Evaluations of Molecular Docking Programs for Virtual Screening

Kenji Onodera,*[,†] Kazuhito Satou, and Hiroshi Hirota

RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan

Received January 29, 2007

Structure-based virtual screening is carried out using molecular docking programs. A number of such docking programs are currently available, and the selection of docking program is difficult without knowing the characteristics or performance of each program. In this study, the screening performances of three molecular docking programs, DOCK, AutoDock, and GOLD, were evaluated with 116 target proteins. The screening performances were validated using two novel standards, along with a traditional enrichment rate measurement. For the evaluations, each docking run was repeated 1000 times with three initial conformations of a ligand. While each docking program has some merit over the other docking programs in some aspects, DOCK showed an unexpectedly better screening performance in the enrichment rates. Finally, we made several recommendations based on the evaluation results to enhance the screening performances of the docking programs.

INTRODUCTION

Virtual screening by molecular docking is increasingly important in drug discovery.[1,2] Such virtual screening is usually carried out in three steps. First, the molecular docking program predicts the most possible complex structures for the complex of a target protein and compound structures from screening libraries. Second, the complexes are scored by the binding energy strengths of the complexes. Finally, ranks are formed according to the docking scores, and the top ranked compounds (Hits) are selected from the virtual screening results. The molecular docking program plays a major role in the virtual screening for docking and scoring, and thus, the docking program assumes the primary responsibility for the goodness of scoring in virtual screening.

Currently, many molecular docking programs are available including DOCK,[3] AutoDock,[4] GOLD,[5] FlexX,[6] Glide,[7] ICM,[8] and Surflex.[9] Newly developed docking programs are published every year. Since many programs exist, it is impossible to try all of them to find a docking program suitable for users' purposes. Thus, evaluations of docking programs are important for the selection of the docking program, and many comparative analyses for docking pose prediction have been reported so far.[10] Evaluations or comparisons of the docking programs for docking pose prediction can be done by visual inspection or measurement of the root-mean-square distances (rmsd) of the ligand heavy atoms between the predicted and the reference coordinates from a crystal structure of the complex.

Since the correct prediction of ligand−protein complexes is usually thought to be the key for correct scoring, we believe that such comparisons are useful for the selection of a docking program for virtual screening. However, the performances on the docking pose prediction do not directly reflect the screening performances in virtual screening.

Several evaluation results for screening performances have been reported. Usually, one or several protein targets were used for such evaluations.[10] Even though some reports used a number of protein targets with several docking programs for docking pose predictions, they used only selected targets and sometimes only selected docking programs to evaluate the virtual screening performances.[2,12,13] The performance of docking programs is known to vary depending on the target proteins.[14,15] Evaluation results with a few target proteins cannot predict the general performance of a docking program. Evaluations with many target proteins are rarely reported, probably because it is more time-consuming to screen a number of compounds from a screening library than to perform docking pose prediction with one or a few ligands.

An evaluation is usually based on the enrichment rates of ligands from nonligands in virtual screening. The performances of the docking pose prediction and virtual screening are probably different, and therefore it is worthwhile to evaluate the docking programs for virtual screening with many target proteins, to determine the general performances of the docking programs.

Currently, the enrichment rate as well as the process time is primarily used to evaluate the screening performance in virtual screening, and good docking programs for virtual screening generate high enrichment in a short process time. However, a docking program does not produce exactly the same results from every docking run, even though the same input files and the same parameter settings are used. This happens due to the use of a random number in the docking algorithm. No report has considered the effect of the random number. If the same results cannot be produced, then is there any impact from the random seed in virtual screening? The enrichment rate may not be the only way to evaluate the screening performances of the docking programs in virtual screening, and the dispersion of the docking results caused by the random number may also be important. Thus, we have evaluated the performances of docking programs in virtual

* Corresponding author phone: +81-3-5452-6238; fax: +81-3-5452-6274; e-mail: onodera.kenji@gmail.com.

[†] Current address: Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro, Tokyo, 153-8505 Japan.

10.1021/ci7000378 CCC: $37.00
Published on Web 06/28/2007

screening with consideration of the usage of random numbers in the algorithms, by repeated prediction of the docking poses and docking scores in the same ligand−protein set, with the same parameter settings of the programs.

When evaluations of docking programs are performed, it is fair to use the best parameter settings in each docking program for each performance evaluation. DOCK, AutoDock, and GOLD each possess 40 or more adjustable parameters. Moreover, some docking programs have multiple docking algorithms (e.g., DOCK) or multiple scoring functions (e.g., GOLD). We believe that there are no ultimate settings that are good for all protein targets or ligands. Otherwise, the developers of the docking program should find the ultimate settings if they really existed. Since there are no ultimate settings, the best parameter settings must be found for each target and for each program, to obtain the best docking results. In addition to the parameter settings, the protein and ligand structure files may be modified to achieve the best results in docking pose prediction.[2,16,17] For example, docking programs have difficulty in predicting the positions of water molecules, and thus, such water molecules are usually removed from the input file. However, some water molecules are essential to form hydrogen bonds, and such molecules are added to the input files specifically. Also, some ligands and proteins need proper protonations for better predictions, according to the molecular interaction states of the complex. The sizes and locations of active sites also affect the predictions. All of these modifications of the input files can be done for known ligand−protein complexes, whereas they can be done only partially for unknown compounds, such as those from screening libraries. It can be more difficult if the target is an orphan receptor, since there are no known ligands for the target, and we cannot use the optimization tactic with the correct ligands to find the best parameter settings and the best modifications. The combinations of parameters, scoring function, docking algorithms, and input file modifications are unlimited, and thus the adjustment becomes a quite difficult task. An expert with one docking program may find better parameter settings for docking program but may not be able to do so for another docking program.[10] Therefore, we used the default settings in each docking program during the evaluations.

Since the computer resource is limited, we need to trade off between the number of docking programs being tested and the thoroughness of the evaluation. In this study, we selected three docking programs: two of the most popular docking programs (DOCK and AutoDock) free for academic users and popular commercial software (GOLD), which has been evaluated for docking pose predictions with many target proteins. The evaluations for virtual screening were performed from various viewpoints with 116 target proteins. Since we did not optimize the parameter settings for the docking programs and the target proteins, as described above, the results might not represent the best performances of the docking programs in virtual screening. We also need to be aware that the evaluation results may differ with different test sets of proteins or ligands, different criteria of a correct docking, or different conditions for performing a docking study.

Table 1. Ligand−Protein Complexes

116 Complexes Analyzed						
1aaq	1abe	1acj	1ack	1acm	1aco	1aec
1aha	1apt	1ase	1atl	1azm	1baf	1blh
1bma	1byb	1cbs	1cbx	1cil	1com	1coy
1cps	1dbb	1dbj	1did	1die	1dwd	1eap
1eed	1epb	1eta	1etr	1fen	1fkg	1fki
1frp	1ghb	1glp	1glq	1hdc	1hdy	1hef
1hfc	1hri	1hsl	1icn	1ida	1igj	1imb
1ive	1lah	1lcp	1lic	1lmo	1lna	1lpm
1lst	1mcr	1mdr	1mmq	1mrg	1mrk	1mup
1nco	1nis	1pbd	1pha	1phd	1rds	1rne
1rob	1slt	1srj	1stp	1tdb	1tka	1tng
1tni	1tnl	1tpp	1tyl	1ukz	1ulb	1wap
1xie	2ada	2ak3	2cgr	2cht	2cmd	2ctc
2dbl	2gbp	2lgs	2mcp	2mth	2phh	2pk4
2plv	2r07	2sim	2yhx	3aah	3cla	3cpa
3gch	3hvt	3ptb	3tpi	4dfr	4phv	5p2p
6abp	6rnt	7tim	8gch			
Failed on Corina Process						
1tph	1trk	1xid	4fab	6rsa		
Failed on DOCK						
1bbp	1ctr	1hyt	1phg	1poc	1snc	1tmn
Failed on GOLD						
1cdg	1dr1	1ldm	4cts	4est		

MATERIALS AND METHODS

All docking experiments were performed using up to 1024 CPUs with Linux clusters (RIKEN Super Combined Cluster). Each node was equipped with 2 CPUs of 3.06 GHz Xeon with 2 or 4 GB of main memory.

Selection of Testing Complexes. A total of 133 ligand−protein complexes were obtained from the Protein Data Bank (PDB) to test the docking programs (Table 1). These complexes were previously selected for another attempt to evaluate a docking program, GOLD.[16] According to the authors of the GOLD evaluation, highly diverse complexes were selected, on the basis of pharmacological interest. In the ligands of the complexes, the number of heavy atoms ranged from 6 to 55, and the number of rotatable bonds was 0−30. These complexes are popular and have been used for many evaluation attempts.[17] It is unavoidable that the complexes have been used to optimize the docking programs, due to their popularity, and it is difficult to know whether the docking programs have been especially optimized for the complexes. Thus, we decided to use the set of complexes to test the docking programs.

Many published evaluations of docking programs used several known ligands for the target proteins.[12,15,18] Since various numbers of known ligands were used, some target proteins can be evaluated with more known ligands. In such cases, the target proteins are not weighted equally when evaluation results are simply averaged. In this study, only the ligand from each complex was assigned as the correct ligand, to maintain the diversity of the target protein data set and the statistical weight.

Preparation of Protein and Ligand Files. The PDB files of the complexes were simply separated into ligand files and protein files. Both the ligands and the proteins were processed to add hydrogen atoms and partial charges (Gasteiger and Marsili) by SYBYL.[19] The evaluations of the docking programs were performed with three initial conformations of a ligand. One of the three input ligand structures was that obtained from the PDB complexes, where the ligand atom coordinates and geometries were exactly those from the crystal structures of the complexes (referred to as PDB). The others were one built with the 3D structure generator, Corina (Molecular Networks GmbH, Erlangen, Germany), from the original ligand structure (referred to as Corina), and one optimized to have an energy minimized structure, using SYBYL with the Corina generated ligand files (referred to as MINI). Corina generates a possible 3D structure of a compound from many known X-ray crystal structure analysis results. When processed with Corina, the generated structures are centered and lack the original conformations. The SYBYL generated structures use the Corina generated files, as described above. Thus, both the Corina and SYBYL generated input ligand structures possess neither the correct positions nor correct conformations of the reference ligands

Screening Library (Decoys). The screening accuracy and the docking scores comparisons were tested with the NCI diversity set[20] (total 1990 entries). The NCI diversity set was selected by the National Cancer Institute (NCI) and was prepared with acceptable conformations of compounds. Here, the compounds within the NCI diversity set were used as decoys. The binding affinities of the compounds are unknown, and most should be nonbinders. However, we need to remind ourselves that there may be unexpected binders in the NCI diversity set. In this case, such compounds in the NCI diversity set may be ranked higher than the correct ligands, even though the docking programs perform perfectly well.

The properties of the compounds were similar between the compounds in the NCI diversity set and the ligands from the complexes. The molecular weight (MW) ranged between 90 and 1297 (av 312), and most (95.9%) fulfilled Lipinski's Rule of Five among the compounds in the NCI diversity set, while the MW ranged from 114 to 776 (av 306), and 91.4% were Rule of Five compatible in the ligands from the complexes selected.

Binding Site Definition. Binding sites are defined to be similar sizes among docking programs, while each docking program requires a binding pocket with a different shape. Basically, the defined binding pockets are areas that cover 5 Å from all heavy atoms of ligands in the complexes. A binding site definition was created for each target protein and each docking program, and the same binding site definition was used for the three input ligands described and for screening the NCI diversity set for the same target protein and the same docking program.

For DOCK, spheres were generated using SPHGEN,[22] and all spheres within 5 Å from all heavy atoms of a ligand were selected as a binding site. For AutoDock, a rectangular solid was assigned, which covered all heavy atoms of a ligand with at least a 5 Å cushion. For GOLD, a minimum size sphere to cover all heavy atoms of a ligand with a 5 Å cushion was assigned as a binding site.

Molecular Docking Programs. The following programs and settings were used for the study.

AutoDock 3.05 is based on a hybrid of a genetic algorithm and an adaptive local search method, named the Lamarkian genetic algorithm.[4] The genetic algorithm mimics the process of natural selection to find solutions. In a docking study application, the ligand conformation represents chromosomes, including ligand translational, orientational, and conformational degrees of freedom, and the individual represents candidate solutions. A random population of individuals is selected based on their fitness, with ongoing mutation and crossover of individual chromosomes. The generation cycle is followed by an adaptive local search in AutoDock. The solutions are given two energy scores, the final docking energy (FDE) and the estimated final energy of binding (EFEB), which include van der Waals and electrostatic interactions, loss of entropy in the ligand, and the number of hydrogen bonds.

DOCK 4.0.1 is based on an incremental construction and random search algorithm.[3] In the incremental construction algorithm, a rigid portion of the ligand (referred to as the anchor in the program) is superimposed onto a user-defined binding site. Then, the ligand conformation is built by adding the remaining parts to the anchor, step by step. The scoring function here is based on the intermolecular terms of the molecular mechanics force field. The energy scoring function used here includes van der Waals and electrostatic interactions. The random search algorithm was not used in this study.

Since DOCK does not have a default setting defined for docking or screening, the parameter settings in a demo supplied with the program were used for the study and were partially modified. The demo settings use the Energy scoring function and the Anchor search method. The modifications made to the demo settings were to set the 'multiple_ligand' and 'multiple_anchor' functions on and to increase the 'configurations_per_cycle' from 10 to 30, the 'maximum_orientations' from 200 to 1000, the 'energy_cutoff_distance' from 4 to 5, and 'maximum_iterations' from 100 to 200.

GOLD 3.0[21] is based on a genetic algorithm as described for AutoDock.[5] GOLD has two scoring functions and several predefined parameter settings. In this study, three combinations of scoring functions and parameter settings were used: Library screening settings with GOLDScore, Standard default settings with GOLDScore, and Standard default settings with ChemScore.

Docking Processes and Screening. All target proteins were tested with three initial ligand conformations with all five program settings listed above. The same docking experiments were repeated 1000 times each. Only the best score solution was selected in each docking run and each initial ligand conformation, although the docking programs proposed several solutions in each docking run. Thus, 1000 docking conformations and scores were obtained from 1000 docking runs in each testing combination, and over 2 million docking runs were performed, using three docking programs with five parameter settings and three initial ligand conformations for the ligand−protein complexes.

Similarly, dockings were performed with five program settings for each of the target proteins against the NCI diversity set, to obtain docking scores for rankings.

rmsd and Rank. The rmsd value was calculated to indicate the difference between the ligand coordinates from the crystal and the conformations predicted by a docking program. When the rmsd value is zero, it means that the solution obtained by the docking program is exactly the same as the ligand coordinates in the crystal structure. In this study, the prediction was classified as successful when the rmsd was 2 Å or smaller.

Table 2. Total Sizes of Defined Docking Sites and Run Times for 116 Ligand−Protein Complexes

	av size (Å^3)	av time (sec)	max. time (sec)	total time (h)
AutoDock	4913	99.9	594.8	9655
DOCK	4913	8.7	160.7	837
GOLD ChemScoreSTD	4973	127.7	516.4	12341
GOLD GOLDScoreLib	4973	9.0	43.6	866
GOLD GOLDScoreSTD	4973	321.2	1513.4	31046

The docking scores from the various docking programs cannot be compared directly, since the base units and the concepts of docking scores differ among the programs. Thus, the docking scores from the predicted ligand−protein complexes were compared to the docking scores for the target protein and the compounds from the NCI diversity set, and the docking scores from the ligand−protein complexes were converted to a ranking in the NCI diversity set by comparing the scores.

RESULTS AND DISCUSSION

A total of 133 ligand−protein complexes were initially processed (Table 1). Among them, five complexes were failed by Corina for generating new ligand conformations. Five and seven complexes were failed during the docking calculations by GOLD and DOCK, respectively. Since the ligand and protein coordinates from the PDB files were used without any modification of the input files, in total, 17 complexes were not used to generate docking results. Finally, 116 out of 133 complexes were processed by all three docking programs. While GOLD and DOCK failed to process some complexes, AutoDock could produce results for all complexes.

Sizes of Binding Sites and Run Times. Although it is fair to compare the performances of docking programs in the same run time,[10] adjustments of the docking program parameters generate new issues. One is that a longer run time is not always necessary to produce better results. Another is that careful adjustments of the parameter settings for docking programs are required to achieve the best results in the given process time. This makes the adjustment process an elaborate task and creates unfairness, since a user can more effectively adjust a familiar program. Thus, we tried not to adjust the parameter settings to equalize the run time, and basically used their default settings, as described in the ‘Materials and Methods’ section.

Docking programs require users to define the docking sites in program-specific shapes. The requirements are a rectangular solid, a set of spheres, and a sphere for AutoDock, DOCK, and GOLD, respectively. Thus, the smallest rectangular solid and sphere to cover a binding site were defined for AutoDock and GOLD, respectively, and a set of spheres within the rectangular solid was defined as the binding site for DOCK.

Three initial ligand conformations were tested 1000 times each for 116 ligand−protein complexes, and thus, a total of 348 000 docking runs were performed with each docking program. Table 2 shows the average sizes of the binding sites in the 116 complexes and the total run times for 348 000 docking runs.

The average binding sites size was similar between DOCK, AutoDock, and GOLD (4913, 4913, and 4973 Å,^{3} respectively) although the shapes of binding sites were different between docking programs. Here, the size of the binding site in DOCK was measured as a minimum rectangle of the surrounding spheres used as the binding site since the spheres can be overlapped and the empty spaces between the spheres can be used for docking. The minimum rectangle surrounding the spheres for DOCK was the same as the binding site for AutoDock. Since different binding site shapes are required by each docking program, the sizes of the binding sites differ. The shape and size differences of the binding sites might interfere with the docking results. However, as mentioned above, adjusting the binding sites is a very difficult task. Since the shape required differs among the docking programs here, we believe that adjustments for fairness are almost impossible in the binding site definition. Thus, no adjustment for the binding site size was done here. Since the concept of minimum binding sites was defined to be as fair as possible in the ‘Materials and Methods’ section, we believe that this method is fair enough to compare the docking results from different docking programs.

GOLD in ‘Library Screening Settings’ with GOLDScore (GOLDScoreLib) and DOCK were the fastest and took about the same amount of time to process the 116 ligand−protein complexes, followed by AutoDock, and GOLD in ‘Standard Default Settings’ with ChemScore (ChemScoreSTD) and GOLDScore (GOLDScoreSTD). ChemScoreSTD was 2.5 times faster than GOLDScoreSTD, and GOLDScoreLib was 36 times faster than GOLDScoreSTD. Again, comparisons of the docking programs under the same conditions are difficult. Although GOLDScoreLib and DOCK took about the same amount of computation time, on average, their maximum computation times were different, and the maximum time in DOCK was four times longer than that in GOLD. Thus, the computation times were similar between those two programs and settings, but the individual ligand−protein complexes were not always processed in the same amount of time.

Performance of Docking Pose Prediction. The rmsd values between the crystal and predicted structures are widely used to confirm whether the correct docking position was obtained by the docking simulation. Usually, an rmsd of 2 Å or smaller is considered as the correct docking position, probably because the resolution in an X-ray crystal structure analysis is often 2 Å, and higher precision than the resolution of the analysis is meaningless. Although the rmsd evaluates all of the atoms in a ligand with the same weight, without considering those involved in molecular interactions and flexible parts in the ligand, visual inspection of the predicted complexes has no standard measure and thus has a higher risk of arbitrary evaluation. Indeed, the visual inspection done by the developers of GOLD classified all ligands with rmsd values less than 2.5 Å as having either a ‘Good’ or ‘Close’ docking pose, which counted as success in their docking runs.[16] Thus, the rmsd is not the best but only an acceptable method for comparing docking programs when a number of ligand−protein complexes are processed.

The docking simulation was repeatedly performed 1000 times for the 116 ligand−protein complexes, with three initial ligand conformations. The percentages of successful prediction (rmsd 2 Å or better) are listed in Table 3. Each category represents the results for 116 000 docking runs with the same docking program, parameter settings, and preparation method

Table 3. Percentages of Ligand−Protein Complexes with rmsd 2 Å or Better

	Corina (%)	MINI (%)	PDB (%)
AutoDock	26.2	27.0	45.4
DOCK	21.6	20.6	28.5
GOLD ChemScoreSTD	45.5	45.3	56.0
GOLD GOLDScoreLib	44.1	44.9	51.4
GOLD GOLDScoreSTD	45.2	46.7	62.5

Table 4. Percentages of Docking Runs That Resulted in Certain Ranges from the Best in Each 1000 Docking Runs

	rmsd 2 Å or better (%)	score 1 point or better (%)	score 10 points or better (%)	rank 100th or better (%)
AutoDock EFEB	67.5	69.1	100	60.9
AutoDock FDE		71.9	99.8	71.5
DOCK	61.3	21.5	83.6	75.1
GOLD ChemScoreSTD	75.7	45.4	94.3	61.6
GOLD GOLDScoreLib	61.0	5.4	56.0	49.7
GOLD GOLDScoreSTD	74.5	40.6	89.7	82.8

of initial ligand conformations for the 116 ligand−protein complexes. The table shows that GOLD was the best in docking pose prediction followed by AutoDock and DOCK, as expected from several published results.[2,11,12] Although better results with GOLD were obtained from parameter settings with longer process times (GOLDScoreSTD was the longest processing time and GOLDScoreLib was the shortest in GOLD), the performances among the GOLD parameter settings were similar.

In view of the initial ligand conformations, slightly better results were obtained when the correct poses and geometry (refereed to as PDB in the table) were provided to the docking programs. However, the ‘PDB’ ligands should produce far better results as compared with the other two initial ligand conformations because the answers were assigned to the ‘PDB’ ligand in each docking run. Thus, all of the docking programs do not simply use the initial ligand conformation supplied for predictions or they could not recognize the correct coordinates of the ‘PDB’ ligand as the answer.

Corina and SYBYL Minimize were used to generate two additional initial ligand conformations for the docking study. Corina is a 3D structure generator that does not draw on input conformations to build 3D structures, even though the input is a 3D structure. Thus, the original ligand conformation was lost by processing with Corina, and the answers were not assigned to the docking study using Corina-generated ligands. Since Corina focuses more on a speedy process, SYBYL Minimize (MINI) was performed with the Corina-generated ligand conformation to find the energetically stable conformations. ‘Corina’ and ‘MINI’ ligands generated similar results in predicting the correct ligand poses based on the rmsd values. In some parameter settings, the ‘Corina’ ligand was better than ‘MINI’ and vice versa.

Focus on Repeated Prediction. The docking runs were repeated 1000 times each in the same set. When repeated, all programs generated some degree of dispersion (Table 4). In the docking pose predictions, 68% ± 10% of the 116 000 docking predictions were within 2 Å from the best rmsd in each ligand−protein complex, with any program. In other words, approximately 30% of the redocking results were quite different (rmsd 2 Å or more) from the best poses predicted, by the docking programs.

As an example of the worst cases, Figure 1 shows the distributions of the docking poses for HIV-1 protease (PDB code: 1AAQ). All settings of the programs resulted in wide distributions of the docking poses, and there were commonly two peaks of correct and inversed ligand angles. The minor peak of GOLDScoreSTD was closer to the reference coordinates of the ligand than the other settings. The docking scores were also widely distributed in 1AAQ (Figure 2). The figure shows the score distribution in GOLDScoreSTD. The docking score is important for the docking program to rank the candidates of ligand conformations and to screen the candidates from a library. Virtual screening ranks candidate compounds based on docking scores. Thus, the relative ranks of the ligands from the complexes were obtained from the docking scores and the docking scores from the NCI diversity set. In the figure, an rmsd of 1.53 Å and a docking score of 84.5 ranked first as compared to the NCI screening results, and an rmsd of 1.28 Å and a docking score of 49.2 ranked 233rd in the NCI screening results (Figure 3). The docking poses, scores, and ranks were widely distributed during the total of 3000 docking runs with three initial ligand conformations for 1AAQ.

The dispersions in the docking scores and the rankings were also examined for the rest of the 116 complexes. Most docking scores were within 10 points from the best docking score predicted by each docking program, when the docking runs were repeated 1000 times (Table 4). However, the docking scores from GOLDScoreLib in the 10 point range were much less than the other docking settings. This is probably due to the fact that the computation time needed by GOLD is longer than those required by the other docking programs and the ‘Library Screening Settings’ were less than the time required to perform local optimization in the GOLD algorithm. When the degrees of the dispersions in ranking were compared among the docking programs and their parameter settings, the results by GOLDScoreLib were clearly more dispersed than the others (Table 4). However, the differences in dispersions between ‘Library Screening Settings’ and ‘Standard Default Settings’ were much smaller in the ranking (49.7% and 82.8% of ranks were within 100th, respectively) than in the docking scores (5.4% and 40.6% of docking scores were within 1 point, respectively) presumably because the docking scores for the compounds in the NCI diversity set were not scored the best, and neither were the ligands from the complexes. In the other programs and their parameter settings, 17.2% (GOLDScoreSTD) to 39.1% (AutoDock) of the ranks were at least 100th lower from the best rank, even though most of the docking scores were within 10 points from the best score. The docking programs find the best docking poses based on the docking scores, and the docking scores should be approximately the same, even though the docking poses are different. Thus, the docking programs are not finding the best results each time since the docking results are dispersed in both the rmsd and docking scores.

A larger ligand has more degrees of freedom in its conformations, and thus, more dispersion in the docking score and rank can be expected. However, DOCK was not that straightforward (Table 5). In general, the difference between the best and the least in rmsd and ranking was

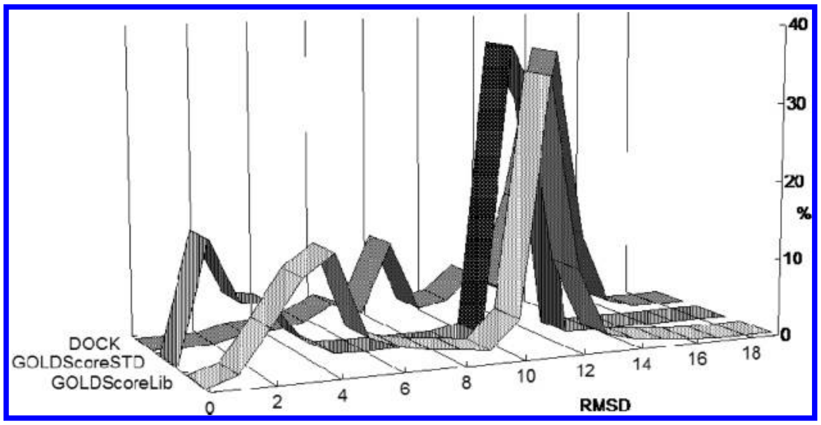

Figure 1. Distribution of docking poses in rmsd for 3000 docking runs for 1AAQ.

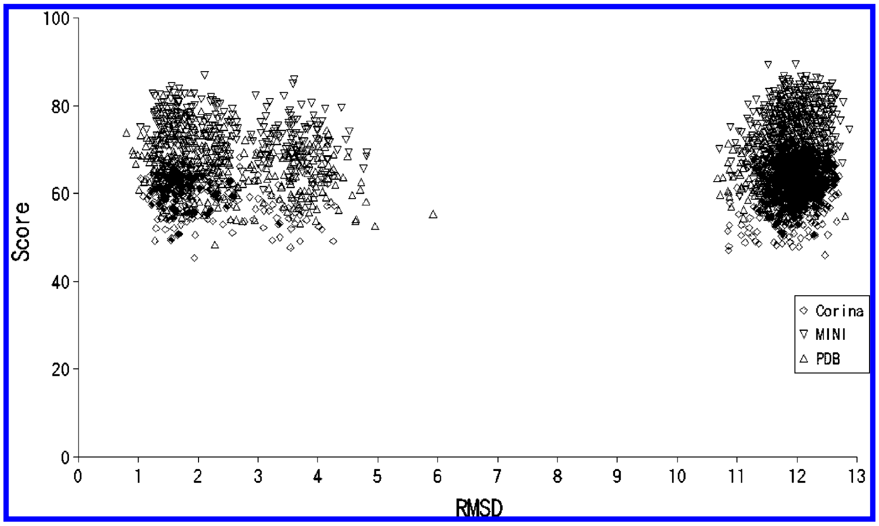

Figure 2. Distribution of docking scores for 3000 dock runs for 1AAQ by GOLDScoreSTD.

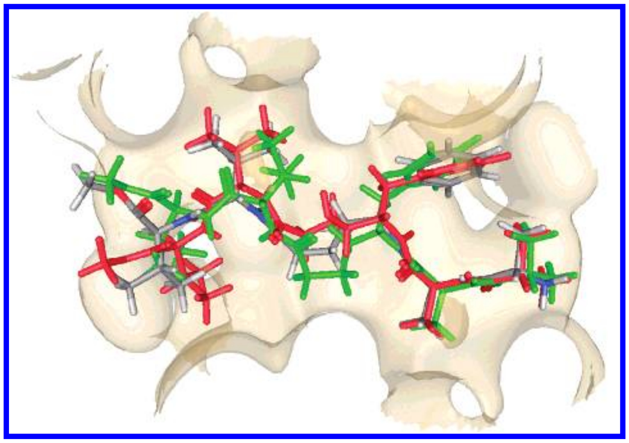

Figure 3. Predicted complex structures for 1AAQ by GOLDScoreSTD. The ligand in green scored 49.2, the ligand in red scored 84.5, and the other ligand was the reference conformation from the crystal structure.

Table 5. Correlation Coefficients of Differences between the Best and the Worst Results in rmsd, Score, and Rank with Size of Binding Site

rmsd	1		
rank	0.7	1	
size of binding site	0.8	0.6	1

correlation with size of binding site	AutoDock EFEB	AutoDock FDE	DOCK	ChemScore STD	GOLDScore Lib	GOLDScore STD
rmsd	0.8	0.8	0.7	0.8	0.8	0.8
score	0.8	0.8	0.2	0.8	0.7	0.8
rank	0.8	0.8	0.4	0.6	0.7	0.6

positively correlated with the sizes of the binding sites. However, the correlation in the docking score and the ranking dispersions with the sizes of the binding sites was much smaller in DOCK, while the correlation in the rmsd dispersion for DOCK was similar to those of the other programs. Thus, DOCK is good at finding the best docking scored conformation consistently.

Screening Performance. Direct comparisons of docking scores are difficult. The units and concepts of docking scores from the various docking programs differ from one another.

The docking score ranges were from minus tens of thousands to a few million kcal/mol for AutoDock, from −100 to 1000 points for DOCK, and from −1 000 000 to 100 points for GOLD, when the NCI diversity set of 1990 structures was used to calculate the docking scores. Over 90% of the docking scores from any docking program tested here were within ±100 points with the NCI diversity set. Even though most of the docking scores from each docking program were within a similar range, such docking scores were mostly ranked in the middle of the NCI diversity set and were not of the top scored structures. In virtual screening, the most important docking score is the top score, since virtual screening is for finding the best compounds and not middle ranked unknown compounds.

Table 6. Percentages of Ligand−Protein Complexes Ranked 100th or Better

	Corina (%)	MINI (%)	PDB (%)
AutoDock EFEB	7.6	8.1	12.0
AutoDock FED	9.5	11.9	13.7
DOCK	41.5	53.8	53.7
GOLD ChemScoreSTD	20.9	23.1	28.4
GOLD GOLDScoreLib	30.1	43.0	36.3
GOLD GOLDScoreSTD	36.9	55.4	45.8

The top 5% is a popular cutoff for success of the screening in the evaluations.[1,10,23] Since we used 1990 compounds from the NCI diversity set as decoys, we performed the comparisons of the screening performances by the percentages of docking runs in which the correct ligands were ranked 100th or better, approximately top 5%. The ranking results surprisingly showed that DOCK was the best at finding the correct ligands, followed by GOLD and AutoDock (Table 6), even when we varied the cutoff of the rank for the success from the top 10 to the top 100 compounds (Figure 4). Modifications to the parameter settings or the input structure files may improve the screening performances in docking programs, and thus, the ratings of docking programs may differ from those obtained. However, it is also true that DOCK has been used for the evaluation of performance in docking pose prediction but has not been thoroughly evaluated for screening performance, since DOCK does not produce appealing results in docking pose predictions, as it scored the worst in the docking pose prediction results discussed above.

The screening performance was not always identical to the docking pose prediction performance for the docking programs. The docking pose prediction performances were similar between the Corina-generated (Corina) and SYBYL Minimize-generated ligand (MINI) conformations. However, the 'MINI' ligand conformations scored better in the screening performances, probably because they received higher scores, due to energetic stability and low internal stringency of the ligand. Thus, the initial ligand conformation affects the docking score, because the docking program does not possess a good energy minimization function.

'Library Screening Settings' in GOLD was 36 times faster than 'Standard Default Settings', and both parameter settings produced similar results in docking pose prediction. In contrast, the screening performance of 'Standard Default Settings' in GOLD was better than that of 'Library Screening Settings'. This is because searches for a local minimum produce better docking scores with longer iteration and computation times.

Target Protein Types. The target proteins from the 116 complexes were classified into 13 types: hydrolases, metalloproteases, aspartic proteases, serine proteases, glycosidases, transferases and kinases, kinases only, lyases, oxidoreductases, immunoglobulins, isomerases, lectins, and viral proteins according to the classification made by Nissink et al.[17] The rmsd-based evaluations revealed that no docking program was significantly superior to GOLD in each complex (data not shown). Thirteen complexes were found solutions with an rmsd of 2 Å or better only by GOLD, and no solution was found by either AutoDock or DOCK alone. The sizes of the binding sites for the complexes that were successfully solved only by GOLD were widely distributed, from 2253 to 7900 Å,[3] and represented the various protein types.

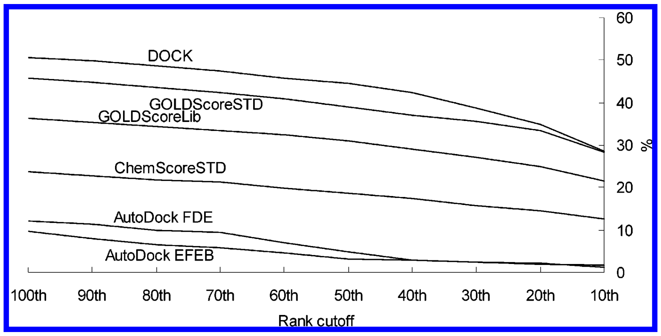

Figure 4. Rank cutoff for success screening and success rate.

Table 7. Size of Binding Site and Percentages of Successful Docking Results

	<3500		3501−6000		>6001	
binding site (Å^3)	2 Å or better (%)	100th or better (%)	2 Å or better (%)	100th or better (%)	2 Å or better (%)	100th or better (%)
AutoDock EFEB	48.1	9.8	26.9	8.2	20.9	11.4
AutoDock FED		11.4		8.4		17.9
DOCK	37.5	45.7	18.0	48.7	10.4	58.7
GOLD ChemScoreSTD	65.2	24.7	48.7	19.3	28.3	26.9
GOLD GOLDScoreLib	69.3	46.1	44.4	47.6	14.3	11.3
GOLD GOLDScoreSTD	62.5	38.0	49.0	49.3	35.8	53.5

When focusing on the target proteins ranked 100th or better by only one docking program or its parameter settings, no docking programs or parameter settings were found to be good at a particular size of binding site or type of target protein. One correct ligand of a complex was ranked in the top 100 only by DOCK alone, while three detected were by ChemScoreSTD alone and nine were identified by GOLDScoreLib alone. This highlights the superiority of GOLD in virtual screening.

All of the docking programs showed better performances on docking pose predictions for smaller binding sites than larger ones (Table 7). However, the screening performance did not show significant differences in term of the sizes of the binding sites for any of the docking programs and their parameter settings, except for GOLDScoreLib. GOLDScoreLib and GOLDScoreSTD showed similar screening performances with binding site sizes smaller than 6000 Å,[3] but GOLDScoreLib showed a lower screening performance for binding sites over 6000 Å^3 probably due to the limitation of maximum iterations. Otherwise, the effects of the binding sites size were similar to the screening performances of all of the docking programs. It seems that the screening performance is better when the binding site size is larger. This probably reflects the fact that a larger ligand can possess more binding atoms to the target protein and can be scored higher, and thus, a larger ligand can be ranked higher.

Some structures of the compounds from the NCI diversity set repeatedly ranked first for different target proteins (Figure 5). One compound (NSC: 109268) was ranked first for 48 out of 116 target proteins by AutoDock. Another (NSC: 131453) was ranked first 37 times by DOCK. The other (NSC: 49789) was ranked first 37 times by GOLD. A total of eight compounds were ranked first more than 10 times by one program. These compounds had molecular weights ranging from 260 to 558, and seven of them fulfilled Lipinski's Rule of Five. Such compounds were never ranked first more than once by the other programs and never shared

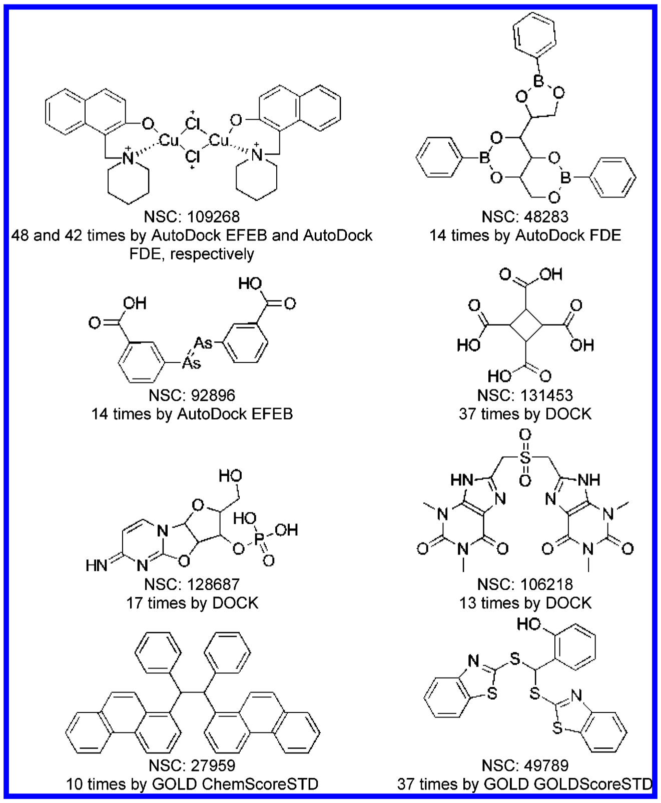

Figure 5. Compounds ranked first frequently.

similar properties. Thus, the compounds were program-specific, and no specific structures needed special attention in all of the docking programs.

CONCLUSIONS

In this paper, we propose two novel criteria for the evaluation of screening performances in virtual screening rather than enrichment of correct ligands. The two criteria are the docking scores of the top ranked compounds and the dispersion in the docking scores when exactly the same docking runs are repeated.

Ideally, the score gaps between the top ranked compounds should be large, for better recognition of active ligands. However, the score differences among the top ranked compounds were much smaller than among the lower ranked compounds in GOLD and DOCK (Figure 6). Only AutoDock showed large score differences between the top ranked compounds and those ranked 50th or worse. It is better to differentiate between the scores of the top ranked compounds, and AutoDock was the best among the tested docking programs. AutoDock also has a feature to perform docking with all 133 protein targets, while GOLD and DOCK failed on some of the targets. Such stable runs in AutoDock are good for users, but it can lead to negative evaluation if the program is forced to calculate scores for unconfident compounds, since only the targets which obtained docking results by all programs were used for the evaluation in this study.

While AutoDock evaluated all dockings, DOCK quit the calculation for potential nonbinders and presented a fixed scored of 1000. By not calculating detailed scores for nonbinders, DOCK could shorten the computation time (Table 2). Thus, DOCK could spend more time for the other compounds than GOLD with library screening settings although the average run times were similar in both two settings.

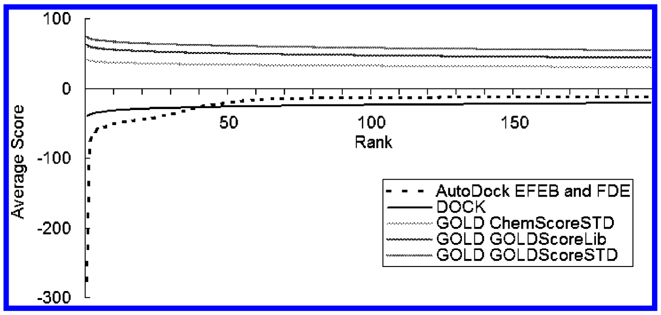

Figure 6. Distribution of average docking scores in the given ranks for 1990 compounds from the NCI diversity set. Positive numbers show stronger interactions in GOLD, while they show weaker interactions in AutoDock and DOCK.

GOLD is only the program featured with modifications of hydrogen positions in the proteins. Such a function was not accounted for in our evaluations since only actives from PDB complexes were used in the study. This time, GOLD wrongfully modified protein structures and made other decoys to be scored higher. Such a feature may be an advantage for proteins which do not have conformational changes but hydrogen positions by binding to other ligands.

The dispersion of the docking scores obtained from repeated docking runs was also observed. As long as a different random seed is used, the docking program will generate different docking results every time, even under the same docking conditions. When docking poses are predicted for only a few ligands, the impact caused by the use of random numbers is not serious since such a prediction can be easily repeated. In virtual screening, it is not preferable to repeat calculations for the same ligands. Users choose to avoid repeating the same docking runs, to reduce the total computation time. In such a situation, the docking program should not generate a large score dispersion in the same docking runs. In the worst case, such a dispersion generated both the top and the worst ranks from the same ligand and target protein set when the docking runs were repeated. GOLD with ‘Standard Default Settings’ had the most time-consuming parameter settings, and it produced the least dispersed docking scores and ranking results.

Such a dispersion was affected by the sizes of the binding sites. Larger binding sites logically generate a greater dispersion of the docking poses and scores, since there are more degrees of freedom for the ligand conformations. Thus, docking predictions are generally less reliable for larger binding sites. The variations in the docking scores positively correlated with the sizes of the docking sites in all of the programs, but this correlation was significantly less in DOCK.

GOLD was the best in docking pose prediction among the tested programs, as expected. However, DOCK was unexpectedly the best in screening performance. DOCK is not usually used for virtual screening probably because of its poorer performance on docking pose prediction. However, it did perform well in screening. The quality of the ranking predictions can be questionable without good predictions of docking poses. The docking pose predictions are usually assessed based on the rmsd, and a value of 2 Å or less is

Table 8. Percentages of Ligand−Protein Complexes Ranked 100th or Better When Only the Best Scores Were Considered in Each Repeated 1000 Docking Runs

	percentage (%)
AutoDock EFEB	17.2
AutoDock FED	20.7
DOCK	61.5
GOLD ChemScoreSTD	35.6
GOLD GOLDScoreLib	64.9
GOLD GOLDScoreSTD	54.6

Table 9. Percentages of Ligand−Protein Complexes Ranked 100th or Better in Two Categories of rmsd Values

	rmsd (Å)	total count	100th or better (%)
AutoDock EFEB	≤2	114378	17.0
	>2	233622	6.1
AutoDock FED	≤2	114378	16.7
	>2	233622	10.0
DOCK	≤2	79697	55.2
	>2	268303	49.1
GOLD ChemScoreSTD	≤2	172901	37.4
	>2	175099	10.4
GOLD GOLDScoreLib	≤2	160426	56.1
	>2	187574	19.4
GOLD GOLDScoreSTD	≤2	176905	51.8
	>2	171095	39.6

considered as a successful docking pose, presumably because the average resolution of the ligand−protein complexes in the PDB is approximately 2 Å (The average resolution of the complexes used in this study was 2.2 Å.). However, hydrogen bonds and other molecular interactions between a ligand and a protein cannot be observed correctly with an rmsd of 2 Å. In fact, all of the docking programs showed different docking scores, even though the docking conformations were similar. If docking programs can assign the correct molecular interactions in that rmsd range, then the docking score should be the same among the successful cases even if the predicted docking poses differ. The GOLD program mentions that “small conformational or positional changes to the ligand can have a large effect on the rescored fitness value as a result of the approximate nature of scoring functions”. Thus, the rmsd based evaluation is questionable in the first place even though no other practical methods are available for the docking pose evaluation as mentioned previously.

DOCK was not the only program predicting correct ligands without good pose prediction. Actually, GOLD and AutoDock also found the correct ligands when rmsd values were more than 2 Å (Table 9). Only the difference was that DOCK could predict the correct ligands with a better ratio. Thus, we believe that an rmsd based performance is not always linked to the screening performance. Especially when the performance with a low enrichment rate is assessed, docking programs are required to have the ability to omit nonbinders from the compound library more than the ability to find correct ligands. The top 5% cutoff may be the case for the low enrichment rate requirement, and DOCK was good at omitting nonbinders.

Usually scoring functions are thought to be responsible for common modeling limitation. In this study, we found that the dispersion in docking scores was also responsible for the limitation as described above. When the same docking runs were repeated, the docking program generates ranges of values in both docking positions and scores. If we extracted only the best score in each repeated docking run, the majority of correct ligands were successfully predicted by DOCK and GOLD (Table 8). Thus, we think that docking theories used in the algorithms are good for the screening, but the common problem is low reproducibility of the same docking results by the programs.

Differences in the docking pose prediction and screening were found in the ligand preparations and the parameter settings. The screening of compounds requires that the initial ligand conformation be energetically minimized, to achieve a higher docking score. Parameter settings with more docking iterations, which require longer computation times, also generate higher scores with careful local minimization. These preparations and parameter settings were not particularly important in the docking pose prediction.

The use of ten or more random conformations of a ligand reportedly generated better predictions of complex structures.[9,10] This fact may be related to the dispersion of the docking results caused by the use of random numbers. A use of ten or more random conformations actually repeats docking runs with the same ligand as many times as the initial conformations prepared. By repeating the docking runs, we could obtain better docking results as described above. Thus, it was no surprise that the use of ten or more initial conformations of a ligand generated better docking results. However, we also knew that the initial conformation of a ligand affected the screening performance of the docking programs. Since energetically minimized conformations could generate better screening results, the use of several minimized conformations of a ligand may improve the docking scores.

Finally, we made recommendations for docking programs for virtual screening. First, parameter settings with more local optimization processes should be used to obtain the best docking scores with the least dispersion. Second, the docking programs should be run several times, as long as time allows, especially when a complex has a larger binding site. Docking programs generate the user-specified number of initial ligand conformations. Users may select more initial generating conformations in the docking program,[24] instead of the second recommendation. Third, several initial ligand conformations should be prepared, if possible, and they must be energetically minimized. The use of more docking runs and more initial ligand conformations will help in finding the global minimum. We also hope that presettings for repeated docking runs will be included in the docking program, for the users’ convenience. In general, virtual screening requires more computational costs for each compound. Even though virtual screening needs more docking runs with many library compounds, the use of rough docking settings makes the results less reliable, unless the calculated binding area is quite small. Thus, we need to select more comprehensive docking settings such as the ‘Standard Default Settings’ in GOLD, to maximize the performances of the docking programs.

We believe that these recommendations can enhance the screening performance of the molecular docking programs currently available in our community. Also, the novel standards for validating the screening performance of the docking programs can be helpful for developing new docking programs or upgrading the current programs. Improvements in the screening performance of the docking program will provide better recognition of novel compounds in drug discovery.

REFERENCES AND NOTES

(1) Jain, A. N. Scoring functions for protein-ligand docking. *Curr. Protein Pept. Sci.* **2006**, *7*, 407−420.

(2) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755−763.

(3) Ewing, T.; 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.

(4) Morris, G. M.; Goodsell, D. S.; Halliday, R.; 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) Jones, G.; Willett, R.; Glen, R.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(6) 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.

(7) 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.

(8) Abagyan, R.; Totrov, M.; Kuznetsov, R. ICM - A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(9) Jain, A. N. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46*, 499−511.

(10) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, *60*, 325−332.

(11) , Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **2004**, *57*, 225−242.

(12) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26*, 11−22.

(13) 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*, 401 -415.

(14) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model*. **2005**, *45*, 1134−1146.

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

(16) 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.

(17) Nissink, J. W.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein-ligand interaction. *Proteins* **2002**, *49*, 457−471.

(18) Knox, A. J.; Meegan, M. J.; Carta, G.; Lloyd, D. G. Considerations in compound database preparation−"hidden“ impact on virtual screening results. *J. Chem. Inf. Model.* **2005**, *45*, 1908−1919.

(19) *SYBYL*, *version 6.92*; Tripos, Inc.: St. Louis, MO, 2004.

(20) DTP − Diversity Set Information. http://dtp.nci.nih.gov/branches/dscb/diversity_explanation.html (accessed Jan 25, 2007).

(21) *GOLD*, *version 3.0*; The Cambridge Crystallographic Data Centre: Cambridge, 2006.

(22) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. Macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(23) Graves, A. P.; Brenk, R.; Shoichet, B. K. Decoys for docking. *J. Med. Chem.* **2005**, *48*, 3714−3728.

(24) Good, A. C.; Cheney, D. L. Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. *J. Mol. Graphics Modell.* **2003**, *22*, 23−30.

CI7000378