Ph4Dock: Pharmacophore-Based Protein−Ligand Docking

Junichi Goto,[†] Ryoichi Kataoka,[†] and Noriaki Hirayama*[,‡]

Computational Science Department, Ryoka Systems Inc., 1−5−2 Irifune, Urayasu, Chiba 279-0012, Japan, and Basic Medical Science and Molecular Medicine, Tokai University School of Medicine, Boseidai, Isehara, Kanagawa 259-1193, Japan

Received August 2, 2004

The development and validation of the program Ph4Dock is presented. Ph4Dock is a novel automated ligand docking program that makes best use of pharmacophoric features both in a ligand and at concave portions of a protein. By mapping of pharmacophores of the ligand to the pharmacophoric features that represent the concaves of the target protein, Ph4Dock realizes an efficient and accurate prediction of the binding modes between the ligand and the protein. To validate the potential of this unique docking algorithm, we have selected 43 reliable crystal structures of protein−ligand complexes. All of the ligands are druglike, and they are varied in nature. The diffraction-component precision index (DPI) originally used in crystallography was applied in this study in order to evaluate the docking results quantitatively. The root-mean-square deviation (rmsd) between non-hydrogen atoms of the ligand in the prediction and experimental results were analyzed using DPI. The rmsd values for 25 structures, consisting of almost 60% of the dataset, are less than three times of the corresponding DPI values. It means that the precision of docking results obtained by Ph4Dock is mostly equivalent to the experimental error in these cases. The present study has demonstrated that Ph4Dock can accurately reproduce the experimentally determined docking modes if the reliable crystal structures are used. Normally the success rate of the docking is judged using rmsd ≤ 2.0 Å as the criterion. The Ph4Dock marked an appreciably good success rate of 86% based on this criterion.

1. Introduction

As the number of protein crystal structures has increased, so too has the interest in using these detailed three-dimensional knowledge for structure-based drug design. The binding sites of these proteins inherently exhibit highly selective recognition of drugs. The protein structures determined by X-ray analysis usually reveal the atomic details regarding these binding sites. If the comprehensive landscape around the binding sites is obtained, we can design molecules that optimally fit the site. Such molecules are expected to be potential drug candidates. The prediction of the binding mode of compounds based on the binding site geometry is a so-called 'docking' problem.

Since the efficient docking technique can be a powerful tool for the computer-aided drug design, many different approaches to solving the docking problems have been proposed. Including the early approach of the DOCK[1] program, currently many programs such as FLExX,[2] AutoDock,[3] and GOLD[4] are available. The GOLD system based on genetic algorithms is one of the most popular programs. The docking problems are not solved yet and none of the currently available programs are perfect in predicting the correct binding modes.[5] Although GOLD achieved a remarkable success rate of 71% in identifying the experimental binding mode, GOLD failed in some cases especially hydrophobic ligands. This failure is inherently due to the algorithm of GOLD because it is normally required that the ligand be hydrogen bonded to the binding site. To increase the success rate and quality of prediction, a new algorithm that overcomes the existing pitfalls should be developed.

Here we describe a novel docking program called Ph4Dock. This program exploits pharmacophoric features both in the ligand and the concave portions of the target protein. In this program pharmacophoric features are defined solely based on electrostatic features. Such simplified pharmacopholic features are useful for a rapid screening of docking modes. All of the possible conformations of the ligand are at first generated. Each conformation of the ligand is assigned an annotation of the pharmacophoric features and a database of annotated conformations is created. The pharmacophoric features of the concave portions are registered as pharmacophore queries. By mapping of the query features of the concaves to the pharmacophoric features of the conformations in the database, appropriate conformations of the ligand are selected and they are aligned in the concaves. This procedure has proved to be very efficient in finding the appropriate binding modes. The prescreened binding modes were refined by the subsequent optimization using the molecular mechanics calculations.

Normally a certain dataset of protein−ligand complexes selected from the Protein Data Bank (PDB)[6] is used in order to evaluate the strengths and weaknesses of the docking algorithms. Most of such complexes are usually selected on the basis of pharmacological interest alone, and much attention has not been paid to the reliability of the experimentally determined structures. To check the docking results in a more rigorous manner, we should select a suite of the reliable structures. If we

* Corresponding author. Tel +81 463 93 1121; e-mail hirayama@is.icc.u-tokai.ac.jp.
[†] Ryoka Systems Inc.
[‡] Tokai University School of Medicine.

10.1021/jm0493818 CCC: $27.50
Published on Web 12/07/2004

use less reliable coordinates or atomic positions with large atomic displacement parameters, we never know whether the docking algorithm can predict the correct binding mode or not. From this point of view, we have selected 43 reliable crystal structures of the protein−ligand complexes. We present here the results of the binding modes for these test structures predicted by Ph4Dock. A comparison between these predictions and experimentally obtained binding modes is made.

2. Computational Methods

In the present study MOE (Molecular Operating Environment)[7] was used as a developing platform of Ph4Dock. All of the algorithms of Ph4Dock were coded by use of MOE's powerful vector language SVL (Scientific Vector Language). A lot of the existing features implemented in MOE were fully employed to realize the functions of Ph4Dock.

Ph4Dock is mainly composed of five steps as follows: conformation search of ligands, concave search, pharmacophore query creation, pharmacophore search, and energy minimization.

2.1. Conformation Search of Ligands. A stochastic conformation search method was employed. This method is similar to the RIPS method[8] which generates new molecular conformations by randomly perturbing the position of each coordinate of each atom in the molecule by some small amount, typically less than 2 Å, followed by energy minimization. In this approach, various conformations are generated stochastically, and they are clustered into unique conformations based on the rmsd values calculated for non-hydrogen atoms of the conformations. The clusters are sorted with their internal potential energy calculated for the conformation that represents the particular cluster. Since the stochastic method is nondeterministic, multiple trials should be made in order to ensure that all possible conformations are covered. Therefore the generation of conformations by the stochastic method will be repeated until the number of the clusters reach to the preset value. The default value is 100, and it has proved to be suitable for docking studies. The generated conformations of ligands are stored in a database file. The generation of good conformations is indispensable for good dockings.

If electrostatic interactions are taken into account in energy minimization, folded conformations are mostly generated by intramolecular electrostatic interactions. To suppress the generation of such conformations, the electrostatic interactions are intentionally ignored during energy minimization. This procedure roughly corresponds to the generation of conformations in a media with a high dielectric constant. Since the hydrogen bond donor and acceptors in extended conformations are preserved for the intermolecular interactions with the protein molecule, this strategy is workable. A similar strategy has been already proposed.[9]

The MMFF94s force field[10] is used throughout the energy evaluation in this program. The missing partial charges are assigned by the rule applied in MMFF94s.

2.2. Concave Search. The three-dimensional structures of proteins are obtained from PDB. Although most of the protein structures in PDB have no coordinates of hydrogen atoms, those coordinates are indispensable for docking studies. Therefore the hydrogen atoms are added in accord with the standard protonation states of acidic and basic residues in proteins and their positions are optimized.

Since the binding site of a small molecule in a protein is not always a deep cleft but in some cases it is a relatively shallow depression, the binding site is designated as 'concave' in this study. The concave is identified as a collection of spheres by use of the modified Delaunay triangulation.[11] The sphere, called 'alpha sphere', is defined by a sphere that directly contacts with three non-hydrogen atoms locating on the surface of the protein. The radius of 1.4 Å is used to represent the sphere of lone-pair active atoms such as nitrogen and oxygen atoms. For the non-lone-pair active atoms such as carbon atoms, the radius of 1.8 Å is used. Dummy atoms with hydrophobic or hydrophilic attributions are placed at the center of the alpha spheres. Then the dummy atoms are clustered with the single-linkage clustering algorithm. Two clusters are merged if there is a pair of dummy atoms within a specified connection distance (default value is 2.5 Å). After clustering, sites with fewer than a specified number of dummy atoms (3 for default) are discarded. If the radius of the bound sphere is smaller than a specified radius (2 Å for default), the dummy atoms are also discarded. Using the above procedure, the concaves located on the surface of the protein are exhaustively searched.

We have checked whether the concave search can cover the actual binding sites using the dataset of 43 protein−ligand complexes. All of the ligands reside in one of the concaves and the size of this concave is the largest one in most cases.

2.3. Pharmacophore Query Creation. The electrostatic interaction energies are calculated between a unit charge on the dummy atoms and the partial charges on the protein atoms. Electrostatic interactions are evaluated by use of a distance-dependent function with the dielectric constant of 1. The dummy atoms are classified into three classes on the basis of their electrostatic interaction energies. The following energy values have been trained for the MMFF94s force field. A dummy atom with energy less than −15 kcal/mol is thought to be in the electronegative environment. It is highly possible that hydrogen-bond acceptors, anions, or hydrogen-bond acceptor/donor groups such as hydroxyl groups are localized in the vicinity of this dummy atom on the surface of the protein. On the other hand, a dummy atom with electrostatic interaction energy greater than 15 kcal/mol is in the electropositive environment that includes hydrogen-bond donors, cations, or hydrogen-bond acceptor/donor groups. The dummy atom with energy between −15 and 15 kcal/mol is deemed to be in the hydrophobic environment. All dummy atoms are then clustered. Electropositive as well as electronegative dummy atoms are clustered with a clustering radius of 0.7 Å while a radius of 1.5 Å is used to cluster hydrophobic dummy atoms. These dummy atoms represent the pharmacophoric features.

To restrict the search of ligand conformations within narrow ranges, the concept of excluded volume is introduced. The excluded volume is defined as an area in protein that cannot be occupied by small molecules such as drugs. If we define spheres at the centers of non-hydrogen atoms that surround the concave, the excluded

volume can be represented by these spheres. The non-hydrogen atom that defines the exclusion sphere is chosen such that its center is located more than 4.5 Å apart from the dummy atoms. A relatively large radius of 2 Å for the exclusion sphere is used to make smooth and continuous surface around the concave.

Based on the phramacophoric features of the dummy atoms and the information about the exclusion area around the concave, a pharmacophore query can be automatically created. Each pharmacophoric feature corresponding to the clustered dummy atoms is then assigned at the center of the cluster with the same radius for clustering. The pharmacophore query thus generated represents the chemical property and shape of a concave. The query typically has several to tens of pharmacophoric features and a few hundreds of exclusion spheres depending on the size and the complexities of the protein surface.

2.4. Pharmacophore Search. If a conformation of the ligand can properly bind with a concave of the protein, the pharmacophores in this conformation should reasonably match the pharmacophoric feature of the concave. In addition the conformation of the ligand should never overlap with the excluded volume around the concave. To find out such conformations, the pharmacophore search is conducted exhaustively of the database that contains multiple conformations of the ligand that generated in the previous step of Ph4Dock. From the practical point of view, the number of pharmacophoric features that should be matched is set to eight for the first trial. If the matching fails, the number of pharmacophoric features to be matched is decreased one by one. The search will stop when three pharmacophoric features of the concave do not match the pharmacophores in the conformation. Then the relevant conformation is judged to be impossible to bind with the concave. The procedure is performed for all of the conformations of the ligand in the database. The conformations appropriately match the pharmacophoric features of the concaves and free from steric clashes with the excluded volume are stored in a database for the following optimization.

2.5. Energy Minimization and Scoring the Results. Energy minimization of the interactions between the protein and the ligand is undertaken in order to optimize the most appropriate position and structure of the ligand in the concave. Since the potential energy calculation is the most time-consuming process, only a part of the protein is used in the calculations. The atoms within a certain cutoff distance from the dummy atoms are included in the minimization. In the first step of rough minimization, a short cutoff distance of 5 Å is used and the atoms of the protein within this distance are included in the calculation. Although the structure of the ligand is optimized, the non-hydrogen atoms of the protein are fixed. In the present study the hydrogen atoms of the protein are optimized because the locations of the hydrogen atoms may significantly change during the docking process. Optimization of side chains and all atoms of the protein can be optionally selected in Ph4Dock. The structures with interaction energies higher than a given threshold value are discarded. In the second step of the optimization, cutoff distance is increased to 9 Å and the same procedure used for the first step is followed. The ligand conformations that reasonably fit the concave are stored in a database together with their interaction energies. The interaction energy is

$$U_{\text{total}} = U_{\text{ele}} + U_{\text{vdw}} + U_{\text{ligand}} + U_{\text{solv}}$$

where U_{ele} and U_{vdw} are the electrostatic and van der Waals interactions between the protein and the conformation of ligand, respectively. U_{ligand} is the conformation energy of the relevant conformation. U_{solv} is the energy due to solvation. U_{total} and its components are stored in the database together with the coordinates of the conformation. Usually multiple conformations for a particular ligand are obtained. The values of U_{total} are sorted, and a given number (10 for default) of conformers for the ligand with the lowest U_{total} values are considered as solutions.

Although various scoring functions are proposed[12] to evaluate the docking results, in the present study only the U_{total} value was used as a scoring function. Even if this scoring function ignores the entropy term of the free energy of binding and did not always give the best solution, reasonable solutions were successfully obtained for the complexes in the test dataset.

Crystallographically determined water molecules are included in the present study since they are assumed to be fractions of the protein molecules themselves. The water molecules, however, may affect the docking results more than a little in some cases. Therefore when we apply docking software to search for potential novel ligands, we should carefully check the role of the water molecules and determine the suitable docking strategy.

2.6. Selection of 43 Protein−Ligand Complexes. To evaluate the power of the algorithm we need a suitable dataset of the complexes. Although a reliable standard dataset is not available, several datasets are proposed for validation of the docking techniques. One such dataset is proposed by the authors of GOLD to evaluate the program. The dataset was recently expanded and released as a complete CCDC/Astex validation set.[13] Complexes included in these dataset were initially selected on the basis of pharmacological interest. The second dataset that we paid attention is one proposed by Wang et al.[5] to evaluate various scoring functions for molecular docking. All of the complexes were selected because their K_i or K_d values have been experimentally measured. It is obvious that the complexes were also selected on the basis of pharmacological interest. For our purpose, the accurate structures are extremely indispensable, because ambiguously determined structures are usually useless to compare the predicted and experimentally determined structures rigorously. At least for the purpose of the validation of the docking algorithm, we believe that strictly selected good structures should be selected. We merged the above two dataset and excluded ineligible data on the following criteria. The structure that contains ligand whose occupancies being less than 1.0 were discarded, because the positions of those atoms are not precisely determined. If the atomic displacements factors of the ligand are not refined or extraordinary large (>50.0 Å^2) the structure was also discarded. The structure whose R_{free} value was not calculated was discarded since we used the R_{free} value as a guide to judge the quality of the

structure. In the structure without R_{free} value, the atomic displacement parameters are usually not refined. The structure whose resolution is worse than 2.5 Å was discarded since the resolution affect the positional error of the atoms significantly. In the current version of Ph4Dock covalently bound ligand is excluded from the scope of target. A structure that contains a decapeptide as a ligand was also excluded from the dataset because, in addition to the size, the peptide is not located in the concave but rather it attaches to the surface of the protein. The problem of this sort is still beyond our scope. Although the complex between streptavidin and biotin (PDB code: 2RTD) is not included in the two datasets, it was added to our dataset. The total number of the structures that fulfill the above conditions is 43.

The dataset consists of a wide range of ligand molecules. The molecular weight varies from 165.1 to 637.7. SlogP that is the index of hydrophobicity proposed by Crippen[14] varies from −7.48 to 5.20. The ranges of molecular weight and SlogP correspond to those observed for most of the clinically available drugs. Therefore these dataset may be especially suitable to evaluate docking programs for drug discovery.

3. Results and Discussion

3.1. A Typical Example: Biotin in Streptavidin. For the complex between streptavidin and biotin, a reliable structure ($R_{free} = 0.236$) was obtained using a high resolution (1.65 Å) data. Streptavidin is a relatively small protein, and the binding site of biotin has a clear shape. Streptavidin shows an extraordinarily high affinity ($K_d \sim 10^{-14}$ M)[15] for biotin. Streptavidin functions as a tetrameric form. The coordinates for a dimer of streptavidin are given in the PDB data 2RTD. Each monomer is designated as chains B and D. Crystallographic results have revealed that both of the binding sites are occupied by the biotin molecule, and the binding modes are essentially the same. We can check the effect of the subtle difference of the binding sites on the docking results. This complex is suitable to evaluate the present algorithm and has been used as a touchstone throughout the development of Ph4Dock. The docking process between chain B and biotin by Ph4Dock will be explained briefly as a typical example of Ph4Dock.

After addition of hydrogen atoms to streptavidin and refinement of their positions, the concaves in the protein were searched. Five concaves searched are shown in Figure 1. The protein is shown in a stick model. Concaves are shown as shaded areas in which dummy atoms are depicted as small spheres. White and red spheres represent hydrophobic and hydrophilic dummy atoms, respectively. In the biggest cancave, the biotin molecule is shown. The shape and volume of the cancave just cover the biotin molecule obtained from X-ray analysis of the complex, and this clearly demonstrates that the concave search has successfully identified the binding site. Judging from their sizes, the other concaves are not suitable for biotin to bind. Therefore in this case, the binding site can be determined unequivocally.

Using the dummy atoms in the concave, pharmacophoric features are defined as shown in Figure 2. Pharmacophoric features shown in spheres represent the chemical characteristics of the particular area of the concave. White, magenta, and yellow spheres represent anionic, cationic, and hydrophobic features, respectively. The gray meshed spheres mean the excluded volume that surrounds the concave. In Figure 2, 10 pharmacophoric features are shown. A pharmacophore query is made based on these pharmacophoric features of the concave.

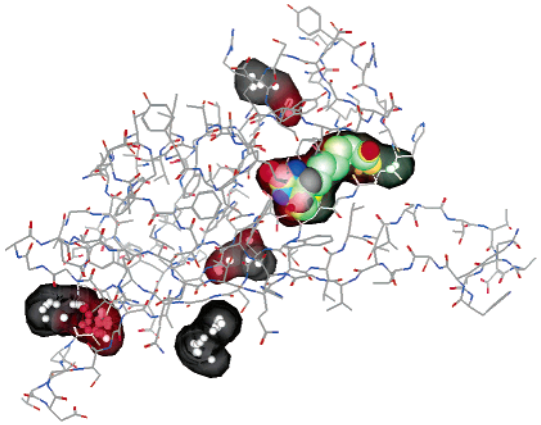

Figure 1. Concave (shaded) in streptavidin. Dummy atoms are shown as small spheres. White and red small spheres represent hydrophobic and hydrophilic, respectively. In the biggest concave region the biotin molecule obtained from X-ray analysis is shown by a space-filling model. It indicates that biotin properly occupies this concave.

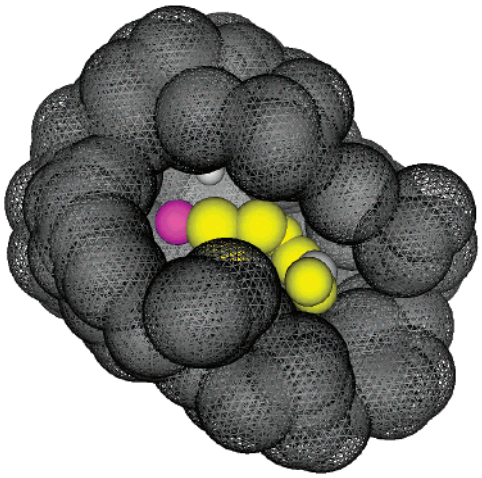

Figure 2. The pharmacophore query obtained for the biggest concave. Pharmacophoric features are shown by solid (features) and meshed (excluded volume) spheres. White, magenta, and yellow spheres represent anionic, cationic, and hydrophobic features, respectively. The gray meshed spheres represent the excluded volume.

The conformations of the biotin molecule whose pharmacophoric features reasonably fit the pharmacophore query of the concave are searched. For this purpose the database that consists of multiple stable conformations of biotin is used. The conformation that best fits to the pharmacophore query is shown in Figure 3. In this case, seven out of 10 pharmacophoric features are matched. In this figure, the pharmacophoric features are drawn by dot clouds. The sizes of the spheres are just the same as those drawn in Figure 2. The position and the structure of each conformation selected are further optimized in the concave by energy minimization. Multiple conformations are optimized to obtain the

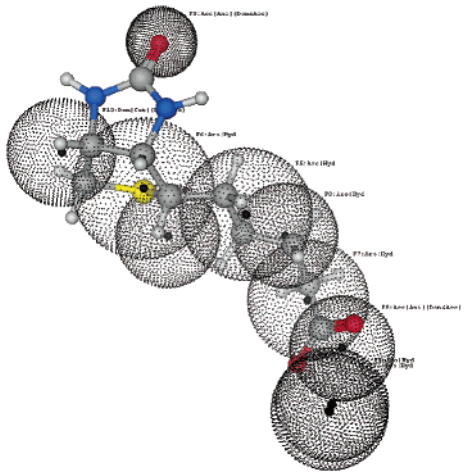

Figure 3. The conformation of biotin best fitted to the pharmacophore query of the largest concave. Seven pharmacophoric features are satisfied by the ligand atoms.

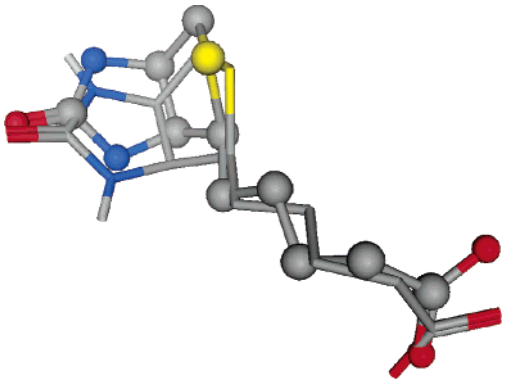

Figure 4. A superposition of the predicted and experimental structures of biotin, shown by ball-and-stick and stick models, respectively.

most appropriate conformation and the best position in the concave. In this example the structure with the lowest U_{total} value best fits the X-ray structure. Figure 4 shows superposed structures of biotin molecules. The structure predicted by Ph4Dock is drawn by a ball-and-stick model and the X-ray result by a stick model. They superpose relatively good with the rmsd value for nonhydrogen atoms being 1.142 Å. Hydrogen bonding patterns and hydrophobic interactions between biotin and the amino acids in the concave are important features to judge the quality of the docking result. Although the structure obtained from the simulation is marginally but measurably different from the X-ray structure, these intermolecular interactions observed in the crystal structure are essentially preserved in the simulated structures. The docking result obtained for this system is entirely satisfactory. This docking problem took about 40 s on an Intel Pentium 4 processor 2.5 GHz.

3.2. Reliability of Docking Results. Although various scoring functions have been discussed,[13] the reliability of the docking results has not been adequately discussed. Relatively subjective criteria have been generally applied to evaluate the docking results so far. Since protein structures show features differing from those of well ordered small molecule structures, the estimation of the standard uncertainty through the inversion of the least-squares full matrix is mostly impracticable. The R factor of the structure and the resolution of the diffraction data are usually used as guides of precision of the crystal structures. These values, however, are not so informative about the precision of the coordinates. Cruickshank introduced[16] the diffraction-component precision index (DPI) to estimate the precision of coordinates obtained by structural refinement of protein diffraction data. The DPI has been shown to provide an estimated standard uncertainty within about 13% of that generated by full-matrix inversion of the unrestrained or restrained normal matrix in at least three cases. The DPI may be 'a good and rough guide' to coordinate precision and can be used to evaluate the reliability of the docking results. Although the formulas presented by Cruickshank is relatively complex, Blow has rearranged the formulas into a more easily usable form.[17] Since DPI is originally proposed as the precision index of atomic coordinates obtained by crystallography, its application to docking problems seems to be too rigorous. DPI, however, can be estimated from the deposited data in the PDB. Therefore DPI may be a useful and objective index to evaluate the quality of the X-ray structure of the complex employed for docking study. In addition DPI is also useful as a simple guide to judge docking quality. The modified formulas of DPI proposed by Blow is

$$\sigma(r,B_{avg}) = 2.2N_{atoms}^{1/2}V_a^{1/3}n_{obs}^{-5/6}R_{free}$$

where N_{atoms} is the number of fully occupied atoms including ordered solvent atoms, V_a the volume of the crystal asymmetric unit, and n_{obs} the number of intensity observations. $\sigma(r,B_{avg})$ corresponds to the approximate standard error of position. This formula was used in this paper.

In the docking study the most useful quantity to judge the docking result is an rmsd between predicted and experimental heavy-atom coordinates of the ligand molecule. Suppose the standard uncertainty of the observed and predicted molecular model is the same in magnitude and equals to σ, the estimated standard uncertainty of the rmsd between the corresponding atoms in the observed and predicted molecule can be approximated to be $\sqrt{2}\,\sigma$. Therefore the magnitude of the rmsd value can be evaluated using the estimated uncertainty, and it gives a reasonable guide to validate the quality of docking results semiquantitatively. In this paper DPI is used as a measure to evaluate the quality of the docking results.

We have shown the docking process of streptavidin−biotin system, only for chain B. The docking of the binding site of chain D and biotin also has been carried out and we compare the two docking results here. In the case of chain D, the docking mode with the minimum U_{total} also showed the minimum rmsd value of 0.577 Å. The rmsd value is significantly smaller than that for chain B. Since the DPI of 2RTD is 0.246 Å, the corresponding estimated error of the rmsd is 0.348 Å. This value indicates that the docking result for chain D essentially agrees with the X-ray result within the experimental error. On the other hand the rmsd value for chain B is much larger. The rmsd value is, however, within a tolerable limit and the docked structure for chain B can be considered as generally correct.

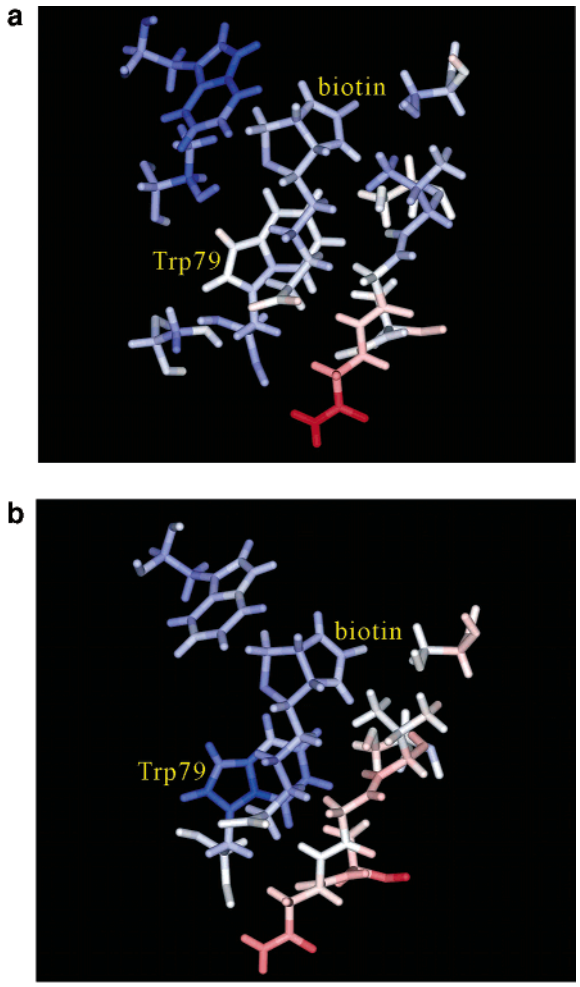

Figure 5. The amino acids around biotin in chains B and D. The residues located within 4.5 Å from biotin are shown in each case. The color of each atom is taken from a gradient that runs from blue for small atomic displacement parameter to red for large one: (a) chain B; (b) chain D.

It is of interest to examine the reasons of the difference of the two docking results. The numbers of non-hydrogen atoms of the protein chain with full occupancy are 773 and 777 in chains B and D, respectively. The average atomic displacement parameters of these atoms in chains B and D are 21.0 and 23.2 Å^2, respectively. The number of atoms with the full occupancy and the average atomic displacement parameter substantially reflect the movement of the protein atoms. Although these parameters seem cancel out each other in this case, the average atomic displacement parameter of non-hydrogen atoms of biotin are appreciably different and they are 17.0 and 15.6 Å^2 for chains B and D, respectively. It indicates that the biotin molecule is more tightly bound to the binding site of chain D. The better rmsd value obtained for the chain D is presumably correlated to the smaller atomic displacement parameters of the ligand in chain D. The amino acids residues located within 4.5 Å from the ligand are shown in Figure 5. The atomic displacement parameters are distinguished by colors. It is noteworthy that the atomic displacement parameters of two Trp residues around the ligand are significantly different. In chain B, the atomic displacement parameters of Trp79 are appreciably large. Trp79 seems be important to hold the chain moiety of biotin at the binding site. The atomic displacement parameters of Trp79 in chain D, however, are substantially small and the atomic displacement parameters of the chain moiety of the ligand is also smaller than the corresponding value in chain B. This comparison indicates that the docking results may significantly depend on the structural characteristics of the binding site. This very case also tells that as far as the evaluation of the docking software concerned we should carefully select the good structures. From a practical point of view we should use as accurate crystal structure as possible to obtain the reliable docking results.

Table 1. Results of Docking Predictions on 43 Complexes[a]

PDB code	resolution (Å)	R_{free}	$2\sqrt{2}\sigma$ (Å)[a]	rmsd (Å)	$rmsd_{min}$ (Å)
1a28	1.80	0.228	0.191	0.580	0.326
1ai5	2.36	0.222	1.137	1.238	min
1aqW	1.80	0.261	0.447	1.165	min
1b58	1.80	0.223	0.583	1.344	min
1b9v	2.35	0.271	1.321	>2.00	
1bcu	2.00	0.212	0.673	0.784	min
1bxo	0.95	0.125	0.045	1.215	1.139
1byg	2.40	0.287	1.510	1.307	1.126
1c1e	1.90	0.294	0.981	>2.00	
1c5c	1.61	0.253	0.600	0.56	min
1c5x	1.75	0.244	1.244	0.678	0.259
1c83	1.80	0.231	0.501	0.79	0.732
1cbs	1.80	0.237	0.617	1.019	min
1ckp	2.05	0.260	0.939	1.167	min
1cvu	2.40	0.235	0.896	1.000	min
1d0l	1.97	0.200	0.602	1.770	min
1d3d	2.04	0.22	0.814	>2.00	
1d3h	1.80	0.185	0.385	1.956	1.949
1d3p	2.10	0.214	0.834	>2.00	
1d4p	2.07	0.231	0.877	0.713	min
1dd7	2.25	0.284	1.307	1.209	1.185
1dg5	2.00	0.243	0.735	1.125	0.792
1ei1	2.30	0.266	1.394	1.381	min
1ejn	1.80	0.240	0.696	0.905	0.842
1f0r	2.10	0.263	1.024	1.405	0.877
1f0s	2.10	0.263	1.007	1.533	1.387
1f3d	1.87	0.218	0.464	0.596	min
1fl3	2.45	0.262	1.137	1.113	min
1kel	1.90	0.258	0.902	1.351	1.284
1lic	1.60	0.225	0.436	1.480	min
1ngp	2.40	0.250	1.162	0.837	0.832
1qcf	2.00	0.257	0.834	0.556	0.325
1qpe	2.00	0.254	0.888	1.013	min
1qpq	2.45	0.253	0.817	0.712	min
1wap	1.80	0.225	0.218	1.251	min
1yee	2.20	0.260	0.399	>2.0	
25c8	2.00	0.292	0.922	>2.0	
2ack	2.40	0.257	1.007	0.918	0.795
2pcp	2.20	0.290	1.301	0.713	min
3erd	2.03	0.248	0.905	0.898	0.299
3ert	1.90	0.262	0.690	0.774	0.613
4lbd	2.40	0.285	1.926	0.906	0.664
2rtd	1.65	0.236	0.696	1.142	min
				0.577	min

[a] $\sigma = \sigma(r, B_{avg})$.

3.3. Experiments on the Dataset of 43 PDB Complexes. The summary of docking results is shown in Table 1. The solution that gives the minimum U_{total} is assumed as the best solution in this study. The rmsd value which was used for the evaluation of the docking precision is derived from the best solution. If the best solution gives the minimum rmsd value, 'min' is indicated in the fifth column, of this table. Otherwise the minimum rmsd value is shown in this column. In the latter case, the best solution did not give the minimum rmsd value, but a solution with the minimum rmsd was obtained from a list of the solutions (10 solutions in default).

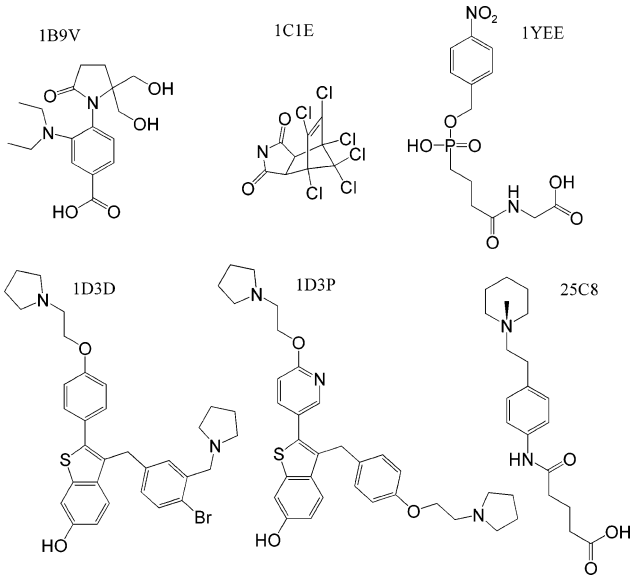

Figure 6. Chemical structures of ligands to which Ph4Dock failed to give reasonable answers.

The rmsd values of 37 structures are 2 Å or less. These structures are considered as almost correctly predicted. The rmsd values of 34 structures are less than 1.5 Å, and they are regarded as correctly predicted. The rmsd values of 25 structures, consisting of almost 60% of the whole dataset, are less than three times of the corresponding estimated error of the observed structures. It means that the docking results attained the level of precision almost comparable to the experimental error in these cases. This strongly suggests that Ph4dock can be practically applicable in the routine drug discovery platform, supposing we have good X-ray structures with which to start. In addition a variety of the ligands that are successfully docked by Ph4Dock demonstrates that the algorithm used in Ph4Dock is suitable to docking problems with a wide range of ligands from hydrophobic to hydrophilic ones.

In six cases, Ph4Dock failed. The docking results for these structures were examined in order to investigate the reasons. Generally it is not clear why Ph4Dock failed in these cases. The chemical structures of these six ligands are shown in Figure 6. No common chemical characteristics are found in these compounds. In the case of 1B9V, Ph4Dock gave a solution in which the carboxyphenyl group almost superposes and the locations of nitrogen atoms are corresponding, but the pyrrolidine ring and the pentylamino group are just flipped. The atomic displacement parameters of the atoms in the diethylamino group are significantly larger than those in the rest of the ligand in the crystal structure. The complex 1C1E contains a ligand with a hexachloro-tricyclo ring system. This ligand is located in a relatively large concave, and there is no specific contact between the ligand and the surrounding amino acids in the crystal structure. It seems that the interactions between the ligand and the protein are weak and they are solely van der Waals ones. Such an environment affects the mobility of the ligand in the concave, and the average atomic displacement parameter takes a relatively high value of 39.5 Å^2. In the cases of 1D3D and 1D3P, the atomic displacement parameters of the benzothiophene rings of the ligands are relatively small. The atomic displacement parameters of two side chains, however, are appreciably large in both ligands especially around their terminals. In addition, one of the side chains in each ligand exposed onto the protein surfaces. Although Ph4Dock could determine the positions of benzothiophene rings, the locations of the side chains are markedly deviated from the X-ray structures and their rmsd values from the corresponding X-ray structures did not drop below 2.0 Å. If the structures of ligands are represented by molecular surfaces, the structures obtained by Ph4Dock virtually superpose on the experimentally determined structures in the cases of 25C8 and 1YEE. The directions of the molecules, however, were just opposite. In the case of 1YEE, the nitro group locates in the inner part of the binding site in the crystal structure, but in the prediction the group locates on the surface of the protein. It is noteworthy that in both cases hydrogen bonds are not playing significant roles to tether the ligands at the corresponding binding sites. The characteristics common to these ligands are that in the crystal structures the terminal parts with carboxylic acids are located in the outer part of the binding sites, and their atomic displacement parameters are significantly larger than those of the remaining parts. In the crystal structure of 1YEE no hydrogen bond between the nitro group and the surrounding amino acids is observed. There is only one hydrogen bond between the amide nitrogen atom and the backbone carbonyl group. In the crystal structure of 25C8 no hydrogen bond is observed between the ligand and the protein. In those problem structures the crystal structures indicate that those ligands are all loosely bound. If the interactions between the ligand and the protein are inherently weak, the evaluation of the proper intermolecular interactions between the ligand and the protein should be essentially difficult. This must be one of the major reasons why Ph4Dock failed in these cases. Although the computation time depends on the complexity of the docking problem, it roughly took between one and 80 min on an Intel Pentium 4 processor 2.5 GHz.

3.4. Other Examples That Demonstrate the Power of Ph4Dock. The high prediction rate achieved in the present study is obviously not only due to the good structures used for the evaluation but also to the high performance of Ph4Dock. GOLD failed to predict the binding modes for a relatively hydrophobic ligand and a complex ligand such as 1ACJ and 1AAQ, respectively. Since both of these structures were not refined by use of R_{free}, the results cannot be validated using DPI. It is, however, interesting to carry out the docking experiments on these structures by use of Ph4Dock. For the docking of 1ACJ, Ph4Dock has given a relatively good answer with the rmsd value of 1.90 Å. Although the amino group is shifted by almost 2.0 Å, the docked molecule as a whole matches the experimentally determined one. Ph4Dock did fairly well in the case of 1AAQ with the rmsd value of 1.17 Å despite the complexity of the ligand. These examples also have proven that Ph4Dock can be applicable to a wide range of protein−ligand docking problems. In the current version of Ph4Dock it is not supposed to treat the covalently bound ligands. In the case of 1LCP in which the ligand coordinates to the Zn ion, the best solution obtained by

a routine application of Ph4Dock had the rmsd value of 1.36 Å. The ligand was located at a reasonable position around the Zn ion for the coordination. This successful prediction implies that Ph4Dock is potentially applicable to this type of problems if a suitable modification to the algorithm is made.

4. Conclusions

We have presented the development and validation of the program Ph4Dock. Ph4Dock is unique in that it employs the prealigned pharmacophore for docking. The full use of the prealigned pharmacophores makes Ph4Dock a versatile and reasonably quick docking program. We have tested the effectiveness of the technique on a dataset of 43 high-quality crystal structures with druglike ligands. We introduced DPI, originally used to evaluate the quality of crystal structure determination, to estimate the docking precision in more rigorous and objective way. Considering the diversity of ligands presented in the test dataset, the success rate of docking attained by Ph4Dock is really impressive. The present study has shown that Ph4Dock can be applied to a wide range of docking problems with different ligand molecules. The present study has also suggested that given a reliable crystal structure of the protein−ligand complex, we are able to predict the docking modes of drug candidates with reasonable accuracy and speed. Although there is a clear tradeoff between accuracy and speed, Ph4Dock is quick and accurate enough and it is substantially applicable to a practical virtual screening.

Acknowledgment. One of the authors (N.H.) is grateful to the Research and Study Program of Tokai University Educational System General Research Organization and the Key Technology Research Promotion Program of New Energy and Industrial Technology Development Organization (NEDO) of Japan for financial support.

References

(1) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des*. **2001**, *15*, 411−428.
(2) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol*. **1996**, *261*, 470−489.
(3) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Hey, R.; Hart, W. E,; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.
(4) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol*. **1997**, *267*, 727−748.
(5) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46* , 2287−2303.
(6) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, F.; Bryce, M. D.; Rogers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*. **1977**, *112*, 535−542.
(7) *MOE* (*Molecular Operating Environment*), version 2004.04; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2004.
(8) Ferguson, D. M.; Raber, D. J. A New approach to probing conformational space with molecular mechanics: Random Incremental Pulse Search. *J. Am. Chem. Soc.* **1989**, *111*, 4371−4378.
(9) Vajda, S.; Kataoka, R.; DeLisi, C.; Margalit, H.; Berzofsky, J. A.; Cornette, J. L. Molecular structure and vaccine design. *Annu. Rev. Biophys. Biophys. Chem*. **1990**, *19*, 69−82.
(10) Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parametrization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.
(11) Edelsbrunner, H.; Facello, M.; Fu, R.; Liang, J. Measuring Proteins and Voids in Proteins. *Proceedings of the 28th Hawaii International Conference on Systems Science*; IEEE Computer Society: Maui, HI, 1995, pp 256−264.
(12) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chmeical Databases.1.Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(13) hhtp:/www.ccdc.cam.ac.uk/products/life_sciences/validate/astex/pdb_entries.
(14) Wildman, S. A.; Crippen, G. M. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(15) Green, N. M. *Methods Enzymol*. **1990**, *184*, 51−67.
(16) Cruickshank, D. W. J. Remarks about protein structure precision. *Acta Crystallogr*. **1999**, *D44*, 583−601.
(17) Blow, D. M. A rearrangement of Cruickshank's formulae for the diffraction-component precision index. *Acta Crystallogr.* **2002**, *D58*, 792−797.

JM0493818