JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Ligand Aligning Method for Molecular Docking: Alignment of Property-Weighted Vectors

Jong Young Joung,[†,‡,⊥] Ky-Youb Nam,[‡,⊥] Kwang-Hwi Cho,[§] and Kyoung Tai No*[,†,∥]

[†]Department of Biotechnology, Yonsei University, Seoul 120-749, Republic of Korea

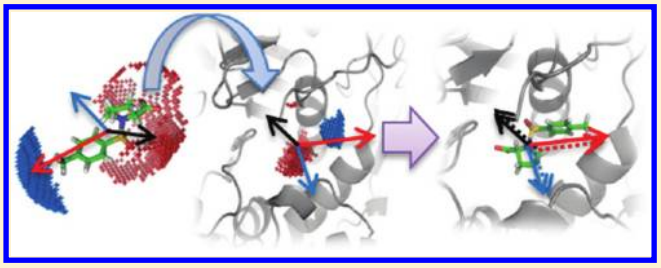
[‡]Bioinformatics & Molecular Design Research Center, Seoul 120-749, Republic of Korea

[§]School of Systems Biomedical Science and CAMD Research Center, Soongsil University, Seoul 156-743, Republic of Korea

[∥]Translational Research Center for Protein Function Control, Seoul 120-749, Republic of Korea

Supporting Information

ABSTRACT: To reduce searching effort in conformational space of ligand docking positions, we propose an algorithm that generates initial binding positions of the ligand in a target protein, based on the property-weighted vector (P-weiV), the three-dimensional orthogonal vector determined by the molecular property of hydration-free energy density. The alignment of individual P-weiVs calculated separately for the ligand and the protein gives the initial orientation of a given ligand conformation relative to an active site; these initial orientations are then ranked by simple energy functions, including solvation. Because we are using three-dimensional orthogonal vectors to be aligned, only four orientations of ligand positions are possible for each ligand conformation, which reduces the search space dramatically. We found that the performance of P-weiV compared favorably to the use of principle moment of inertia (PMI) as implemented in LigandFit when we tested the abilities of the two approaches to correctly predict 205 protein−ligand complex data sets from the PDBBind database. P-weiV correctly predicted the alignment of ligands (within rmsd of 2.5 Å) with 57.6% reliability (118/205) for the top 10 ranked conformations and with 74.1% reliability (152/205) for the top 50 ranked conformations of Catalyst-generated conformers, as compared to 22.9% (47/205) and 31.2% (64/205), respectively, in the case of PMI with the same conformer set.



INTRODUCTION

Protein−ligand docking plays a crucial role in the rational drug discovery process, high throughput virtual screening, and lead optimization. The ability to correctly dock ligands into protein binding sites is fundamental to the identification of native binding modes of ligands.[1] To predict correct binding conformations, the following should be considered: (a) the generation of an initial ligand-protein conformation near the native conformation in ligand-protein conformation space, and (b) a scoring function that has good correlation with binding free energy of the complex in aqueous solution.

The determination of the native structure of the ligand-protein complex by exploring the very complicated ligand-protein conformational space with computational methods is a very time-consuming and difficult task. This is especially true when the exploration starts far from the native complex structure in the conformational space. Therefore, it is important to generate a starting point as near to the native structure as possible in the conformational space to eventually obtain a conformation similar to the native structure. To locate such a starting point, which consists of position, orientation, and conformation of a ligand, currently used docking methods adopt incremental construction approaches (FlexX[2]), genetic algorithms (GOLD,[3] AutoDock[4]), systematic incremental search techniques (Glide[5]), shape-based algorithms (DOCK[6]), and Monte Carlo simulations (LigandFit[7]).[8] More advanced searching methods have been developed by combining the merits of several methods mentioned above. In AutoDock 3.0, a genetic algorithm is combined with a local search method to efficiently explore possible solutions near the initial starting conformation.[4] In the studies of Gilson et al., a hybrid optimization algorithm, which combined ideas from several optimizers to efficiently find a low energy conformation, was suggested to improve searching efficiency.[9,10] If the initial state of a ligand is well-defined, the searching cost to find reliable binding modes in docking can be reduced. To define the initial state that can lead to the optimal binding mode of a given ligand, molecular recognition in the complex interaction system of proteins and small molecules should be considered.

The phenomenon of ligand binding to its receptor proteins involves molecular recognition in the aqueous environment. Therefore, solvation effects play a significant role in the molecular binding process.[11] The conformation of ligands should be determined by their interaction with water molecules as well as with the receptors. This has been verified by several

Received: October 20, 2011
Published: April 3, 2012

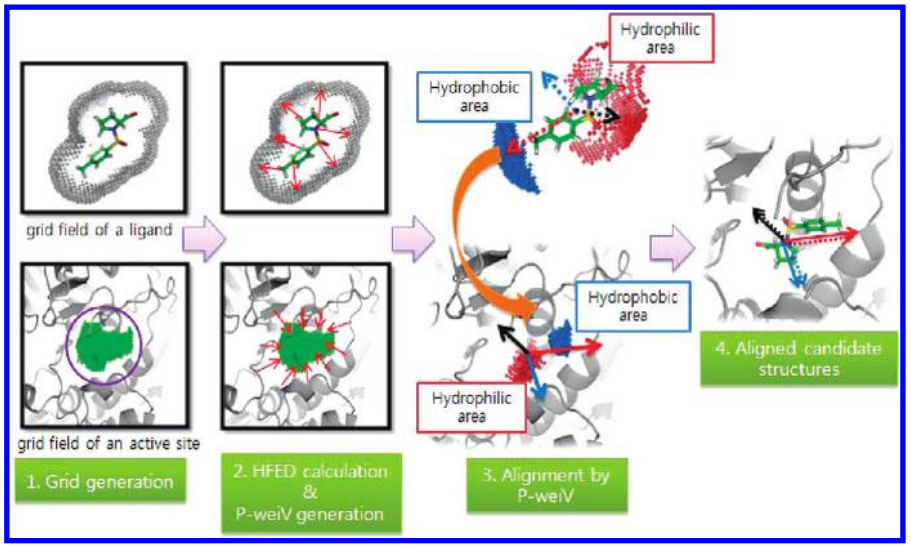


Figure 1. Scheme of the P-weiV alignment process to find the initial binding poses. (1) Grids are generated around a ligand and within the active site of a protein. (2) Hydration free energy is calculated and assigned to each grid point. P-weiV is generated from this property-field by principal axis analysis. (3) A ligand is positioned into an active site by superimposing P-weiVs with each other. (4) An aligned structure with superimposed P-weiVs.

studies, which deliberated the effects of explicit water molecules in ligand−receptor binding. Young et al. showed the importance of the thermodynamic contribution of explicit water molecules in enclosed regions like the protein binding site.[12] They confirmed that the solvation effects are the principal thermodynamic driving force for the binding of small molecule ligands to their target protein. Suresh et al. have proved that the structural water molecule on the binding interface of HIV protease with inhibitors mediated hydrogen bond between carbonyl groups of inhibitors and amide/amine groups in the flap region of the binding site.[13]

Solvation effects are considered not only in explicit ways. In most cases, implicit solvation is employed. In several docking algorithms that consider implicit solvation effects to determine the correct binding mode, approximate solvation energy is involved in the scoring functions. Many currently used scoring functions include solvation energy in their scoring scheme.[14−18]

In many docking algorithms, solvation effects have been approximated by the implicit solvation model. Usually these docking algorithms involve implicit solvation effects in the scoring functions to discriminate the correct binding mode among the pool of generated docking modes.[14−18] Shoichet et al. improved docking results in screening the Available Chemical Directory database[19] for several enzymes by including solvation effects on ligands.[11] The empirical scoring function in Autodock 3.0 showed that the contribution of binding free energy of solvation is more dominant than that of hydrogen bonds.[4] Gilson et al. applied the finite difference Poisson−Boltzmann method to calculate electrostatic hydration energies of charged groups of molecules and evaluated the contribution of solvation.[20] As demonstrated by these studies, the inclusion of solvation effects has improved the accuracy of predicting correct binding modes in docking simulations.

In this paper, we propose a novel algorithm to consider solvation effects to determine the position of the ligand. The algorithm generates initial binding modes of the ligand in active sites of the target protein by considering the solvation effect. The solvation effect is represented by the hydration free energy density (HFED) model, which calculates the hydration free energy of a target molecule in a surrounding grid space. The initial binding modes are selected using three-dimensional orthogonal vectors derived from HFED of ligands and active sites.[21,22] Because we are using three-dimensional orthogonal vectors to be aligned, there are only four possible orientations of a ligand for each ligand conformation, which reduces the search space dramatically.

MATERIALS AND METHODS

Hydration Free Energy Density (HFED). The hydration free energy density of the ligands and the proteins were calculated using the HFED method,[21,22] which was developed to describe the contribution of electrostatic interactions to reproduce the hydration free energy of small organic molecules, including amino acids. In the HFED method, hydration free energy of a molecule is expressed by

$$\Delta G_{\mathrm{hyd}} = \Delta G_{\mathrm{inter}} + \Delta G_{\mathrm{cav}} = \sum_{k}^{Q} \sum_{j}^{m} c_j(\Delta l, R_{\mathrm{shell}}) h_j^0(r_{ik}) + C_s S_s + \mathrm{const} \quad (1)$$

where $\Delta G_{\mathrm{inter}}$ is the Gibbs free energy of interaction between a molecule and its solvent environment, which is expressed as the linear combinations of four basis functions h_j^0. Each basis function implies physical properties of interactions between solutes and solvents. These basis function vectors $\vec{X} = \{h_1^0, h_2^0, h_3^0, h_4^0\}$ are

$$\vec{X} = \left\{ \left| \sum_{i=1}^{N_A} \frac{q_i}{r_{ik}} \right|, \sum_{i=1}^{N_A} \frac{q_i^2}{r_{ik}}, \sum_{i=1}^{N_A} \frac{\alpha_i}{r_{ik}^3}, \sum_{i=1}^{N_A} \frac{\alpha_i}{r_{ik}^6} \right\} \quad (2)$$

where q_i is an atom-centered net atomic charge of ith atom that is calculated by the modified partial equalization of orbital electronegativity (MPEOE) method.[23−27] The atomic polarizability, α_i, is calculated by the charge-dependent effective atomic polarizability (CDEAP) method.[28] The variable r_{ik} is a distance between a grid point and an atomic center.

dx.doi.org/10.1021/ci200501p | J. Chem. Inf. Model. 2012, 52, 984−995

The cavitation free energy, ΔG_{cav}, is expressed as the solvent accessible surface area:

$$\Delta G_{cav} = C_s S_s + \text{const} \tag{3}$$

where S_s and C_s are the number of grid points and a coefficient, respectively.

The coefficients in eq 1 were optimized to fit the experimental hydration free energy.[21]

Defining the Property-Weighted Vector (P-weiV). The hydrophobic stabilization is the main driving force of the binding between a protein and a ligand in aqueous solution.[29] Therefore, at an early stage in the binding the ligand locates and orients to maximize hydrophobic stabilization. The contact area between the protein and the ligand is also maximized to increase hydrophobic interaction. In order to describe the degree of hydrophobic stabilization due to ligand binding, the hydration free energy density tensor (HFED tensor), developed by Son et al.[30] was introduced.

Property-weighted vector (P-weiV) was devised to account for the contribution of hydration free energy into the prediction of binding modes. The P-weiVs are the principal components derived from the HFED tensor of a molecule. To generate P-weiVs, a weighted covariance matrix is constructed with hydration free energy density values as weights.[31] Given that the weighted covariance matrix is the mean error tensor[32] with the HFED, the HFED tensor can be described with this matrix. We can write the weighted covariance matrix:

$$S = \begin{pmatrix} S_{xx} & S_{xy} & S_{xz} \\ S_{yx} & S_{yy} & S_{yz} \\ S_{zx} & S_{zy} & S_{zz} \end{pmatrix} \tag{4}$$

$$S_{jk} = \frac{\sum_{i=1}^{N} w_i (q_{ij} - \overline{q_j})(q_{ik} - \overline{q_k})}{\sum_{i=1}^{N} w_i} \qquad j, k = x, y, z \tag{5}$$

where **S** is the weighted covariance matrix with elements S_{jk}, the weighted covariance among the x, y, and z coordinates, w_i is the hydration free energy density at ith grid point, q_{ij} and q_{ik} represent the j- and k-components of the positional vector of the ith grid point, $\overline{q_j}$, respectively, and N is the number of grid points of a given molecule. The terms $\overline{q_j}$ and $\overline{q_k}$ are the center of Cartesian coordinates j and k, respectively, and are obtained as follows

$$\overline{q_j} = \frac{1}{N} \sum_{i=1}^{N} q_{ij} \qquad j = x, y, z \tag{6}$$

The weighted covariance matrix represents the spatial distribution of hydration properties surrounding a molecule.[30] The eigenvectors of the property-weighted covariant matrix, P-weiVs, are derived by diagonalization of **S** through a unitary transformation.

$$\mathbf{S} = R^{\dagger} S^{l} \mathbf{R} \tag{7}$$

S^l and **R** are the diagonalized covariance matrix and the unitary transform matrix, respectively. Finally, we obtain three eigenvectors, which are property-weighted vectors. Among the three P-weiVs, the largest P-weiV represents the principal axis with the largest variance of the distribution of hydration free energy density surrounding a molecule. Each set of P-weiVs is generated for both the ligand and the binding site of the protein. Both sets of P-weiVs are aligned to give an initial orientation for the ligand posing in the binding pocket in the early stage of protein−ligand binding. (Figure 1)

Protein−Ligand Complex Data Set. To evaluate the accuracy of the P-weiV alignment method, a total of 210 protein−ligand complexes from the PDBBind database (version 2007) were used as a test data set.[33] The PDBBind database contains protein−ligand crystal structures and their experimental binding affinities, of which 1300 structures are classified as a refined set from among 3214 protein−ligand complexes. Here, 210 data in the core set, which is a nonredundant data set of the refined set, was used to develop the P-weiV method and validate its reliability. This data set has been widely used for developing other docking algorithms and testing their accuracies.[34]

Ligand Preparation: Generation of Multiple-Conformers. Because the alignment with P-weiV uses rigid ligand−rigid protein pairs, the prediction result of alignment depends on the initially given conformation of a ligand. To consider the conformational diversity of ligands in the alignment, Catalyst[35] and Vconf[36] were used to generate multiple conformers. Among 210 data in the core data set of PDBBind, multiple conformers were successfully generated in 205 and 198 cases by Catalyst and Vconf, respectively. Due to missing atom types in both programs, neither could generate multiple conformers of all of the given molecules.

Describing the Binding Pocket. The binding sites were defined as grid boxes similar to Autodock. The binding site was detected using a cocrystallized ligand in the complex structure. First of all, the center of the Cartesian coordinates of the ligand crystal structure was determined. From this center point, a 1000 Å^3 (10 Å × 10 Å × 10 Å) grid box was generated to cover the binding site fully, with a grid interval of 0.5 Å. Any grid points inside the van der Waals surface of the atoms of the protein were eliminated. The remaining grid points were used for defining the binding pocket of a protein and generating the hydration free energy field in the binding pocket. And then in following stage, the alignment using P-weiVs, input ligand conformers were aligned matching their centers of coordinates on the center of the remaining grid points and rotated to fit P-weiVs of ligands to P-weiVs of an active site.

Locating a Ligand in the Binding Pocket of a Protein. It is assumed that at the early stage of protein−ligand binding, the binding free energy is gained by the elimination of interface between the nonpolar surface of a ligand and water molecules (hydrophobic collapse) when a ligand is fitted to a binding pocket.[37] One can satisfy this assumption by aligning the P-weiVs of the ligand and of the binding pocket, which mimics the matching the surfaces of the ligand and of the protein. The alignment was performed with the following procedure.

(i) A ligand is moved into an active site by matching the center of the Cartesian coordinates of a ligand and the center of Cartesian coordinates of a grid field in the binding pocket.

(ii) Then the first eigenvector (P-weiV) of the ligand is rotated to align its orientation with the first eigenvector of the binding pocket. The rotational matrix that is derived from the rotation between these two different vectors is applied to rotate the coordinates of a ligand in the binding pocket.

(iii) In addition to the first eigenvector, the second eigenvector is also considered in positioning the ligand.

dx.doi.org/10.1021/ci200501p | J. Chem. Inf. Model. 2012, 52, 984−995

To rotate the ligand conformation according to the second eigenvector, the rotational matrix that comes from matching the second eigenvector between the ligand and the binding pocket is also applied.

As a result, the alignment, which is the translation followed by the two rotations using both the first and second eigenvectors of the ligand and the binding pocket, can produce four aligned conformations of the ligand, which are the initial binding modes. Therefore, for given N multiple conformers of one protein−ligand complex, a total of $4N$ candidates of initial binding modes are generated with this method.

With the P-WeiV aligned ligand in the binding pocket, a translational searching was performed. Energy calculations were performed at the 125 grid points of the 5 Å × 5 Å × 5 Å box that centered at the center of binding pocked and the lowest energy position among the grid points was selected.

Calculation of Relative Binding Energy. To evaluate the reliability of binding modes from the alignment, a scoring function was needed to find a defined binding position among the initial binding positions that are the products of P-weiV alignments. The relative binding energy of each generated docking pose was obtained after the alignment of each of the first and second P-weiVs to rank the $4N$ binding modes. In order to rank the $4N$ binding modes of a ligand of a given protein−ligand complex, the relative binding energy was calculated using the force field as described below. The binding energy was expressed as

$$\Delta E_{\text{binding}} = \Delta E_{\text{ele}}^{\text{P-L}} + \Delta E_{\text{vdw}}^{\text{P-L}} + \Delta E_{\text{conf}}^{\text{L}} + \Delta G_{\text{hyd}}^{\text{P-L}} \quad (8)$$

$\Delta E_{\text{ele}}^{\text{P-L}}$, $\Delta E_{\text{vdw}}^{\text{P-L}}$, $\Delta E_{\text{conf}}^{\text{L}}$, and $\Delta G_{\text{hyd}}^{\text{P-L}}$ represent the changes in the electrostatic interaction, in van der Waals interactions, in the conformation energy of the ligands, and in the hydration free energy upon binding, respectively. The electrostatic, *van der Waals* interaction, and conformational energy were calculated using the consistent valence force field (CVFF).[38] The hydration free energy was calculated using the HFED model. Contributions of hydration free energy from both the ligand and the protein and the overall hydration free energy values on each grid point were summed to give the hydration effect upon binding. In previous work, this type of energy function has been validated to effectively determine the relative stability of different conformations of amino acids in an aqueous system.[39]

Comparison to Validate the Accuracy of the P-weiV Method. To evaluate the accuracy of the P-weiV method, the method was compared with the alignment method using principal moment of inertia (PMI), as implemented in LigandFit, which uses initial alignment using the PMI of a

Table 1. Data Sets from the Multiple-Conformer Generators Catalyst and Vconf and the Average Numbers of Multiple-Conformers Generated

	number of data points used by the algorithms	average number of generated multiple conformers[c]
Catalyst_254[a]	205	127
Catalyst_50	205	37
Vconf_500[b]	198	101
Vconf_100	198	36

[a]The maximum number of conformers generated by Catalyst was 254 or 50, as indicated by the data set name. [b]The number of searching steps in Vconf was 100 or 500, as indicated by the data set name. [c]The average numbers are rounded off.

Table 2. Number of Ligands with Correct Binding Modes (rmsd <2.5 Å) in the Top-Ranked List for Each Multiconformer Data Set

	Correct Binding Rate in Ranked List															
	top 1 correct (%)		top 5 correct (%)		top 10 correct (%)		top 50 correct (%)		top 100 correct (%)		top 150 correct (%)		all correct (%)	
data set	P-weiV	PMI	P-weiV	PMI	P-weiV	PMI	P-weiV	PMI	P-weiV	PMI	P-weiV	PMI	P-weiV	PMI
Catalyst_254	66 (32.2)	19 (9.3)	107 (52.2)	40 (19.5)	118 (57.60)	47 (22.9)	156 (76.1)	66 (32.2)	161 (78.5)	71 (34.6)	166 (81.0)	76 (37.1)	172 (83.9)	80 (39.0)
Catalyst_50	66 (32.2)	19 (9.3)	107 (52.2)	40 (19.5)	118 (57.60)	47 (22.9)	156 (76.1)	66 (32.2)	161 (78.5)	71 (34.6)	166 (81.0)	76 (37.1)	172 (83.9)	80 (39.0)
Vconf _500	56 (28.3)	19 (9.6)	88 (44.4)	27 (13.6)	100 (50.5)	41 (20.7)	115 (58.1)	48 (24.2)	123 (62.1)	50 (25.3)	130 (65.7)	54 (27.3)	136 (68.7)	58 (29.3)
Vconf _100	5 (26.8)	19 (9.6)	91 (46.0)	29 (14.6)	101 (51.0)	36 (18.2)	120 (60.6)	48 (24.2)	126 (63.6)	49 (24.7)	129 (65.2)	50 (25.3)	129 (65.2)	51 (25.8)

dx.doi.org/10.1021/ci200501p | J. Chem. Inf. Model. 2012, 52, 984−995

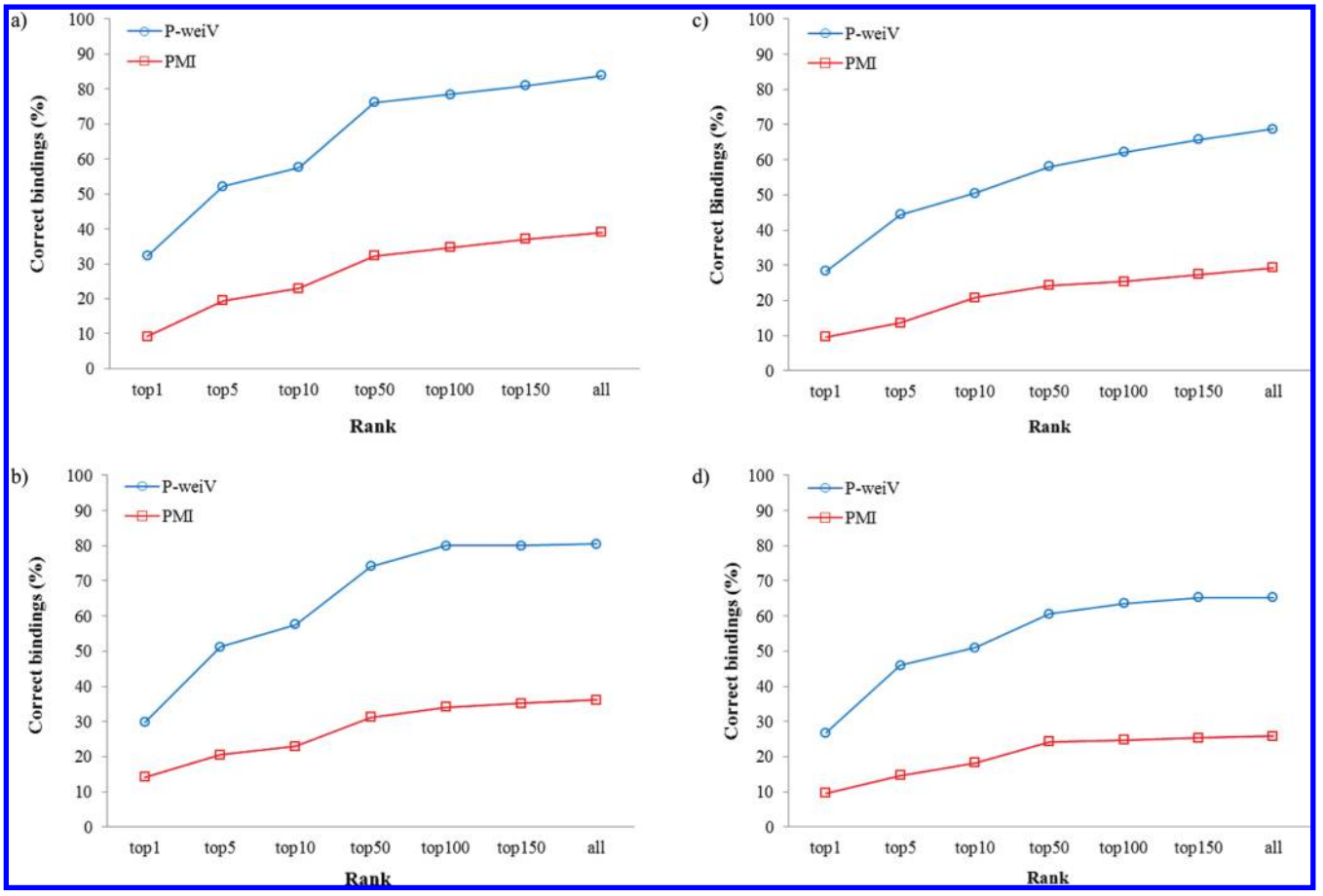


Figure 2. Percentage of correct binding modes (rmsd less than 2.5 Å) about four multiple conformer data sets: (a) Catalyst_254, (b) Catalyst_50, (c) Vconf_500, and (d) Vconf_100.

ligand and grids of an active site. The active site region was filled with grid points. The shape feature of a collection of grid points of a binding site and a ligand were characterized by the shape matrix. PMIs derived from these shape matrices were aligned to position the ligand over the site grid. The comparison to the PMI alignment method between two different collections of points was intended solely to check the shape discrepancy, without other molecular properties, such as hydration free energy density. The overall alignment procedure was similar to that of the P-weiV method. By using these PMIs in place of the P-weiVs in our method, each conformer generates four aligned binding modes in a protein−ligand complex. For each of these binding modes, the binding energy calculation was performed using the same binding energy function described by eq 8. Thus, we can compare the accuracies of PMIs and P-weiVs without the influence of the scoring function used. The root mean squared deviation (rmsd) between the crystal structure and the predicted binding mode of the ligand were used to compare the accuracy of both the methods. The numbers of correct binding modes (rmsd of heavy atoms $\leq$2.5 Å)[40] among highly ranked ligand conformations were counted for comparing the two methods.

RESULTS AND DISCUSSION

To evaluate the performance of the P-weiV method against the PMI method, four data sets were prepared. We generated two conformer data sets by setting the maximum number of conformers at 254 (Catalyst_254) and at 50 (Catalyst_50) using Catalyst. Two other conformer data sets were generated by setting the maximum number of searching steps to 500 (Vconf_500) and to 100 (Vconf_100) using Vconf (Table 1). The performance of the P-weiV and the PMI method for four conformer data sets were compared (Table 2; see Tables S1−4 for more details). The P-weiV method shows better reliability in generating correct binding modes than the PMI method for these four conformer data sets (Figure 2). The P-weiV method correctly predicted the binding modes with rmsd $\leq$2.5 Å for 57.6% of the top 10 ranked binding modes, while the PMI method only correctly predicted 22.9%. With the smaller size of the Catalyst-generated multiple conformer set, the P-weiV method and the PMI method showed similar accuracies. P-weiV identified binding modes with higher accuracy than the PMI method in overall rmsd ranges with multiconformer data sets using Vconf (Tables S3 and S4 of the Supporting Information). Therefore, based on these results, it is apparent that the alignment with orientations derived from hydration free energy density performed better than shape-based alignment in reproducing known binding modes for a protein−ligand data set.

Even when the numbers of multiple conformers in data sets were reduced by using each algorithm (the Catalyst_50 and Vconf_100 sets), the predictabilities of correct binding modes were similar to the larger data sets (the Catalyst_254 and Vconf_500 sets). Although it is difficult to directly compare its performance with those of currently available docking

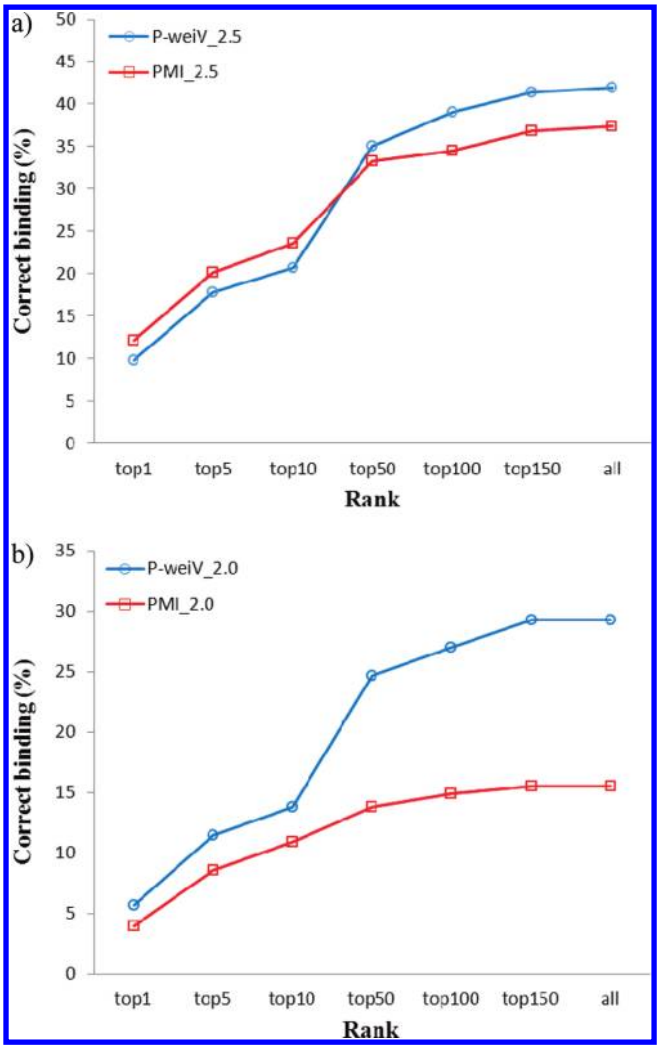


Figure 3. Comparison of correct binding rates of results from P-weiV combined with translational search and PMI. Comparison of results of correct binding with (a) RSMD standard 2.5 Å and (b) rmsd standard 2.0 Å.

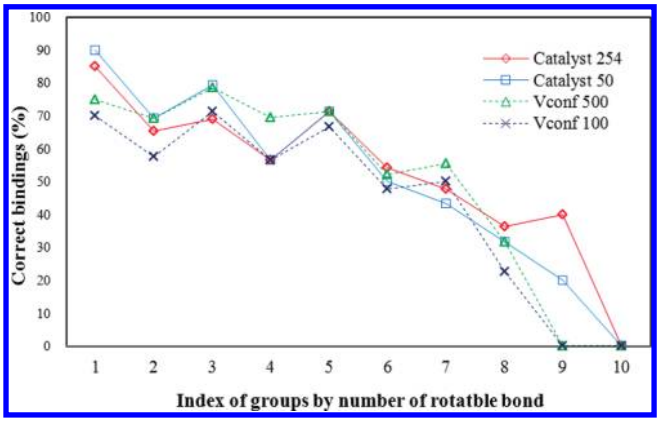


Figure 4. Percentage of correct binding modes of the data sets Catalyst_254 (red diamonds), Catalyst_50 (blue squares), Vconf_500 (green triangles), and Vconf_100 (purple *x*'s) by P-weiV are compared to the number of rotatable bonds, which are shown in Table S6 in the Supporting Information.

programs, the potential efficiency of the P-weiV method can be inferred by considering the computational costs of those programs. Glide adopts hierarchical filters to search for the proper location of ligands in the active sites. To consider ligand flexibility, Glide rotates flexible bonds attached to a rigid core, resulting in 120−670 conformers for a data set of 796 protein−ligand complexes.[5] GOLD uses a genetic algorithm (GA) for conformational searching that applies genetic operations 100 000 times to subpopulations containing 100 individuals.[3] AutoDock uses a more advanced genetic algorithm, the Lamarckian genetic algorithm, to evaluate energy functions with up to 1.5×10^6 operations.[4] LigandFit, which is implemented in Discovery Studio, assigns a random conformation for input ligands. Rigid body minimization is employed as each input ligand is flexibly docked to the corresponding protein with a maximum of 10 000 trials.[7] The P-weiV method showed similar performance of finding initial docking poses using small data sets. From these results, it was validated that P-weiV can assess the correct binding modes with reduced computational efforts in generating multiple conformations of ligands in the initial stage of docking process. If the P-weiV method combined with optimization processes be used in virtual screening works, it would search virtual hits from the large scale of chemical libraries using reduced computational cost.

In the results of the alignments combined with translational search, the rates of correct binding with 2.0 Å rmsd of P-weiV method are better than PMI. And when candidates were selected in top 50 ranks and above, the rates of correct binding with 2.5 Å rmsd of P-weiV method are higher than PMI (Figure 3 and Table S5 of the Supporting Information). Since the translational search used in these trials is a kind of minimization of the initial alignment, the structural deviation of initial binding modes having a rmsd of over 2.0 Å is decreased after translational search. Therefore the success rates of P-weiV with 2.0 Å rmsd, which is the general standard of correct binding in docking methods, are higher than PMI method.

The results of the P-weiV method were analyzed by the number of rotatable bonds of ligands (Figure 4). As the number of rotatable bonds increases, the accuracy of the P-weiV method decreases. The groups of index number 6 and above have several sets of ligands with numbers of rotatable bond together (see Table S6 for more details). This is a common problem in the other docking methods,[5] because the conformational space to be searched increases exponentially as the number of rotatable bonds in the ligand increases. However, the P-weiV method needs to consider only four orientations for each conformer while other docking programs consider many modes. Even though P-weiV considers a limited number of conformations, 63.9% (106 out of 166) of ligands in the Catalyst_254 data set having 1−10 rotatable bonds fall within the correctly predicted range of rmsd of less than 2.5 Å from the top 10 binding modes. Similarly, 65.6% (109 out of 166) of ligands in the Catalyst_50 data set and 67.9% (108 out of 159) of ligands in Vconf_500 data set, both with the same number of rotatable bonds have rmsd of less than 2.5 Å. In the Vconf_100 data set, about 60.4% (96 out of 159) of ligands are predicted as correct binding modes. For more flexible ligands in the four data sets, the proportion of ligands that have rmsd within 2.5 Å decreases. For the multiconformer data sets of ligands with less than 10 rotatable bonds, P-weiV showed potential in identifying reliable binding modes efficiently.

Some examples from the Catalyst_254 data set that have at least one correct binding mode within the top 10 ranking are

dx.doi.org/10.1021/ci200501p | J. Chem. Inf. Model. 2012, 52, 984−995

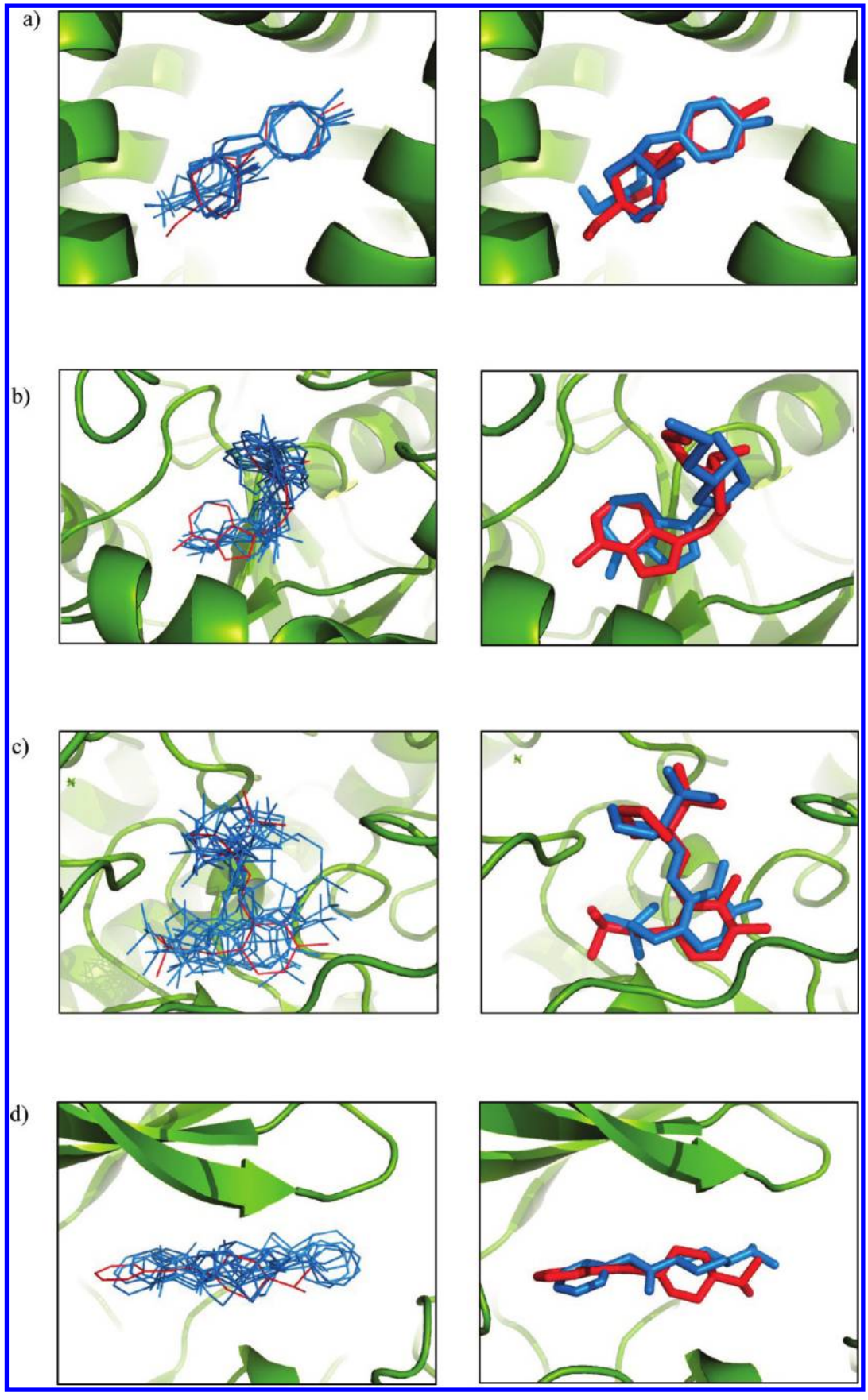


Figure 5. continued

dx.doi.org/10.1021/ci200501p | J. Chem. Inf. Model. 2012, 52, 984−995

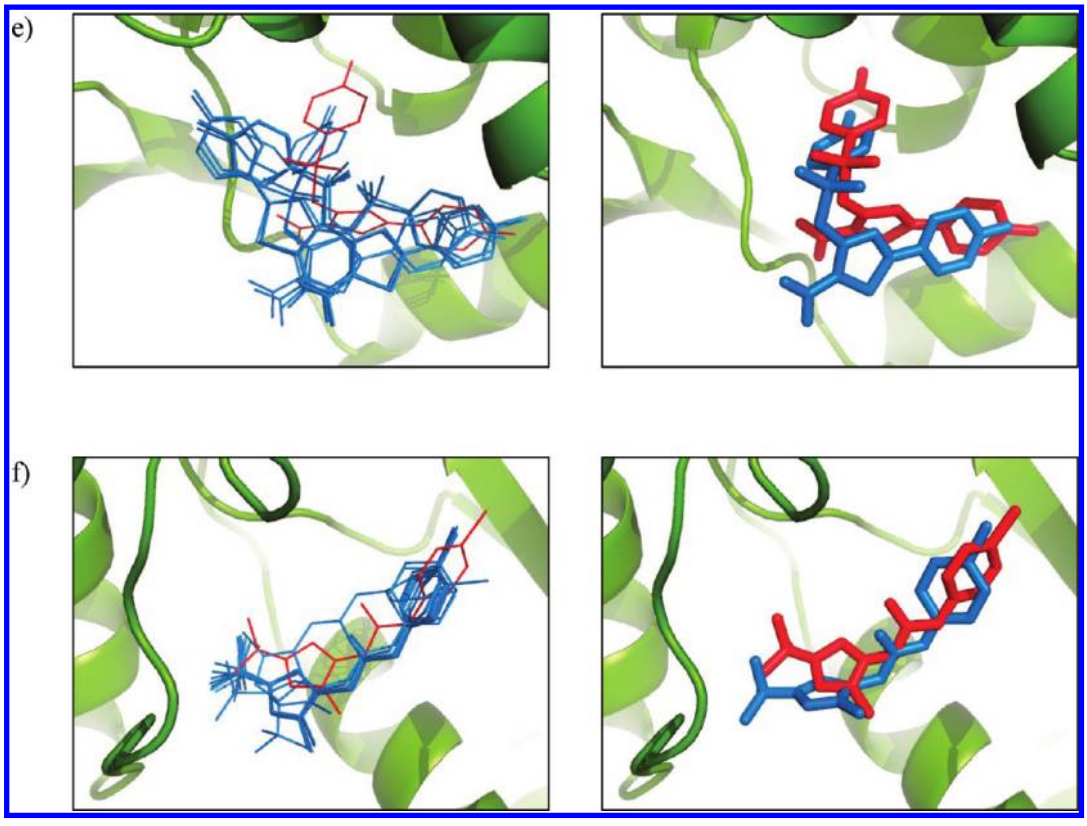


Figure 5. Examples whch have at lease one correct binding mode within the top 10 rank from Catalyst_254 data set. The left side shows aligned structures (blue) with rmsd ≤2.5 Å in the active site of a receptor (green) with a crystal ligand (red). The right side shows that a comparison of the best aligned structure (blue) and the crystal structure (red). The protein−ligand complexes are (a) human estrogen receptor alpha ligand-binding domain, PDB 2B1V, rmsd 0.58−1.43 Å; (b) MTA/AdoHcy nucleosidase, PDB 1Y6Q, rmsd 1.01−4.48 Å; (c) Diakylglycine decarboxylase, PDB 1M0Q, rmsd 1.28−4.81 Å; (d) catalytic subunit of cAMP-dependent protein kinase, PDB 1Q8T, rmsd 1.11−5.56 Å; (e) hepatitis C virus RNA-dependent RNA polymerase, PDB 2D3Z, rmsd 2.12−6.20 Å; and (f) p38 MAP kinase, PDB 1KV1, rmsd 1.17−2.22 Å.

listed in Figure 5. In the first pictures of each example, all of the top 10 ranking binding modes are shown together with the crystal ligand structure (red). In the second pictures, only the best predicted pose with lowest rmsd among 10 binding modes is presented with the crystal structure (red). From these results, we recognized that the relatively hydrophilic (or hydrophobic) areas of ligands with the relatively low (or high) hydration free energy could be aligned with similar orientation to areas with similar properties in the binding sites. (See Figure S1 of the Supporting Information for more detail) These results imply that the P-weiVs could be good identifiers in approximately matching sub-structures of ligands to the proper locations in the binding pocket.

In some cases, hydrogen bonding between a protein and a ligand is not accurately described by the P-weiVs (Figure 6). Since directional interaction such as the hydrogen bond[41] was not included in the hydration free energy calculation, positional and orientational deviation between a binding mode and a crystal structure occurred in cases where hydrogen bonds are more significant. Moreover, the scoring function representing binding energy does not have an energy component for hydrogen bonding. When docking poses were scored according to this binding energy, some of the candidates with good poses accompanied by well-matched hydrogen bonding may not have been ranked highly. If the effects of hydrogen bonds on the surrounding field are incorporated into the calculation of hydration free energy and the scoring function, the reliability of the P-weiV method would increase for cases in which the hydrogen bonding interaction is critical. However, the P-weiV method does not accurately express the effect of hydrogen bonding interactions because HFED does not include explicit HB term in the model but HB contribution is included implicitly in the HFED model during parameters optimization with experimental data. Since P-weiV method generates initial binding modes, the details of the distance and angles of the HB should be obtained through the postrefinement procedure. With the initial binding mode generated with P-weiV, simple energy minimization and simulated annealing ware carried out as postrefinement procedure. Through the postrefinement procedure we can obtain the binding structure that has small geometric standard deviation compared with the crystal structure (see figure S2), especially for the ligand-protein complex that gives wrong prediction only with P-weiV alignment procedure as described in Figure 6.

The P-weiV method was also used to align ligands within the active sites of metalloproteins (Table S7 of the Supporting Information). Metal ions are important to the enzymatic reactions of metalloproteins. The catalytic metal cations in metalloproteins are usually "hard" metal ions with low polarizability that are difficult to deform. The hard metal ions in enzymes are the alkali metal ions Na^+ and K^+, the alkaline earth ions Mg^{2+}, Ca^{2+}, Mn^{2+}, Cr^{3+}, and Co^{3+}, and the ions

dx.doi.org/10.1021/ci200501p | J. Chem. Inf. Model. 2012, 52, 984−995

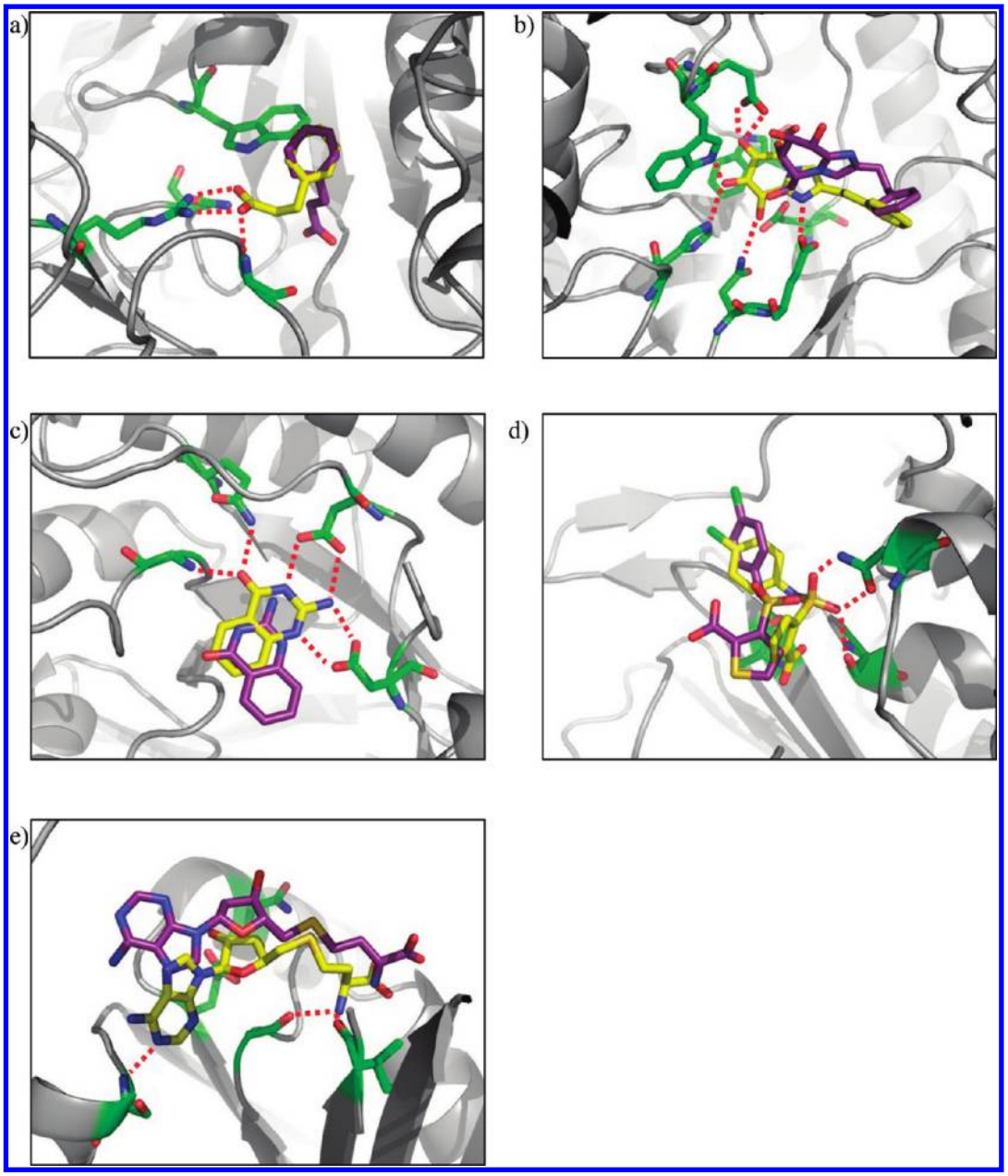


Figure 6. Examples of aligned structures (purple) illustrating local hydrogen bonds (red dotted lines) that are not located correctly and the crystal structures of ligands (yellow) that form hydrogen bonds with neighboring residues in the active sites: (a) *E. coli* aspartate aminotransferase with hydrocinnamic acid, PDB 1TOI, where the carboxylic group orientation is perpendicular to the hydrogen bond donors; (b) beta-glycosidase with phenethyl substituted glucoimidazole, PDB 2CER, where the aligned glycoimidazole is beside its position in the crystal structure; (c) tRNA-guanine transglycosylase with 2-aminoquinazolin-4(3*H*)-one(AQO), PDB 1S39, where the orientation of AQO is slightly perpendicular to the original structure due to the hydrogen bond between the hydroxyl and amine group of AQO not being specified; (d) AmpC beta-lactamase with 3-[(4-chloroanilino) sulfonyl] thiophene-2-carboxlyic acid, PDB 1L2S, where the hydrogen bond by the oxygens of the sulfuric group is weakly present; and (e) human histamine methlytransferase with *s*-adenosyl-L-homocysteine (SAH), PDB 1JQD, where the overall orientation of SAH is well aligned but slightly above the hydrogen bond partners.

Zn^{2+} and Fe^{3+}.[42] The original HFED method was developed to predict hydration free energy for organic molecules and did not consider these metal ions.[21,22] In order to include metal ions in the hydration free energy calculation, their partial net atomic charges were replaced by formal charges and their polarizability was ignored due to low polarizability of these metal ions. We analyzed 71 protein−ligand complexes generated by the Catalyst multiple-conformer generator and 66 complexes by the Vconf generator. The P-weiV method was able to predict binding modes more correctly than the PMI method for the metalloprotein complexes from the Catalyst and Vconf data sets (Figure 7). Within the top 10 ranked binding modes from the metalloproteins data set, the P-weiV method showed almost 2-fold higher accuracy relative to the PMI method. In Catalyst-generated conformer sets, P-weiV correctly identified binding modes 54.9% and 52.1% in the large set and small set, respectively, while the PMI method only correctly predicted 28.2% and 23.9%. In the Vconf-generated conformer sets, the results were substantially similar to the results from the Catalyst-generated sets.

dx.doi.org/10.1021/ci200501p | *J. Chem. Inf. Model.* 2012, 52, 984−995

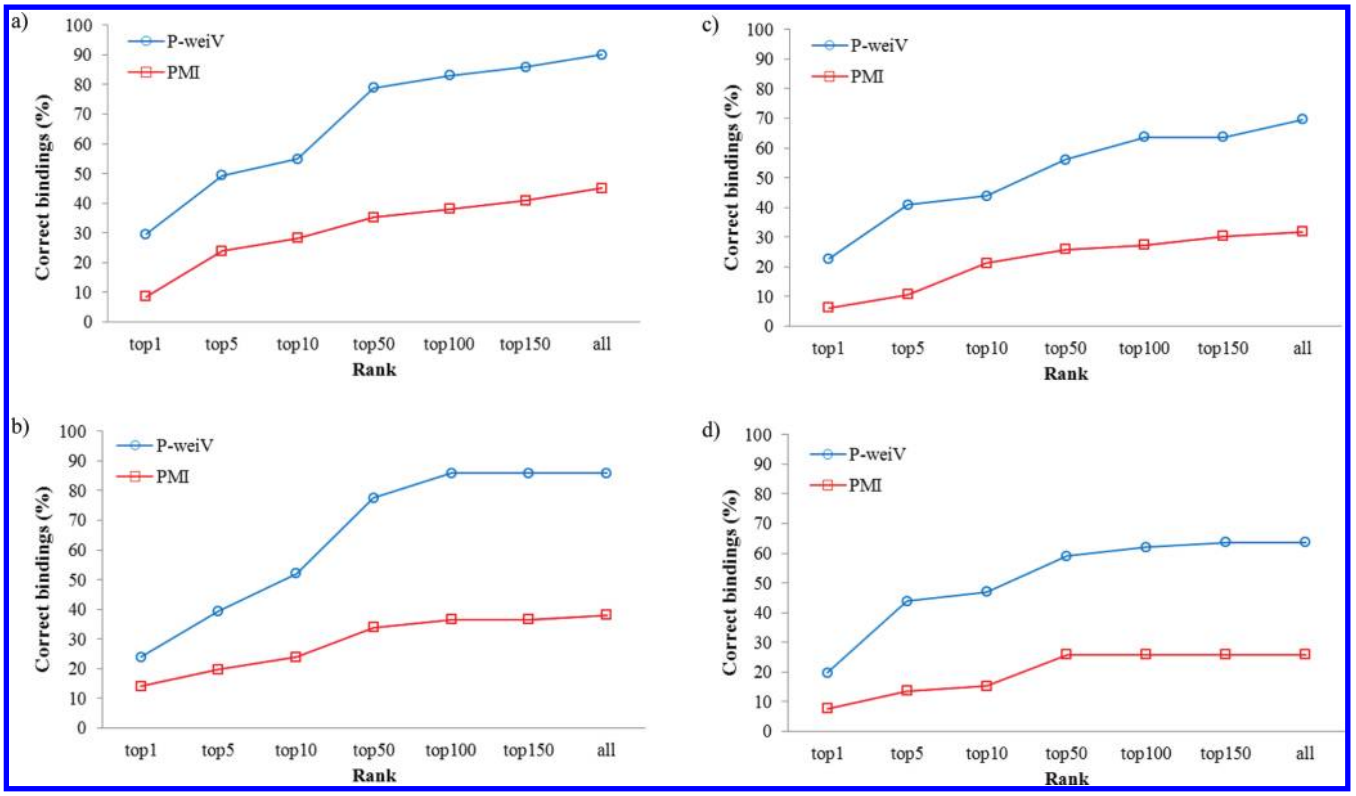


Figure 7. Comparison of fraction of correct binding modes among candidates from P-weiV and PMI with metalloprotein complexes: (a) Catalyst_254 data set, (b) Catalyst_50 data set, (c) Vconf_500 data set, and (d) Vconf_100 data set.

CONCLUSIONS

We present a molecular alignment algorithm, called the P-weiV method, to predict initial binding modes of ligands according to the vectors reflecting hydration free energy density. The P-weiV method reliably gives initial orientations of ligands during the docking process. The spatial vectors used to suggest binding orientations of ligands are the orthogonal axes, meaning the largest gradient of hydration free energy density. The binding modes were generated by aligning the orthogonal vectors of the ligand with those of the receptor. Those binding modes were then ranked according to the scoring function of relative binding energy. In order to validate the accuracy of the P-weiV method, four kinds of multiple conformer data sets of ligands were prepared using Catalyst and Vconf. To validate the advantage of this method compared with shape-based positioning of ligands, which is the currently available algorithm among docking softwares, a PMI-based alignment method was used to generate binding modes with the same data set. For all prepared data sets, the P-weiV method exhibited better accuracy than the PMI method. The P-weiV method produces more correct binding modes (approximately 30% more) than the PMI method in top 10 ranked lists in all four types of data sets. Furthermore, compared with different numbers of multiple-conformer data sets, the P-weiV method showed consistent performance in a data set with reduced numbers of conformers. These comparisons with different sizes of data sets indicate that a reduction computational cost could be expected in virtual screening using large numbers of chemical libraries to discover hit compounds. Although the P-weiV method has the potential to predict acceptable binding modes more efficiently, further work, including development of a revised hydration free energy model and a scoring function incorporating hydrogen bonds and interactions with catalytic metal ions, is required to improve reliability to assess correct binding modes.

ASSOCIATED CONTENT

Supporting Information

Tables S1−S7 and Figures S1 and S2 as mentioned in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ktno@yonsei.ac.kr. Telephone: 82-2-393-9551. Fax: 82-2-393-9554.

Author Contributions

⊥These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This research was supported by a grant from the Translational Research Center for Protein Function Control (2010-0001931), the National Research Foundation, the Ministry of Education, Science and Technology, and the Korea Healthcare Technology Research and Development Project, Ministry for Health, Welfare & Family Affairs (A085105, A100096), Republic of Korea. We thank Professor Sung Kwang Lee at Hannam University and Dr. Chang Joon Lee for helpful discussion. We also acknowledge VeraChem LLC for the license for Vconf.

REFERENCES

(1) Fuhrmann, J.; Rurainski, A.; Lenhof, H. P.; Neumann, D. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *J. Comput. Chem.* **2010**, *31*, 1911−1918.

(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) 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.

(4) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, 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.

(5) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(6) 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.

(7) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289−307.

(8) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.

(9) Chang, C.-E.; Gilson, M. K. Free Energy, Entropy, and Induced Fit in Host-Guest Recognition: Calculations with the Second-Generation Mining Minima Algorithm. *J. Am. Chem. Soc.* **2004**, *126*, 13156−13164.

(10) David, L.; Luo, R.; Gilson, M. K. Ligand-receptor docking with the Mining Minima optimizer. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 157−171.

(11) Shoichet, B. K.; L., A. R.; Kuntz, I. D. Ligand solvation in molecular docking. *Proteins* **1999**, *34*, 4−16.

(12) Young, T.; Abel, R.; Kim, B.; Berne, B. J.; Friesner, R. A. Motifs for molecular recognition exploiting hydrophobic enclosure in protein-ligand binding. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104*, 808−813.

(13) Suresh, C. H.; Vargheese, A. M.; Vijayalakshmi, K. P.; Mohan, N.; Koga, N. Role of structural water molecule in HIV protease-inhibitor complexes. *A QM/MM Study* **2008**, *29*, 1840−1849.

(14) Arora, N.; Bashford, D. Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions. *Proteins* **2001**, *43*, 12−27.

(15) Huang, S. Y.; Zou, X. Inclusion of solvation and entropy in the knowledge-based scoring function for protein-ligand interactions. *J. Chem. Inf. Model.* **2010**, *50*, 262−273.

(16) Korb, O.; Stutzle, T.; Exner, T. E. Empirical scoring functions for advanced Protein-Ligand docking with PLANTS. *J. Chem. Inf. Model.* **2009**, *49*, 84−96.

(17) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins.* **2003**, *52*, 609−623.

(18) Zoeteay, V.; Grosdidieray, A.; Cuendeta, M.; Michielin, O. Use of the FACTS solvation model for protein-ligand docking calculations. Application to EADock. *J. Mol. Recog.* **2010**, *23*, 457−461.

(19) Guner, O. F.; Hughes, D. W.; Dumont, L. M. An integrated approach to three-dimensional information management with MACCS-3D. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 408−414.

(20) Gilson, M. K.; Honig, B. The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J. Comput.-Aided Mol. Des* **1991**, *5*, 5−20.

(21) In, Y.; Chai, H. H.; No, K. T. A partition coefficient calculation method with the SFED model. *J. Chem. Inf. Model.* **2005**, *45*, 254−263.

(22) No, K. T.; Kim, S. G.; Cho, K. H.; Scheraga, H. A. Description of hydration free energy density as a function of molecular physical properties. *Biophys. Chem.* **1999**, *78*, 127−145.

(23) No, K. T.; Grant, J. A.; Scheraga, H. A. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 1. Application to neutral molecules as models for polypeptides. *J. Phys. Chem.* **1990**, *94*, 4732−4739.

(24) No, K. T.; Grant, J. A.; Jhon, M. S.; Scheraga, H. A. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. 2. Application to ionic and aromatic molecules as models for polypeptides. *J. Phys. Chem.* **1990**, *94*, 4740−4746.

(25) Park, J. M.; No, K. T.; Jhon, M. S.; Scheraga, H. A. Determination of net atomic charges using a modified partial equalization of orbital electronegativity method. III. Application to halogenated and aromatic molecules. *J. Comput. Chem.* **1993**, *14*, 1482−1490.

(26) Park, J. M.; Kwon, O. Y.; No, K. T.; Jhon, M. S. Determination of Net Atomic Charges Using a Modified Partial Equalization of Orbital Electronegativity Method. IV. Application to Hypervalent Sulfur- and Phosphorus-Containing Molecules. *J. Comput. Chem.* **1995**, *16*, 1011.

(27) Suk, J. E.; No, K. T. Determination of Net Atomic Charges Using a Modified Partial Equalization of Orbital Electronegativity Method: V. Application to Silicon-Containing Organic Molecules and Zeolites. *Bull. Korean Chem. Soc.* **1995**, *16*, 915.

(28) No, K. T.; Cho, K. H.; Jhon, M. S.; Scheraga, H. A. An Empirical Method To Calculate Average Molecular Polarizabilities from the Dependence of Effective Atomic Polarizabilities on Net Atomic Charge. *J. Am. Chem. Soc.* **1993**, *115*, 2005.

(29) Koehl, P.; Delarue, M. Polar and nonpolar atomic environments in the protein core: Implications for folding and binding. *Proteins* **1994**, *20*, 264−278.

(30) Son, S. H.; Han, C. K.; Ahn, S. K.; Yoon, J. H.; No, K. T. Development of three-dimensional descriptors represented by tensors: Free energy of hydration density tensor. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 601−609.

(31) Todeschini, R.; Gramatica, P. 3D-modelling and prediction by WHIM descriptors. Part 5. Theory development and chemical meaning of WHIM descriptors. *Quant. Struct.-Act. Relat.* **1997**, *16*, 113−119.

(32) Grafarend, E.; Kelm, R. Point and interval estimations especially of point errors, in multidimensional least-squares adjustment. *Bulletin Géod.* **1972**, *104*, 165−184.

(33) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.

(34) Meier, R.; Pippel, M.; Brandt, F.; Sippl, W.; Baldauf, C. ParaDockS: A framework for molecular docking with population-based metaheuristics. *J. Chem. Inf. Model.* **2010**, *50*, 879−889.

(35) Smellie, A.; Teig, S. L.; Towbin, P. Poling: Promoting conformational variation. *J. Comput. Chem.* **1995**, *16*, 171−187.

(36) Chang, C. E.; Gilson, M. K. Tork: Conformational Analysis Method for Molecules and Complexes. *J. Comput. Chem.* **2003**, *24*, 1987−1998.

(37) Kyte, J. The basis of the hydrophobic effect. *Biophys. Chem.* **2003**, *100*, 193−203.

(38) Friedrich Ritschl, M. F.; Fiedler, K.; Khler, J. E. H.; Kubias, B.; Meisel, M. An Extension of the Consistent Valence Force Field (CVFF) with the Aim to Simulate the Structures of Vanadium Phosphorus Oxides and the Adsorption of n-Butane and of 1-Butene on their Crystal Planes. *Z. Anorg. Allg. Chem.* **2002**, *628*, 1385−1396.

(39) Nam, K.-Y.; Cho, D. H.; Paek, K.; No, K. T. Investigation of some amino acids conformations at the interface of binary mixture using the solvation free energy density model. *Chem. Phys. Lett.* **2002**, *364*, 267−272.

(40) Cavasotto, C. N.; Abagyan, R. A. Protein Flexibility in Ligand Docking and Virtual Screening to Protein Kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.
(41) Kollman, P. A. A theory of hydrogen bond directionality. *J. Am. Chem. Soc.* **1972**, *94*, 1837−1842.
(42) Glusker, J. P. Structural Aspects of Metal Liganding to Functional Groups in Proteins. In *Advances in Protein Chemistry*; Academic Press: New York, 1991; Vol. *42*, pp 1−76.