Lessons in Molecular Recognition. 2. Assessing and Improving Cross-Docking Accuracy

Jeffrey J. Sutherland,*,† Ravi K. Nandigam,‡ Jon A. Erickson,§ and Michal Vieth§

Discovery Informatics and Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, and School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907

Received July 16, 2007

Docking methods are used to predict the manner in which a ligand binds to a protein receptor. Many studies have assessed the success rate of programs in self-docking tests, whereby a ligand is docked into the protein structure from which it was extracted. Cross-docking, or using a protein structure from a complex containing a different ligand, provides a more realistic assessment of a docking program's ability to reproduce X-ray results. In this work, cross-docking was performed with CDocker, Fred, and Rocs using multiple X-ray structures for eight proteins (two kinases, one nuclear hormone receptor, one serine protease, two metalloproteases, and two phosphodiesterases). While average cross-docking accuracy is not encouraging, it is shown that using the protein structure from the complex that contains the bound ligand most similar to the docked ligand increases docking accuracy for all methods ("similarity selection"). Identifying the most successful protein conformer ("best selection") and similarity selection substantially reduce the difference between self-docking and average cross-docking accuracy. We identify universal predictors of docking accuracy (i.e., showing consistent behavior across most protein−method combinations), and show that models for predicting docking accuracy built using these parameters can be used to select the most appropriate docking method.

INTRODUCTION

Methods for predicting binding modes of small organic molecules to protein receptors are widely used within drug discovery efforts. Docking of a ligand is typically achieved by generating a number of orientations (or poses) of a ligand within the active site, and scoring of poses to identify one or more that closely approximate the bioactive conformation determined by X-ray crystallography. Docking algorithms are also used for identifying putative binders from virtual chemical databases and for estimating the binding affinity of protein−ligand complexes.[1−3]

Understanding binding modes of ligands is important for lead optimization. While X-ray crystallography continues to deliver structures at an increasing rate,[4] it is rarely possible to determine structures for all ligands of interest within a lead optimization effort. Docking methods are widely used in this instance, and establishing their limitations should be a key priority for any researcher employing them.

A large number of studies have compared docking algorithms on a range of protein benchmark sets (refs 5−8 are some recent examples). The accuracy of docking algorithms is known to vary with the properties of protein active sites,[6−8] and trends established using multi-protein benchmark sets may not be observed for the target of interest. Furthermore, most published studies assess methods by docking ligands into the protein from which they were extracted (self-docking). In real applications, one is always performing cross-docking: using a protein structure cocrystallized with a different ligand. As most docking algorithms employ rigid protein structures, there is no explicit allowance for ligand-induced conformational changes. Cross-docking tests reveal significantly decreased accuracy compared to typical self-docking accuracies achieved by leading programs[9,10] (ca. 70% of first-ranked poses within 2 Å RMSD of the X-ray pose for self-docking[2]). To our knowledge, no docking algorithm that allows full protein flexibility has been shown to yield significantly better results in large cross-docking tests.

We have expanded on our previous cross-docking benchmark,[9] including a broader range of protein targets of pharmaceutical interest. Full cross-docking (i.e., docking every ligand to every protein structure of a given target) was performed with CDocker,[11] Fred,[12] and Rocs[13] (the latter is a ligand-based method; "docking" refers to using X-ray conformations of ligands as templates). A number of strategies for the practical improvement of docking accuracy are examined. While overall cross-docking accuracy is not encouraging, our results indicate that significant improvements can be realized, substantially bridging the gap between self- and cross-docking accuracy.

METHODS

Benchmark Set. The cross-docking benchmark set described in this work attempts to represent a range of human proteins of therapeutic interest: the kinases CDK2 and MAPK14 (p38a), the nuclear hormone receptor ESR1 (estrogen receptor alpha), the serine protease F2 (thrombin), the matrix metalloproteases MMP8 and MMP13, and the phosphodiesterases PDE4B and PDE5A. The targets were selected to exemplify the major druggable proteomic classes[14] (where X-ray structures are available), and where a sufficient

* Corresponding author phone: (317) 655-0833; e-mail sutherlandje@lilly.com.
† Discovery Informatics, Eli Lilly and Company.
‡ School of Chemical Engineering.
§ Lilly Research Laboratories, Eli Lilly and Company.

10.1021/ci700253h CCC: $37.00
Published on Web 10/23/2007

Table 1. Average (Standard Deviation) of Molecular Properties for Data Sets[a]

	MW	Rotbond	ClogP
CDK2	362 (90)	4.8 (2.0)	2.5 (2.2)
MAPK14	386 (76)	4.6 (4.3)	4.0 (2.1)
ESR1	341 (111)	3.7 (4.5)	4.4 (2.0)
F2	436 (93)	7.9 (2.9)	2.0 (1.6)
MMPX[a]	410 (108)	8.2 (2.1)	2.5 (1.5)
PDEX[a]	378 (90)	5.1 (2.6)	1.5 (2.7)

[a] MW, molecular weight; Rotbond, number of rotatable bonds; ClogP, calculated logP. [b] In many figures, we combine MMP8 and MMP13 as MMPX and PDE4B and PDE5A as PDEX due to their small size and to minimize figure clutter; each pair of targets is closely related in sequence (62% identity for MMPX; 85% for PDEX).

number of structures exist in the Protein Data Bank (PDB)[15] for evaluating cross-docking accuracy. A threshold of 2.5 Å resolution was selected in order to help ensure use of higher quality data; however, for the metalloproteases and phosphodiesterases, the threshold was increased to 3.5 Å in order to have a better representation and number of co-complex structures. As the range of molecular properties suggests (Table 1), some of these targets are anticipated to be challenging for docking algorithms.

The benchmark set includes multiple complexes of the same protein−ligand pair (e.g., estradiol bound to ESR1) and complexes involving low-affinity, nondruglike ligands (e.g., ATP binding to the kinases). Apoprotein structures (i.e., containing no ligand) are excluded, given notably low docking accuracy reported in previous studies.[9,10] To our knowledge, the benchmark set does not include any covalent complexes (prevalent for F2).

To simplify the evaluation of additional docking programs, we have selected 10 representative structures as an alternative to generating the full matrix (i.e., all proteins × all ligands) of docking results. When more than 10 structures were available for a given protein, the structures were selected using a spread-selection algorithm applied to the $N \times N$ matrix of pairwise active site root-mean-square deviations (RMSD) of the proteins, where N is the number of structures available. This subset of 10 is not discussed further in this article, beyond noting that docking accuracy calculated over all protein−ligand pairs is nearly identical to docking accuracy calculated using the 10 protein × all ligand docking experiment. The PDB codes of the complexes are listed in the Supporting Information; the data sets will be supplied upon request.

Ligand Preparation. Bond orders were added automatically to ligands in PDB format (without CONNECT records) using Maestro,[16] and saved as SDF files. Bond orders were examined and manually fixed where appropriate. In order to remove any bias in docking calculations, SDF files were converted to Smiles strings (detecting chirality from the three-dimensional (3D) arrangement of atoms in the SDF file), and converted back to 3D structures in SDF format using Corina.[17] A standard scheme was applied for assigning formal charges (deprotonated acids, protonated amines, etc.).

Creating Ensembles of Protein Conformations. All docking calculations were automated with our in-house docking package, CDocker.[11] We have implemented the ability to store multiple protein conformations, which allows the work described in this article to be easily automated.

Generation of protein conformation ensembles involves the following operations, starting from a protein−ligand complex for a target: (1) First is the identification of similar sequences in the Protein Data Bank (PDB) using the sequence comparison program Blast[18] (proteins with 95% or higher sequence identity over 3/4 of the input protein sequence). (2) Next is harmonization of atom and residue numbering in the hits compared to the input structure, and mutation/rebuilding of residues either missing or different in the Blast hits, compared to the input structure. Missing side-chain heavy atoms are added using the SCWRL program,[19] followed by their minimization using the CHARMM19 force field[20] with the remainder of the protein fixed. In practice, the input and Blast-retrieved structures rarely differ within the active site region of proteins, so there was no rebuilding/minimization of active site heavy atoms. (3) Last, superposition of Blast-retrieved hits (protein and ligand atoms) onto the input structure is accomplished with the program LSQMAN[21] or by using sequence-matched superposition of atoms (see below). Docking calculations performed with the multiconformer proteins obtained with this algorithm yield docking accuracies within 4% of those obtained using the original PDB structures (results not shown).

Protein Refinement. A protein refinement procedure has been implemented in CDocker, whereby protein atoms in the active site region are minimized in the field of several ligands in their bioactive conformation, extracted from complexes of the same protein. The BLOCK module in CHARMM[20] is used to scale the protein−ligand interactions by the inverse of the number of ligands, and the ligand−ligand interactions to 0. The ligand heavy atoms are fixed, as are protein atoms outside the active site. Mass-weighted harmonic restraints are applied to backbone heavy atoms in the active site region (scale factor of 1). Thus, this procedure adapts side chains to the average field of the other ligands in their bioactive conformation. For metalloenzymes, the ion and atoms within 3 Å are fixed, due to established limitations of force fields for ion−residue interactions.[22]

Template ligands were selected in the following manner: (1) ligands having Daylight similarity[23] less than 0.85 compared to the docked ligand were identified from X-ray structures involving the same protein (using ligands with similarity above 0.85 would provide an artificial test); (2) if more than five ligands were available, five were selected using an in-house program that performs spread selection on Daylight fingerprints.

Docking Algorithms. The docking algorithms CDocker and Fred and the ligand overlay tool Rocs were evaluated on the benchmark set (we distinguish the algorithm CDocker from the package CDocker; the latter is used to automate all calculations, including running of the commercial programs Fred and Rocs; henceforth, all references to CDocker concern the docking algorithm). All three algorithms are described in detail elsewhere, and are only briefly described here.

CDocker performs simulated-annealing molecular dynamics simulations of small molecules in proteins, using the CHARMM program and the CHARMm force field,[24] which includes parameters for small molecules.[11] Strongly repulsive/attractive potentials at small separation of atoms are linearly truncated, allowing energy barriers to be overcome in short

simulations; the softening of potentials is reduced in concert with temperature. Scoring consists of adding the usual intraligand energy terms (bonded and nonbonded contributions) and protein−ligand nonbonded interactions. A commercial version of CDocker (that we have not evaluated) is now available.[24]

Fred and Rocs are commercially available programs[25] that allow convenient command-line operation and fast docking/overlay of molecules (versions and command-line parameters are provided in the Supporting Information). Both programs require as input an ensemble of ligand conformations, which we generate with Omega.[26] Using Fred,[12] each ligand conformation is rigidly optimized using shape and chemical complementarity of protein and ligand, followed by consensus scoring using a number of scoring functions. In this work, the docked results are ranked by the Chemgauss2 scoring function. Using Rocs,[13] each ligand conformation is rigidly optimized using shape and chemical (or "color") complementarity of docked ligand and reference ligand in a specified conformation. With reference to Rocs, the term "docking" indicates that the reference template is the ligand extracted from the complex which would otherwise supply protein coordinates for docking. The usefulness of subsequently minimizing Rocs overlays within the active site is tested in a method identified as "Rocs-Opt". While Rocs overlays are scored using that program's combo score, final scoring for minimized overlays uses the CHARMm energy function.

For all methods, 40 poses were retained for assessing the ability of programs to generate a pose within 2 Å RMSD of the bioactive conformation (i.e., sampling accuracy).

Calculation of Standard Errors. For values with continuous distributions, standard errors are calculated from the relationship

$$\text{standard error} = \text{standard deviation of property}/\sqrt{N}$$

For success/failure properties (i.e., docking within 2.0 Å RMSD or not), standard errors may be calculated using the standard deviation from Bernoulli trials:

$$\text{standard error} = \sqrt{(P(1-P)/N)}$$

where P is the probability of success and N is the number of samples. It is not necessary to estimate this quantity from bootstrapping;[27] both approaches give identical results.

RESULTS

Uncertainty in RMSD Calculations from the Superposition of Ligands. For the evaluation of cross-docking accuracy, it is necessary to place each pair of structures into a common reference frame (i.e., the complex containing the ligand that will be docked, and the complex from which the protein used in docking is selected). This is accomplished by the superposition of complete protein domains or active sites from each structure. Our internal organization of complexes uses a selected template for each protein, superposing other structures onto the template using the program LSQMAN[21] (a structure-based superposition algorithm which uses 3D positions of C-α atoms in the complete protein domain). Alternatives include superposition of C-α, C-α + C-β, or all heavy atoms matched between structures by sequence, using the complete domain or the active site only, etc. To quantify the uncertainty arising from the choice of one method over others, protein structures were superimposed using two approaches and the coordinates of (X-ray) ligands compared. The first approach superposes both protein domains onto a template using LSQMAN, while the second superposes one structure onto the other using all protein heavy atoms within 5 Å of the ligand (i.e., the active site). For each pair of structures, a RMSD value is calculated by comparing the coordinates of ligand heavy atoms from each approach. The average over all data sets is 0.3 Å, and varies substantially by target (Figure 1). The large variation for MAPK14 arises from superposing DFG-out onto DFG-in conformers; it represents the target in this benchmark with the largest ligand-induced fit effects.

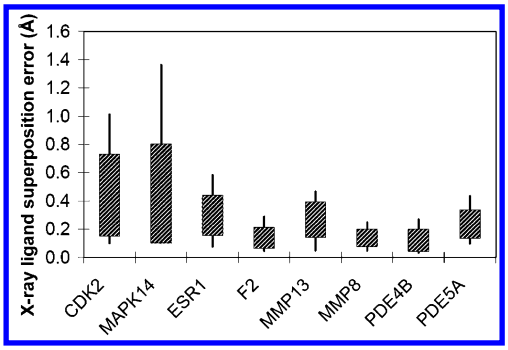

Figure 1. Range of RMSD values obtained by comparing the coordinates of ligand heavy atoms (X-ray conformation), following the superposition of a pair of structures onto a common template using the program LSQMAN,[21] vs superposition of a pair of structures onto each other by minimizing the RMSD of all active site heavy atoms between the two structures. The thin bars indicate 5 and 95 percentiles, while the thick bars indicate the range of values within 1 standard deviation from the average. The number of structure pairs for each target, in parentheses, is CDK2 (6723), MAPK14 (591), ESR1 (737), F2 (5207), MMP8 (279), MMP13 (126), PDE4B (192), and PDE5A (72).

Defining Successfully Docked Structures. A large body of work comparing docking methods (mostly in self-docking tests) uses the ligand heavy atom RMSD threshold of 2.0 Å for defining successful docking. For cross-docking evaluations, some studies use a threshold of 2.5 Å in recognition of uncertainties arising from protein superposition.[28] Because of the well-established precedent of using 2.0 Å, we retain this definition. Given the uncertainty resulting from protein superposition quantified above, and the fact that the binding mode can be properly conveyed even for RMSD values above 2.0 Å (especially for larger ligands),[27] most of the analysis was also carried out using a threshold of 2.5 Å (included in the Supporting Information). For cross-docking, the difference in accuracy using either threshold is mostly confined to 10%. In using the term "docking accuracy" or "docked successfully", we consider only the first-ranked pose.

Comparisons of Self- and Cross-Docking. The programs CDocker, Fred, and Rocs were used to dock ligands from each data set in self-docking (docking a ligand back into the cocrystallized protein; Table 2) and cross-docking (docking a ligand into a protein from a different complex; Table 3) experiments. Discussion of self-docking results excludes the application of Rocs/Rocs-Opt, which are clearly artificial tests (using the X-ray conformation of a ligand to identify the most similar conformation in an ensemble of conformers of the same ligand), but which may serve to

Table 2. Self-Docking Results: Percent of Poses within 2.0 Å of X-ray Pose

	CDK2	MAPK14	ESR1	F2	MMP8	MMP13	PDE4B	PDE5A
no. of structures	82	25	24	72	14	7	13	9
				First-Ranked Pose				
CDocker	41	36	79	33	21	43	38	22
CDocker-P[a]	32	29	67	24	21	43	15	22
Fred	50	40	42	31	7	17	15	11
Fred-P[a]	44	20	29	28	0	33	0	11
Rocs[b]	95	92	67	87	71	100	100	100
Rocs-Opt[b]	95	92	83	72	64	100	100	78
				Lowest-RMS Pose				
CDocker	67	68	83	56	50	71	85	44
CDocker-P[a]	77	63	88	47	64	86	54	78
Fred	77	64	67	51	29	67	62	44
Fred-P[a]	77	52	46	51	21	50	62	33
Rocs[b]	100	100	83	99	86	100	100	100
Rocs-Opt[b]	100	100	83	99	86	100	100	100

[a] CDocker-P and Fred-P employed protein refinement prior to docking (Methods). [b] Self-docking for Rocs/Rocs-Opt uses the X-ray conformation of a ligand to score an Omega conformer ensemble of the same ligand; docking success rate calculated only for cases where the program produced a solution (see Supporting Information Table S.4).

Table 3. Cross-Docking Results: Percent of Poses within 2.0 Å of X-ray Pose

	CDK2	MAPK14	ESR1	F2	MMP8	MMP13	PDE4B	PDE5A
no. of structures[a]	82	25	24	72	14	7	13	9
⟨A-site RMS⟩[b]	1.1	1.3	0.8	0.3	0.4	0.8	0.3	0.6
				First-Ranked Pose				
CDocker	13	9	62	19	37	29	18	17
CDocker-P[c]	24	12	74	18	21	12	12	14
Fred	29	11	24	20	11	8	8	1
Fred-P[c]	34	12	26	16	10	14	12	6
Rocs	16	11	25	16	13	19	28	14
Rocs-Opt	21	12	27	15	15	19	31	13
				Lowest RMSD Pose				
CDocker	40	29	86	38	63	36	53	36
CDocker-P[c]	64	40	90	40	55	36	63	47
Fred	58	30	55	44	20	28	53	38
Fred-P[c]	69	34	46	37	19	28	49	39
Rocs	32	18	30	29	18	25	51	18
Rocs-Opt	34	19	30	31	19	25	53	18

[a] The number of cross-dockings equals [(no. of structures)(no. of structures − 1)]; docking success rate calculated only for cases where the program produced a solution (see Supporting Information Table S.4). [b] Average active site RMSD calculated over all pairs of protein structures (using heavy atoms within 5 Å of ligand); this is a measure of active site flexibility or ligand-induced fit. [c] CDocker-P and Fred-P employed protein refinement prior to docking (Methods).

highlight the artificial nature of self-docking tests. CDocker performs well on ESR1 in self- and cross-docking, whereas Fred is moderately successful for CDK2. For both self- and cross-docking, there is a need to improve on the low sampling accuracy reflected by the limited percentage of ligands having any pose within 2 Å RMSD of the X-ray conformation. Of note, the reasonable self-docking accuracies (using the first-ranked pose) for CDK2 and ESR1 are two of three target−method combinations where a pose within 2 Å is generated in ca. 80% of cases or more. As expected for self-docking, refining a protein using the X-ray structures of other ligands (methods CDocker-P and Fred-P) decreases docking accuracy, compared to using the unmodified protein structure.

From the target dependence of self-docking accuracy, it is clear that docking studies performed for any target should be preceded by validation studies on that target. CDocker and Fred are more accurate (ca. 60% and 40%, respectively) on a subset of PDBBind-2003[29] containing 281 complexes with druglike ligands, a set similar to the typical diverse benchmark used in many self-docking tests. Our previous work[9] indicates similar accuracy for the programs FlexX[30] and Gold.[31]

The results from the cross-docking experiments (Table 3) show an expected drop in accuracy in comparison to self-docking. As others have observed,[6,8] an algorithm that performs better on one protein class may not be the best choice for other classes (e.g., CDocker is superior to Fred on ESR1 and MMPs, while the opposite is true for the kinases CDK2 and MAPK14). On average, ligand-based superposition using Rocs (with or without refinement in the active site) is somewhat inferior to docking. There is no method which emerges as generally superior from this perspective.

For CDocker and Fred, correlations of self- and cross-docking accuracy are moderate (Figure 2A). While low accuracy in self-docking tests can be useful to identify targets where cross-docking will yield poor results, the results underscore the difficulty of drawing quantitative conclusions about docking algorithms using self-docking tests, especially

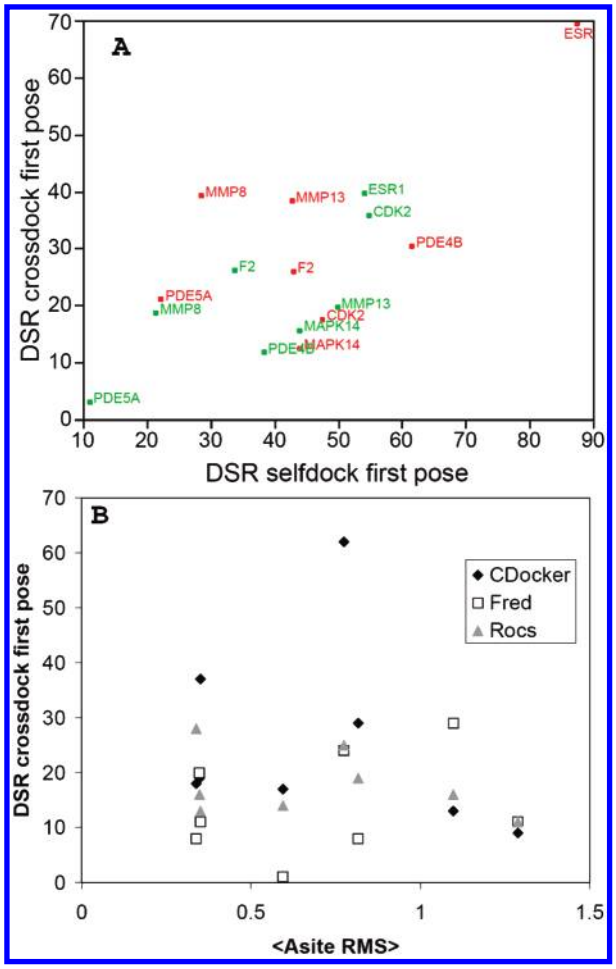

Figure 2. Docking success rate (DSR) at 2.0 Å for self-docking vs cross-docking using CDocker (red; $R^2 = 0.48$) and Fred (green; $R^2 = 0.67$) (A); cross-docking DSR vs average active site RMSD calculated over all pairs of protein structures for each target (B).

those performed on diverse benchmark sets that include few examples of the protein of interest. The average active site RMSD for proteins (a measure of protein flexibility, or ligand-induced fit) of a given target is at best a weak predictor (the rank−order correlation coefficient, i.e., Spearman's rho, vs active site RMSD is 0.57 for CDocker after removing ESR1, 0.29 for Fred, and −0.40 for Rocs) (Figure 2B). The relationship between docking accuracy and protein flexibility appears to be more significant within a protein class, as demonstrated by our previous work on proteases[9] and kinases (unpublished results).

Strategies for Improving Cross-Docking Accuracy. The induced fit effect, absent from self-docking tests, is recognized as a major challenge for cross-docking experiments due to protein structure changes ranging from side-chain movements to rearrangements involving backbone atoms (e.g., DFG-in vs DFG-out in MAPK14 structures). The target dependence of induced fit effects is highlighted by the variation in the average active site RMSD calculated between all pairs of structures for each target (Table 3). We have evaluated a number of strategies for minimizing this effect, within the standard fixed-protein docking protocol.

A simple strategy for implicitly accommodating induced fit consists of refining the coordinates of the protein structure using the X-ray conformations of other ligands complexed to the same protein (Methods). Improvements in docking

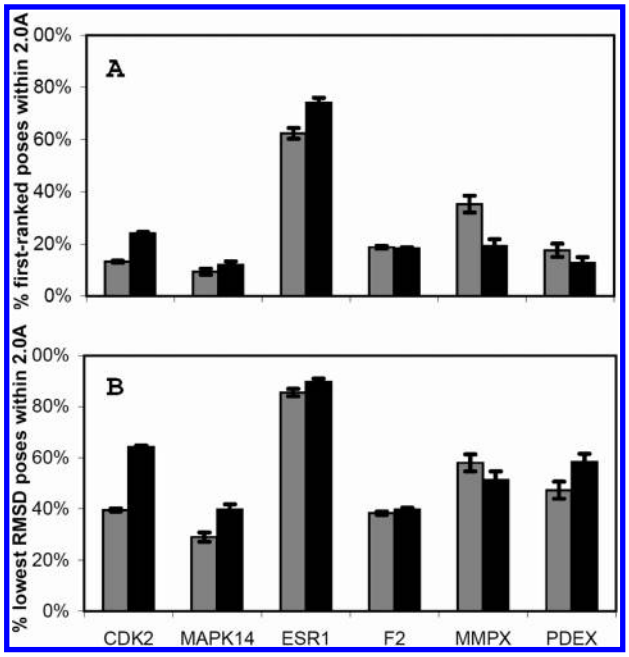

Figure 3. Use of protein refinement with CDocker (black bars), vs no protein refinement (gray bars); best ranked pose (A); lowest RMSD pose (B). Error bars denote the standard error of the average.

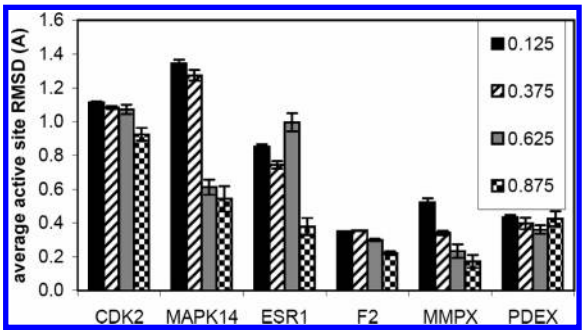

Figure 4. Average ± standard error of active site RMSD vs ligand similarity for pairs of protein−ligand complexes, measured using Daylight fingerprints. Ligand similarities are divided into four bins, with bin centers at 0.125 (black), 0.375 (hashed), 0.625 (gray), and 0.875 (checkered).

accuracy obtained from this scheme are method- and target-dependent, with a larger improvement for CDocker than for Fred. This could be attributed to the force-field dependence of CDocker, where protein refinement significantly improves accuracy for CDK2 and ESR1. Furthermore, the decreased accuracy observed for the first-ranked pose on the MMP and PDE sets is either smaller or reversed when considering the lowest RMSD pose (Figure 3). This suggests that protein refinement increases the ability of CDocker to generate a low RMSD pose, with a corresponding improvement in docking accuracy when scoring correctly ranks the pose.

Another strategy to account for induced fit effects in docking relies on the comparison of ligands. It is reasonable to postulate that highly similar ligands will induce similar protein conformations to a greater extent than dissimilar ligands. This is observed when comparing ligand similarity and active site RMSD, with some variation in the strength of the relationship (Figure 4); the absence of a significant relationship for the PDEs may reflect their higher active site rigidity. There are notable exceptions to this trend, such as 54 pairs of CDK2−ATP complexes having an average

Table 4. Influence of Protein Selection: Percentage of Poses within 2.0 Å of X-ray Pose for Similarity Selection vs Best Selection of Protein Structure

	CDK2	MAPK14	ESR1	F2	MMP8	MMP13	PDE4B	PDE5A
no. of structures	82	25	24	72	14	7	13	9
				Similarity Selection[a]				
CDocker	28	28	71	31	36	29	31	33
Fred	40	20	33	35	7	33	15	0
Rocs	65	32	50	62	29	67	69	33
Rocs-Opt	66	40	63	51	29	50	62	22
				Best Selection[b]				
CDocker	38	21	83	32	54	67	33	38
Fred	44	25	38	33	23	20	25	13
Rocs[c]	41	25	43	39	31	40	58	25
Rocs-Opt[c]	46	25	43	33	31	40	67	25

[a] Similarity selection uses the protein structure from the complex containing the ligand most similar to the docked ligand; complexes having bound ligands with Daylight similarity ≥0.95 compared to the docked ligand are excluded. [b] Best selection uses the same protein structure for all docking, selected as that which maximizes docking accuracy. [c] For Rocs ligand-based superposition, best selection means always using the most successful template ligand for all superpositions; the best protein is indicated for each method in Table S.6 in the Supporting Information.

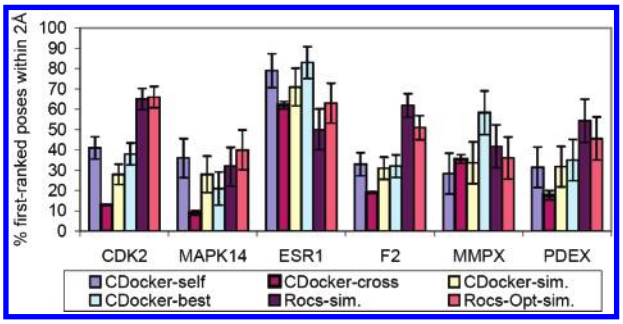

Figure 5. Docking accuracy (% of first-ranked poses within 2 Å) as a function of target. The low self-docking accuracy for MMPX arises from poor MMP8 results (21%). The assessment of accuracy (especially self, best, and similarity selection) is more uncertain for MMPX and PDEX, given the smaller number of complexes available. See text for description of methods.

RMSD greater than 1.5 Å (mostly arising from the flexible glycine rich loop; weighing RMSD calculations by *B*-factors may result in a more meaningful measure).

Motivated by this observation, cross-docking accuracy was assessed for a selection protocol in which the protein structure used for docking is taken from the complex with bound ligand most similar to the docked ligand (excluding complexes with bound ligands having Daylight similarity greater than 0.95 compared to the docked ligand). We refer to this approach as "similarity selection". An alternative approach consists of identifying the most successful protein conformer for docking all ligands, and using it for all cross-docking ("best selection").

Several observations emerge from this analysis (Table 4; Figure 5): (1) As previously reported,[9,10] average cross-docking accuracy is significantly lower than self-docking accuracy. (2) It is encouraging that cross-docking accuracy can be made to approach self-docking accuracy by either identifying the best protein structure and using it for all docking (best selection), or automatically selecting the protein conformation bound to the ligand most similar to the one that will be docked (similarity selection); differences between similarity selection and best selection are not statistically significant. This indicates that improvements in self-docking that yield more reliable predictions for these targets may allow reliable cross-docking. (3) Although the average overlay accuracy of Rocs did not surpass the average cross-docking accuracy, using the X-ray conformation of the ligand most similar to that which will be docked (i.e., similarity selection) generally produces results that are either similar to or much better than those of the best docking strategy considered here. This is an important observation, given that ligand-based overlay methods are used less frequently than docking methods when protein X-ray structures are available. To our surprise, Rocs remains superior/competitive with CDocker and Fred to similarities as low as ca. 0.25, although the threshold exhibits some data set dependence (Figure 6).

An improved result for ligand-based overlay methods vs docking methods may not be observed in cases where there are very few protein structures (and therefore few X-ray ligands) to serve as references. Nonetheless, even for targets such as MMP13 (seven structures) and PDE5A (nine structures), Rocs-based overlays were as successful as docking. Refining Rocs overlays with minimization in the active site increases accuracy for some targets (MAPK14, ESR1), but decreases it for others (F2, MMPs, PDEs); the differences are not statistically significant. An intermediate strategy which constrains Rocs overlays during minimization might yield improvements over Rocs alone.

Factors Influencing Docking Accuracy. The relationship between docking accuracy and ligand properties such as molecular weight (MW), calculated logP (ClogP), rotatable bond counts (Rotbond), hydrogen-bond donors (Hdon), and acceptors (Hacc) have been noted in comparisons of docking algorithms.[7] To determine the impact of molecular properties on docking accuracy, logistic regression was used for modeling the probability of correctly docking (within 2 Å); a one-parameter plus intercept model was generated for each target−method−property combination. In addition to the molecular properties noted above, three metrics comparing the docked ligand to the X-ray complex ligand were computed: Daylight similarity (simDY), the ratio of heavy atoms in the smaller and larger ligands (simNA), and the ratio of heavy atoms in the docked and X-ray ligands (NAratio). The latter two properties are chosen to account for similarities in ligands' volumes, and whether the docked ligand exceeds the volume of the X-ray ligand. In addition, we consider the resolution of the X-ray complex.

For each property, Table 5 indicates the *P*-value of the corresponding model; a *P*-value less than 0.05 indicates a

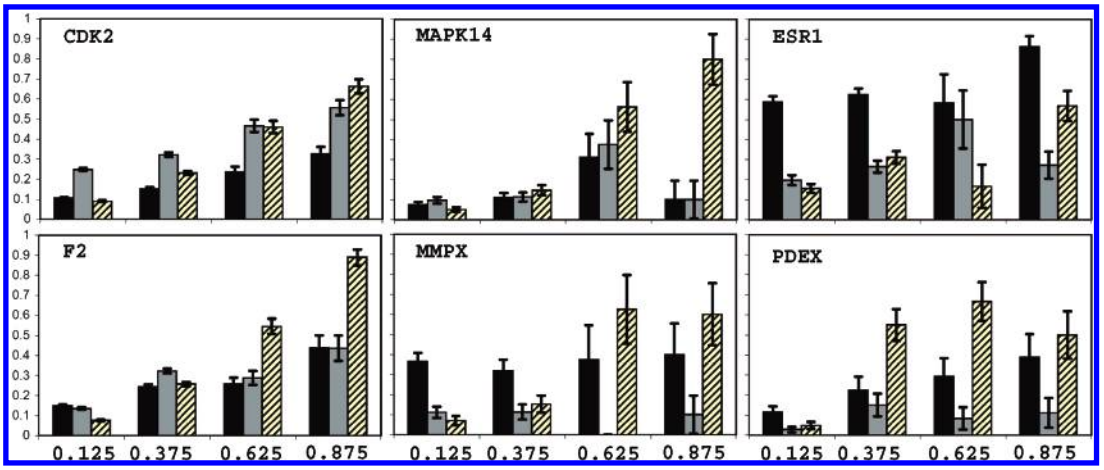

Figure 6. Probability of docking within 2 Å as a function of similarity between the docked ligand and the X-ray complex ligand for CDocker (black), Fred (gray), and Rocs (hashed). Ligand similarities have been binned into ranges 0−0.25, 0.25−0.5, 0.5−0.75 and 0.75−1, with bin centers shown on the chart. There are few pairs of ligands in the higher similarity bins for some targets, reflected by the larger standard error bars. Number of cross-dockings, in parentheses, for bins 1−4 are CDK2 (4113, 2084, 266, 178), MAPK14 (376, 198, 16, 10), ESR1 (276, 220, 12, 44), F2 (3221, 1632, 194, 64), MMPX (134, 72, 8, 10) and PDEX (146, 40, 24,18).

Table 5. Significance of Molecular Properties for Explaining Docking Accuracy by Logistic Regression[a]

	CDK2	MAPK14	F2	ESR1	MMPX	PDEX
			CDocker			
simDY	<0.0001	0.04	<0.0001	<0.0001	0.95	<0.0001
simNA	<0.0001	0.14	<0.0001	<0.0001	0.18	0.06
NAratio	*0.0004*	*0.10*	0.82	*0.22*	0.005	*0.03*
resolution	0.02	0.68	*0.92*	0.68	0.15	*0.74*
MW	*<0.0001*	*0.16*	*0.02*	0.02	<0.0001	*0.03*
ClogP	<0.0001	0.84	*<0.0001*	0.14	<0.0001	*0.02*
Hacc	*<0.0001*	*0.00*	0.92	0.02	*<0.0001*	*0.32*
Hdon	*<0.0001*	*<0.0001*	0.65	0.02	*<0.0001*	0.10
Rotbond	*0.01*	*<0.0001*	0.95	0.62	*<0.0001*	*<0.0001*
			Fred			
simDY	<0.0001	0.1	0.05	<0.0001	*0.62*	0.09
simNA	<0.0001	0.26	<0.0001	<0.0001	0.11	0.22
NAratio	*<0.0001*	*0.0005*	*<0.0001*	*<0.0001*	*0.22*	0.63
resolution	0.11	*0.83*	0.005	<0.0001	0.51	0.34
MW	*<0.0001*	*0.0005*	*<0.0001*	0.08	*0.56*	*0.56*
ClogP	<0.0001	0.71	*<0.0001*	*<0.0001*	0.22	0.07
Hacc	*<0.0001*	*<0.0001*	*<0.0001*	0.001	*0.31*	*0.007*
Hdon	*<0.0001*	*<0.0001*	<0.0001	0.19	*0.05*	*0.79*
Rotbond	*<0.0001*	*<0.0001*	*<0.0001*	0.06	*0.5*	*0.57*
			Rocs			
simDY	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
simNA	<0.0001	<0.0001	<0.0001	<0.0001	0.10	<0.0001
NAratio	*0.0001*	*0.001*	*0.01*	*<0.0001*	*0.03*	*0.17*
resolution	0.35	0.12	0.03	0.45	*0.01*	0.83
MW	*<0.0001*	0.53	<0.0001	*0.02*	*0.02*	*0.01*
ClogP	<0.0001	0.17	*<0.0001*	*0.12*	*0.31*	*0.08*
Hacc	*<0.0001*	*0.28*	<0.0001	*0.16*	0.30	*<0.0001*
Hdon	*<0.0001*	*0.21*	<0.0001	0.02	*0.83*	*0.04*
Rotbond	*<0.0001*	*0.16*	*<0.0001*	*<0.0001*	0.57	0.73

[a] Values in the table are *P*-values indicating the significance of each molecular property in a single parameter (plus intercept) logistic regression model, where the value fit is 1 for successful docking and 0 otherwise; *P*-values ≤0.05 indicate with 95% certainty that the effect of the parameter does not arise by chance. Italicized values indicate a decrease in docking accuracy with increasing value of the parameter. simDY (Daylight similarity of the docked ligand to the X-ray ligand), simNA (ratio of heavy atoms between the smallest of the docked/X-ray ligand and the largest), and Rotbond (number of rotatable bonds) have a significant and consistent impact on docking accuracy for most targets and methods (see text).

significant relationship between a property and docking accuracy. We examine the impact of molecular properties as a function of target, to avoid wrongly attributing to a ligand molecular property an effect that arises from differences in the protein active sites; such differences are best assessed by validating a docking strategy for each target for which it will be used.

Several properties have a consistent behavior across most targets and algorithms. For CDocker, simDY and simNA both generally indicate increasing docking accuracy with increasing values; NAratio, MW, ClogP, Hacc, and Hdon are significant for many targets, but do not have a consistent effect on docking accuracy (i.e., correlated with increased docking accuracy for some targets, decreased for others).

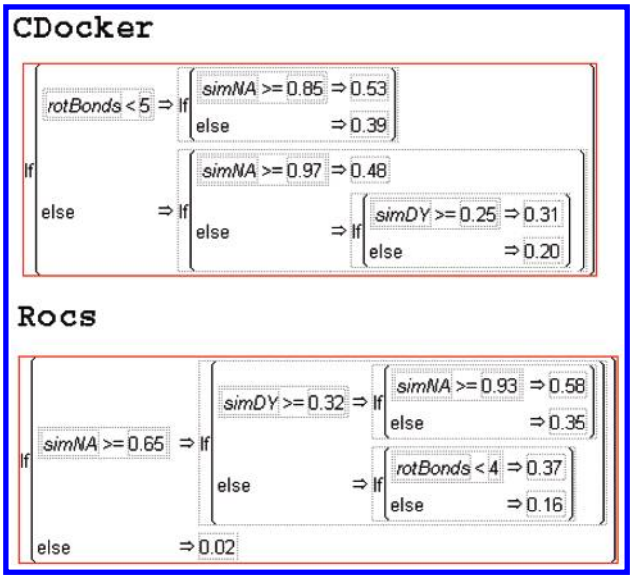

Figure 7. Recursive partitioning models trained on 828 CDK2, ESR1, and F2 ligand−protein pairs. The probability of docking within 2 Å is indicated at each terminal node.

Rotbond is significant in 4/6 cases; increasing rotatable bonds always corresponds to decreasing docking accuracy. For Fred and Rocs, a similar trend emerges: simDY, simNA, NAratio, and Rotbond are consistent predictors of docking accuracy; for Fred, MW is also a consistent predictor. In short, simDY, simNA, Rotbond, and to a lesser extent NAratio appear to be general indicators of the likelihood of success regardless of target and docking algorithm.

The common practice of choosing structures for docking based on their resolution appears to be unfounded; there is no relationship between resolution and docking accuracy observed in this work. This stands in contrast to results of Cole et al.,[27] who have observed increasing self-docking accuracy with improving resolution. The difference may be explained by reasoning that experimental uncertainty in protein coordinates (X-ray) matters less for cross-docking, due to differences in coordinates between the actual structure and that used for docking.

Choosing a Docking Method Automatically. We have implemented within our CDocker package the ability to define multiconformation proteins, and to use other algorithms besides CDocker (presently Fred, Rocs) for docking. This allows the selection of a docking algorithm and the selection of a protein conformation to vary automatically with each ligand docked. For protein conformation selection, a reasonable choice appears to be the similarity selection strategy outlined above. In order to choose an algorithm automatically, 276 protein−ligand pairs (excluding pairs from the same X-ray structure) were selected from the CDK2, ESR1, and F2 sets; 12 ligands from each set were selected randomly and docked to 23 proteins (12 ligands corresponds to half of ESR1 ligands, and these were docked to each of the 23 protein structures taken from different complexes; for CDK2 and F2, the 23 proteins were selected randomly). Two recursive partitioning (or decision tree) models were derived from the training data using the statistical package JMP,[32] and were used for predicting whether the ligand was docked within 2 Å (class 1) or not (class 0) by CDocker and Rocs. The variables used include simDY, simNA, and Rotbond (Figure 7).

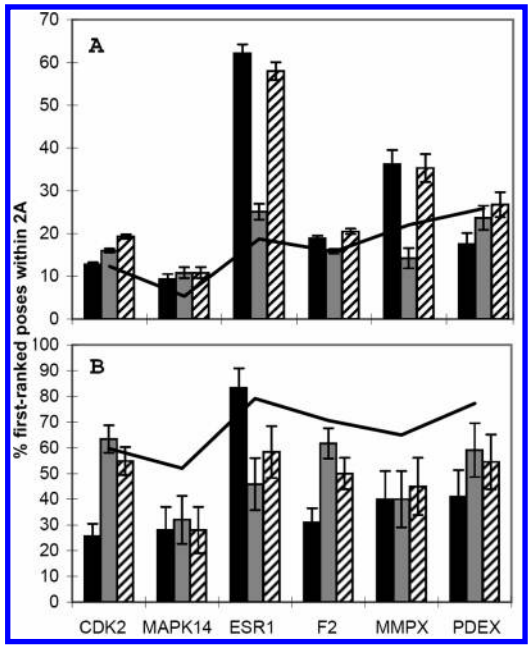

Figure 8. Comparison of automatic selection of a docking algorithm based on recursive partitioning models (hashed bars) to CDocker (black bars) and Rocs (gray bars). Panel A uses all protein−ligand pairs for a target (excluding pairs from the same X-ray structure), and panel B uses the similarity selection strategy: choosing the protein−ligand complex that contains the most similar ligand, compared to the docked ligand. Values for CDocker and Rocs differ from those in Figure 5 due to the exclusion of training ligands (see text). The line series indicates the percent of automatic method selections that uses Rocs.

The remaining 10 960 protein−ligand pairs involving nontraining ligands were scored with the CDocker and Rocs models, and the algorithm which gives the highest score for a particular pair selected. This strategy either modestly surpasses the best of CDocker and Rocs, or modestly falls short, whether considering the average over all protein−ligand pairs or the similarity selection strategy (see above); differences are statistically significant only for average CDK2 (automatic better) and similarity selection ESR1 (CDocker better) (Figure 8). For similarity selection, it is possible that a docking algorithm that compares more favorably with Rocs would allow this strategy to be more successful.

A further refinement of this strategy would quantify the relationship between model scores and the probability of docking within 2 Å, allowing for confidence of predictions to be reported with results. As with the model-building reported above, it may be preferable to use a more diverse set of proteins, to avoid training models that simply identify classes of protein more amenable to docking rather than identifying general trends.

CONCLUSIONS

We have examined the performance of several docking strategies for binding mode prediction of small organic molecules. The focus has been placed on cross-docking accuracy, since this is what is asked of a docking program applied in a drug discovery setting. Barring a small number of exceptions, average cross-docking accuracy was low for the methods examined.

A small number of general indicators of docking accuracy were identified. One of these, the Daylight similarity of the

docked ligand compared to the ligand from the X-ray complex, allows for the selection of protein structure such that the docked ligand is maximally similar to the X-ray complex ligand (similarity selection). This approach yields results approaching self-docking accuracy, although identifying the most successful protein structure and using it for all docking is moderately more accurate.

When applying the similarity selection strategy, we have found that ligand-based overlays obtained with Rocs are generally more accurate than either of the docking algorithms examined. This appears to contradict the typical preference given to docking algorithms when X-ray structures are available for docking.

The need to include explicit protein flexibility is widely accepted, and its inclusion in docking is an area of active development.[3,28,33−37] While small test sets (typically <25 cross-dockings) and anecdotal successes are promising, it will be necessary to evaluate such programs in large tests similar to those described in this article in order to assess whether they outperform some of the simple "pseudoflexibility" approaches adapted to rigid docking discussed in this article.

Obtaining accurate binding modes from docking is important at the lead optimization stage of drug development. We hold the view that docking accuracy below 50% is of little value for binding mode prediction. Some of the strategies outlined in this article allow for the identification of ligands likely to be docked successfully, even if the average ligand is not. With X-ray determination of protein−ligand complexes costing tens of thousands of dollars per structure, docking algorithms that are accurate for binding mode prediction would be highly desirable, even if they consume many hours of CPU time per ligand.

ACKNOWLEDGMENT

We thank Rick Higgs for helpful discussions on standard error calculations, and members of the Global Computational Chemistry Group for suggestions and review of the manuscript.

Supporting Information Available: Tables indicating parameters used for Rocs and Fred, docking accuracy at 2.5 Å, the percentage of protein−ligand pairs for which docking programs produced a solution, and the most successful protein for cross-docking. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3* (11), 935−949.

(2) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49* (20), 5851−5855.

(3) Lengauer, T.; Lemmen, C.; Rarey, M.; Zimmermann, M. Novel technologies for virtual screening. *Drug Discovery Today* **2004**, *9* (1), 27−34.

(4) Scapin, G. Structural biology and drug discovery. *Curr. Pharm. Des.* **2006**, *12* (17), 2087−2097.

(5) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On evaluating molecular-docking methods for pose prediction and enrichment factors. *J. Chem. Inf. Model.* **2006**, *46* (1), 401−415.

(6) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47* (3), 558−565.

(7) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56* (2), 235−249.

(8) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49* (20), 5912−5931.

(9) Erickson, J. A.; Jalaie, M.; Robertson, D. H.; Lewis, R. A.; Vieth, M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. *J. Med. Chem.* **2004**, *47* (1), 45−55.

(10) Murray, C. W.; Baxter, C. A.; Frenkel, A. D. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J. Comput.-Aided Mol. Des.* **1999**, *13* (6), 547−562.

(11) Wu, G.; Robertson, D. H.; Brooks, C. L., III; Vieth, M. Detailed analysis of grid-based molecular docking: A case study of CDOCKER−A CHARMm-based MD docking algorithm. *J. Comput. Chem.* **2003**, *24* (13), 1549−1562.

(12) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68* (1), 76−90.

(13) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48* (5), 1489−1495.

(14) Swindells, M. B.; Overington, J. P. Prioritizing the proteome: identifying pharmaceutically relevant targets. *Drug Discovery Today* **2002**, *7* (9), 516−521.

(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(16) *Maestro*, version 7.5; Schrodinger, Inc.: New York, 2006.

(17) *Corina*, version 3.2; Molecular Networks GmbH: Erlangen, Germany, 2005.

(18) Altschul, S. F.; Madden, T. L.; Schaffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25* (17), 3389−3402.

(19) Canutescu, A. A.; Shelenkov, A. A.; Dunbrack, R. L., Jr. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* **2003**, *12* (9), 2001−2014.

(20) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(21) *LSQMAN*, version 9.6.2; Uppsala Software Factory: http://xray.bmc.uu.se/usf/ (accessed Aug 31, 2007).

(22) Lamoureux, G.; Roux, B. Absolute Hydration Free Energy Scale for Alkali and Halide Ions Established from Simulations with a Polarizable Force Field. *J. Phys. Chem. B* **2006**, *110* (7), 3308−3322.

(23) *Daylight*, version 4.8.1; Daylight Chemical Information Systems: Aliso Viejo, CA, 2003.

(24) *Discovery Studio*; Accelrys: San Diego, CA, 2007.

(25) *OpenEye applications*, version 2.1; OpenEye Scientific Software, Inc.: Santa Fe, NM, 2005.

(26) Bostrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J. Mol. Graphics Modell.* **2003**, *21* (5), 449−62.

(27) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R., Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, *60* (3), 325−32.

(28) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337* (1), 209−25.

(29) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48* (12), 4111−4119.

(30) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261* (3), 470−489.

(31) 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* (3), 727−748.

(32) *JMP*, version 5.1.1; SAS Institute: Cary, NC, 2004.

(33) 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* (7), 1739−1749.

(34) Kamper, A.; Apostolakis, J.; Rarey, M.; Marian, C. M.; Lengauer, T. Fully automated flexible docking of ligands into flexible synthetic receptors using forward and inverse docking strategies. *J. Chem. Inf. Model.* **2006**, *46* (2), 903−911.

(35) Sherman, W.; Day, T.; Jacobson, M. P.; Friesner, R. A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J. Med. Chem.* **2006**, *49* (2), 534−553.

(36) Mizutani, M. Y.; Takamatsu, Y.; Ichinose, T.; Nakamura, K.; Itai, A. Effective handling of induced-fit motion in flexible docking. *Proteins* **2006**, *63* (4), 878−891.

(37) Meiler, J.; Baker, D., ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility. *Proteins* **2006**, *65* (3), 538−548.

CI700253H