respectively.

Heuristic Selection Score. We introduce a heuristic selection score $d(m)$ for atom correspondence m. The heuristic selection score $d(m)$ consists of three scores: d_{neighbor} score for atom types of bonded atoms, d_{ec} score for the extended connectivity, and d_{topology} score for difference of topological distance.

The d_{neighbor} score is the difference in the numbers of neighbor bonded atoms of specific atom types, which can be defined as

$$d_{neighbor}(a,b) = \sum_{e \in \text{ all the atom types}} |N_A(e,a) - N_B(e,b)| \quad (8)$$

where $N_A(e,a)$ is the number of e-type atoms connected to focused atom a for the molecule A. The previous section and Figure 1A describe atom types in detail. Figure 7 shows an example of the d_{neighbor} score. Hadagone has employed a similar score with d_{neighbor} score.[17]

The d_{ec} score is a difference in the extended connectivity value, which was first introduced by Morgan for the unique description of chemical structure,[30] which can defined as follows

$$d_{ec}(a,b) = |EC_A^r(a) - EC_B^r(b)| \quad (9)$$

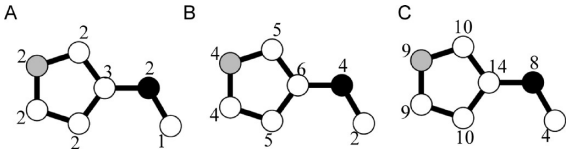


Figure 8. Example of the extended connectivity values for the d_{ec} score: (A) round 0; (B) round 1; and (C) round 2.

where $EC_A^r(a)$ is r-round extended connectivity value for atom a in molecule A. The 0-round value $[EC_A^0(a)]$ is the number of connected heavy atoms with atom a (degree of chemical graphs). For the next round ($r = 1$), the value of $EC_A^1(a)$ is the sum of the assigned EC_A^0 values for the atoms connected to atom a. Herein, we employed a two-round extended connectivity ($r = 2$). Figure 8 shows an example of the EC values. Chen and Robien have employed a similar score with d_{ec} score to select starting atom pairs.[10]

The d_{topology} score is a difference in topological distances between two close atom pairs, which can be defined as

$$d_{topology}(a_i,a_j,b_i,b_j) = \begin{cases} |T_A(a_i,a_j) - T_B(b_i,b_j)| & \text{if} \min[T_A(a_i,a_j), T_B(b_i,b_j)] \le \alpha \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where topological distance $T_A(a_i,a_j)$ is the number of bonds in the shortest path for two atoms a_i,a_j in molecule A. In this study, we employ $\alpha = 4$. Topological distances for all the pairs of atoms are calculated by the Floyd–Warshall algorithm.[31] The score d_{topology} is used only to determine the topologically constrained MCS (TD-MCS and TC-MCS) using $\theta > 0$.

The total selection score d is defined as the sum of these three scores with weights as

$$\begin{aligned} d(m) &= d(\{(a_1,b_1),(a_2,b_2), ..., (a_{|m|},b_{|m|})\}) \\ &= \sum_{i=1}^{|m|} [w_{neighbor} \cdot d_{neighbor}(a_i,b_i) + w_{ec} \cdot d_{ec}(a_i,b_i)] \\ &+ w_{topology} \sum_{i=1}^{|m|} \sum_{j=i+1}^{|m|} d_{topology}(a_i,a_j,b_i,b_j) \end{aligned} \quad (11)$$

where w_{neighbor}, w_{ec}, and w_{topology} are weight parameters for the score d_{neighbor}, d_{ec}, and d_{topology}, respectively. In this study, we employ an equal weighting: $w_{\text{neighbor}} = w_{\text{ec}} = 1$ and $w_{\text{topology}} = 0$ for D-MCS and C-MCS and $w_{\text{neighbor}} = w_{\text{ec}} = w_{\text{topology}} = 1$ for TD-MCS and TC-MCS. These values for weights have room for improvement because we did not systematically test other values.

Similarities between Two Molecules and Agreement Measure for Two Correspondences. To evaluate similarities between two molecules A and B by atom correspondence m, the following Tanimoto (Jaccard) coefficient is employed:

$$\text{similarity}(V_A, V_B, m) = \frac{|m|}{|V_A| + |V_B| - |m|} \quad (12)$$

where $|m|$ is the number of corresponding atom pairs for correspondence m and $|V_A|$ and $|V_B|$ are the number of heavy atoms in molecule A and B, respectively. This similarity values depend on the type of MCS and the algorithm used to determine

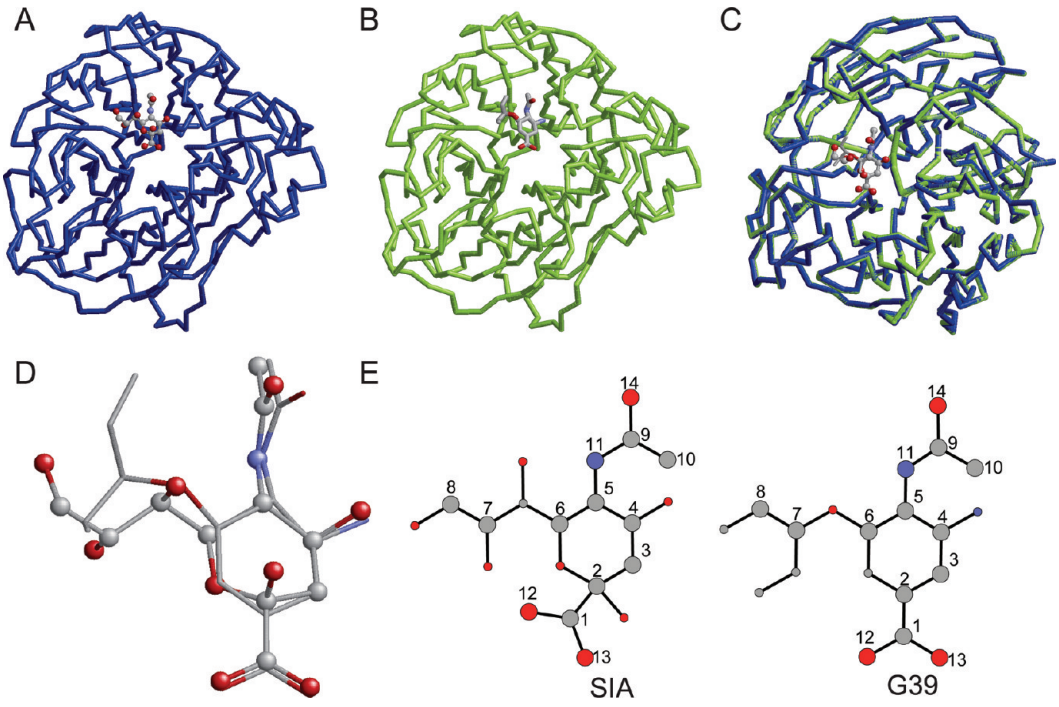


Figure 9. Construction of a disconnected MCS using two 3D structures of protein−ligand complexes. (A) Complex structure of neuraminidase and SIA, PDB code: 1mwe. (B) Complex structure of neuraminidase and G39, PDB code: 2qwh. (C) Complex structures with bound ligands superimposed for their protein structures. (D) Superimposed ligand structures. Ball and stick model shows SIA, while wire frame model indicates G39. (E) 3D atom correspondence derived from the superimposed ligand structures. This correspondence is the same as the MCS shown in Figure 1D.

correspondence m. As shown in Figure 1, the number of corresponding atom pairs $|m|$ for the D-MCS is often larger than that for the C-MCS. Additionally, similarities by the exact algorithms are equal or larger than that by the heuristic build-up algorithms.

A following measure is introduced to evaluate the agreement between two atom correspondences m_1 and m_2:

$$\text{agreement}(m_1, m_2) = \frac{|m_1 \cap m_2|}{|m_1| + |m_1| - |m_1 \cap m_2|} \qquad (13)$$

where $|m_1 \cap m_2|$ is the number of common atom pairs shared by two correspondences m_1 and m_2. This agreement value is also based on the concept of Tanimoto (Jaccard) coefficient.

Exact Atom Correspondences Generated by Maximum Clique Algorithms. To evaluate the performance of our build-up algorithm, we compare our results to those using the exact algorithms. We implemented two maximum-clique algorithms: Bron−Kerbosch algorithm for the D-MCS and TD-MCS[6] and c-clique algorithm for the C-MCS and TC-MCS.[7,8] These algorithms often require larger computational costs than heuristic algorithms but have been shown to generate correct MCSs.

The problems to find D-MCS and TD-MCS are transformed into maximum clique problems for a vertex product graph. Vertices in the product graph correspond to atom pairs with the same atom types (the set P defined in eq 1). Edges in the product graph are classified into c-edges (connected edges) and d-edges (disconnected edges). Vertices (a_i,a_j) and (b_i,b_j) are connected by the c-edge if bond (a_i,a_j) = bond (b_i,b_j) = 1, whereas they are connected by the d-edge if bond (a_i,a_j) = bond (b_i,b_j) = 0. The maximum clique which consists of the c- or d-edge corresponds to the D-MCS and can be enumerated by the Bron−Kerbosch algorithm.[6] A c-clique is defined as a clique which consists of the c- or d-edge, if it is a connected graph of c-edges. The maximum c-clique corresponds to the C-MCS and can be enumerated by the modified Bron−Kerbosch algorithm, which was proposed by Koch[7] and corrected by Cazals and Karande.[8]

The topologically constrained MCS (TD- and TC-MCS) can also be calculated by the maximum clique algorithms, if the definition of the d-edge is modified. When the TD- and TC-MCS are calculated, the d-edge between vertices (a_i,a_j) and (b_i,b_j) is generated if bond (a_i,a_j) = bond (b_i,b_j) = 0 and $|T_A(a_i,a_j) - T_B(b_i,b_j)| \leq \theta$. Thus, TD- and TC-MCSs can be exactly calculated by the Bron−Kerbosch algorithm and the maximum c-clique algorithm, respectively.

3D Atom Correspondences Generated by Superimposed 3D Structures of Molecules. Agreement of an atom correspondence with that by the exact algorithm does not guarantee that the correspondence has biologically meaning. Hence, we introduce a "3D correspondence" for a biologically correct atom correspondence. This correspondence is determined from superimposed 3D structures of two molecules. If the two molecules can bind to the same protein, then atoms located at the same 3D position on the protein structure are defined as a corresponding atom pair.

To generate representative superimposed molecules, the following procedure is employed. First, we generate clusters of protein chains registered in the Worldwide Protein Data Bank (wwPDB),[32] using the single-linkage clustering method with 95% sequence identity as the threshold. Second, we select protein chains from each cluster with nonredundant bound molecule (ligand) types. Third, all the protein chains are superimposed into a representative chain, using the program MATRAS;[29] bound molecules are also translated and rotated together with their proteins. Finally, we obtain superimposed 3D structures of molecules.

Table 1. Data Set of the Superimposed Ligand−Protein Structures[a]

	descriptions	PDB code	N_{lig}	avg N_{hv}	min N_{hv}	max N_{hv}	similarity(%)
HIV	HIV protease	1hsg A(MK1)	166	43.3	7	80	18.3
THR	thrombin	1ad8 A (MDL)	154	29.2	8	55	21.3
CDK2	cyclin-dependent protein kinase 2	2wih A(P48)	154	24.6	9	36	23.7
CAH2	carbonic anhydrase 2	1oq5 A(CEL)	116	19.7	8	28	29.8
NEU	neuraminidase	2qwh A(G39)	15	21.3	20	28	49.3
total			605	29.9	7	80	22.4

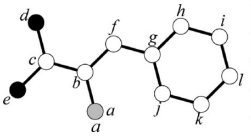
[a] PDB code: PDB code for the representative complex structure. Three letters in the parentheses are ligand ID used in PDB. N_{lig}: number of superimposed ligands, and avg N_{hv}, min N_{hv}, max N_{hv}: average, maximum, and minimum number of heavy atoms, respectively. Similarity (%): average similarity value using the exact C-MCS correspondence.

From these superimposed 3D structures, 3D atom correspondences are generated by collecting superimposed atoms with the same atom types that satisfy the condition of disconnected common substructure. We employ the following greedy algorithm. First, for two superimposed structures, we measure geometric distances between the positions of the same type of atoms and then sort these atom pairs by distance. The largest tolerant difference in the geometric distance is 2 Å. Second, we prepare an empty list of atom pairs, and the closest atom pairs are added to the list one by one, if they satisfy the conditions of a disconnected common substructure (defined in eqs 1−4). Figure 9 shows a example of a 3D atom correspondence.

For the evaluation, we used the ligands of the HIV protease, which is a homo dimeric protein; its 3D structure has a 2-fold symmetry. Two transformations are feasible to superimpose two 3D structures with 2-fold symmetries. Therefore, two superimposed 3D structures for each ligand in the HIV data set are generated. These two 3D structures provide two 3D correspondences for each molecular pair. Between the two 3D correspondences, the correspondence with a larger number of corresponding atom pairs is selected for the correct standard.

Data Set of the Superimposed Ligand−Protein Structures. From the wwPDB database[32] downloaded on December 1, 2010, we generated 28 889 clusters by the procedure described in the previous subsection. Among these clusters, we selected 5 clusters which contain more than 100 unique bound ligands in their 3D structures: HIV protease (166 ligands), thrombin (154 ligands), cyclin-dependent protein kinase 2 (154 ligands), carbonic anhydrase 2 (116 ligands), and neuraminidase (15 ligands). Then the target proteins of these ligands are superimposed into their representative proteins. Table 1 summarizes the PDB codes of representative proteins and the averaged number of heavy atoms in the data set. The molecules in the carbonic anhydrase 2 data set are restricted to 28 heavy atoms, to apply the exact D-MCS algorithm, which requires large computational costs.

Generated Equivalent Atom Correspondences for 3D Atom Correspondences. To calculate the agreement value for the 3D atom correspondence, the symmetries of the molecules are considered. Some molecules have equivalent atom sets that cannot be distinguished from their physicochemical properties. For example, the six carbon atoms in a benzene molecule are equivalent, and two oxygen atoms in carboxyl group are equivalent. To enumerate these equivalent atoms, we identified all the permutations of the atoms that have a connection table identical to the original table.[33] Figure 10 shows an example of a permutation. To calculate the agreement value between the build-up and 3D correspondences, the following procedure is implemented. First, all the permutations for two molecules being compared



	a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**
ρ1	a	b	c	**e**	**d**	f	g	h	i	j	k	l
ρ2	a	b	c	d	e	f	g	**j**	**k**	**h**	**i**	l
ρ3	a	b	c	**e**	**d**	f	g	**j**	**k**	**h**	**i**	l

Figure 10. Three atom permutations ρ1, ρ2, and ρ3 for a phenylalanine molecule. White, gray, and black circles correspond to carbon, nitrogen, and oxygen atoms, respectively.

are generated. Then using these two permutations and the 3D atom correspondence, we generate the correspondences equivalent to the given 3D correspondence. Finally, we calculate the maximum agreement value for the build-up correspondence and for all of the generated equivalent 3D correspondences. This permutation process is necessary only to compare 3D correspondences but not to compare the exact correspondence because the exact algorithms can enumerate all correspondences with the largest number of atom pairs, including equivalent atom permutations.

Implementation and Availability. We implemented our build-up algorithm as a C source code on the Linux platform. We call our program "kcombu" (Kemical compound COMparison using Build-Up algorithm). The kcombu program package also includes the exact maximum clique algorithms. We have developed a Web server (http://strcomp.protein.osaka-u.ac.jp/kcombu) where users can input a ligand query to search for similar ligands registered in the wwPDB database. Academic users can download the source code from the server. Additionally, the data set of superimposed molecular structure can be downloaded from the server.

RESULTS

Comparison with the Exact Atom Correspondences. We calculated five types of MCS using the build-up algorithm for all combination of molecules in each data set. The exact C-MCS was calculated for all the five sets, however, the exact TD-MCS was calculated for only the cyclin-dependent protein kinase 2 (CDK2), carbonic anhydrase 2 (CAH2), and neuraminidase (NEU) data sets. The exact D-MCS was calculated for only the CAH2 and NEU data sets (Table 2). It is because the exact TD-MCS and D-MCS algorithms require an extremely long computation time for pairs of large molecules. The build-up algorithm can output K different atom correspondences, we only used the best correspondence with the lowest heuristic distance score in the evaluation. The exact algorithms output all the correspondences with the maximum number of corresponding atom pairs. We calculated all the agreement values for one build-up correspondence

versus all the exact correspondences. Among these values, we employed the largest agreement value for the evaluation.

Table 2 and Figure 11 summarize the average agreement values between build-up and exact correspondences. The agreement value of the C-MCS for the total data set is 89.6%. Considering the large computational costs for the exact algorithm shown in the next subsection, we think the agreement is satisfactory. The agreement value of C-MCS is the highest in the same data set, that of D-MCS is the lowest, and that of TD-MCS is between those for C-MCS and D-MCS. These agreement values negatively correlate with their sizes of search space shown in Figure 3. As the search space becomes larger, the build-up algorithm more often fails to find the correct correspondence.

The agreement values for C-MCS vary with the data set; the neuraminidase set has the highest value (100%), and the HIV protease set has the lowest value (80.9%). The poor results for the HIV protease set may be because the HIV protease data set has more dissimilar pairs of larger molecules, as shown in Table 1. Figure 12 shows a clear positive correlation between the agreement and the similarity. Most of the molecular pairs with more than 40% similarity have almost 100% agreement with the exact correspondence. In contrast, some of the molecular pairs with less than 40% have nearly 0% agreement.

Figure 13 shows the effect of the parameter K in the build-up algorithm on the agreement value with the exact correspondence. The agreement value with our default value $K = 40$ is much higher than those for the simple greedy algorithm ($K = 1$). We also found that the agreement value almost converges when the parameter $K = 40$.

Computation Times of the Build-Up and the Exact Algorithms. Next, we compared computation times of the exact and build-up algorithms. In the calculation, we used one core of Intel Xeon processor X5355 (2.56 GHz) in a Linux environment. Figure 14 plots the computation time versus the number of heavy

Table 2. Average Agreement Values (%) between Build-up and Exact Correspondences

	HIV	THR	CDK2	CAH2	NEU	Total
C-MCS	80.9[a]	92.7	91.9	97.5	100.0	89.6
TD-MCS ($\theta = 0$)	ND	ND	86.4	97.7	100.0	ND
TD-MCS ($\theta = 1$)	ND	ND	77.3	95.8	97.5	ND
TD-MCS ($\theta = 2$)	ND	ND	75.1	95.4	97.2	ND
D-MCS	ND	ND	ND	78.8	89.0	ND

[a] The molecule pair GGX and GGV is not included because the calculation of its exact C-MCS required more than one month.

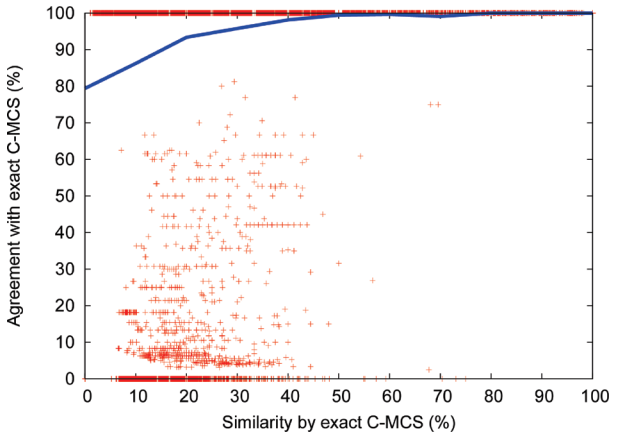


Figure 12. Plot of agreement values between the build-up and exact correspondence for C-MCS versus molecular similarities based on the exact correspondence. Blue line indicates the average agreement values.

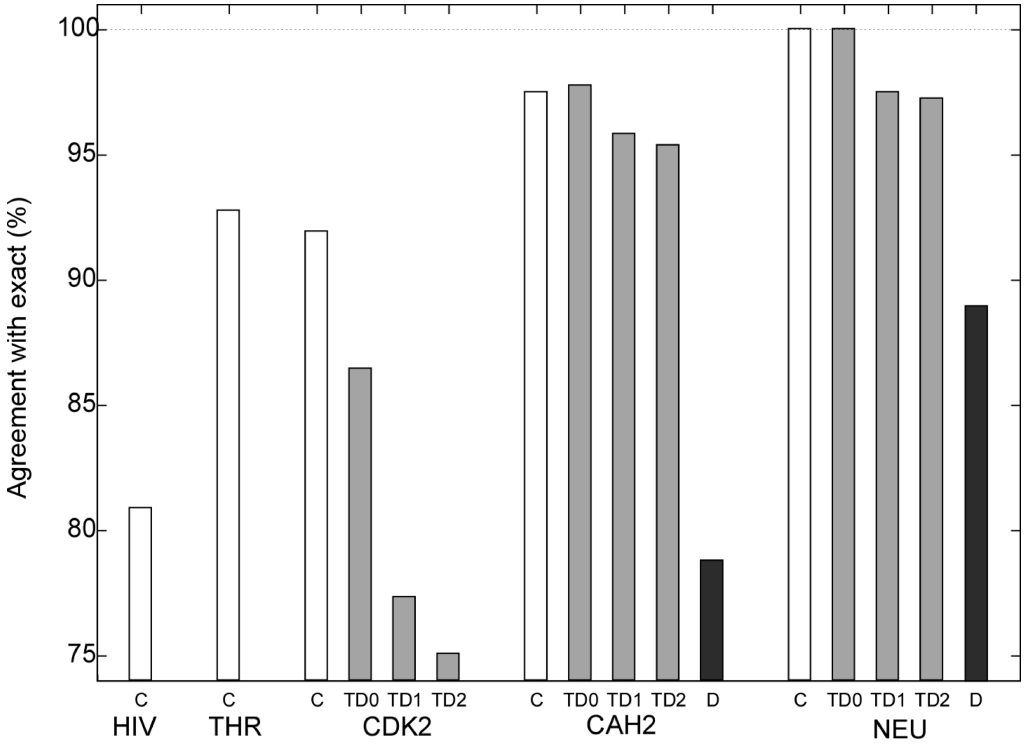


Figure 11. Average agreement values between the build-up and exact correspondences for the five data sets. C and D correspond to C-MCS and D-MCS, respectively. TD0, TD1, and TD2 correspond to TD-MCS using $\theta = 0-2$, respectively.

dx.doi.org/10.1021/ci2001023 |J. Chem. Inf. Model. 2011, 51, 1775–1787

atom pairs (product of number of heavy atoms for each molecule, $|V_A| \times |V_B|$). Figure 14A and B corresponds to the results of the exact maximum clique and build-up algorithms, respectively. For the both algorithms, the D-MCS (red points) generally requires more time than the C-MDS (blue points). The computation time for TD-MCS (green points) is almost between those for D-MCS and C-MCS.

Both exact and build-up algorithms require small computation time to compare a pair of small molecules. To compare a pair of molecules where numbers of heavy atoms are about 20 ($|V_A| \times |V_B| = 400$), the exact algorithms for C-, TD-, and D-MCS require 0.01, 0.1, and 10 s, respectively. However, the computation times of exact algorithm increase rapidly with $|V_A| \times |V_B|$. Especially, the exact D-MCS algorithm (red points in Figure 14A) shows the clear linear relationship between the logarithm of the computation time and the number $|V_A| \times |V_B|$. It indicates that the computation time of exact D-MCS algorithm increases almost exponentially with $|V_A| \times |V_B|$ (0.0000361*exp[0.0269*$|V_A| \times |V_B|$]). Hence, the exact algorithm cannot calculate the D-MCS for pairs of large molecules in a reasonable time. For example, calculating two HIV protease inhibitors "BAY" and "IPF", where numbers of heavy atoms are 80 and 66, respectively, may require 10^{49} years. The exact algorithm for C-MCS (blue points in Figure 14A) requires a much shorter computation time than that for D-MCS, and a clear exponential relationship is not observed. However, for some pairs of larger molecules, the exact C-MCS algorithm can also require a much longer computation time. For example, the exact C-MCS comparison between the molecules "A79" and "AI", which have 58 heavy atoms, requires 362 687 s (= 4.2 days), whereas the build-up C-MCS comparison for these molecules requires 0.21 s. The exact C-MCS comparison between the molecules "GGX" and "GGV", which have 60 and 63 heavy atoms, respectively, was not finished after a month.

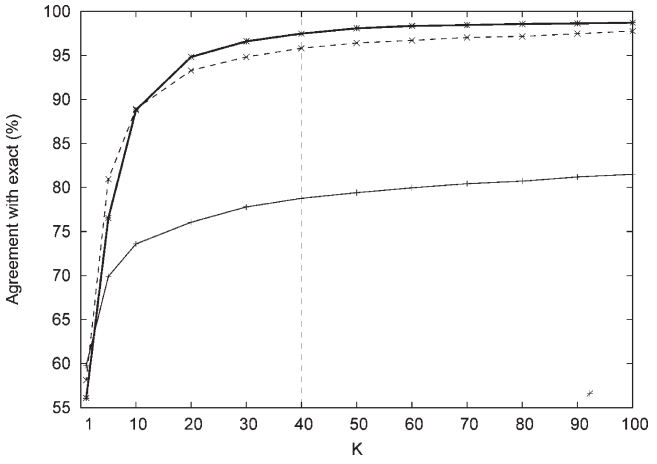


Figure 13. Plot of agreement values between the build-up and the exact correspondences versus the parameter K in the build-up algorithm. The CAH2 data set was used for the evaluation. Bold, thin, and dotted lines correspond to C-MCS, D-MCS, and TD-MCS using $\theta = 1$, respectively.

The build-up algorithms do not show an exponential relationship between the computation time and the number of heavy atoms (Figure 14B). Typically, the build-up algorithms require less than one second. The longest computation times for the build-up D-MCS and C-MCS algorithms are 3.27 and 0.27 s, respectively. These results demonstrate the advantage of the build-up algorithm with respect to the computational speed; this advantage compensates for the inaccuracy of the build-up algorithm, shown in the previous section.

Comparison with the 3D Atom Correspondences. Next, we compared the build-up correspondences with the 3D atom correspondences derived from the superimposed 3D structures of molecules. For the five data sets, we calculated eight types of MCS using the build-up algorithm: C-MCS, D-MCS, TC-MCS, and TD-MCS with $\theta = 0-2$. For the reference, we calculated the average agreement value for the exact C-MCS algorithm with the 3D correspondence. Because the exact algorithm outputs multiple correspondences, we employed the average agreement value for the evaluation of exact C-MCS algorithm. Table 3 and Figure 15 show the average agreement values for the total data set. D-MCS exhibits the worst agreement, and C-MCS is much better than D-MCS. We speculate that D-MCS often includes inconsistencies in the topological distance, which leads to the low agreement with the 3D correspondences. The agreement value of exact C-MCS is slightly lower than that of build-up MCS because the exact MCS was evaluated by the average agreement value among generated multiple correspondences, whereas the build-up MCS was evaluated by one correspondence with the best heuristic score. The agreement values of TC-MCS are almost similar to that of C-MCS, indicating that the topological constraint did not improve the performance of C-MCS. In contrast, TD-MCS exhibits much better agreement than D-MCS, and it also exhibits

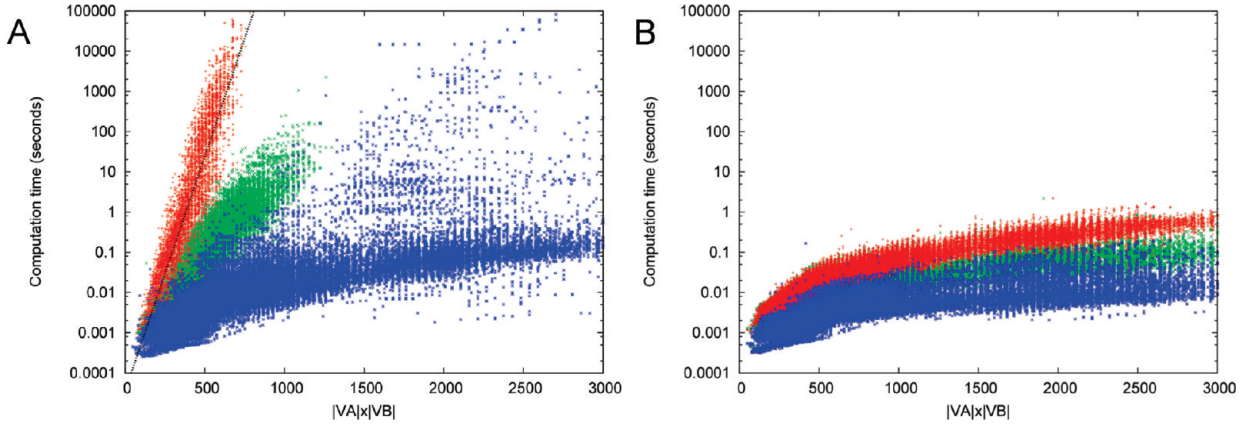


Figure 14. Plot of the computation time to calculate MCS versus the product of the number of heavy atoms $|V_A| \times |V_B|$. Red, blue, and green points correspond to D-MCS, C-MCS, and TD-MCS using $\theta = 1$, respectively. (A) Plot for the exact maximum clique algorithm. (B) Plot for the build-up algorithm.

dx.doi.org/10.1021/ci2001023 |J. Chem. Inf. Model. 2011, 51, 1775–1787

significantly better agreement than C-MCS. Among the three θ values, $\theta = 1$ provides the best results. Figure 16 shows one of the molecule pairs in the thrombin data set and demonstrates the correspondence of TD-MCS with $\theta = 1$ agrees well with the 3D correspondence.

The agreement values depend on the similarities of the 3D correspondences. Table 4 shows the average agreement values for molecule pairs when the similarities based on 3D atomic correspondences are more than 40%. Compared to Table 3, the values in Table 4 are much higher. These results show that molecule pairs with a larger number of superimposed atoms tend to have better agreement values between the build-up and the 3D correspondences. Figure 17 graphically summarizes the relationship between the molecular similarity and the agreement with the 3D correspondence. Figure 17A plots the agreement versus molecular similarity based on the 3D correspondence, while Figure 17B plots molecular similarity based on the build-up algorithm (TD-MCS with $\theta = 1$). Both plots indicate that the agreement with 3D correspondence is correlated to the molecular similarity, although the similarity based on 3D correspondence has a better correlation than that based on the build-up correspondence. These observations suggest that molecular similarity using the build-up atom correspondences is a good indicator of its accuracy. The results herein confirm that molecular pairs with 40% similarities can be correctly matched for more than half of the atoms.

Table 3. Average Agreement Values for All the Molecule Pairs for the Build-Up Algorithms

	HIV	THR	CDK2	CAH2	NEU	total
N_{pair}[a]	13 695	11 781	11 781	6670	105	44 032
C-MCS	24.7	24.2	14.0	49.5	70.1	25.6
C-MCS (exact)	24.0	22.9	12.5	47.4	67.9	24.3
TC-MCS ($\theta = 0$)	24.7	24.0	13.9	50.2	69.1	25.6
TC-MCS ($\theta = 1$)	24.6	24.3	13.9	50.3	70.1	25.7
TC-MCS ($\theta = 2$)	24.7	24.2	13.9	49.6	70.1	25.5
TD-MCS ($\theta = 0$)	24.7	25.0	16.2	58.4	83.9	27.8
TD-MCS ($\theta = 1$)	26.4	26.2	15.9	57.1	88.2	28.4
TD-MCS ($\theta = 2$)	26.2	26.8	15.7	55.4	87.3	28.1
D-MCS	16.2	18.1	11.3	41.6	73.7	19.4

[a] N_{pair}: number of ligand pairs.

DISCUSSION

The build-up algorithm introduced in this study is quite powerful. It can calculate C-MCS atom correspondences with 89.6% agreement with the exact correspondence using a much shorter computation time. The computation times for exact algorithms increase exponentially with number of atoms, whereas those for the build-up algorithms do not increase rapidly (Figure 14). The lower computational cost allows the build-up algorithm to search

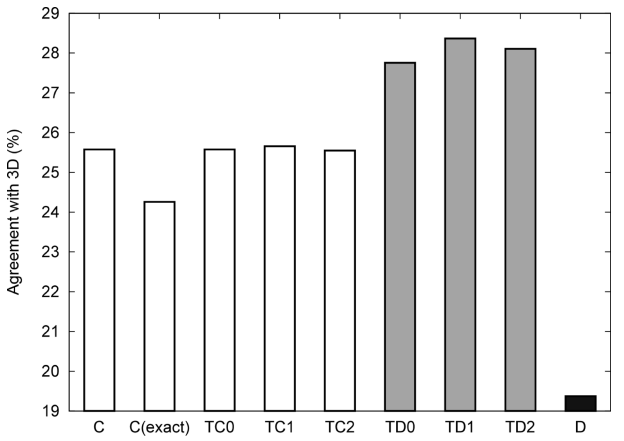


Figure 15. Average agreement values between the build-up and the 3D correspondences for the total data set. C and D correspond to C-MCS and D-MCS, respectively. TC0, TC1, and TC2 correspond to TC-MCS using $\theta = 0-2$, respectively. TD0, TD1, and TD2 correspond to TD-MCS using $\theta = 0-2$, respectively.

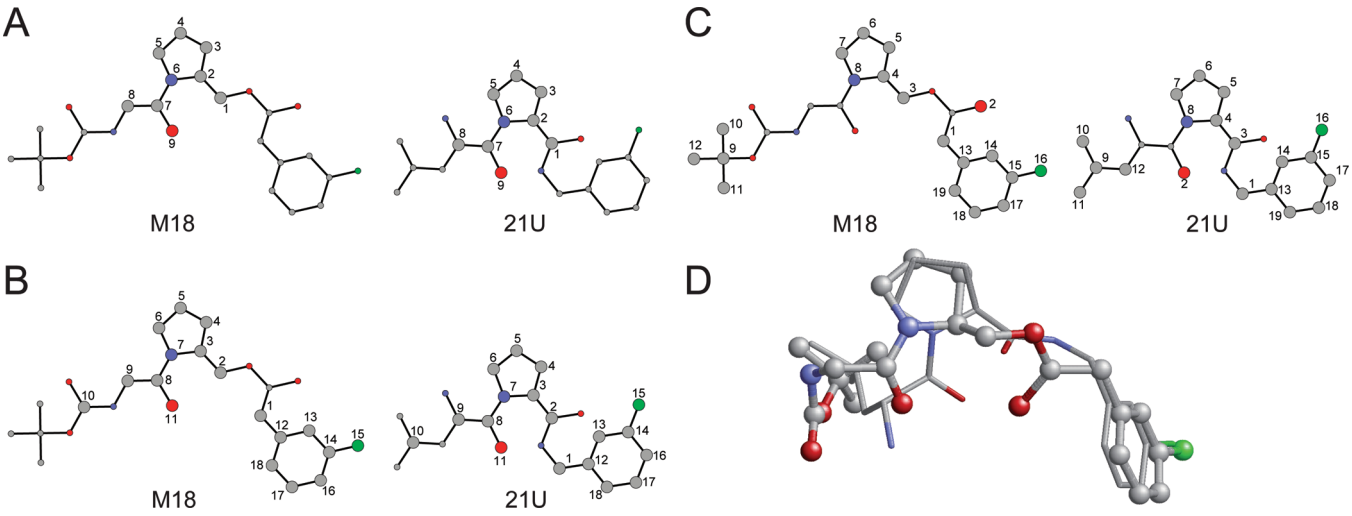


Figure 16. Atom correspondences for two molecules M18 and 21U bound to thrombin. Green circles correspond to chlorine atoms. (A) Build-up correspondence by C-MCS. Agreement with the 3D correspondence is 27.3%. (B) Build-up correspondence by TD-MCS with $\theta = 1$. Agreement with the 3D correspondence is 60.9%. (C) 3D correspondence by superimposed 3D structures of the molecules. (D) Superimposed 3D structures for M18 (PDB code: 3egk H) and 21U (PDB code: 2zgb H). The ball and stick model indicates the M18 molecule, while the wire frame model denotes the 21U molecule.

dx.doi.org/10.1021/ci2001023 |J. Chem. Inf. Model. 2011, 51, 1775–1787

larger chemical database, without drastically increasing computation time.

The topologically constrained disconnected MCS (TD-MCS) is introduced in this study, which generates superimposable disconnected common substructures. The topological constraint is defined using the topological distance, which has been often employed to reduce the search space[10] and evaluate the connection similarity between two atoms.[17] Similar concepts to the TD-MCS have been proposed so far. Takahashi et al. proposed the disconnected MCS for the edges weighted by the topological path length.[21] Rarey and Dixon introduced the topology maintaining matching in the feature tree method (FTrees).[22] Marialke et al. proposed the maximum embedded common subgraph (EMCS) or the gapped alignment, which was similar to our TD-MCS with $\theta = 5$.[14] However, they did not fairly evaluate the performances with and without the topological constraint.

The evaluation using the 3D correspondence shows that C-MCS is much better than D-MCS, and moreover, TD-MCS is significantly better than C-MCS. However, there are two limitations to TD-MCS. First, for the larger molecules, such as the HIV data set, C-MCS may provide better results than TD-MCS. It is because the search space of C-MCS is smaller than that of TD-MCS, and the searching ability of the build-up method is limited. Second, some dissimilar molecular pairs do not satisfy the condition of topological distance with small θ values. For example, the 3D correspondence in Figure 16C does not satisfy the topological condition with $\theta < 9$. Figure 18 shows histograms of the largest difference of topological distance for the 3D correspondence; 88% of similar molecular pairs satisfy the condition of TD-MCS with $\theta = 2$, but only 62% of dissimilar molecular pairs satisfy the condition. To superimpose dissimilar molecules, information regarding atom types and bond connections is insufficient. We may have to use other molecular similarities, such as 3D molecular alignments or 3D pharmacophores.

We employ 3D atom correspondence as the correct standard, assuming that biological similar activities of the molecules compared can be explained by their similar affinities to the single target protein. Therefore, this standard works mainly for virtual screening to find inhibitors for a single target protein or to induce reactant and product molecules that bind to the same enzyme. Of course, sufficient numbers of complex 3D structures are

Table 4. Average Agreement Values for Molecule Pairs with High 3D Similarities[1]

	HIV	THR	CDK2	CAH2	NEU	Total
N_{pair}[b]	1421	1331	496	1292	93	4633
C-MCS	70.9	73.9	77.5	82.9	75.9	75.9
C-MCS (exact)	71.2	72.9	76.5	80.6	74.7	75.0
TC-MCS ($\theta = 0$)	70.5	73.5	77.5	82.3	75.3	75.5
TC-MCS ($\theta = 1$)	70.8	74.0	77.5	83.1	76.0	75.9
TC-MCS ($\theta = 2$)	70.8	73.9	77.5	82.9	76.0	75.9
TD-MCS ($\theta = 0$)	72.8	76.7	81.9	84.9	86.6	78.5
TD-MCS ($\theta = 1$)	74.9	79.2	79.8	85.3	90.3	79.9
TD-MCS ($\theta = 2$)	74.9	80.0	78.5	83.8	89.3	79.5
D-MCS	47.3	58.8	55.1	64.8	76.5	56.9

[a] Similarities by the 3D atomic correspondences are restricted to more than 40%. [b] N_{pair}: number of ligand pairs.

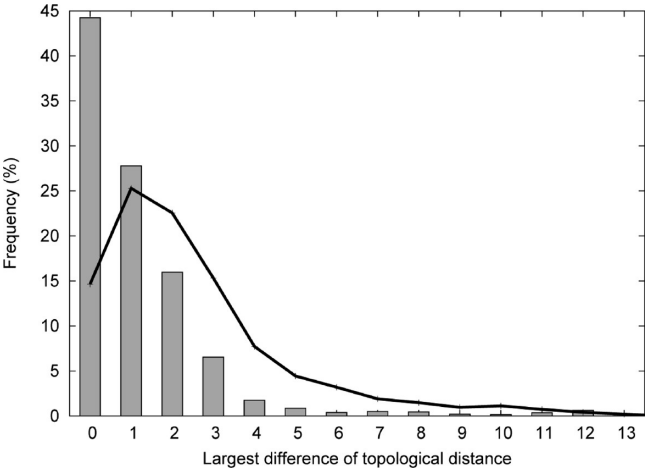


Figure 18. Histograms of the maximum difference of topological distance for the 3D correspondences. All five data sets are used in the calculation. Boxes and the black line correspond to the frequency for molecular pairs with more than 40% similarities and that with less than 40% similarities, respectively.

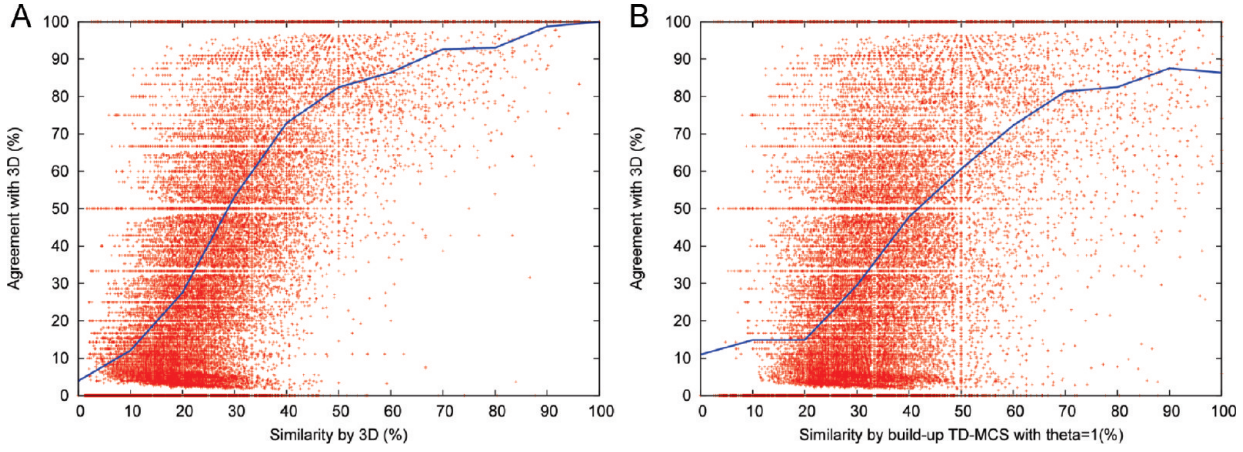


Figure 17. Plot of agreement values between the build-up and 3D correspondence for TD-MCS with $\theta = 1$ versus molecular similarities. (A) Agreement values versus similarities based on 3D correspondences. (B) Agreement values versus similarities based on the build-up TD-MCS correspondences with $\theta = 1$. Blue lines are the average agreement values.

dx.doi.org/10.1021/ci2001023 |J. Chem. Inf. Model. 2011, 51, 1775–1787

necessary for the evaluation, but the recent rapid increase in the 3D structure data in wwPDB makes this approach feasible for more target proteins. We found that a small amount of similar molecules bind in different (more than 2 Å) positions. These cases correspond to the points at the bottom right points in Figure 17B. We speculate that some molecules may bind to the target protein using multiple binding conformations or that small conformational changes in the protein affect the binding conformation of molecules. However, because the ratio of these molecular pairs is very small; we think this is not a serious problem to evaluate atom correspondences. Bostrom et al. have reported similar results.[34]

In this study, we compared our program with the program of the exact algorithm implemented by ourselves. It is because we want to compare the performances of algorithms for the same MCS problem under the same classifications of atoms and bonds. Some of the heuristic programs proposed so far are available,[13,16] however, we cannot fairly compare their performances using our agreement values, because these programs employ their own classification of atoms and bonds. Both our two correct standard correspondences are also based on our atom classification (shown in Figure 1A); atoms with different atom types are not matched in our correct correspondences. Therefore, to compare with other MCS programs, we have to employ other performance tests, such as virtual screening tests or 3D superimposition tests using root-mean-square deviation (rmsd).

The virtual screening test to extract active compounds from the database is more popular in previous studies. Instead of the screening test, we employed the atom correspondence test for the following two reasons. First, generating the correct atom correspondence is more basic requirement than extracting active molecules, and it is necessary for the molecular alignment. Second, the correct atom correspondences can be evaluated by a pair of molecules, whereas a virtual screening test requires numerous numbers of active and inactive molecules. However, this study clearly demonstrates that our build-up algorithm is fast enough to search large molecular database. We plan to evaluate our program using a virtual screening test in the near future.

Although this study employs the simple atom classification based only on heavy atoms, this classification can be further improved. We observed that atoms with different element names can be superimposed into the same position of target proteins. For example, Figure 9D shows that the oxygen ring atom in SIA is superimposed onto the carbon ring atom in G39; the oxygen atom of right side of hydroxyl group of SIA is superimposed onto the nitrogen atom of amide group. Figure 16D shows similar superimpositions of atoms with different element names. Thus, our atom classification may be too strict; the physical properties of atoms, such as hydrogen-bond donors/acceptors may be more important than the element name. We plan to improve our atom classification and to implement matching scores between atoms in the near future.

CONCLUSION

In this study, we presented the build-up algorithm that is more powerful than the simple greedy algorithm for the MCS problem and implemented it as the program kcombu. The topologically constrained disconnected MCS (TD-MCS) introduced in our study agrees well the correct standard correspondences taken from the 3D complex structures. Our program kcombu will be available through the Web server. Its fast computation makes it a useful tool for similar molecular searches against large chemical databases. It is our hope that other researchers will use our program. We also encourage them to develop better programs based on our kcombu program.

AUTHOR INFORMATION

Corresponding Author
E-mail: kawabata@protein.osaka-u.ac.jp.

ACKNOWLEDGMENT

We thank Yusuke Miyata for use of the beta version of the kcombu program and the useful comments. We also thank Takashi Kosada, Yu Takano, and Haruki Nakamura for their assistance to create our Web server in Osaka University. This work was supported by Grant-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science.

REFERENCES

(1) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods ? *J. Comput.-Aided Mol. Des.* **2002**, *7*, 903–911.

(2) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 1–30.

(3) Raymond, J. W.; Willett, P. Maximum common substructure isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521–533.

(4) Barnard, J. M. Substructure searching methods: old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.

(5) McGregor, J. J. Backtrack search algorithms and the maximal common subgraph problem. *Software Pract. Exper.* **1982**, *12*, 23–34.

(6) Bron, C.; Kerbosch, J. Algorithm 457. Finding all cliques of an undirected graph. *Commun. ACM* **1973**, *16*, 575–577.

(7) Koch, I. Enumerating all connected maximal common subgraphs in two graphs. *Theor. Comput. Sci.* **2001**, *250*, 1–30.

(8) Cazals, F.; Karande, C. An algorithm for reporting maximal *c*-cliques. *Theor. Comput. Sci.* **2005**, *349*, 484–490.

(9) Garey, M. R.; Johnson, D. S. *Computers and intractability: A guide to the theory of NP-completeness*; W.H.Freeman and Company: New York, 1979; p 202.

(10) Chen, L.; Robien, W. MCSS: A new algorithm for perception of maximal common substructures and its application to NMR spectral studies. 1. The algorithm. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 501–506.

(11) Raymond, J. W.; Gardiner, E. J.; Willett, P. RASCAL: Calculation of graph similarity using maximum common edge subgraphs. *Comput. J.* **2002**, *45*, 631–644.

(12) Raymond, J. W.; Gardiner, E. J.; Willett, P. Heuristics for similarity searching of chemical graphs using a maximum common edge subgraph algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305–316.

(13) Hattori, M.; Okuno, Y.; Goto, S.; Kanehisa, M. Development of a chemical structure comparison for integrated analysis of chemical and genetic information in the metabolic pathway. *J. Am. Chem. Soc.* **2003**, *125*, 11853–11865.

(14) Marialke, J.; Korner, R.; Tietze, S.; Apostolakis, J. Graph-based molecular alignment (GMA). *J. Chem. Inf. Model.* **2007**, *47*, 591–601.

(15) Cao, Y.; Jiang, T.; Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **2008**, *24*, i366–i374.

(16) Rahman, S. A.; Bashton, M.; Holliday, G. L.; Schrader, R.; Thornton, J. M. Small molecule subgraph detector (SMSD) toolkit. *J. Chemoinf.* **2009**, *1*, 12.

(17) Hagadone, T. R. Molecular substructure similarity searching: efficient retrieval in two-dimensional structure databases. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 515–521.

(18) Berglund, A. E.; Head, R. D. PZIM: A method for similarity searching using atom environments and 2D alignment. *J. Chem. Inf. Model.* **2010**, *50*, 1790–1795.
(19) Brown, R. D.; Jones, G.; Willett, P. Matching two-dimentional chemical graphs using genetic algorithms. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 63–70.
(20) Wang, T. Zhou. EMCSS: a new method for maximal common substructure search. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 828–834.
(21) Takahashi, Y.; Suekawa, M.; Sasaki, S. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Model.* **1992**, *32*, 639–643.
(22) Rarey, M.; Dixon, J. S. Feature tree: a new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471–490.
(23) Chen, Q.; Higgs, R. E.; Vieth, M. Geometric accuracy of three-dimensional molecular overlays. *J. Chem. Inf. Model.* **2006**, *46*, 1996–2002.
(24) Jones, G.; Gao, Y.; Sage, C. R. Elucidating molecular overlays from pairwise alignments using a genetic algorithm. *J. Chem. Inf. Model.* **2009**, *49*, 1847–1855.
(25) Gibson, S.; Scheraga, H. A. Revised algorithms for the build-up procedure for predicting protein conformation by energy minimization. *J. Comput. Chem.* **1987**, *8*, 826–827.
(26) Park, B. H.; Levitt, M. The complexity and accuracy of discrete state models of protein structure. *J. Mol. Biol.* **1995**, *249*, 493–507.
(27) Le, Q.; Pollastri, G.; Koehl, P. Structural alphabets for protein structure classification: a comparison study. *J. Mol. Biol.* **2009**, *387*, 431–450.
(28) Mizuguchi, T.; Go, N. Comparison of spatial arrangements of secondary structure elements in proteins. *Protein Eng.* **1995**, *8*, 353–362.
(29) Kawabata, T.; Nishikawa, K. Protein structure comparison using the Markov transition model of evolution. *Proteins* **2000**, *41*, 108–122.
(30) Morgan, H. L. The generation of a unique machine description for chemical structures— A technique developed at chemical abstracts service. *J. Chem. Doc.* **1965**, *5*, 107–113.
(31) Korte, B; Vygen, J. *Combinatorial Optimization: Theory and Algorithm*; Springer-Verlag: Berlin and Heidelberg, Germany, 2008; p 158.
(32) Berman, H; Henrick, K; Nakamura, H; Markley, J. L. The worldwide protein data bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res.* **2007**, *35*, D301–D303.
(33) Masinter, L. M.; Sridharan, N. S.; Carhart, R. E.; Smith, D. H. Applications of artificial intelligence for chemical inference. XIII. Labeling of objects having symmetry. *J. Am. Chem. Soc.* **1974**, *96*, 7714–7723.
(34) Brostrom, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion? *J. Med. Chem.* **2006**, *49*, 6716–6752.