JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

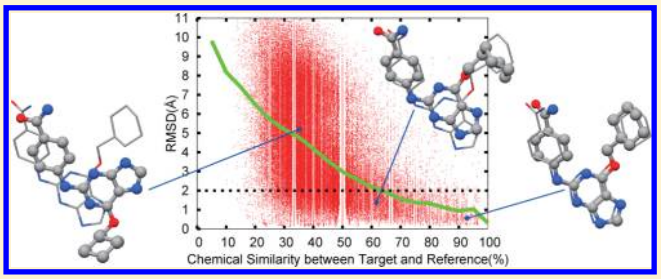
3D Flexible Alignment Using 2D Maximum Common Substructure: Dependence of Prediction Accuracy on Target-Reference Chemical Similarity

Takeshi Kawabata* and Haruki Nakamura

Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

Supporting Information

ABSTRACT: A protein-bound conformation of a target molecule can be predicted by aligning the target molecule on the reference molecule obtained from the 3D structure of the compound−protein complex. This strategy is called "similarity-based docking". For this purpose, we develop the flexible alignment program *fkcombu*, which aligns the target molecule based on atomic correspondences with the reference molecule. The correspondences are obtained by the maximum common substructure (MCS) of 2D chemical structures, using our program *kcombu*. The prediction performance was evaluated using many target-reference pairs of superimposed ligand 3D structures on the same protein in the PDB, with different ranges of chemical similarity. The details of atomic correspondence largely affected the prediction success. We found that topologically constrained disconnected MCS (TD-MCS) with the simple element-based atomic classification provides the best prediction. The crashing potential energy with the receptor protein improved the performance. We also found that the RMSD between the predicted and correct target conformations significantly correlates with the chemical similarities between target-reference molecules. Generally speaking, if the reference and target compounds have more than 70% chemical similarity, then the average RMSD of 3D conformations is <2.0 Å. We compared the performance with a rigid-body molecular alignment program based on volume-overlap scores (*ShaEP*). Our MCS-based flexible alignment program performed better than the rigid-body alignment program, especially when the target and reference molecules were sufficiently similar.



INTRODUCTION

Three-dimensional (3D) structural alignment (superposition, overlay) of chemical compounds can elucidate their corresponding atoms of active conformations in 3D space.[1] If multiple compounds with the same biological activities are known, then the superimposition of their 3D structures will reveal common chemical features (pharmacophores) that interact with the target proteins.[2] This is the basis for 3D-QSAR[3] and CoMFA[4] studies. In principle, these studies can be applied to 3D structures predicted by computer programs. Moreover, the 3D structure of the compound on the receptor protein displays the interacting atom pairs between the compound and the receptor, and thus it should provide information for compound modifications to obtain higher activity. If the experimental 3D structure data of a compound (called a "reference" or "template" molecule) with its receptor protein are available, then the protein-bound structure of another compound (called a "target" molecule) can be predicted by aligning the conformation of the target compound on the known 3D structure of the reference. Therefore, 3D alignments with the known protein-bound 3D structure are often called "similarity-based docking" ("guided docking", "template-based docking").[5] The only difference between the standard 3D structural alignment and the similarity-based docking is that the latter can use the 3D structures of receptor proteins. Applications of similarity-based docking are expanding, due to the increasing amounts of 3D structure data for compound−protein complexes. Recently, similarity-based methods for modeling both ligand and receptor proteins were proposed.[6,7]

Many approaches for 3D structural alignment have been reported, summarized by Lemmen and Langauer.[1] These approaches can be classified in terms of the freedom of optimization and objective functions. For the freedom of optimization, the approaches can be roughly classified into two classes: rigid-body alignment and flexible alignment. The rigid-body alignment methods align molecules simply by translation and rotation, without any conformational changes.[8−16] Since the compounds that bind to receptor proteins are usually quite flexible, the rigid-body methods often use multiple conformations for one compound, otherwise their prediction accuracies would be limited. Therefore, good conformer-generating programs such as *OMEGA*[17] and *balloon*[18] are required for rigid-body superposition. In contrast, the flexible superposition methods change the molecular conformations to fit the reference, and so they basically need only a single conformation for a compound.[19−29] The flexible alignment algorithm is more complicated than the rigid-body algorithm and is similar to that of docking programs using molecular mechanical potential

Received: January 5, 2014
Published: June 4, 2014

dx.doi.org/10.1021/ci500006d | *J. Chem. Inf. Model.* 2014, 54, 1850−1863

energies. This is the reason why some flexible superposition programs were developed by modifying existing flexible docking programs.[19−22]

The objective functions used in the alignment methods can be roughly classified into volume overlapping and point-matching function. The volume overlapping functions measure the 3D shape similarity by the volume overlap of two atomic shapes.[8−12,19−22,26,27] Most of them employ Gaussian distribution functions to represent the atomic shapes, because the overlap integral of Gaussian distribution functions can be easily obtained in an analytical form.[9−12] Grant and Pickup rigorously formalized the volume of Gaussian distribution functions,[10,30] and their program *ROCS* seems to be the most popular rigid-body superposition program using the volume overlapping function.[10] Electrostatic potentials are also included in the form of the Gaussian distribution function in many studies.[8,9,11,12] Most of the methods using the Gaussian distribution function employ gradient-based optimizations to search for the optimal pose and conformer.[10,11]

On the other hand, the point-matching function measures the number of matches of 3D points. The distances between the matched points are expected to be compatible.[13−16,23−26,28,29] For the representative 3D points, the centers of atoms are most frequently used. The center of a chemical group or the hypothetical positions of receptor atoms is also employed. Matching the pharmacophore patterns can be regarded as optimizing the point-matching function, because these patterns are usually described by the geometry of several 3D points. Point matches can be determined by various methods such as the clique detection algorithm of the distance compatibility graph,[14,15,26] the geometric hashing algorithm,[13] and the 2D maximum common substructure.[6,16,28,29] Most of these point-matching algorithms consider the points as a rigid-body; however, some of the methods are designed to tolerate differences in the point−point distances.[14,15] Raymond and Willett proposed a 3D-maximum common substructure (3D-MCS) using a smoothed bounded distance matrix and determined the upper-bound distance by the shortest bond length path (topological distance).[15] This restraint is quite similar to the topological distance constraint of TD-MCS, as discussed below. One advantage of the point-matching functions is their simple rigid superimposition. Namely, if matches with at least three point pairs are available, then the least-squares fitting algorithm quickly generates the optimal rigid superposition of the target and reference molecules.[31] Several flexible transformation algorithms have been developed to superimpose the matched point pairs. Most of them employed gradient-based optimization to minimize the sum of the squared distances of the matched points[23,24,28,29] The *FlexS* program performs flexible superposition by building fragments of matched points in a step-by-step fashion.[26] Some methods use both volume overlapping and point matching. For example, the rigid-body superposition program *ShaEP* generates many poses by matching "field-graph" vertices and optimizes these poses by Newton methods using volume-overlapping scores.[12]

In this study, we propose a new flexible superposition program with a point-matching function. Our approach can be regarded as an extension of the method proposed by Marialke et al.[28,29] Before 3D superimposition, the atom matches are determined by the 2D-MCS. Next, the pose and the conformation of the target molecule are transformed to fit the matched atoms in the reference molecule. This approach has several advantages. First, the calculation of atom matches by 2D-MCS has a smaller computational cost than that with a 3D structure. Second, 2D-MCS is free from the variability of input 3D conformations predicted by conformation generating programs. Third, flexible transformation using atomic correspondences can be performed effectively, because a torsion-angle stamping procedure can be applied.[28,29] Although our approach is similar to the method proposed by Marialke et al., the details are more refined. We employed our program, *kcombu* to calculate the corresponding atom pairs. It can calculate many types of matching, such as connected, disconnected, and topologically constrained disconnected MCSs.[32] Our previous study showed that topologically constrained MCS (TD-MCS) agreed well with the 3D correspondence obtained by the complex 3D structures, suggesting that TD-MCS could lead to well-superimposed 3D conformations. We also implemented a modification process for correct chiralities and a ring-stamping process to enhance the performance. We evaluated the performances of our program using hundreds of 3D compound structures with several different types of MCS and atomic classifications.

Another novel point of this study is that we studied the dependence of the prediction accuracy on the target-template molecular similarity using a large amount of 3D structural data. As with the homology modeling of protein structure,[33,34] the prediction accuracy of similarity-based docking might be affected by the similarities between the target and reference molecules. However, the relationship between the accuracies and the ligand similarities has not been studied comprehensively. Therefore, we prepared about 130,000 pairs of 3D complexes with identical receptor proteins and different ligands and performed the similarity-based dockings for all of the compound pairs of the data set. We then checked the prediction accuracies and the similarities between the target and reference molecules.

MATERIALS AND METHODS

Outline of the 3D Transformation. Our program, *fkcombu* (flexible superposition program based on *kcombu*), flexibly transforms the target molecule onto the 3D structure of the reference molecule. We assume that the 3D conformations of both the reference and the target molecules are available as inputs. The conformation of the reference should be experimentally determined, whereas that of the target molecule is provided by a conformer generating program. The use of the 3D structure of a receptor protein of the reference complex is optional. It only affects the crashing energy E_{prot}, explained below. The prediction procedure consists of the following four steps. First, one-to-one atom correspondence is obtained by calculating the MCS, which is performed using the *kcombu* program based on the build-up algorithm.[32] Second, the stamping operation is applied to the target molecule, to paste the torsion angles of the reference molecule onto the corresponding bonds in the target molecule. Third, rigid-body superimposition is performed to superimpose the corresponding atom pairs, using the standard RMSD optimization algorithm.[31] Finally, a steepest-descent minimization is performed for the initial random conformations generated by randomly changing torsion angles. Number of the initial random conformations was set to 10. The random changing of torsion angle is not applied to the first initial conformation. Among the 10 minimized conformations, the conformation with the lowest total potential energy E (defined in eq 17) is chosen as the predicted conformation. The final minimization step is the most time-consuming. The details of each step are described in the following subsections.

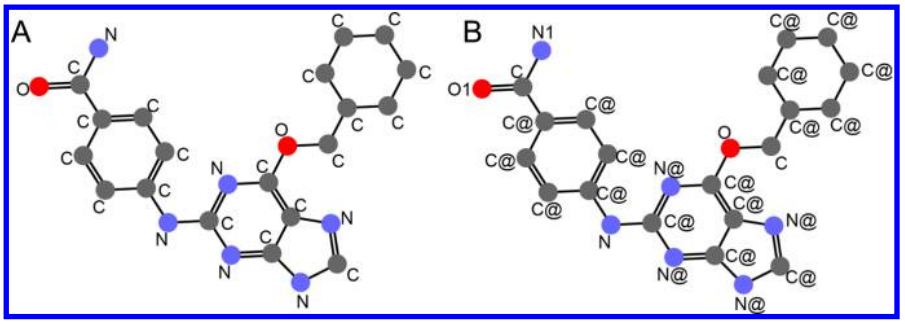


Figure 1. Atom types of the CDK2 ligand N41. (A) Element-based classification. (B) Kcombu-default classification.

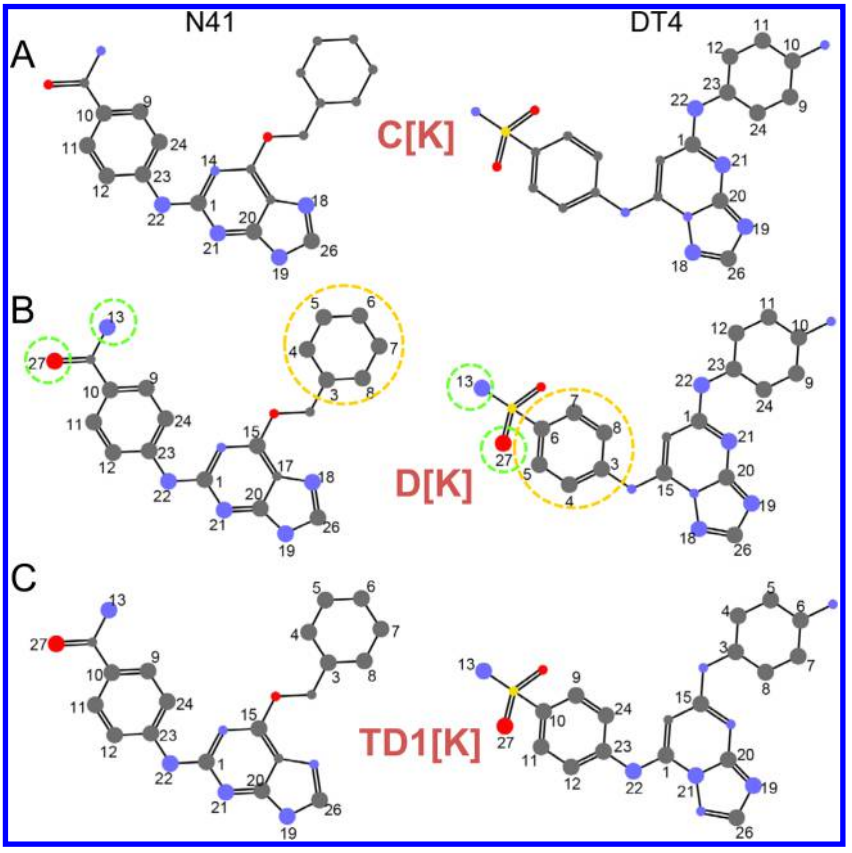


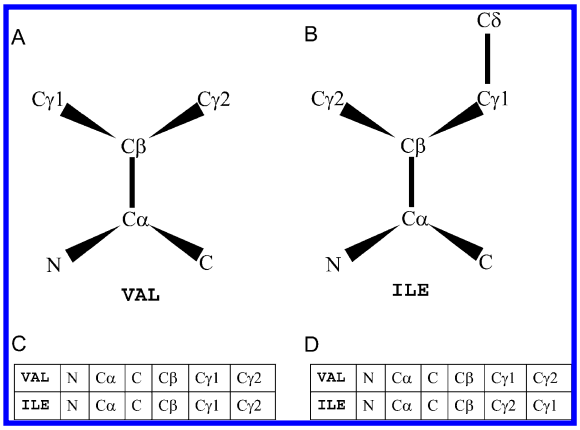
Figure 2. Maximum common substructures of the CDK2 ligands N41 and DT4, using the kcombu-default classification (denoted as [*K*]). (A) C-MCS; The number of corresponding atoms is 13, and the Tanimoto coefficient is 31.0%. (B) D-MCS; The number of corresponding atoms is 22, and the Tanimoto coefficient is 66.7%. (C) TD-MCS with $\theta = 1$; The number of corresponding atoms is 21, and the Tanimoto coefficient is 61.8%. The 3D structures predicted using these MCSs are shown in Figure 12.

One-to-One Atomic Correspondence Using Maximum Common 2D Substructure. We employed one-to-one atomic correspondences obtained by the MCS of 2D chemical structures calculated by the *kcombu* program, based on the build-up heuristic algorithm.[32] The build-up algorithm is a greedy algorithm taking N_{keep} correspondences for each building step. The parameter N_{keep} was denoted as K in our previous report. We employed N_{keep} = 100 in this study. The MCS is defined as a maximum substructure present in two molecules with the same atom types and bond connections. Thus, it does not depend on the 3D coordinates of the molecules. The simplest classification of atom types is the element-based one, as shown in Figure 1A. Our previous study employed the slightly more detailed atomic classification to reduce the computational costs without losing sensitivity. A core of the atom type is also the element name, but the character '@' is added for ring structures, and the character '1' is added for oxygen and nitrogen atoms bonded to one heavy atom.[32] This atom classification is called "kcombu-default" in this study (Figure 1B). Both classifications ignore hydrogen atoms and do not distinguish the bond orders (single, double or triple). It is because the bond orders cannot be distinguished correctly without knowing positions of hydrogen atoms, which often lack in X-ray crystal structures.

Various types of MCS can be calculated by the *kcombu* program, including connected MCS (C-MCS), disconnected MCS (D-MCS), and topologically constrained disconnected MCS (TD-MCS). The C-MCS should be a connected graph, or the atoms of the C-MCS must belong to one connected component. Figure 2A is an example of C-MCS. The D-MCS is not imposed on this connectedness constraint. It allows atom pairs with several connected components and thus can represent a wider variety of correspondences. However, D-MCS some-

times generates a correspondence that cannot be realized by superimposition in 3D space, as shown in Figure 2B. Therefore, TD-MCS was introduced to avoid these unrealistic correspondences. TD-MCS is a disconnected MCS allowing only the θ difference in the topological distance of the corresponding atomic pairs. Figure 2C is an example of TD-MCS with $\theta = 1$. We previously showed that TD-MCS had the best agreement with the 3D correspondence among the various MCS types.[32]

2D-MCS generally provides a good 3D superposition; however, it sometimes fails to detect the correct correspondence of chiral molecules. If two of the same groups are attached to one tetra-coordinated carbon, then these two groups cannot be distinguished in 2D structure. However, if the molecule is chiral, then they can be distinguished in 3D space. Examples of valine and isoleucine are shown in Figure 3A,B. The two carbon atoms



VAL	N	Cα	C	Cβ	Cγ1	Cγ2
ILE	N	Cα	C	Cβ	Cγ1	Cγ2

VAL	N	Cα	C	Cβ	Cγ1	Cγ2
ILE	N	Cα	C	Cβ	Cγ2	Cγ1

Figure 3. Atomic correspondences with chiral consistencies for valine and isoleucine molecules. Atom names were obtained from the PDB database notations. (A) Valine. (B) Isoleucine. (C) Atomic correspondence with inconsistent chiralities. The chirarities around the Cβ atoms are different. (D) Atomic correspondence with consistent chiralities.

Cγ1 and Cγ2 in valine are both equivalent in the 2D structure. This means that both atomic correspondences, (Cγ1-Cγ1, Cγ2-Cγ2) and (Cγ1-Cγ2, Cγ2-Cγ1), satisfy the conditions for 2D-MCS. However, in 3D space, (Cγ1-Cγ2, Cγ2- Cγ1) can generate perfect 3D superposition, whereas (Cγ1-Cγ1, Cγ2-Cγ2) cannot. To solve this problem, we introduced the following procedure, after calculating the MCS in the 2D structure by the *kcombu* program. At first, the chiral consistencies for all pairs of tetrahedral coordinates were checked. If chiral inconsistencies were found, then equivalent atomic correspondences were generated by all permutation of the atoms with the same connection relationship. The chiral consistencies for each generated equivalent correspondence were examined. If a new correspondence had consistent chiralities,then the original one was replaced by this new correspondence.

System To Be Optimized. Let us define the variables for describing a system to be optimized, as shown in Figure 4. The arrays of 3D vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, $\{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_M\}$, and $\{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_L\}$ correspond to the positions of the heavy atoms in the target molecule, the reference molecule, and the receptor protein molecule of the reference complex, respectively. The numbers N, M, and L are the numbers of atoms in these molecules, respectively. The maximum common substructure search provided the corresponding atom pairs of the i-th target atom and the $r(i)$-th reference atom. The optimization was performed

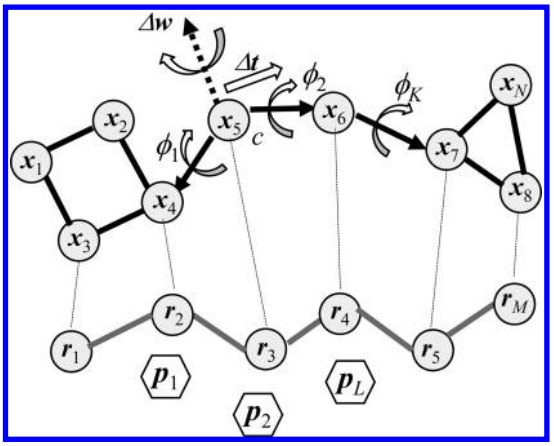


Figure 4. A schematic view of the system to be optimized. The arrays of 3D vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, $\{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_M\}$, and $\{\boldsymbol{p}_1, \boldsymbol{p}_2, ..., \boldsymbol{p}_L\}$ correspond to the positions of the heavy atoms in the target molecule, the reference molecule, and the receptor protein molecule, respectively. The conformation was changed using three variables: the translation vector $\Delta\boldsymbol{t}$, the rotational vector $\Delta\boldsymbol{w}$, and the K torsion angles $\{\Delta\phi_1, \Delta\phi_2, ..., \Delta\phi_K\}$ for rotational bonds.

only for the atom positions of the target molecule $\{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\}$, and those of the reference compound and the receptor protein molecules remained fixed.

For the rotational and torsion angle transformations of the target molecule, the center atom c of the target molecule must be chosen. The center atom is determined by using the sum of topological distances. The topological distance $T(i,j)$ is the number of bonds on the shortest path between the i-th and j-th atoms. The sum $S(i)$ of the topological distance from the i-th atom to any other atoms was calculated as follows:

$$S(i) = \sum_{j=1}^{N} T(i, j) \tag{1}$$

The atom with the lowest sum $S(i)$ is assigned as the center atom c, as shown in Figure 4.

Next, we introduced the rotational bonds of the target molecule between all heavy atoms, that are not included in any ring structures. The number of rotational bonds is K. Figure 5 shows several variables of k-th bond: $z(k)$, $o(k)$, $\mathbf{v}_k$, and M_k. They are used for the stamping and the steepest-descent minimization described in the following subsections. The two terminal atoms of the k-th rotational bond are called $z(k)$ and $o(k)$. As shown in

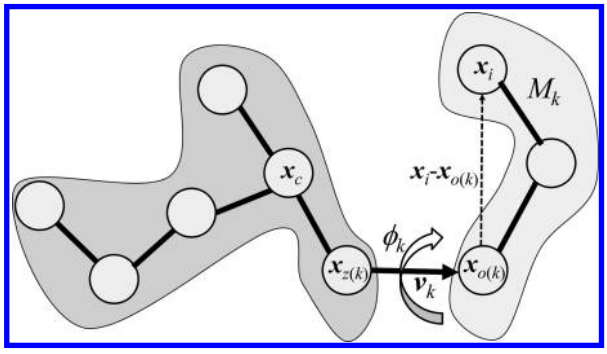


Figure 5. Definition of the atoms around the k-th rotational bond. The atom x_c is the center atom. The two terminal atoms of the k-th rotational bond are called $z(k)$ and $o(k)$. The atomic group M_k is defined as the atoms affected by the dihedral angle on bond k.

dx.doi.org/10.1021/ci500006d | J. Chem. Inf. Model. 2014, 54, 1850−1863

Figure 5, the discrimination of the atoms $z(k)$ and $o(k)$ is determined by their topological distance from the center atom c. The $z(k)$-th atom is closer to the center atom than $o(k)$-th atom: $T(c,z(k)) < T(c,o(k))$. We define the directional vector $\boldsymbol{v}_k$ for the k-th rotational bonds as follows:

$$\boldsymbol{v}_k = \frac{\boldsymbol{x}_{o(k)} - \boldsymbol{x}_{z(k)}}{\| \boldsymbol{x}_{o(k)} - \boldsymbol{x}_{z(k)} \|} \tag{2}$$

The vector $\boldsymbol{v}_k$ is used for the rotation of atoms around the k-th rotational bond, explained below. Next, we define the atom group M_k as the atoms with positions that are affected by the change of dihedral angle of the bond k, as shown in Figure 5. The M_k group is determined as a set of atoms i with a topological distance from the $o(k)$-th atoms is shorter than that from the $z(k)$-th atom: $T(i,o(k)) < T(i,z(k))$.

For the optimization described in the next section, the rotational dependencies of torsion angles must be summarized by the topological distances. The indexes of rotation bonds $\{1, 2, ..., K\}$ were sorted by the increasing order of the topological distance $T(c,o(k))$ for the optimization described in the next section. All of the rotational torsion angles are described by the vector $\boldsymbol{\Phi} = (\phi_1, \phi_2, ..., \phi_K)^T$.

Stamping. Stamping is a procedure proposed by Marialke et al. to transform the conformation of the target molecule, by copying the local conformation of the template molecule.[28,29] Although they only employed torsion angle stamping, we additionally employed ring structure stamping.

Torsion angle stamping is quite simple. The rotational bond k is chosen and checked to verify whether it satisfies the following three conditions (Figure 6). (1) The target atoms, $y(k)$ and $z(k)$,

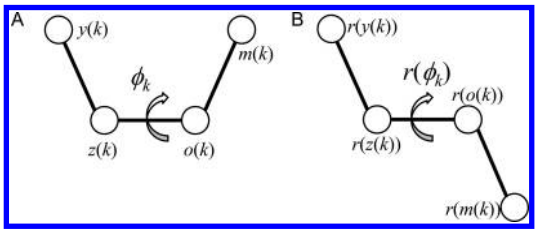


Figure 6. Stamping procedure of the torsion angle. (A) The four atoms to be stamped of the target molecule. (B) The corresponding four atoms of the reference molecule.

are covalently connected, and atoms, $o(k)$ and $m(k)$, are also connected. (2) The four target atoms, $y(k)$, $z(k)$, $o(k)$, and $m(k)$, have corresponding atoms in the reference molecules, $r(y(k))$, $r(z(k))$, $r(o(k))$, and $r(m(k))$, respectively. (3) The two target atoms, $z(k)$ and $o(k)$, and their corresponding reference atoms, $r(z(k))$ and $r(o(k))$, are not included in any ring structures. If these three conditions are satisfied, then the following transforming procedure is applied:

$$\boldsymbol{x}_i := R[(r(\phi_k) - \phi_k)\boldsymbol{v}_k](\boldsymbol{x}_i - \boldsymbol{x}_{o(k)}) + \boldsymbol{x}_{o(k)} \quad \text{for } \boldsymbol{x}_i \in M_k \tag{3}$$

where the angle ϕ_k is the torsion angle for the target atoms $y(k)$, $z(k)$, $o(k)$, and $m(k)$, and the angle $r(\phi_k)$ is the corresponding torsion angle for the four reference atoms, $r(y(k))$, $r(z(k))$, $r(o(k))$, $r(m(k))$. The matrix $R[(r(\phi_k) - \phi_k)\boldsymbol{v}_k]$ is a rotation matrix for the rotation axis $\boldsymbol{v}_k$ and the angle $r(\phi_k) - \phi_k$.

We also implemented a stamping process for nonplanar ring structures, such as cyclohexane or sugars. Figure 7 shows an example of cyclohexane for changing a target "boat" conformation into a reference "chair" conformation. Detailed procedures for the nonplanar ring stamping are described in the Supporting Information.

Steepest-Descent Minimization. The optimization was performed using the steepest-descent method, in which the conformation of the molecule was gradually transformed using the negative direction of the first derivative of the total energy. The conformation was adapted using three variables: the translation vector $\Delta\boldsymbol{t}$, the rotational vector $\Delta\boldsymbol{w}$, and the K torsion angles $\Delta\boldsymbol{\Phi} = (\Delta\phi_1, \Delta\phi_2, ..., \Delta\phi_K)$ for rotational bonds. These variables were iteratively changed using the force $\boldsymbol{F}$, the torque $\boldsymbol{T}$, and the gradient by the torsion angles $\boldsymbol{U}$ as follows:

$$\Delta\boldsymbol{t} = \lambda\alpha\frac{\boldsymbol{F}}{\| \boldsymbol{F} \|} \tag{4}$$

$$\Delta\boldsymbol{w} = \lambda\beta\frac{\boldsymbol{T}}{\| \boldsymbol{T} \|} \tag{5}$$

$$\Delta\boldsymbol{\Phi} = \lambda\gamma K\frac{\boldsymbol{U}}{\| \boldsymbol{U} \|} \tag{6}$$

where parameters α, β, and γ are magnitude parameters. In this study, we employed $\alpha = 0.1$ Å, $\beta = 10°$, and $\gamma = 10°$. The step size λ ranged from 0 to 1, determined using the golden linear search in each step. The iteration described in eqs 4−6 did not depend on any atomic masses. In this sense, this is rather similar to Brownian dynamics, although it did not use random force and torque.[35] The force $\boldsymbol{F}$, the torque $\boldsymbol{T}$, and the gradient of the torsion angles $\boldsymbol{U}$ are defined as follows:

$$\boldsymbol{F} = -\sum_{i=1}^{N}\frac{\partial E}{\partial \boldsymbol{x}_i} \tag{7}$$

$$\boldsymbol{T} = -\sum_{i=1}^{N}(\boldsymbol{x}_i - \boldsymbol{x}_c) \times \frac{\partial E}{\partial \boldsymbol{x}_i} \tag{8}$$

$$\boldsymbol{U} = -\frac{\partial E}{\partial \boldsymbol{\Phi}} = \left(-\frac{\partial E}{\partial \phi_1}, -\frac{\partial E}{\partial \phi_2}, ..., -\frac{\partial E}{\partial \phi_K}\right)^T \tag{9}$$

where E is the potential energy, described in the next subsection. The detailed descriptions of the first derivative of the energy, by the position of the target atom $\boldsymbol{x}_i$ and the torsion angle ϕ_k, are summarized in the Appendix section. The three transformations defined by eqs 4−6 are not independent. In other words, if the order of the three transformations is changed, then different conformations are generated even if the three deviations $\Delta\boldsymbol{t}$, $\Delta\boldsymbol{w}$, and $\Delta\boldsymbol{\Phi}$ are identical. Therefore, we arbitrarily decided to use the following order of transformations. First, the transformation using the translation vector $\Delta\mathbf{t}$ and the rotational vector $\Delta\boldsymbol{w}$ was performed using the following equation:

$$\boldsymbol{x}_i := R[\Delta\boldsymbol{w}](\boldsymbol{x}_i - \boldsymbol{x}_c) + \boldsymbol{x}_c + \Delta\boldsymbol{t} \tag{10}$$

where $R[\Delta\boldsymbol{w}]$ is the rotation matrix derived from the rotational vector $\Delta\boldsymbol{w}$. Second, the transformation by the k-th rotational torsion angle was performed in according to the numeric order of the angle (from $k = 1$ to $k = K$):

$$\boldsymbol{x}_i := R[\Delta\phi_k\boldsymbol{v}_k](\boldsymbol{x}_i - \boldsymbol{x}_{o(k)}) + \boldsymbol{x}_{o(k)} \quad \text{for } \boldsymbol{x}_i \in M_k \tag{11}$$

where $\boldsymbol{v}_k$ is the unit bond directional vector defined in eq 2, and $o(k)$ is the atom number of the bond-end atom for the torsion angle ϕ_k. Note that the indexes of the rotation bonds $\{1,2, ..., K\}$ were sorted by the increasing order of the topological distance from the center atom c.

dx.doi.org/10.1021/ci500006d | J. Chem. Inf. Model. 2014, 54, 1850−1863

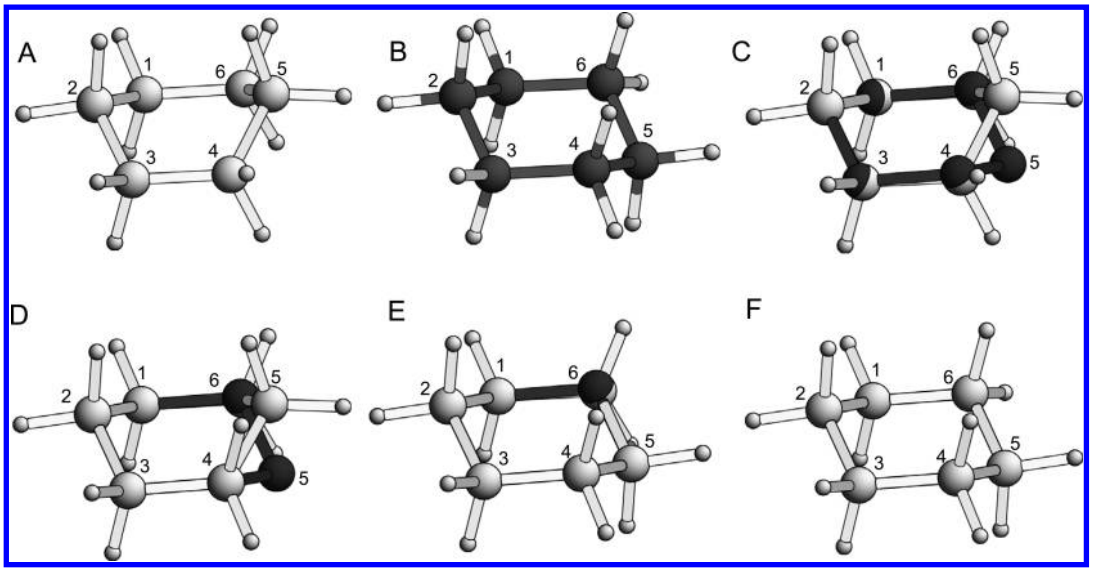


Figure 7. An example of the ring stamping process for cyclohexane conformations. (A) Target "boat" conformation. (B) Reference "chair" conformation. (C) A reference conformation is superimposed on the target, using the first three atoms (1, 2, 3). (D) The atoms connected to the fourth atom are transformed to superimpose three atoms (3, 4, 5). (E) The fifth atoms and the atoms connected to the fifth atom are transformed to superimpose three atoms (4, 5, 6). (F) The sixth atom and the atoms connected to the sixth atom are transformed to superimpose three atoms (5, 6, 1).

Table 1. Four Data Sets of Superimposed 3D Structures of Compounds

				no. of heavy atoms			no. of rings (SSSR)			average similarity(%)	
	receptor protein	reference PDB	Nligand	avg	min	max	avg	min	max	TD1[E]	fingerprint
NEU	neuraminidase	2qwf (G20)	17	22.2	20	28	1.1	1	2	60.6	45.3
THR	Thrombin	1ad8 (MDL)	167	31.1	9	55	3.2	0	6	43.1	30.5
CAH2	carbonic anhydrase 2	3hkn (MFS)	193	20.7	8	56	2.0	0	6	41.5	22.2
CDK2	cyclin-dependent protein kinase 2	2w05 (FRT)	259	24.1	9	36	3.2	2	8	39.4	23.4

Potential Energy To Be Minimized. The potential energy is composed of three energies: the atom matching energy E_{match}, the self-crashing energy E_{self}, and the protein-crashing energy E_{prot}. The energy for matching corresponding atom pairs, E_{match}, is defined as follows:

$$E_{\text{match}}(\{\boldsymbol{x}_i\}, \{\boldsymbol{r}_{r(i)}\}) = \sum_{i=1, r(i)\neq\phi}^{N} h(\boldsymbol{x}_i, \boldsymbol{r}_{r(i)}) \tag{12}$$

where $r(i)$ is the index of the corresponding atom of the reference molecule for the i-th atom of the target molecule. The harmonic matching energy $h(\boldsymbol{x}, \boldsymbol{r})$, between two atom positions $\boldsymbol{x}$ and $\boldsymbol{r}$, is defined as follows:

$$h(\boldsymbol{x}, \boldsymbol{r}) = \frac{1}{2}(\boldsymbol{x} - \boldsymbol{r})^2 \tag{13}$$

The energy for crashing within target molecules, E_{self}, is defined as follows:

$$E_{\text{self}}(\{\boldsymbol{x}_i\}) = \sum_{i=1}^{N}\sum_{j>i}^{N} s(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{14}$$

The sigmoid crashing energy $s(\boldsymbol{x}, \boldsymbol{p})$, between two atom positions $\boldsymbol{x}$ and $\boldsymbol{p}$, is defined as follows:

$$s(\boldsymbol{x}, \boldsymbol{p}) = \frac{\exp[-\alpha_{\text{sig}}(\|\boldsymbol{x} - \boldsymbol{p}\| - R(\boldsymbol{x}) - R(\boldsymbol{p}) + \tau)]}{1 + \exp[-\alpha_{\text{sig}}(\|\boldsymbol{x} - \boldsymbol{p}\| - R(\boldsymbol{x}) - R(\boldsymbol{p}) + \tau)]} \tag{15}$$

where $R(\boldsymbol{x})$ is the van der Waals radius of atom $\boldsymbol{x}$, α_{sig} is a steepness parameter, and τ is a parameter for tolerance. In this study, we employed α_{sig} =10 and τ = 1.0 Å. Similarly, the energy for crashing between the target molecule and the receptor protein molecule of the reference complex E_{prot} is defined as follows:

$$E_{\text{prot}}(\{\boldsymbol{x}_i\}, \{\boldsymbol{p}_m\}) = \sum_{i=1}^{N}\sum_{m=1}^{L} s(\boldsymbol{x}_i, \boldsymbol{p}_m) \tag{16}$$

In our *fkcombu* program, the receptor protein of the reference complex only affected this crashing energy, E_{prot}. In the following Results section, we discuss the prediction performance with and without the energy E_{prot}. The total energy E is defined as the sum of these three potential energies:

$$\begin{aligned} E(\{\boldsymbol{x}_i\}, \{\boldsymbol{r}_{r(i)}\}, \{\boldsymbol{p}_m\}) &= w_{\text{m}}\cdot E_{\text{match}}(\{\boldsymbol{x}_i\}, \{\boldsymbol{r}_{r(i)}\}) + w_{\text{s}}\cdot E_{\text{self}}(\{\boldsymbol{x}_i\}) \\ &\quad + w_{\text{p}}\cdot E_{\text{prot}}(\{\boldsymbol{x}_i\}, \{\boldsymbol{p}_m\}) \end{aligned} \tag{17}$$

where the parameters w_{m}, w_{s}, and w_{p} are the weights for the energies E_{match}, E_{self}, and E_{prot}, respectively. In this study, we employed $w_{\text{m}} = w_{\text{s}} = w_{\text{p}} = 1$.

Data Set of Superimposed Compound−Protein Structure. We prepared four data sets of superimposed conformations of compounds: neuraminidase (NEU), thrombin (THR), carbonic anhydrase 2 (CAH2), and cyclic dependent protein kinases 2 (CDK2). The 3D structural data were downloaded from the PDB on July 3, 2013. All the protein sequences in the

dx.doi.org/10.1021/ci500006d | J. Chem. Inf. Model. 2014, 54, 1850−1863

Table 2. Success Rates (%) for All Target-Reference Pairs

abbreviation	method	ac[a]	E_{prot}	NEU	THR	CAH2	CDK2	all
Npair[b]				272	27722	37056	66822	131872
C[K]	C-MCS	K		70.6	10.1	15.7	8.8	11.1
C[E]	C-MCS	E		64.7	14.8	16.0	9.6	12.6
Cp[K]	C-MCS	K	E_{prot}	69.9	11.6	15.8	11.7	13.0
Cp[E]	C-MCS	E	E_{prot}	64.7	15.7	16.3	11.8	14.0
TD 1[K]	TD-MCS θ = 1	K		79.0	19.5	25.6	14.5	18.8
TD 1[E]	TD-MCS θ = 1	E		69.9	20.2	27.9	15.8	20.3
TD 1p[K]	TD-MCS θ = 1	K	E_{prot}	83.8	20.3	26.3	16.4	20.1
TD 1p[E]	TD-MCS θ = 1	E	E_{prot}	73.5	24.2	28.9	17.3	22.1
TD 2[K]	TD-MCS θ = 2	K		81.6	21.0	25.4	14.2	18.9
TD 2[E]	TD-MCS θ = 2	E		68.8	20.2	26.3	15.4	19.6
TD 2p[K]	TD-MCS θ = 2	K	E_{prot}	86.0	21.8	25.9	15.6	20.0
TD 2p[E]	TD-MCS θ = 2	E	E_{prot}	71.3	25.1	28.7	16.6	21.9
D [K]	D-MCS	K		75.7	17.2	22.0	10.9	15.5
D [E]	D-MCS	E		66.9	19.8	24.2	12.0	17.2
D p[K]	D-MCS	K	E_{prot}	79.0	17.7	22.4	11.8	16.2
D p[E]	D-MCS	E	E_{prot}	68.0	20.4	24.6	12.8	17.8
ShaEP_20	c = 20	–		80.9	16.7	24.2	11.7	16.4
ShaEP_50	c = 50	–		80.9	21.7	24.5	13.7	18.5
ShaEP_100	c = 100	–		81.2	22.6	24.7	14.0	18.9
DOCK				47.8	14.4	13.8	10.6	12.4

[a]Atomic classification. K: kcombu-based, E: element-based. [b]Number of target-reference pairs.

PDB were clustered by the single linkage clustering algorithm with 95% sequence identities. Within each protein cluster, only complexes with unique compounds were chosen. Compound structures with metal ions or with more than 100 heavy atoms were removed from the data set. We also compared each PDB structure with its ideal coordinates in the SDF file, downloaded from the LigandExpo server (http://ligand-expo.rcsb.org). Compound PDB structures that lacked heavy atoms or had inconsistent chiral atoms were removed from the data set. The proteins and their bound compounds were superimposed by the protein structure comparison program MATRAS[36] against the reference protein structure. The data set details are described in Table 1. The PDB_IDs of the four data sets are given in the Supporting Information. For the initial configuration of the target molecule, the ideal coordinates, in the SDF format downloaded from the LigandExpo server, were employed.

Calculation Procedures for ShaEP and DOCK. To compare our flexible point-matching program, we also performed the prediction using a rigid-body volume-overlapping program, *ShaEP*, which was developed by Vainio et al.[12] It is a rigid-body superposition program using the volume-overlapping score and the electrostatic potential field score. For the *ShaEP* program, we prepared three conformation sets of target molecules generated by the *balloon* program,[18] which was also developed by Vanio and Johnson. Conformers were generated with different options for the number of conformers (-c 20, -c 50, and -c 100). The number of generated conformers for each molecule was not the same as the number given by the option -c. The average numbers of generated conformations for one molecule are 19.7, 50.7 and 90.8, for the options -c 20, -c 50, and -c 100, respectively. The average computing times for generating conformations for one molecule were 10.11, 33.84, and 75.78 s, for the options -c 20, -c 50, and -c 100, respectively. Hydrogen atoms and partial charges for the target and reference molecules were also generated using the *balloon* program.

To compare our program with one of the standard molecular docking program, we chose the *DOCK* 6.6, which is one of the first and still widely used molecular docking programs.[37−39] It mainly uses interaction energies between a ligand molecule and a receptor protein to generate a bound pose of the ligand and does not use any reference ligand molecule. We employed the ligand-flexible docking method with rigid protein structures. Basically, a docking calculation was performed following the tutorial document. Adding hydrogen atoms and partial charges for receptor proteins and ligand molecules was done by UCSF Chimera 1.8.1.[40] The *dms* program was used to generate a molecular surface for each receptor protein. The *sphgen* program was used to create probe spheres on the molecular surface. Spheres found within 10 Å of any atoms in the reference ligand molecule were selected as a binding site. The *grid* program was used to generate the grid for potential energy. The anchor-and-grow algorithm was employed to generate binding pose of the ligand molecule on the receptor protein. All the parameters for the calculations were taken from the sample input file, except the parameter "pruning_conformer_score_cutoff". We set the parameter = 100.0 instead of the default value = 25.0, otherwise the program sometimes failed to generate a conformation. Receptor and ligand molecules for the data set of carbonic anhydrase 2 (CAH2) were carefully prepared, because most of the ligands in the data set bound to the Zn^{2+} ion in the receptor protein by the sulfonamide group of the ligand. The Zn^{2+} ion at the active site was kept in the receptor protein. One of hydrogen atom in the ligand sulfonamide group ($-SO_2NH_2$) was removed, so that the sulfonamide group was negatively charged ($-SO_2NH$).

Implementation and Availability. We have implemented our flexible alignment program *fkcombu* as a component of the *kcombu* program package. It is written in C for the Linux platform. The source code of the *kcombu* program package can be downloaded from our Web server (http://strcomp.protein.osaka-u.ac.jp/kcombu). Additionally, the data set of superimposed molecular structures can be downloaded from the server. The PDB_IDs of the four data sets are given in the Supporting Information. In the calculation, we used one core of

an Intel Xeon processor X5650 (2.66 GHz) in a Linux environment.

RESULTS

Success Rates of the Entire Data Set. To test the performance of the program *fkcombu*, we predicted the conformations of a target compound for all of the target-reference pairs in the data set. Various settings of the *fkcombu* program were examined, such as the types of MCS (C-MCS, TD-MCS with $\theta = 1$, TD-MCS with $\theta = 2$, and D-MCS), the atomic classifications (element-based, kcombu-default), and the inclusion and the exclusion of the protein-crashing energy E_{prot}.

The accuracy of each predicted conformer was measured by the RMSD from the correct pose of the target molecule. Equivalent atomic correspondences between the predicted conformer and the correct pose were generated using the symmetries of the molecules (see Figure10 in the ref 32). Among them, the correspondence with the smallest RMSD was chosen to calculate the accuracy. The success rate [%] for each method is defined as the percentage of conformers with an RMSD $\leq$ 2.0 Å.

The success rates are summarized in Table 2 and Figure 8. These predictions are regarded as "cross-docking", because the

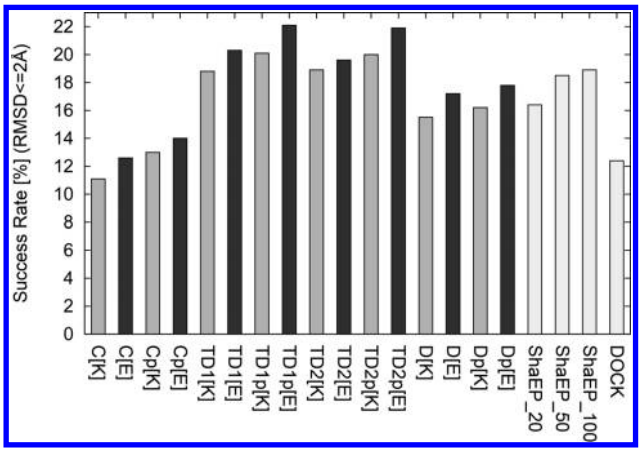


Figure 8. Success rates [%] of predictions for all data sets. The success rate is defined as the percentage of conformers with RMSD $\leq$2.0 Å. Abbreviations of the methods are shown in Table 2. Predictions with the identical pairs are excluded.

energy E_{prot} is calculated using the reference protein, not using the target protein. We excluded predictions with identical pairs, where the target molecule was used as the reference molecules. The rates for rigid-body volume-overlapping alignment program, *ShaEP* and that for the molecular docking program DOCK are also shown. We found that the type of atomic correspondence largely affected the success rates. The success rates of TD-MCS were better than those of C-MCS and D-MCS, whereas the rates of TD-MCS with $\theta = 1$ and with $\theta = 2$ were similar. These tendencies are consistent with our previous report about the agreements with 3D correspondences.[32] Surprisingly, the success rates for the simple element-based atom types were 0.7−2.0% higher than those for the kcombu-default atom types considering the ring and edge structures.

The crashing potential energy E_{prot} with the receptor protein improved the success rate by 0.6−2.3%. This shows that the 3D structures of receptor proteins can improve the performance, although Marialke et al. reported that the combination of the docking program with the 3D chemical structural alignment (HomDock) did not appreciatively improve the performance.[29]

Larger numbers of conformers led to the better performances of *ShaEP*. In fact, *ShaEP* with $c = 100$ was better than that of $c =$ 20 and 50. The success rate of *ShaEP* with $c = 100$ was similar to those of TD-MCS without the energy E_{prot}. The success rate for *DOCK* was similar to those of C-MCS, lower than D-MCS and TD-MCS.

We also compared with the *ROCS* program[10] by generating *fkcombu* predictions for the data sets of Sutherland et al.[41] The success rates of the data sets are summarized in Table S4. The rates of *ROCS* and *fkcombu* were quite similar.

We also examined more a rigorous threshold RMSD value for the success rate. Figure S1 summarizes the success rates defined as the percentages of conformers with RMSD $\leq$ 1.0 Å. The differences in the success rates with RMSD $\leq$ 1.0 Å between *fkcombu* and *ShaEP* (Figure S1) were more significant than those with RMSD $\leq$ 2.0 Å (Figure 8), although the values of the success rates with RMSD $\leq$1.0 Å were very low (2−5%). The underlying reason is probably that the flexible superposition program *fkcombu* has better ability to fit the target to the reference than the rigid superposition program *ShaEP*.

Correlation between Prediction Accuracy and Target-Reference Chemical Similarity. We examined the correlations between prediction accuracy and target-reference chemical similarity for the various methods. We prepared four Tanimoto coefficients by maximum common substructures (C-MCS and TD-MCS $\theta = 1$ using the element-based and the kcombu-default atomic classifications), and a 2D fingerprint Tanimoto coefficient. The fingerprint coefficient was calculated by the program *babel*, in the chemical toolbox OpenBabel[42] with the option −xfFP2. The correlation coefficients between the RMSDs and the chemical similarities are summarized in Table S1. These values of similarity-based methods (except *DOCK*) ranged from −0.245 to −0.568, indicating that they had statistically significant negative correlations. The Tanimoto coefficients of the 2D fingerprints had slightly weaker correlations than those of MCSs had. It is reasonable that the predictions based on C-MCS had higher correlations with the Tanimoto coefficients of C-MCS. In the same manner, predictions based on TD-MCS had higher correlations with the Tanimoto coefficients of TD-MCS. Interestingly, the *ShaEP* predictions had significant correlations with the Tanimoto coefficients based on MCS (from −0.380 to −0.422), although the *ShaEP* program did not use any atomic MCSs during its superimposing calculation. The *DOCK* had much weaker correlations (ranged from −0.031 to −0.165) than similarity-based methods had.

Figure 9 shows a plot of RMSDs and Tanimoto coefficient for the *fkcombu* prediction based on TD-MCS with $\theta = 1$ and element-based atom types. The average RMSD values are plotted as the green line. A similar RMSD−Tanimoto plot for the *ShaEP* prediction with $c = 100$ is shown in Figure S2, and that for the *DOCK* prediction is shown in Figure S3. The curves for average RMSD values of the various methods are summarized in Figure10. All of the curves except *DOCK* are quite similar, although the TD-MCS curves are slightly lower than the other two curves. Generally speaking, the average RMSD values of similarity-based docking methods are <2.0 Å when the target-references similarities are more than 70%. In contract, the RMSD value of *DOCK* is not strongly dependent on the target-reference similarities.

Success Rates for Similar and Dissimilar Target-Reference Pairs. We also calculated the success rates for two

dx.doi.org/10.1021/ci500006d | J. Chem. Inf. Model. 2014, 54, 1850−1863

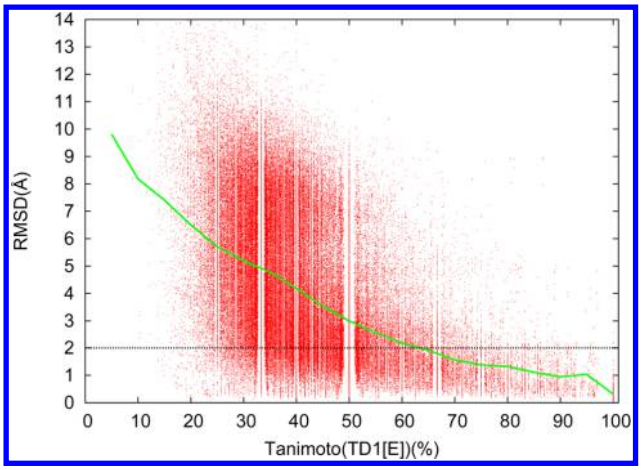


Figure 9. A plot of prediction accuracy (RMSD) versus chemical similarity between target-reference molecules (Tanimoto coefficient of TD1[E]) for the prediction of TD1[E] (*fkcombu* prediction using TD-MCS with θ = 1 and element-based atom types). The correlation coefficient is −0.481. The green line is the average prediction accuracy.

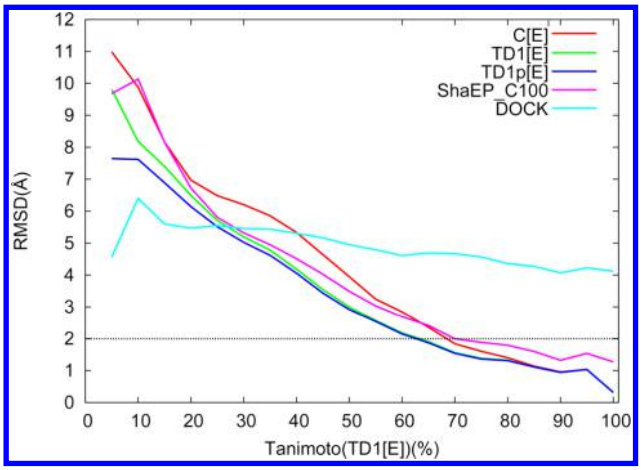


Figure 10. Average prediction accuracies (RMSD) of the four methods against the chemical similarites between the target-reference molecules. Abbreviations of the methods are shown in Table 2.

divided data sets: dissimilar and similar target-reference pairs. We chose a threshold value of 70% for the dissimilar and similar boundaries, because it roughly corresponds to the average RMSD value = 2.0 Å (shown in Figure 10). The Tanimoto coefficient with TD-MCS with θ = 1 and the element-based atom types was employed for the similarity. Table 3 and Figure 11 show the success rates for similar pairs, while Table S2 and Figure S3 show those for the dissimilar pairs. The predictions with identical pairs were excluded to calculate the success rates. The success rates for the similar sets were 4−8 times higher (shown in Table 3 and Figure 11). They ranged from 62 to 85% except *DOCK*, although the ratio of similar molecular pairs was small (3.5%). Interestingly, the *fkcombu* program works much better than the *ShaEP* program for the similar data set. In fact, the differences in the success rates were more than 10%. The success rate of *DOCK* for the similar pairs (24.0%) was much lower than those of similarity-based methods, although it was two times higher than those for dissimilar pairs (11.9%). Since 96.5% of the molecular pairs are classified as "dissimilar", the success rates for the dissimilar set are similar to those for the entire set (shown in Table S2 and Figure S3). Namely, they were around 15%.

DISCUSSION

Topologically Constrained MCS Is Better than Connected and Disconnected MCSs. Our results clearly demonstrated that TD-MCS leads to more accurate prediction than C-MCS and D-MCS. As far as we know, this is the first direct demonstration that TD-MCS is better than the conventional MCSs to predict the protein-bound conformations, although several related reports have been published. Marialke et al. proposed the maximum embedded common subgraph (EMCS), which is similar to TD-MCS with θ = 5.[28,29] However, they did not clearly show the difference in the accuracies between their new isomorphism and other conventional MCSs. In our previous report, we also found that TD-MCS has better agreement with the 3D correspondences, suggesting indirectly that TD-MCS may provide more accurately predicted conformations.[32]

Higher accuracies of TD-MCS were always observed in each of the four data sets (Table 2) and in both the similar and dissimilar pairs (Tables 3 and S2). The predictions for the molecule N41, based on DT4, are shown in Figure 12, and their atomic correspondences are shown in Figure 2. The failures of C-MCS may be due to the smaller numbers of corresponding atoms, which do not cover the entire molecules. D-MCS usually generates the largest number of corresponding atoms, but the corresponding atoms often cannot be superimposed well in 3D space. This may occur because some corresponding atom pairs have different topological distances, as shown in Figure 2B.

Element-Based Atomic Classification Is Better. We found that the atomic classification also affects the prediction accuracy. The simple element-based classification is better than the kcombu-default classification considering ring and terminal structures, as shown in Figures 8 and 11. This tendency is robustly observed for most of the MCS types and for the dissimilar and similar data sets, except for the neuraminidase data set (NEU), as shown in Tables 2, 3, and S2.

Figure 13 shows an example of the similar target-reference pair, 21Z and X02. These molecules are quite similar except for their left sides where 21Z has a ring structure and X02 has an aliphatic group. For this case, the element-based classification provides a larger correspondence including the ring-aliphatic matches, which leads to a better prediction. Figure S5 indicates a dissimilar target-reference pairs, A07 and L0F. The structures of these two molecules are quite different, but they share a similar overall shape and two characteristic neighboring nitrogen atoms in the five-membered ring. Even for these dissimilar pairs, the element-based classification also provides a larger correspondence including ring-aliphatic matches, and thus better prediction. Generally speaking, larger numbers of corresponding atoms tend to produce better predictions.

On the contrary, the NEU data set had an inverse trend. For the NEU data set, the success rates of the kcombu-default classification were better than those of the element-based classification for all types of MCS. Figure S6 shows the prediction of the NEU ligand ZMR, based on the reference molecule ABX. The kcombu-default classification provides a better prediction.

The number of rings in the data set may explain why one of the classifications is better. Table 1 indicates that most of the compounds in the NEU data set have a single ring, whereas the molecules in other data sets (THR, CAH2, and CDK2) contain various numbers of rings. If we superimpose all of the atoms of the two molecules with different numbers of rings, as in the cases shown in Figures 13 and S5, then the correspondences between

Table 3. Success Rates (%) for Similar Target-Reference Pairs

abbreviation	method	ac[a]	E_{prot}	NEU	THR	CAH2	CDK2	all
Npair[b]				82	1736	1028	1719	4565
C[K]	C-MCS	K		98.8	67.1	59.6	81.7	71.5
C[E]	C-MCS	E		98.8	80.7	68.6	83.0	79.1
Cp[K]	C-MCS	K	E_{prot}	98.8	70.3	59.5	85.7	74.2
Cp[E]	C-MCS	E	E_{prot}	98.8	82.0	69.4	86.3	81.1
TD 1[K]	TD-MCS $\theta = 1$	K		100.0	78.6	69.7	86.2	79.8
TD 1[E]	TD-MCS $\theta = 1$	E		100.0	83.9	78.3	87.0	84.1
TD 1p[K]	TD-MCS $\theta = 1$	K	E_{prot}	100.0	78.7	69.6	87.6	80.4
TD 1p[E]	TD-MCS $\theta = 1$	E	E_{prot}	100.0	85.3	78.6	88.0	85.1
TD 2[K]	TD-MCS $\theta = 2$	K		100.0	78.5	70.0	86.2	79.8
TD 2[E]	TD-MCS $\theta = 2$	E		98.8	83.9	76.8	87.1	83.8
TD 2p[K]	TD-MCS $\theta = 2$	K	E_{prot}	100.0	78.7	70.1	87.4	80.4
TD 2p[E]	TD-MCS $\theta = 2$	E	E_{prot}	98.8	85.2	78.8	88.1	85.1
D [K]	D-MCS	K		100.0	74.4	70.0	85.7	78.1
D [E]	D-MCS	E		98.8	81.6	78.7	86.7	83.2
D p[K]	D-MCS	K	E_{prot}	100.0	74.9	69.9	86.6	78.6
D p[E]	D-MCS	E	E_{prot}	98.8	82.9	78.9	87.4	84.0
ShaEP_20	$c = 20$	–		96.3	58.9	72.3	57.7	62.1
ShaEP_50	$c = 50$	–		98.8	71.7	73.7	69.0	71.6
ShaEP_100	$c = 100$	–		98.8	72.3	74.7	71.5	73.0
DOCK				53.7	30.9	19.1	18.7	24.0

[a]Atomic classification. K: kcombu-based, E: element-based. [b]Number of target-reference pairs.

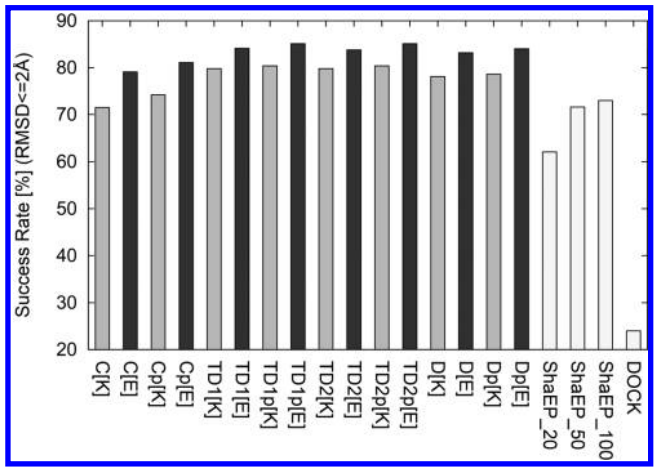


Figure 11. Success rates [%] of predictions for the similar target-reference pairs, with Tanimoto similarity (TD1[E]) ≥ 70%. Abbreviations of the methods are shown in Table 3. Predictions with the identical pairs are excluded.

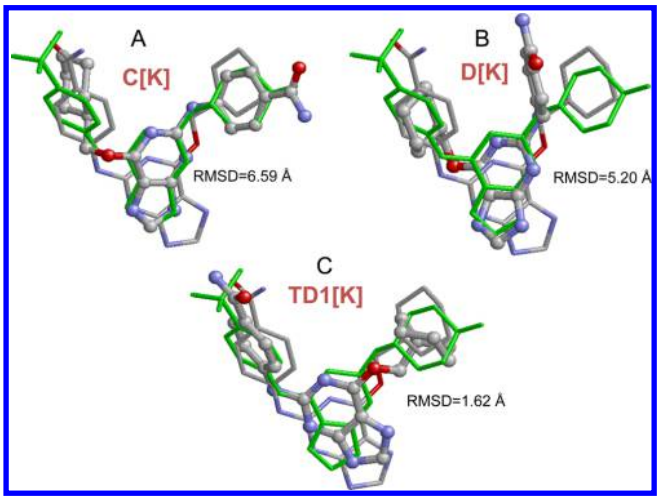


Figure 12. Predicted conformations of the CDK2 target ligand N41 using the reference compound DT4, showing that the types of MCS affect the prediction accuracies. The MCSs for these predictions are shown in Figure 2. A predicted conformation of N41 is shown as a ball-and-stick model, the correct conformation of N41 obtained from the PDB entry 1oiy is drawn by a wireframe model with CPK colors, and the reference conformation of DT4 taken from the PDB entry 2c6l is depicted by a green wireframe model. (A) The predicted conformation is based on the C-MCS (Figure 2A, C[K]), and its RMSD from the correct target conformation is 6.59 Å. (B) The predicted conformation is based on the D-MCS (Figure 2B, D[K]), and its RMSD from the correct target conformation is 5.20 Å. (C) The predicted conformation is based on TD-MCS with $\theta = 1$ (Figure 2C, TD1[K]), and its RMSD from the correct target conformation is 1.62 Å.

ring and nonring atoms are unavoidable. For these cases, the element-based classification may have an advantage. For comparing molecules that have only one ring, such as the case shown in Figure S5, discriminating the ring and nonring atoms may generate a better correspondence.

The better performance of the simple element-based classification may be controversial, although this does not indicate that the element-based classification is the best. There may be other possible atom classifications, such as Tripos MOL2 atom-type definition, H-bond donor/acceptor classifications employed by many pharmacophore models,[2] and partial charge-based classifications.[22] However, it is important to note that physical properties of two ligand compounds are not always conserved, as shown in Figures 13, S5, and S6. Aromatics and aliphatic atoms or hydrophobic and hydrophilic atoms are sometimes located in the same position on the receptor protein. The conservation of physical properties may be related to noncovalent bonds determined by the 3D structure of the compound−protein complex. It is expected that ligand atoms forming bonds (such as hydrogen bonds and salt bridges) with protein atoms are more conservative in physical properties than nonbonded ligand atoms.

dx.doi.org/10.1021/ci500006d | J. Chem. Inf. Model. 2014, 54, 1850−1863

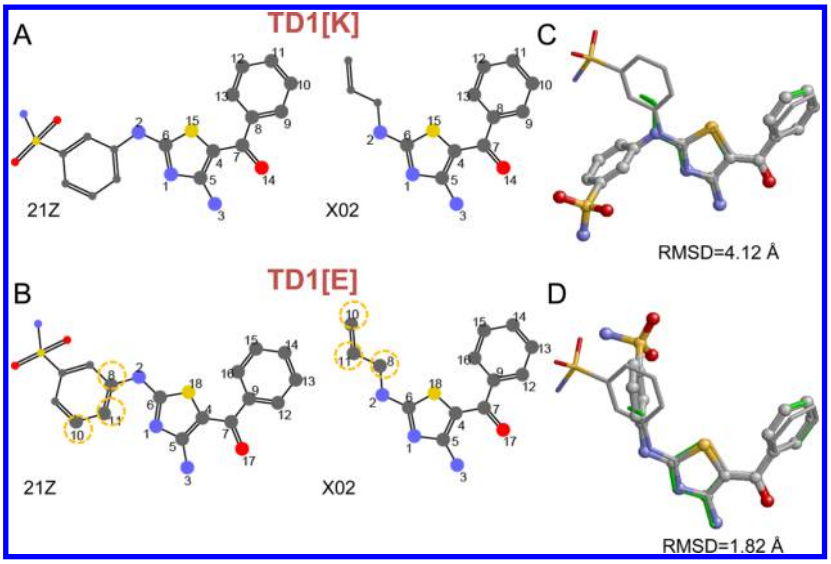


Figure 13. Atom correspondences and predicted conformations of the CDK2 target ligand 21Z using the reference compound X02, showing that the prediction by the element-based classification is better than that by the kcombu-default classification. TD-MCS with θ = 1 was used. (A) Correspondence by the kcombu-default atom classification (TD1[K]). The number of corresponding atoms is 15, and the Tanimoto coefficient is 53.6%. (B) Correspondence by the element-based classification (TD1[E]). The number of corresponding atoms is 18, and the Tanimoto coefficient is 72.0%. (C and D) The predicted conformation of 21Z (ball-and-stick model), correct conformation of 21Z obtained from the PDB entry 3rni (wireframe model with CPK color) and the reference conformation of X02 obtained from the PDB entry 3qqk (wireframe model, colored green). The predicted conformation (C) is based on the kcombu-default classification, and its RMSD from the correct 08Z conformation is 4.12 Å. The predicted conformation (D) is based on the element classification, and its RMSD from the correct 08Z conformation is 1.82 Å.

Dependence of Prediction Accuracy on Target-Reference Chemical Similarity. Many researchers have recognized that the prediction accuracy of similarity-based docking depends on the target-reference ligand similarity. However, comprehensive research like the current analysis has never been reported. Sutherland et al. reported the docking by "similarity selection" using Daylight similarity substantially increased the success rates of *CDocker*,[43] *Fred*,[44] and *ROCS*,[10] where the protein structure from the complex that contained the bound ligand most similar to the target ligand was used.[41] They reported that when applying the "similarity selection" strategy, the *ROCS* were generally more accurate than either of the docking programs (*CDocker* and *Fred*). Marialke et al. also reported that the choice of the most similar template leads to an improved success rate, but they did not provide a quantitative analysis of the accuracies and similarities.[29] Dalton and Jackson reported performance tests for protein−ligand homology-modeling, using 82 target-reference complex structure pairs.[7] They reported that the ligand-modeling accuracy was strongly dependent on the target-template ligand structural similarity, rather than the target-template protein sequence similarity. However, their data set was not large enough to extract general tendencies. In addition, since they employed a template complex in which the protein was not identical to that of the target, the dependencies on the ligand similarities were not purely extracted.

Our analysis using a large number of 3D conformations of various ligands revealed that the average prediction error (RMSD) was <2 Å, if the Tanimoto similarity between the target and reference compounds was more than 70%, as shown in Figure 9. This tendency was observed not only in our flexible superposition program but also in the rigid-body superposition program *ShaEP* (Figures 10 and S2). Therefore, it may be a general trend observed in all of the similarity-based docking programs. We hope that this finding will be helpful to estimate the accuracies of predicted conformations. The RMSD values of the docking program *DOCK* slightly correlated with the chemical similarity, however its correlation was weak.

Comparison with the Rigid-Body Superposition Method. We compared our flexible point-matching program *fkcombu* with the rigid-body superimposition method *ShaEP*, using *balloon* as the conformer generator. The success rates of the *fkcombu* program were almost comparable or slightly better than those of the *ShaEP* program. Several characteristic differences were observed. First, the success rate of the *ShaEP* program strongly depends on the number of prepared conformations, whereas the *fkcombu* program only needs one conformation for the target molecule. Second, the *fkcombu* program displays a better performance advantage for similar target-reference molecular pairs, as shown in Table 3 and Figure 13. This advantage was remarkable for the identical pair, where the target molecule was used as the reference molecule. This test is called a "self-docking" experiment. The success rates of the self-docking are summarized in Table S3. The success rate of the self-docking for the *fkcombu* program with the TD1[E] option was 99.8% (635/636), whereas that for the *ShaEP* program with the option *c* = 100 was only 87.1% (554/636). The *fkcombu* failed for self-docking of only one compound, L86 (PDB code: 1nm6) in the THR data set, which has a 19-membered ring. Similarly, the *fkcombu* program was superior to the program *ROCS* for the self-docking test, as shown in Table S5, although we did not confirm that *fkcombu* was superior to *ROCS* for the similar target-reference pairs. Third, the *fkcombu* program can provide more accurate predictions with RMSD $\leq$ 1.0 Å, as shown in Figure S1, although the absolute value of the rate is quite small (3−5%).

Comparison with the Standard Docking Method. We also compared our program with one of the standard docking programs, *DOCK* 6.6. We found that the prediction accuracy of *DOCK* correlates with chemical similarity much weaker than the similarity-based methods (*fkcombu* and *ShaEP*) did. Consequently, the success rate of *DOCK* for the similar pairs

dx.doi.org/10.1021/ci500006d | *J. Chem. Inf. Model.* 2014, 54, 1850−1863

increased less than those of the similarity-based methods did. Figure 11 indicates that similarity-based docking methods are clearly superior to the *DOCK* program for the similar target-reference pairs. For the dissimilar pairs, the difference of success rates between *DOCK* and similarity-based methods were rather small (Figure S4 and Table S2), although success rates of all the methods were low (about 10−20%). In summary, we strongly recommend similarity-based docking methods, if a 3D structure of a similar compound is available. If not available, the predictions are difficult for both standard docking methods and similarity-based methods. More elaborated docking methods, such as receptor-flexible docking and molecular dynamics simulation may be required for correct predictions without similar reference ligand structures.

Computation Time. The computation times of *fkcombu* and *ShaEP* are summarized in Table 4. In general, 1−4 s were required for one superposition. Superposition with the protein crashing energy E_{prot} required additional computation time (about 2 s). The computation times of the *ShaEP* program were somewhat proportional to the number of prepared conformations. The computation times of *fkcombu* with E_{prot} were almost the same as those of *ShaEP* with 100 conformations. However, the rigid-body superposition method required additional costs for preparing the multiple conformations for one molecule. The program *balloon* took 75.78 s on average to generate 100 conformations for one molecule. This was much longer than the time required for 1 vs 100 superpositions by *ShaEP*. If the costs for generating conformations are considered, then the *fkcombu* program is superior to the rigid-body superposition program, in terms of computation time.

Table 4. Average Computation Times to Predict the Conformation of One Target Molecule

	method	ac[1]	E_{prot}	avg time (s)
TD 1[E]	TD-MCS $\theta = 1$	E		1.28
TD 1p[E]	TD-MCS $\theta = 1$	E	E_{prot}	3.60
ShaEP_20	$c = 20$	−		0.84
ShaEP_50	$c = 50$	−		2.10
ShaEP_100	$c = 100$	−		3.79

[1]Atomic classification. E:element-based.

CONCLUSION

In this study, we developed the flexible alignment program *fkcombu* based on 2D-MCS. The performance test showed that the most accurate predictions are provided by the program *fkcombu* using TD-MCS and element-based atomic classification with the structures of receptor proteins. The prediction accuracy depends on the target-reference chemical similarity. The average RMSD of the 3D predictions is <2.0 Å, if the similarity is more than 70%. We recommend similarity-based docking methods, if a 3D structure of a similar compound is available. When only the dissimilar reference is available or more accurate predictions are required, then the combinations with molecular simulation programs are expected to improve the accuracy. The *fkcombu* program can generate reasonably good initial conformations for docking calculations and molecular dynamics simulations. To get more reference structures, complex structures of homologous proteins to the target protein can be used as the reference, although the prediction accuracy may decrease with the dissimilarity between the homologue and the target protein.

In comparison with the rigid-body methods, the *fkcombu* program has an advantage in terms of total computational costs, including the preparation of 3D conformers, because only a single conformation of one molecule is required for our program. Our program performed significantly better than the program *ShaEP*, especially when the target and the reference molecules were sufficiently similar. In addition, our program is quite fast, and it can be applied not only to pairwise 3D structural alignments but also to one vs many structural comparisons. Our program is expected to assist wide range of researchers who want to predict the protein-bound conformations of compounds.

APPENDIX

Derivative of the Energy by the Target Atom Positions

For the steepest-descent search, the first derivatives of the total energy by the target atom position and the rotational angle are necessary. First, we show the derivatives by the target atom position. The three energies E_{match}, E_{self}, and E_{prot} are represented by two types of functions h and s. Therefore, we first show the derivative of the functions h and s by atom positions:

$$\frac{\partial h(\boldsymbol{x}, \boldsymbol{r})}{\partial \boldsymbol{x}} = \frac{1}{2}\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{r})^2 = \boldsymbol{x} - \boldsymbol{r} \tag{18}$$

$$\frac{\partial s(\boldsymbol{x}, \boldsymbol{p})}{\partial \boldsymbol{x}} = \frac{\partial s(\boldsymbol{x}, \boldsymbol{p})}{\partial \|\boldsymbol{x} - \boldsymbol{p}\|}\cdot\frac{\partial \|\boldsymbol{x} - \boldsymbol{p}\|}{\partial \boldsymbol{x}} = -\alpha_{sig} s(\boldsymbol{x}, \boldsymbol{p})[1 - s(\boldsymbol{x}, \boldsymbol{p})]\cdot\frac{(\boldsymbol{x} - \boldsymbol{p})}{\|\boldsymbol{x} - \boldsymbol{p}\|} \tag{19}$$

The derivatives of three energies E_{match}, E_{self}, and E_{prot} by the position of i-th target atom $\boldsymbol{x}_i$ are described by eq 18 and 19, as follows:

$$\frac{\partial}{\partial \boldsymbol{x}_i}E_{match}(\{\boldsymbol{x}_i\}, \{\boldsymbol{r}_{r(i)}\}) = \sum_{j=1, r(j)\neq\phi}^{N}\frac{\partial h(\boldsymbol{x}_j, \boldsymbol{r}_{r(j)})}{\partial \boldsymbol{x}_i} = \begin{cases} \boldsymbol{x}_i - \boldsymbol{r}_{r(i)} & \text{if } r(i) \neq \phi \\ \boldsymbol{0} & \text{otherwise} \end{cases} \tag{20}$$

$$\frac{\partial}{\partial \boldsymbol{x}_i}E_{self}(\{\boldsymbol{x}_i\}) = \sum_{j, j\neq i}^{N}\frac{\partial s(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \boldsymbol{x}_i} \tag{21}$$

$$\frac{\partial}{\partial \boldsymbol{x}_i}E_{prot}(\{\boldsymbol{x}_i\}, \{\boldsymbol{p}_m\}) = \sum_{m=1}^{M}\frac{\partial s(\boldsymbol{x}_i, \boldsymbol{p}_m)}{\partial \boldsymbol{x}_i} \tag{22}$$

Derivative of Energy by Rotational Torsion angle

Next, we describe derivatives by the rotational torsion angle.[23,45,46] A basic principle to calculate the derivative by the target rotational angle is that deviation vector $\Delta\boldsymbol{x}$ for the rotating point $\boldsymbol{x}$ around the rotational axis $\boldsymbol{w}$ with the rotational angle $\Delta\phi$ is described by the following equation:

$$\Delta\boldsymbol{x} = \Delta\phi\boldsymbol{w} \times \boldsymbol{x} \Rightarrow \frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}\phi} = \boldsymbol{w} \times \boldsymbol{x} \tag{23}$$

where the length of vector $\boldsymbol{w}$ is 1. Geometric positions of these vectors are shown schematically in Figure 14. Using eq 23, the derivative of position of i-th target atom $\boldsymbol{x}_i$ by the rotational angle ϕ_k is described as follows:

dx.doi.org/10.1021/ci500006d | J. Chem. Inf. Model. 2014, 54, 1850−1863

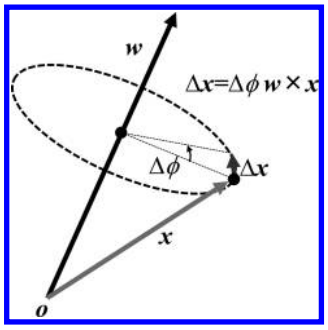


Figure 14. A schematic view of the deviation vector $\Delta \boldsymbol{x}$, for the rotating point $\boldsymbol{x}$ around the rotational axis $\boldsymbol{w}$ with the rotational angle $\Delta\phi$.

$$\frac{\partial \boldsymbol{x}_i}{\partial \phi_k} = \begin{cases} \boldsymbol{v}_k \times (\boldsymbol{x}_i - \boldsymbol{x}_{o(k)}) & \text{if } \boldsymbol{x}_i \in M_k \\ \boldsymbol{0} & \text{otherwise} \end{cases} \quad (24)$$

The derivatives of the three energies by the rotational angle ϕ_k are described using the eqs 18, 19, and 24:

$$\frac{\partial}{\partial \phi_k} E_{\text{match}}(\{\boldsymbol{x}_i\}, \{\boldsymbol{r}_{r(i)}\}) = \sum_{i=1, r(i)\neq\phi}^{N} \frac{\partial h(\boldsymbol{x}_i, \boldsymbol{r}_{r(i)})}{\partial \boldsymbol{x}_i} \cdot \frac{\partial \boldsymbol{x}_i}{\partial \phi_k}$$
$$= \sum_{i=1, r(i)\neq\phi}^{N} \begin{cases} (\boldsymbol{x}_i - \boldsymbol{r}_{r(i)}) \cdot \boldsymbol{v}_k \times (\boldsymbol{x}_i - \boldsymbol{x}_{o(k)}) & \text{if } \boldsymbol{x}_i \in M_k \\ \boldsymbol{0} & \text{otherwise} \end{cases} \quad (25)$$

$$\frac{\partial}{\partial \phi_k} E_{\text{self}}(\{\boldsymbol{x}_i\}) = \sum_{i=1}^{N} \sum_{j>i}^{N} \frac{\partial s(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \boldsymbol{x}_i} \cdot \frac{\partial \boldsymbol{x}_i}{\partial \phi_k}$$
$$= \sum_{i=1}^{N} \sum_{j>i}^{N} \begin{cases} \frac{\partial s(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \boldsymbol{x}_i} \cdot \boldsymbol{v}_k \times (\boldsymbol{x}_i - \boldsymbol{x}_{o(k)}) & \text{if } \boldsymbol{x}_i \in M_k \text{ and } \boldsymbol{x}_j \notin M_k \\ \frac{\partial s(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\partial \boldsymbol{x}_j} \cdot \boldsymbol{v}_k \times (\boldsymbol{x}_j - \boldsymbol{x}_{o(k)}) & \text{if } \boldsymbol{x}_i \notin M_k \text{ and } \boldsymbol{x}_j \in M_k \\ \boldsymbol{0} & \text{otherwise} \end{cases} \quad (26)$$

$$\frac{\partial}{\partial \phi_k} E_{\text{prot}}(\{\boldsymbol{x}_i\}, \{\boldsymbol{p}_m\}) = \sum_{i=1}^{N} \sum_{m=1}^{M} \frac{\partial s(\boldsymbol{x}_i, \boldsymbol{p}_m)}{\partial \boldsymbol{x}_i} \cdot \frac{\partial \boldsymbol{x}_i}{\partial \phi_k}$$
$$= \sum_{i=1}^{N} \sum_{m=1}^{M} \begin{cases} \frac{\partial s(\boldsymbol{x}_i, \boldsymbol{p}_m)}{\partial \boldsymbol{x}_i} \cdot \boldsymbol{v}_k \times (\boldsymbol{x}_i - \boldsymbol{x}_{o(k)}) & \text{if } \boldsymbol{x}_i \in M_k \\ \boldsymbol{0} & \text{otherwise} \end{cases} \quad (27)$$

■ ASSOCIATED CONTENT

Supporting Information

The supporting Information contains a detailed procedure for the stamping of non-planar ring conformations, five tables and six figures. Table S1 reports correlation coefficient between RMSD and chemical similarities. Table S2 and S3 report success rates for dissimilar target-reference pairs and for self-docking, respectively. Table S4 and S5 report success rates using the dataset of Sutherland et al.[41] for cross-docking and self-docking, respectively. Figure S1 shows success rates using the more rigorous threshold. Figure S2 and S3 show plots of the prediction accuracy versus the chemical similarity for the prediction of *ShaEP* and UCSF DOCK, respectively. Figure S4 shows success rates of predictions for the dissimilar target-reference pairs. Figure S5 and S6 show atom correspondences and predicted conformations for different atom classifications. The text file "NEU_CAH2_THR_CDK2_datasets.txt" contains PDB_IDs of the four datasets. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

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

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank Mitsuhito Wada and Yoichi Murakami for providing useful comments on the beta version of the *fkcombu* program. We also thank Narutoshi Kamiya and Yu Takano for their advices about interactions between the zinc ion and the sulfonamide group. This work was supported by a Grant-in-Aid for Scientific Research (C) from Japan Society for the Promotion of Science, and the Platform for Drug Discovery, Informatics, and Structural Life Science (PDIS).

■ REFERENCES

(1) Lemmen, C.; Lengauer, T. Computational methods for the structural alignment of molecules. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 215−232.

(2) Sippl, W. Pharmacophore identification and pseudo-receptor modeling. In *The Practice of Medicinal Chemistry*; Wermuth,C. G., Ed.; Academic Press: Burlington, MA, 2008; Chapter 28, pp 572−586.

(3) Langer, T.; Bryant, S. D. 3D quantitative structure-property relationships. In *The Practice of Medicinal Chemistry*; Wermuth,C. G., Ed.; Academic Press: Burlington, MA, 2008; Chapter 28, pp 587−604.

(4) Cramer, R. D. C.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(5) Fradera, X.; Mestres, J. Guided docking approaches to structure-based design and screening. *Curr. Top. Med. Chem.* **2004**, *4*, 687−700.

(6) Brylinsk, M.; Skolnick, J. FINDSITE$^{\text{LHM}}$: A threading-based approach to ligand homology modeling. *PLoS Comp.Biol.* **2009**, *5*, e1000405.

(7) Dalton, J. A. R.; Jackson, R. M. Homology-modelling protein-ligand interactions: allowing for ligand-induced conformational change. *J. Mol. Biol.* **2010**, *399*, 645−661.

(8) Kearsley, S. K.; Smith, M. S. An alternative method for the alignment of molecular structures: maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **1990**, *4*, 615−633.

(9) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.

(10) Grant, J. A.; Gallardo, M. A.; Pickup, R. T. A fast method of molecular shape comparison: a simple application of a Gaussian description of molecular shape. *J. Comput. Chem.* **1996**, *17*, 1653−1666.

(11) Mestres, J.; Rohrer, D. C.; Maggiora, G. M. MIMIC: a molecular-field matching program. Exploiting applicability of molecular similarity approaches. *J. Comput. Chem.* **1997**, *18*, 934−954.

(12) Vanio, M. J.; Puranen, J. S.; Johnson, M. S. ShaEP: Molecular overlay based on shape and electrostatic potential. *J. Chem. Inf. Model.* **2009**, *49*, 492−502.

(13) Kinnings, S. L.; Jackson, R. M. LigMatch: a multiple structure-based ligand matching method for 3D virtual screening. *J. Chem. Inf. Comput. Sci.* **2009**, *49*, 2056−2066.

(14) Brint, A. T.; Willett, P. Upperbound procedures for the identification of similar three-dimensional chemical structures. *J. Comput.-Aided Mol. Design* **1989**, *2*, 311−320.

(15) Raymond, J. W.; Willett, P. Similarity searching in databases of flexible 3D structures using smoothed bounded distance matrices. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 908−916.

(16) Brylinski, M. Nonlinear scoring functions for similarity-based ligand docking and binding affinity prediction. *J. Chem. Inf. Model.* **2013**, *53*, 3097−3112.

(17) Hawkins, P. C. D.; Skillman, G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer generation with OMEGA: Algorithm and validation using high quality structures from the protein databank and Cambridge structure database. *J. Chem. Info. Model.* **2010**, *50*, 572−584.
(18) Vanio, M. J.; Johnson, M. S. Generating conformer ensembles using a multiobjective genetic algorithm. *J. Chem. Inf. Model.* **2007**, *47*, 2462−2474.
(19) Fradera, X.; Knegtel, R. M. A.; Mestres, J. Similarity-driven flexible ligand docking. *Proteins* **2000**, *40*, 623−636.
(20) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.
(21) Guosheng, W.; Vieth, M. SDOCKER: a method utilizing existing X-ray structures to improve docking accuracy. *J. Med. Chem.* **2004**, *47*, 3142−3148.
(22) Fukunishi, Y.; Nakamura, H. A new method for in-silico drug screening and similarity search using molecular dynamics maximum volume overlap (MD-MVO) method. *J. Mol. Graphics Modell.* **2009**, *27*, 628−636.
(23) Hurst, T. Flexible 3D searching: the directed tweak technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.
(24) Moock, T. E.; Henry, D. R.; Ozkaback, A. G.; Alamgir, M. Conformational searching in ISIS/3D databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184−189.
(25) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.
(26) Lemmen, C.; Langauer, T. Time-efficient flexible superposition of medium-sized molecules. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 357−368.
(27) Labute, P.; Williams, C. Flexible alignment of small molecules. *J. Med. Chem.* **2001**, *44*, 1483−1490.
(28) Marialke, J.; Korner, R.; Tietze, S.; Apostolakis, J. Graph-based molecular alignment (GMA). *J. Chem. Inf. Model.* **2007**, *47*, 591−601.
(29) Marialke, J.; Tietze, S.; Apostolakis, J. Similarity based docking. *J. Chem. Inf. Model.* **2008**, *48*, 186−196.
(30) Grant, J. A.; Pickup, R. T. A Gaussian description of molecular shape. *J. Phys. Chem.* **1995**, *99*, 3503−3510.
(31) Kerney, C. F. F. Quaternions in molecular modeling. *J. Mol. Graphics Modell.* **2007**, *25*, 595−604.
(32) Kawabata, T. Build-up algorithm for atomic correspondence between chemical structures. *J. Chem. Inf. Model.* **2011**, *51*, 1775−1787.
(33) Chothia, C.; Leak, A. M. The relation between divergence of sequence and structure in proteins. *EMBO J.* **1986**, *5*, 823−826.
(34) Marti-Renom, M. A.; Stuat, A. C.; Fiser, A.; Sanchez, R.; Melo, F.; Sali, T. Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Struct.* **2000**, *29*, 291−325.
(35) Gabdouline, R. R.; Wade, R. C. Brownian dynamics simulation of protein-protein diffusional encounter. *Methods (Amsterdam, Neth.)* **1998**, *14*, 329−341.
(36) Kawabata, T.; Nishikawa, K. Protein structure comparison using the Markov transition model of evolution. *Proteins* **2000**, *41*, 108−122.
(37) Ewing, T. J. A.; Makino, S.; Pegg, S.; Skillman, G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecular database. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.
(38) Moustakas, D. T.; Lang, P. N.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijimans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20*, 601−619.
(39) Lang, P. T.; Brozell, S. R.; Mukherjee, S.; Pettersen, E. F.; Meng, E. C.; Thomas, V.; Rizzo, R. C.; Case, D. A.; James, T. L.; Kuntz, I. D. DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA* **2009**, *15*, 1219−1230.
(40) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605−1612.
(41) Sutherland, J. J.; Nandigam, R. K.; Erickson, J. A.; Vieth, M. Lessons in molecular recognition. 2. Assessing and improving cross-docking accuracy. *J. Chem. Inf. Model.* **2007**, *47*, 2293−2302.
(42) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(43) Wu, G.; Robertson, D. H.; Brooks, C. L., III; Vieth, M. Detailed analysis of grid-based molecular docking: A case study ot CDOCKER - A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24*, 1549−1562.
(44) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−90.
(45) Noguti, T.; Go, N. A method of rapid calculation of a second derivative matrix of conformational energy for large molecules. *J. Phys. Soc. Jpn.* **1983**, *52*, 3685−3690.
(46) Abe, H.; Braun, W.; Noguti, T.; Go, N. Rapid calculation of first and second derivatives of conformational energy with respect to dihedral angles for proteins. General recurrent equations. *Comput. Chem.* **1984**, *8*, 239−247.