Articles

Comparative Evaluation of 11 Scoring Functions for Molecular Docking

Renxiao Wang, Yipin Lu, and Shaomeng Wang*

Department of Internal Medicine and Comprehensive Cancer Center, University of Michigan Medical School and Department of Medicinal Chemistry, University of Michigan College of Pharmacy, 3316 CCGC Building, 1500 E. Medical Center Drive, Ann Arbor, Michigan 48109-0934

Received September 4, 2002

Eleven popular scoring functions have been tested on 100 protein−ligand complexes to evaluate their abilities to reproduce experimentally determined structures and binding affinities. They include four scoring functions implemented in the LigFit module in Cerius2 (LigScore, PLP, PMF, and LUDI), four scoring functions implemented in the CScore module in SYBYL (F-Score, G-Score, D-Score, and ChemScore), the scoring function implemented in the AutoDock program, and two stand-alone scoring functions (DrugScore and X-Score). These scoring functions are not tested in the context of a particular docking program. Instead, conformational sampling and scoring are separated into two consecutive steps. First, an exhaustive conformational sampling is performed by using the AutoDock program to generate an ensemble of docked conformations for each ligand molecule. This conformational ensemble is required to cover the entire conformational space as much as possible rather than to focus on a few energy minima. Then, each scoring function is applied to score this conformational ensemble to see if it can identify the experimentally observed conformation from all of the other decoys. Among all of the scoring functions under test, six of them, i.e., PLP, F-Score, LigScore, DrugScore, LUDI, and X-Score, yield success rates higher than the AutoDock scoring function. The success rates of these six scoring functions range from 66% to 76% if using root-mean-square deviation ≤2.0 Å as the criterion. Combining any two or three of these six scoring functions into a consensus scoring scheme further improves the success rate to nearly 80% or even higher. However, when applied to reproduce the experimentally determined binding affinities of the 100 protein−ligand complexes, only X-Score, PLP, DrugScore, and G-Score are able to give correlation coefficients over 0.50. All of the 11 scoring functions are further inspected by their abilities to construct a descriptive, funnel-shaped energy surface for protein−ligand complexation. The results indicate that X-Score and DrugScore perform better than the other ones at this aspect.

Introduction

Molecular docking has been a focus of attention for many years. Generally speaking, today's flexible docking programs, such as DOCK,[1] AutoDock,[2] FlexX,[3] and GOLD,[4] are able to predict protein−ligand complex structures with reasonable accuracy and speed. These docking programs find their most important applications in virtual database screening approaches in which hundreds of thousands of molecules are docked into the binding pocket on the target molecule to identify plausible binders. When used prior to experimental screening, they can be considered as powerful computational filters to reduce labor and cost. With a dramatic increase in the number of "druggable" biological targets in this postgenomic era, molecular docking will undoubtedly continue to play an important role in drug discovery.

Molecular docking is basically a conformational sampling procedure in which various docked conformations are explored to identify the right one. In today's docking programs, this sampling procedure is based on either genetic algorithm, Monte Carlo simulation, simulated annealing, distance geometry, or other miscellaneous methods. But no matter what kind of method is applied, conformational sampling must be guided by a scoring function (or energy function) that is used to evaluate the fitness between the protein and the ligand. The final docked conformations are also usually selected according to their scores. Apparently, accuracy of the scoring function has a major impact on the quality of molecular docking results.

A widely spread concept is that the major weakness of today's docking programs lies not in sampling methods but in scoring functions. As a matter of fact, considerable efforts have been devoted to the development of computational methods for describing protein−ligand interactions. During the past decade, a number of approaches have been reported. They can be roughly grouped into three categories: force field methods,[1−2,4] empirical scoring functions,[3,5−11] and knowledge-based potentials.[12−16] All of these scoring functions have been

* To whom correspondence should be addressed. Phone: (734) 615-0362. Fax: (734) 647-9647. E-mail: shaomeng@med.umich.edu.

10.1021/jm0203783 CCC: $25.00
Published on Web 05/08/2003

validated on various sets of protein−ligand complex structures, and thus, in principle, all of them could be implemented in a docking program. So it is very intriguing and also important to know which scoring functions generally perform better than the others. Actually, several comparative studies of various scoring functions have already been published.[17−20] In these studies, typically a number of docking programs and scoring functions are tested on selected targets. The results are judged by docking accuracy, scoring accuracy, or the ability to identify known active compounds from a random pool. The ultimate goal of these studies is to find the best docking/scoring combination that can be applied reliably to some specific targets.

These studies certainly represent one valid approach for evaluating scoring functions in a molecular docking context. They are especially useful when one has an immediate interest in seeking active compounds for a particular target through virtual database screening because they point out what docking/scoring combination may give the most promising results. But a potential drawback in these studies is that they emphasize more on the overall performance of a complicated procedure in which the docking algorithm and the scoring function are coupled together. If a certain docking/scoring combination fails, it is not always clear which one should be blamed: the docking algorithm, the scoring function, or both. Therefore, scoring functions *themselves* are not fairly compared in this way. Even when all of the scoring functions are tested in the context of one docking program, there is still another problem: the docking program is typically run with a “default” set of parameters, which does not guarantee an adequate sampling of possible docked conformations. If the conformational sampling is biased at the first place, all the subsequent scoring function evaluations may be invalid. Therefore, ideally one should explore various settings of the docking program, especially those parameters controlling conformational sampling, to examine how they affect the final docking/scoring results. Unfortunately, this issue has never been addressed adequately in previous docking/scoring studies.

The objective of our study is to conduct a fair evaluation of various scoring functions in the context of molecular docking. Our central idea is to isolate the conformational sampling procedure from the scoring procedure so that all of the scoring functions can be compared on the same ground. To achieve this, an ensemble of docked conformationals of each ligand molecule is generated by using the AutoDock program. Considerable efforts are made to ensure that this conformational ensemble achieves diversity rather than focuses on a few energy minima. Then, all of the scoring functions under test are applied to score the conformational ensemble. We have tested 11 popular scoring functions on a wide spectrum of 100 protein−ligand complexes. The performance of each scoring function is evaluated by how well it reproduces the experimentally determined structures and binding affinities and how well it constructs a descriptive, funnel-shaped energy surface for protein−ligand complexation. The strength and weakness of these scoring functions are discussed. Consensus scoring, as a practical strategy for improving docking accuracy, is also explored.

Methods

Preparation of the Test Set. The test set used in this study is constructed from 230 protein−ligand complexes used in our previous work.[11] All of these complexes have crystal structures and experimentally measured K_i or K_d values. In this study, we consider only complex structures with resolution better than 2.5 Å, which are 172 in total. Each complex is then subjected to an exhaustive conformational sampling procedure (described in detail in the next section). One hundred complexes have passed this procedure and are included in the final test set (see Table 1). Forty-three different types of proteins are presented in this test set. Molecular weights of ligand molecules range from 122 to 913. Numbers of rotatable single bonds (rotors) in ligand molecules range from 0 to 20. Dissociation constants of these complexes range from 1.49 to 10.15 (in $-\log K_d$ or $-\log K_i$ units), spanning nearly nine orders of magnitudes. All ligand molecules bind to their target proteins noncovalently.

Coordinates of all the complexes are downloaded from the Protein Data Bank.[21] For the convenience of computation, each complex is split into a protein molecule, which is saved in PDB format, and a ligand molecule, which is saved in Mol2 format. Metal ions, if residing on the protein−ligand interface, are left with the protein. All of the other organic and inorganic cofactors, as well as all the water molecules, are removed. Hydrogen atoms are added to both the protein and the ligand. Atom types and bond types of the ligand molecule are inspected and corrected manually. The protein is assigned AMBER united-atom charges, while the ligand is assigned MMFF94 charges. All of the above work is done by using the SYBYL software (version 6.8)[22] on an SGI Octane2 graphics workstation.

Conformational Sampling Procedure. The AutoDock program (version 3.0)[2] is employed to generate an ensemble of docked conformations for each ligand molecule. This program uses a genetic algorithm (GA) for conformational sampling. Each GA run outputs a single docked conformation as the final result. Since a conformational ensemble is desired, 100 individual GA runs are performed to generate 100 docked conformations for each ligand. Each GA run is performed with a population of 100 chromosomes, a crossover ratio of 0.80, a mutation ratio of 0.20, and an elitism ratio of 0.10. During docking, all the rotatable single bonds in the ligand, i.e., sp^3-sp^3 and sp^3-sp^2, are allowed to rotate except those whose rotations do not result in different conformations, such as the ones connecting a terminal $-CH_3$ group. Flexibility in cyclic parts of the ligand is neglected. Searching steps for translation, rotation, and torsions are set to 0.5 Å, 15°, and 15°, respectively. The size of the docking box is 30 Å × 30 Å × 30 Å, which is centered at the experimentally observed position of the ligand. This box is large enough to enclose the largest binding pocket observed in the entire test set. Grid spacing inside the docking box is 0.25 Å. Initial conformations of the ligand are generated randomly in the docking box. Other miscellaneous parameters are assigned the default values given by the AutoDock program. The protein structure is kept fixed during docking.

Since this conformational ensemble forms the basis for all subsequent scoring function evaluations, we expect it to depict the conformational space of the ligand (with respect to the protein) as completely as possible rather than focus on a few energy minima that are particularly favored by AutoDock. To achieve this goal, the following criteria have been applied to monitor the quality of the final conformation ensemble generated by AutoDock: (i) Root-mean-square deviation (rmsd) values (calculated by using the experimentally observed bound conformation as the reference) of all the docked conformations should spread throughout a wide range, e.g., 0−15 Å. (ii) The number of distinctive conformational clusters (counted by AutoDock using a clustering criterion of 2.0 Å) should fall between 30 and 70. This further ensures the diversity of the ensemble. (iii) A number of conformations should be close enough to the experimentally observed conformation (rmsd ≤ 2.0 Å). This ensures a proper sampling of the global minimum

Table 1. The 100 Protein−Ligand Complexes Used in This Study (Part a) and Grouped by Interaction Type (Parts b−d) and Those Discarded during Conformational Sampling (Part e)

PDB code	−log K_d	resoln (Å)	description	PDB code	−log K_d	resoln (Å)	description
1bbz	5.82	1.65	ABL tyrosine kinase/peptide ligand	1fmo	8.64	2.20	phosphotransferase/inhibitor PKI(5−24)
2xim	2.28	2.30	D-xylose isomerase/xylitol	2pk4	4.32	2.25	plasminogen kringle 4/aminocaproic acid
4xia	1.54	2.30	D-xylose isomerase/D-sorbitol	1inc	8.00	1.94	porcine pancreatic elastase/benzoxazinone inhibitor
8xia	2.95	1.90	D-xylose isomerase/D-xylose	4sga	3.27	1.80	proteinase A/Ace-Pro-Ala-Pro-Phe
1fkb	9.70	1.70	FK506 binding protein/rapamycin	5sga	2.85	1.80	proteinase A/Ace-Pro-Ala-Pro-Tyr
1fkf	9.40	1.70	FK506 binding protein/FK506	5p21	5.32	1.35	ras p21 protein/GPPNP
1hvr	9.51	1.80	HIV-1 protease/XK263	1rbp	6.72	2.00	retinol binding protein/retinol
1tet	6.20	2.30	IGG1 monoclonal FAB fragment/CTP3	1rgk	4.31	1.87	ribonuclease T1/2′-AMP
1abe	6.52	1.70	L-arabinose binding protein/L-arabinose	1rgl	4.43	2.00	ribonuclease T1/2′-GMP
1abf	5.42	1.90	L-arabinose binding protein/D-fucose	1rnt	5.18	1.90	ribonuclease T1/2′-GMP
1apb	5.82	1.76	L-arabinose binding protein/D-fucose	6rnt	2.37	1.80	ribonuclease T1/2′-AMP
1bap	6.85	1.75	L-arabinose binding protein/L-arabinose	1b5g	8.00	2.07	serine protease/peptide mimetic inhibitor
5abp	6.64	1.80	L-arabinose binding protein/D-galactose	1ba8	9.00	1.80	serine protease/peptide mimetic inhibitor
6abp	5.64	1.67	L-arabinose binding protein/L-arabinose	1bb0	8.36	2.10	serine protease/peptide mimetic inhibitor
7abp	5.54	1.67	L-arabinose binding protein/D-fucose	1yyy	5.09	2.10	serine protease/CVS1695
8abp	4.00	1.49	L-arabinose binding protein/D-galactose	1zzz	5.13	1.90	serine protease/CVS1694
9abp	8.00	1.97	L-arabinose binding protein/D-galactose	2sns	6.70	1.50	staphylococcal nuclease/2′-deoxy-3′,5′-diphosphothymidine
1e96	5.22	2.40	RAC/P67phox	1sre	4.00	1.78	streptavidin/HABA
1add	6.74	2.40	adenosine deaminase/1-deazaadenosine	1tlp	7.56	2.30	thermolysin/phosphoramidon
2ak3	3.86	1.90	adenylate kinase isoenzyme-3/AMP	4tln	3.72	2.30	thermolysin/Leu-NHOH
1adb	8.40	2.40	alcohol dehydrogenase/CNAD	5tln	6.37	2.30	thermolysin/benzylmalonyl-L-alanyl-glycine-*p*-nitroanilide
9aat	8.22	2.20	aspartate aminotransferase/pyridoxal-5′-phosphate	7tln	2.47	2.30	thermolysin/CH2CO-Leu-OCH3
1bzm	6.03	2.00	carbonic anhydrase I/sulfonamide drug	1tmn	7.47	1.90	thermolysin/*N*-(1-carboxy-3-phenyl)-L-Leu-Trp
1cbx	6.35	2.00	carboxypeptidase A/L-benzylsuccinate	2tmn	5.89	1.60	thermolysin/N-phosphory-L-leucinamide
2ctc	3.89	1.40	carboxypeptidase A/L-phenyl lactate	3tmn	5.90	1.70	thermolysin/Val-Trp
3cpa	4.00	2.00	carboxypeptidase A/glycyl-L-tyrosine	1a46	5.70	2.12	thrombin/beta-strand mimetic inhibitor
1cla	5.28	2.34	chloramphenicol acetyltransferase/chloramphenicol	1a5g	10.15	2.06	thrombin/peptide inhibitor
3cla	4.94	1.75	chloramphenicol acetyltransferase/chloramphenicol	1bcu	5.00	2.00	thrombin/proflavin
4cla	5.47	2.00	chloramphenicol acetyltransferase/chloramphenicol	1d3d	9.09	2.04	thrombin/benzo[B]thiophene inhibitor
2csc	3.36	1.70	citrate synthase/D-malate	1d3p	7.39	2.10	thrombin/benzo[B]thiophene inhibitor
5cna	2.00	2.00	concanavalin A/a-Me-D-mannopyranoside	1etr	7.41	2.20	thrombin/MQPA
1af2	3.10	2.30	cytidine deaminase/uridine	1ets	8.22	2.30	thrombin/NAPAP
1dhf	7.40	2.30	dihydrofolate reductase/folate	1sta	5.35	2.00	transthyretin/3,3′-diiodo-L-thyronine
1dr1	5.57	2.20	dihydrofolate reductase/biopterin	4tim	2.16	2.40	triosephosphate isomerase/2-phosphoglycerate
1drf	7.44	2.00	dihydrofolate reductase/folate	6tim	3.21	2.20	triosephosphate isomerase/glycerol-3-phosphate
1ela	6.35	1.80	elastase/TFA-LYS-PRO-ISO	7tim	5.40	1.90	triosephosphate isomerase/phosphoglycolohydroxamate
7est	7.60	1.80	elastase/TFAP	1bra	1.82	2.20	trypsin mutant/benzamidine
3fx2	9.30	1.90	flavodoxin/riboflavin monophosphate	1ppc	6.16	1.80	trypsin/NAPAP
2gbp	7.40	1.90	galactose binding protein/galactose	1pph	6.22	1.90	trypsin/3-TAPAP
1hsl	7.30	1.89	histidine binding protein/histidine	1tng	2.93	1.80	trypsin/aminomethylcyclohexane
2cgr	7.27	2.20	KAPPA fab fragment/antigen GAS	1tnh	3.37	1.80	trypsin/4-fluorobenzylamine
2qwb	2.74	2.00	neuraminidase/sialic acid	1tni	1.70	1.90	trypsin/4-phenylbutylamine
2qwc	3.55	1.60	neuraminidase/neu5ac2en	1tnj	1.96	1.80	trypsin/2-phenylethylamine
2qwd	4.85	2.00	neuraminidase/4-amino-neu5ac2en	1tnk	1.49	1.80	trypsin/3-phenylpropylamine
2qwe	7.48	2.00	neuraminidase/4-guanidino-neu5ac2en	1tnl	1.88	1.90	trypsin/t-2-phenylcyclopropylamine
2qwf	5.67	1.90	neuraminidase/ligand G20	3ptb	4.50	1.70	trypsin/benzamidine
2qwg	8.40	1.80	neuraminidase/ligand G28	1bhf	4.38	1.80	tyrosine kinase P56LCK/ACE-IPA-GLU-GLU-ILE
1mnc	9.00	2.10	neutrophil collagenase/hydroxamate	2xis	5.82	1.71	xylose isomerase/xylitol
1exw	3.90	2.40	palmitoyl protein thioesterase/hexadecylsulfonyl fluoride				
1bxo	10.00	0.95	penicillopepsin/phosphonate inhibitor				
1apt	9.40	1.80	penicillopepsin/pepstatin analogue				
1apw	8.00	1.80	penicillopepsin/IvaValValDfo- *N*-methylamide				

(b) Protein−Ligand Interactions Dominated by Hydrophilic Factors, 44 in Total

1abe, 1abf, 1adb, 1add, 1af2, 1apb, 1bap, 1e96, 1fmo, 1hsl, 1rgk, 1rgl, 1rnt, 1tet, 1yyy, 1zzz,
2ak3, 2csc, 2gbp, 2qwb, 2qwc, 2qwd, 2qwe, 2qwf, 2sns, 2xim, 2xis, 3cpa, 3fx2, 3ptb, 4tim, 4xia,
5abp, 5cna, 5p21, 6abp, 6rnt, 6tim, 7abp, 7tim, 8abp, 8xia, 9aat, 9abp

(c) Protein−Ligand Interactions Having Mixed Factors, 32 in Total

1a46, 1a5g, 1apt, 1apw, 1b5g, 1ba8, 1bb0, 1bhf, 1bra, 1bxo, 1bzm, 1cbx, 1dhf, 1dr1, 1drf, 1etr,
1mnc, 1ppc, 1pph, 1sre, 1tlp, 1tng, 1tnh, 2ctc, 2pk4, 2qwg, 2tmn, 3tmn, 4sga, 4tln, 5sga, 5tln

(d) Protein−Ligand Interactions Dominated by Hydrophobic Factors, 24 in Total

1bbz, 1bcu, 1cla, 1d3d, 1d3p, 1ela, 1ets, 1exw, 1fkb, 1fkf, 1hvr, 1inc, 1rbp, 1tha, 1tmn, 1tni,
1tnj, 1tnk, 1tnl, 2cgr, 3cla, 4cla, 7est, 7tln

(e) Protein−Ligand Complexes Discarded during the Conformational Sampling Procedure, 72 in Total

1a94, 1anf, 1apu, 1apv, 1bai, 1bll, 1bxq, 1cps, 1csc, 1ctt, 1dih, 1eed, 1elc, 1epo, 1epp, 1fq4,
1fq5, 1hbv, 1hew, 1hpv, 1htf, 1htg, 1hvi, 1hvj, 1hvk, 1hvl, 1hvs, 1l82, 1l83, 1l86, 1l87, 1ldm,
1mdq, 1mfc, 1mfe, 1ppk, 1ppl, 1ppm, 1pso, 1rne, 1snc, 1tmt, 2dri, 2er6, 2er7, 2er9, 2ifb, 2msb,
2rnt, 3csc, 3er3, 4dfr, 4er1, 4er2, 4er4, 4est, 4gr1, 4hvp, 4phv, 4tmn, 5acn, 5enl, 5er2, 5hvp,
5tim, 5tmn, 6enl, 6tmn, 7acn, 7hvp, 8acn, 8cpa

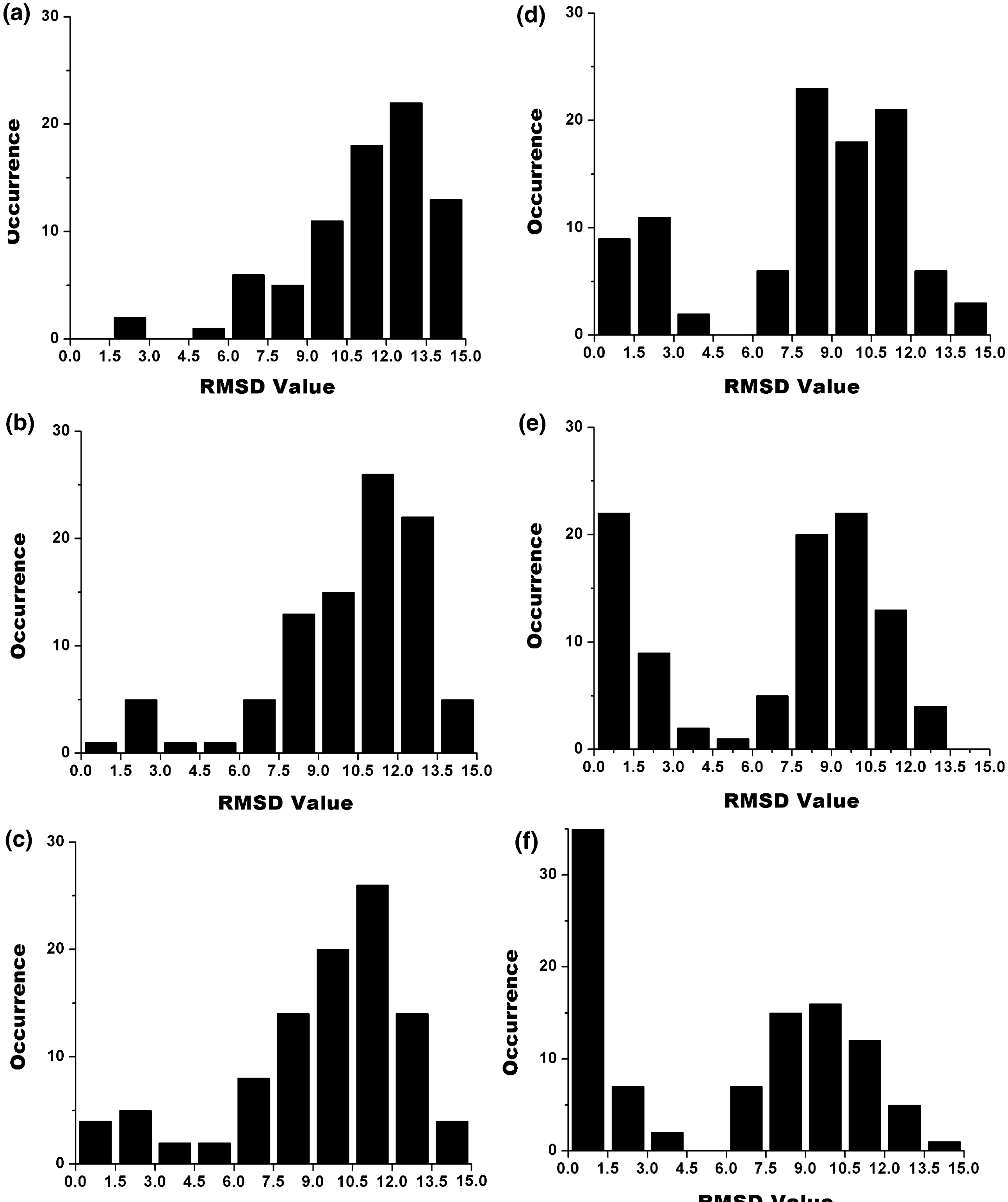

Figure 1. The rmsd distributions observed in the final conformational ensemble of PDB entry 1BXO when (a) GA generation = 1, (b) GA generation = 10, (c) GA generation = 20, (d) GA generation = 30, (e) GA generation = 50, and (f) GA generation = 100.

on the energy surface and its vicinity area. Applying these criteria, we find that the quality of the final conformational ensemble is mainly determined by the length of each GA run, which is set by the *ga_num_generations* parameter in AutoDock. Typically, if GA runs are too short, most of the conformations will be still very close to their initial states, and the conformational sampling will be far from complete. In contrast, if GA runs are too long, most of the conformations will have started to converge to certain clusters, and thus, the diversity of the final conformational ensemble will become low. Figure 1 gives one example of how the length of each GA run affects the rmsd distribution in the final conformational ensemble. As one can interpret from there, controlling the length of each GA run is the key to obtaining the conformational ensemble meeting our requirements.

But the problem is that a fixed value of *ga_num_generations* is not likely to satisfy all the protein−ligand complexes included in our test set. Therefore, we have adopted an inter-

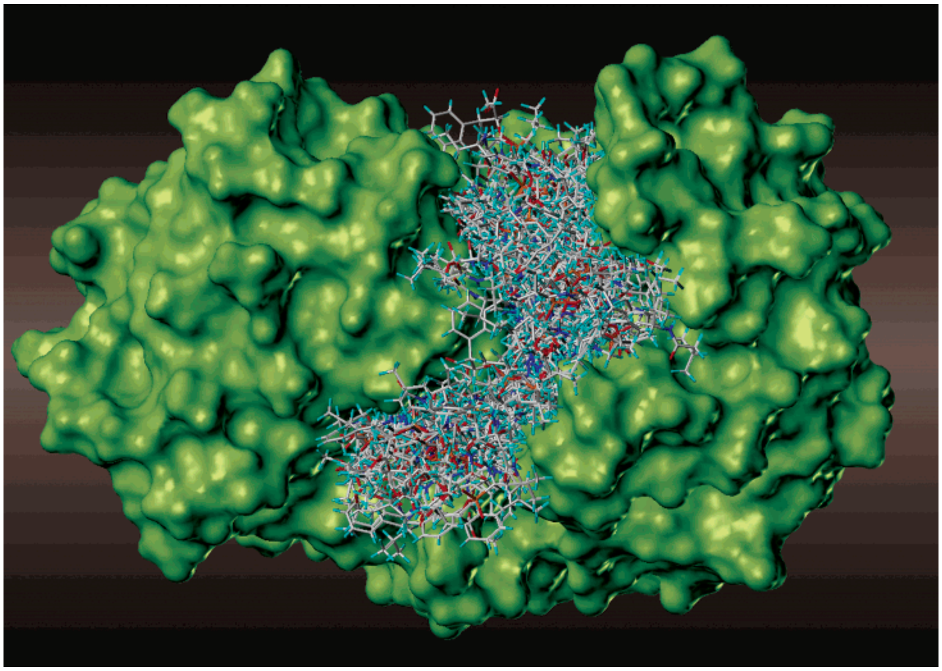

Figure 2. Conformational ensemble of the ligand molecule generated by AutoDock (PDB entry 1BXO).

active procedure to find a proper value of *ga_num_generations* for each complex. We give the *ga_num_generations* parameter an initial value of 50, run the docking program, and then examine the results. We will accept the resulting conformational ensemble if its quality meets our requirements. If not, we will decrease or increase the value of *ga_num_generations* by an increment of 10 and rerun the docking program. This procedure is repeated until a satisfactory conformational ensemble is obtained. In fact, a major part of our effort is spent at this step. To limit the computational time to an acceptable level, the maximum value of *ga_num_generations* that we will try is 200. For some complexes, a satisfactory conformational ensemble is not obtained even at this level of computation. Typically, the ligands in these cases are large flexible molecules, such as oligopeptides, and therefore may need even more extensive conformational sampling. These complexes, 72 in total, are not included in our final test set. They are also listed in Table 1. For all of the successful ones, 100 in total, we then add the experimentally observed bound conformation of the ligand to the 100 AutoDock generated docked conformations. This further ensures the completeness of the conformational ensemble because AutoDock may not have generated exactly the same conformation. This conformation should not be missed because it represents the true global minimum and is probably the most important spot on the energy surface. The total number of docked conformations of each ligand thus becomes 101. These conformations usually cover the entire binding pocket and its vicinity area. One example is illustrated in Figure 2.

As described above, the AutoDock program has been used in our study as the conformation generator. Ideally, we should use a totally independent program that can generate all the possible docked conformation of a given ligand molecule. It is possible to generate all the rigid conformations of a ligand molecule, but rigid docking is not our interest. To the best of our knowledge, there is no program that can efficiently generate all of the flexible docked conformations of a ligand molecule. Therefore, we must borrow an existing docking program for this purpose. Among all the docking programs available to us, AutoDock is the closest one to our requirements. It allows the user to control the conformational sampling procedure conveniently so that we can obtain the desired conformational ensembles. In addition, all the docked conformations generated by AutoDock are selected through GA procedures and are further minimized on-the-fly during docking. Thus, they represent local minima rather than random spots on the protein–ligand interaction energy surface, and this is exactly what we need for the subsequent scoring function evaluations. It needs to be mentioned that other flexible docking programs, if they have the equivalent features, could be used in our study as well.

Scoring Procedure. Eleven scoring functions have been tested in our study, including the scoring function implemented in the AutoDock program, four scoring functions from the LigFit module in Cerius2, version 4.6 (LigScore, PLP, PMF, and LUDI),[23] four scoring functions from the CScore module in SYBYL, version 6.8 (F-Score, G-Score, D-Score, and ChemScore),[22] and another two stand-alone scoring functions (DrugScore, version 1.2, and X-Score, version 1.0). They can be roughly grouped into three categories: (i) force field based methods, i.e., AutoDock, G-Score and D-Score, (ii) empirical scoring functions, i.e., LigScore, PLP, LUDI, F-Score, ChemScore, and X-Score, and (iii) knowledge-based potentials of mean force, i.e., PMF and DrugScore. All of these 11 scoring functions are briefly described in the Appendix. Specific parameter/options used in our study when applying these scoring functions are also described there.

The conformational ensemble of each ligand generated from the conformational sampling procedure is saved in a multi-Mol2 file. It is loaded into a spreadsheet in SYBYL to apply the four scoring functions implemented in SYBYL. DrugScore and X-Score also accept this file as valid input for computation. This multi-Mol2 file is then translated into SDF format by using the Babel program (version 1.6) to apply the four scoring functions implemented in Cerius2. After all the computations are completed, a Perl script is written to extract the scoring results from SYBYL, Cerius2, DrugScore, X-Score, and AutoDock and to compile them into one table. During this process, we change the signs of some score functions to ensure that a lower score always indicates a higher binding affinity. All of the results, together with the test set itself, can be downloaded from http://sw16.im.med.umich.edu/software/xtool/.

Results and Discussions

As described above, we have tested 11 popular scoring functions on a large variety of protein–ligand complexes. These scoring functions are either commercially available or can be obtained from their original authors. All of them have been widely used in structure-based drug design, and thus, a comparative evaluation of these

Table 2. Success Rates of 11 Scoring Functions under Different rmsd Criteria

	success rate (%)				
scoring function[a]	rmsd ≤1.0 Å	rmsd ≤1.5 Å	rmsd ≤ 2.0 Å	rmsd ≤2.5 Å	rmsd ≤3.0 Å
Cerius2/PLP	63	69	76	79	80
SYBYL/F-Score	56	66	74	77	77
Cerius2/LigScore	64	68	74	75	76
DrugScore	63	68	72	74	74
Cerius2/LUDI	43	55	67	67	67
X-Score	37	54	66	72	74
AutoDock	34	52	62	68	72
Cerius2/PMF	40	46	52	54	57
SYBYL/G-Score	24	32	42	49	56
SYBYL/ChemScore	12	26	35	37	40
SYBYL/D-Score	8	16	26	30	41

[a] Scoring functions are ranked by their success rates at rmsd ≤ 2.0 Å.

Table 3. Success Rates of 11 Scoring Functions When Considering Multiple Conformations

	success rate (%) when considering		
scoring function[a]	only the best conformation	the best two conformations	the best three conformations
Cerius2/PLP	76	87	88
SYBYL/F-Score	74	89	90
Cerius2/LigScore	74	78	82
DrugScore	72	82	86
Cerius2/LUDI	67	80	85
X-Score	66	78	79
AutoDock	62	74	78
Cerius2/PMF	52	59	64
SYBYL/G-Score	42	58	66
SYBYL/ChemScore	35	47	51
SYBYL/D-Score	26	45	56

[a] Scoring functions are ranked by their success rates when only the best-scored conformation of each ligand is considered.

scoring functions will interest many researchers in this field. It needs to be emphasized though that the scoring functions implemented in SYBYL and Cerius2 may not always accurately reproduce their original approaches. Therefore, all our evaluations on these scoring functions are valid only to themselves and should not be extended to their original approaches.

Docking Accuracy. The most straightforward method for evaluating a scoring function in terms of docking accuracy is to inspect how closely the best-scored (or the lowest-energy) docked conformation predicted by this scoring function resembles the one observed in the experimental complex structure. Here, we define that a prediction is successful if the rmsd value of the best-scored conformation is less than or equal to 2.0 Å from the experimentally observed conformation. This is the default criterion used throughout this paper unless specified. Success rates of all 11 scoring functions tested in our study are listed in Table 2. If using the AutoDock scoring function (success rate = 62%) as reference, one can see that six scoring functions, i.e., PLP, F-Score, LigScore, DrugScore, LUDI, and X-Score, give better results (success rates ranging from 66% to 76%) while the other four scoring functions, i.e., PMF, G-Score, ChemScore, and D-Score, do not (success rates ranging from 26% to 52%). Success rates of all 11 scoring functions under other rmsd criteria (1.0−3.0 Å) are also listed in Table 2. It is not surprising that the success rates of all the scoring functions drop under a tighter criterion and increase under a looser criterion. However, rankings of these scoring functions generally do not change during this process. Notably, PLP, F-Score, LigScore, and DrugScore perform the best in this test. Their success rates are all above 70% with rmsd ≤ 2.0 Å and can still stay above 50% even with rmsd ≤ 1.0 Å. Considering the remarkable diversity presented in the test set, the performance of these scoring functions is very impressive.

When analyzing the results from a molecular docking job, one may also want to examine other conformations rather than only the best-scored one. In our study, the 101 conformations of each ligand are clustered by using an rmsd criterion of 2.0 Å. The best-scored conformation in each cluster is then selected as the representative of its cluster. In other words, this step selects the best-scored yet nonduplicate conformations in the ensemble. Success rates of all of the 11 scoring functions are then recalculated by considering the best two conformations or the best three conformations of each ligand. The results are summarized in Table 3. From there, one can see that the success rates of almost all of the scoring functions will improve considerably if the second or even the third best-scored conformation is taken into account. This can also be interpreted as that when the true conformation is missed as the very best one in binding score, it will probably appear as the second or the third best one. If considering the best three conformations in each case, five scoring functions, i.e., PLP, F-Score, LigScore, DrugScore, and LUDI, have success rates higher than 80%. So it is a good idea for a docking program to output multiple docked conformations for analysis.

To further evaluate these scoring functions, another test we have conducted is to classify the 100 complexes into subsets according to the chemical nature of their protein−ligand interactions and then to check the success rate of each scoring function for these subsets. The classification is aided by using X-Score (see Appendix): For any given protein−ligand complex, if the contribution of the H-bond term in X-Score is 50% larger than the hydrophobic term, it is classified as the "*hydrophilic*" type. If the contribution of the hydrophobic term is 50% larger than the H-bond term, it is classified as the "*hydrophobic*" type. Otherwise, the complex is considered to have mixed hydrophilic and hydrophobic factors in the protein−ligand interaction and thus is classified as the "*mixed*" type. Note that X-Score is used for this classification process because it is the only one with open source codes, so we can analyze the hydrophobic and the hydrophilic terms conveniently. All three subsets are listed in Table 1. The success rates of all of the 11 scoring functions on these three subsets are summarized in Table 4. Generally speaking, higher success rates are observed for the hydrophilic subset. Seven scoring functions, i.e., PLP, F-Score, LigScore, DrugScore, LUDI, X-Score, and AutoDock, achieve success rates above 70%. This is not surprising because all of these scoring functions have sufficient consideration of hydrogen bonding. When the hydrophobic factor in protein−ligand interactions takes a larger share, such as the mixed subset and the hydrophobic subset, some of these scoring functions perform less satisfactorily, such as DrugScore, LUDI, X-Score, and AutoDock. This is also not surprising, since

Table 4. Success Rates of 11 Scoring Functions on Different Subsets of Complexes

	success rate (%)			
scoring function[a]	overall (100)	hydrophilic (44)	mixed (32)	hydrophobic (24)
Cerius2/PLP	76	77	78	71
SYBYL/F-Score	74	75	75	71
Cerius2/LigScore	74	77	75	67
DrugScore	72	73	81	58
Cerius2/LUDI	67	75	66	54
X-Score	66	82	59	46
AutoDock	62	73	53	54
Cerius2/PMF	52	68	44	33
SYBYL/G-Score	42	55	34	29
SYBYL/ChemScore	35	32	34	42
SYBYL/D-Score	26	23	28	29

[a] Scoring functions are ranked by their overall success rates.

hydrophobic interactions are nonspecific and nondirectional and thus are more difficult to be characterized. What is surprising is that certain scoring functions, i.e., PLP and F-Score, are able to maintain their success rates across all three subsets. Scoring functions of this kind are definitely more welcome in molecular docking applications.

In this docking test, the six relatively successful scoring functions, compared to the AutoDock scoring function, are all empirical scoring functions except DrugScore. They typically have well-balanced considerations of polar and nonpolar, enthalpic and entropic factors in protein−ligand binding. Another common feature shared by these scoring functions is that they are all calibrated with various sets of protein−ligand complexes. The slightly inferior performance of LUDI and X-Score in this test can be understood because, unlike the other four, they are originally developed to reproduce the binding affinities of protein−ligand complexes rather than their structures. For example, both LUDI and X-Score use very simple distance and angular functions in their equations, which are based more on chemical intuition rather than a statistical analysis of a large number of experimental structures. The atomic radii used by them are also borrowed from other existing force fields rather than being independently derived. Moreover, we point out that the hydrophobic term in these two scoring functions needs to be largely improved because the overall performance of these two scoring functions are pulled back by their relatively poor performance in the hydrophobic and the mixed subsets. Once these aspects are fully optimized, LUDI and X-Score will probably catch up with other more successful scoring functions in this docking test.

DrugScore, which is a knowledge-based potential of mean force approach, also performs very well (success rate = 72%). PMF approaches are different from other scoring methods by deriving potentials through interpreting inverse Boltzmann distributions from a large number of experimental structures. However, DrugScore uses an equation combining pairwise potentials and molecular surface based potentials. The introduction of molecular surfaces is supposed to capture the hydrophobic effect more effectively, which is a common practice witnessed in empirical scoring functions. Thus, the boundary between DrugScore and empirical scoring functions is actually blurred. In comparison, the PMF approach by Muggue et al. yields a lower success rate (52%) in this test. According to this approach, protein−ligand interactions are expressed as a sum of pure distance-dependent pairwise potentials. Our opinion is that pairwise potentials may not be as effective as surface-based algorithms for describing the hydrophobic effect in protein−ligand binding. The observation that Muggue's PMF approach performs more poorly than DrugScore for the hydrophobic and the mixed subsets seems to support this remark.

Generally speaking, force field based scoring functions, i.e., AutoDock (success rate = 62%), G-Score (success rate = 42%), and D-Score (success rate = 26%), are less successful in this test. One frequently overlooked fact is that classical force fields are typically not developed for describing intermolecular interactions. Therefore, truncating the noncovalent part of a force field and then applying it to protein−ligand binding, such as D-Score, is not expected to give very good results, although it was almost the standard practice in early years. After some special reparametrization, the performance of force field based scoring functions can definitely be improved, such as what has been seen in the case of AutoDock and G-Score. However, the hydrophobic effect still cannot be adequately formularized in a force field equation. One can see in Table 4 that without exception all of the three force field based scoring functions perform more poorly for the hydrophobic subset and the mixed subset. Although some approaches have emerged in which a separate PB/SA or GB/SA term is introduced into a force field equation to compute solvation energy, they are usually coupled with extensive molecular dynamic samplings and have not been widely applied to molecular docking studies. Another practical problem associated with force field based scoring functions is the computation of the so-called electrostatic interaction energy. To compute this energy, atom-centered partial charges must be assigned to both the protein and the ligand. Theoretical derivation of such a charge distribution in the solvent still remains a problem, especially for a large flexible molecule like a protein. The state-of-the-art solution for proteins is several sets of "template" charges derived from model systems. For ligands, there is a wide spectrum of schemes ranging from very simple empirical methods to high-level ab initio calculations. But this leads to another potential problem: should not the atomic charges on the protein and the ligand be derived by the same method so they can match each other? Yet another one is the dielectric constant. The binding pocket is more or less shielded from the bulk solvent, and thus, the electrostatic microenvironment inside it is supposed to be different from that of the bulk solvent. People have been using two, four, eight, or a distance-dependent dielectric constant to compute the electrostatic interactions between the complex. As described in the Methods section, we use AMBER unit-atom charges for proteins, MMFF94 charges for ligands, and a distance-dependent dielectric constant in our computation. This scheme is a reasonable one. Applying other schemes may improve the performance of AutoDock and D-Score in our test, but we do not expect a major improvement by doing so. It is interesting to note that some force field based scoring functions, such as G-

Score, have chosen to avoid this electrostatic term by using other alternatives.

The poor performance of ChemScore (success rate = 35%) in this test is not so easy to understand because it has an equation very similar to those of F-Score, LUDI, and X-Score. It is definitely not because something has gone wrong in the application of this scoring function because other scoring functions, such as F-Score, in SYBYL/CScore perform well under exactly the same procedure. More details about this scoring function, for example, how it is programmed in SYBYL, need to be revealed before a fair comment can be made.

Consensus Scoring. Combining multiple scoring functions, known as consensus scoring, has been demonstrated in various virtual database screening studies to be an effective way for getting improved hit rates.[24,25] These studies inspire us to investigate if consensus scoring also works for molecular docking, i.e., identifying the correct bound conformation of a given ligand from many computer-generated decoys. In our study, we have tested all the possible double and triple combinations of the six relatively successful scoring functions, i.e., F-Score, LigScore, PLP, DrugScore, LUDI, and X-Score. The "rank-by-rank" strategy described in our previous study[26] is adopted here to perform consensus scorings because the results given by these six scoring functions have different units. Each scoring function involved in the given consensus scoring scheme is applied to rank all the conformations of the ligand. The final rank of a certain conformation is its average rank received from all the scoring functions. For example, assume a consensus scoring scheme combines scoring functions A and B. If a certain conformation is ranked at the fourth position among all 101 candidates by scoring function A and ranked at the sixth position by scoring function B, its final rank is (4 + 6)/2 = 5. In our study, the best possible rank is 1 and the worst one is 101. The best-ranked conformation of each ligand is then compared to the experimentally observed one to calculate the success rate of each consensus scoring scheme.

Success rates of all the consensus scoring schemes tested in our study are summarized in Table 5. Compared to individual scoring functions, whose success rates range from 66% to 76%, double scoring schemes produce success rates between 76% and 80%, while triple scoring schemes produce success rates between 80% and 84%. So it is clear that consensus scoring is also generally more effective than single scoring for molecular docking tasks. Another observation is that which scoring functions are actually included in the consensus scoring scheme seems to be less crucial. All of the double scoring schemes give approximately the same level of success rates and so do all of the triple scoring schemes. This again confirms our previous speculation[26] that the nature of consensus scoring is multiple sampling, which explains why consensus scoring has been seen repeatedly to outperform single scoring in various studies. It is well-known that repeated measurements reduce the noise in data collection and thus give more converged and more accurate results. In a statistical sense, the effectiveness of this practice is determined by the number of repeated measurements. This is exactly what one can see in Table 5. On average, the success rates of triple scoring schemes are higher than those of double scoring schemes.

Table 5. Success Rates of Various Consensus Scoring Schemes[a]

consensus scoring scheme	success rate (%)
double scoring	
DrugScore + LigScore	80
DrugScore + F-Score	79
DrugScore + LUDI	79
LigScore + PLP	79
LigScore + F-Score	79
LigScore + X-Score	78
DrugScore + PLP	78
LigScore + LUDI	77
PLP + X-Score	77
PLP + LUDI	77
DrugScore + X-Score	77
PLP + F-Score	76
triple scoring	
LigScore + DrugScore + F-Score	84
LigScore + DrugScore + PLP	84
LigScore + DrugScore + LUDI	83
LigScore + PLP + LUDI	82
DrugScore + PLP + F-Score	82
DrugScore + PLP + X-Score	82
LigScore + DrugScore + X-Score	81
LigScore + PLP + F-Score	80
LigScore + PLP + X-Score	80
DrugScore + PLP + LUDI	80

[a] Since F-Score, LUDI, and X-Score have very similar equations and thus may be less complementary to one other, we do not allow any two of them to appear simultaneously in one consensus scoring scheme.

In conclusion, although consensus scoring does not provide a better understanding of protein−ligand interactions, our results demonstrate that it is still a practical strategy for obtaining more reliable results in molecular docking studies. Therefore, we recommend that consensus scoring should be applied whenever possible. One thing to keep in mind is that each individual scoring function in a consensus scoring scheme should itself be a good one. A blind combination of some arbitrary scoring functions does not necessarily lead to better results.

Binding Affinity Prediction. Predicting the correct binding mode of a ligand is only one aspect of molecular docking. In a practical application, such as virtual database screening, large numbers of molecules are docked onto the target and then the top "hits" are selected according to their binding scores. Therefore, an equally important aspect of a scoring function is how well it can predict real binding affinities.

We have examined all of the 11 scoring functions to see the correlations between their scores and the experimentally measured binding affinities of the 100 protein−ligand complexes in the test set. Note that for virtual database screening approaches, this correlation does not have to be linear. As long as a scoring function can provide the correct rankings of candidate molecules, it will work perfectly. The Spearman correlation coefficient (R_s),[27] which calculates the correlation between two sets of rankings, is a proper quantitative measurement for this purpose.

$$R_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

Table 6. Correlations between Binding Scores and Experimentally Determined Binding Affinities Given by 11 Scoring Functions

	Spearman correlation coefficient (r_s) based on	
scoring function[a]	the experimentally observed conformations	the best-scored conformations
X-Score	0.660	0.698
Cerius2/PLP	0.592	0.607
DrugScore	0.587	0.601
SYBYL/G-Score	0.569	0.531
SYBYL/D-Score	0.475	0.488
SYBYL/ChemScore	0.431	0.435
Cerius2/LUDI	0.430	0.456
Cerius2/PMF	0.369	0.367
Cerius2/LigScore	0.363	0.418
SYBYL/F-Score	0.283	0.253
AutoDock	0.141	0.423

[a] Scoring functions are ranked by correlation coefficients that are calculated by using the experimentally observed conformation of each ligand.

Here, d_i is the ranking difference of the ith complex under two criteria. In our case, one criterion is the experimental binding affinity and the other one is the binding score computed by a scoring function. In theory, the Spearman correlation coefficient falls between −1 and +1, where +1 corresponds to a perfect correlation, −1 corresponds to a perfect inverse correlation, and zero corresponds to total disorder. All of the 11 scoring functions in our study have been applied to calculate the binding score of each complex in the test set using its experimental structure. The Spearman correlation coefficients given by these scoring functions are summarized in Table 6.

In Table 6, one can see that the performance of these scoring functions in this test is generally less encouraging than their performance in the previous docking test. Among all the scoring functions, X-Score gives the best agreement between its scores and the experimental binding affinities with a correlation coefficient of 0.66. PLP, DrugScore, and G-Score rank at the second, third, and fourth places, respectively, with correlation coefficients ranging between 0.57 and 0.59. The score−affinity correlations given by these four scoring functions are shown in Figure 3. All of the other scoring functions give very poor score−affinity correlations. In Table 6, we also list the correlation coefficients yielded by each scoring function when the best-scored conformation of each ligand is used instead of the true conformation in computation. All of these four scoring functions, i.e., X-Score, PLP, DrugScore, and G-Score, are able to maintain their level of accuracy in this

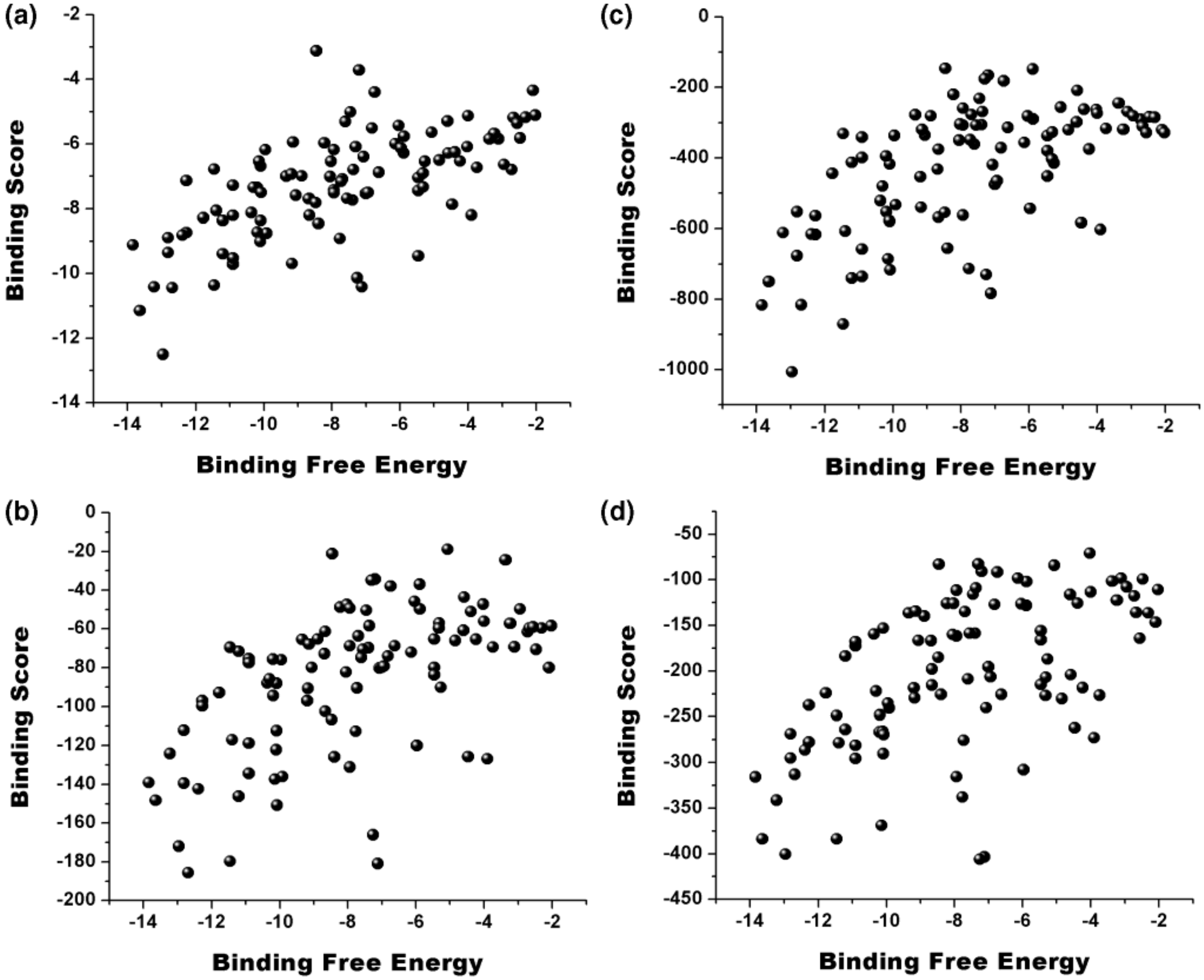

Figure 3. Correlations between experimentally determined binding free energies (kcal/mol) and binding scores of 100 protein−ligand complexes given by (a) X-Score (R_s = 0.660), (b) PLP (R_s = 0.592), (c) DrugScore (R_s = 0.587), and (d) G-Score (R_s = 0.569).

situation. This is logical because these scoring functions are able to identify the experimentally observed conformation as the best-scored one for a large number of the complexes in the test set. Even when the observed conformation is missed as the best-scored one in some cases, as we have discussed earlier, its score is usually not far from the best-scored one and thus will not significantly affect the final score−affinity correlation of the entire test set. These correlation coefficients computed by using best-scored conformations may provide an even more realistic picture of the performance of those scoring functions because in a real molecular docking application the "true" bound conformations of ligand molecules are unknown.

The better performance of X-Score in this test can be understood because it is calibrated to reproduce the binding affinities of a variety of protein−ligand complexes, some of which are included in the test set used in this study. PLP, DrugScore, and G-Score should be appreciated for their performance because they are originally developed as "docking functions" rather than "scoring functions". Comparing the results in Tables 2 and 6, one can see that some other scoring functions perform unsatisfactorily in scoring even though they are very good at docking, such as F-Score and LigScore. Therefore, as a practical strategy for improving the hit rates in virtual database screening, one can use a good "docking function", such as PLP and F-Score, to obtain binding modes of all the ligand molecules first and then apply a good "scoring function", such as X-Score, to rerank them. However, scoring function developers should pay more attentions to the scoring functions that are good at both docking and scoring. Our results suggest that this is possible, since some scoring functions, such as X-Score, PLP, and DrugScore, have already demonstrated reasonable compromises between docking and scoring.

Here, we have described the performance of 11 scoring functions in reproducing the binding affinities of a variety of protein−ligand complexes. Another way to test a scoring function for its ability to predict binding affinities is to dock a library of molecules onto selected target proteins and then to examine how well this scoring function ranks the known active compounds to the top. This type of test has been seen in several previous studies.[17−19,24,25] We have not conducted this test in our study because it is unaffordable to perform the exhaustive conformational sampling as we have described earlier for a large number of ligand molecules.

Funnel-Shaped Energy Surface. We have also tested all of the 11 scoring functions for their abilities to construct a funnel-shaped energy surface of protein−ligand complexation, one aspect that has not been adequately discussed before. The concept of a funnel-shaped energy surface was originally proposed in protein folding studies and has been widely accepted nowadays.[28] It is reasonable to speculate that receptor−ligand complexation is actually similar because, regardless of its initial position, the ligand can always find the binding pocket on the receptor and bind in a unique way. It will be very difficult to understand this process if the energy surface of receptor−ligand complexation is not funnel-shaped but rugged. In our opinion, besides pinning out the correct location of the global minimum, an ideal scoring function should also be able to give a descriptive funnel-shaped energy surface that does not have many false minima to impair the efficiency of conformational sampling. For molecular docking tasks, this feature may be just as important as the first one.

Table 7. Correlations between Binding Scores and rmsd Values Given by 11 Scoring Functions

	cumulative occurrence of Spearman correlation coefficient (R_s)				
scoring function[a]	≥0.00	≥0.20	≥0.40	≥0.60	≥0.80
X-Score	99	96	77	53	19
DrugScore	93	87	73	46	21
AutoDock	97	86	71	42	12
Cerius2/PLP	94	84	67	39	13
SYBYL/D-Score	92	83	67	39	9
Cerius2/PMF	89	79	61	38	21
Cerius2/LUDI	99	90	66	37	8
SYBYL/F-Score	97	93	72	34	9
SYBYL/G-Score	93	79	56	28	6
Cerius2/LigScore	86	71	49	26	4
SYBYL/ChemScore	86	67	41	16	1

[a] Scoring functions are ranked by the occurrence of R_s values larger than or equal to 0.60.

In our study, the 101 docked conformations of each ligand molecule represent 101 spots on the protein−ligand complexation energy surface. Note that the shape of this energy surface is defined by the scoring function. The question is how to detect if a funnel exists on this energy surface. Here, we examine the correlation between the rmsd values and the binding scores of all the docked conformations. Assuming that the funnel bottom corresponds to the experimental complex structure, one would expect that a lower score is associated with a smaller rmsd value and vice versa. Of course, this kind of rmsd−score correlation alone may not sufficiently define a funnel-shaped energy surface because of the multidimensional nature of an energy surface. But it must be one of the necessary features of a funnel-shaped energy surface. In our study, we have calculated this rmsd−score correlation for each ligand by applying each of the 11 scoring functions. For each ligand, all of the 101 docked conformations are considered in the calculation. The Spearman correlation coefficient (R_s) is again adopted as a quantitative measurement. The results are summarized in Table 7. For the convenience of analysis, cumulative occurrences of R_s values are provided.

In Table 7, one can see that X-Score gives the best rmsd−score correlations among all of the scoring functions under our test. It gives R_s values better than or equal to 0.60 for 53% cases. Considering that X-Score does not always identify the experimental conformation as the best-scored one (success rate = 66%), this rate is even more impressive. Another strong competitor is DrugScore, which is almost equally good at giving rmsd−score correlations. It is interesting to note that F-Score and LigScore, which perform well in our docking test, yield relatively poor results in this test. Figure 4 shows one example of how different scoring functions can produce various rmsd−score correlations for the same protein−ligand complex. In this particular case (PDB entry 1CBX), all of the six scoring functions are able to identify the experimentally observed conformation (or a really close one) as the best-scored one. But

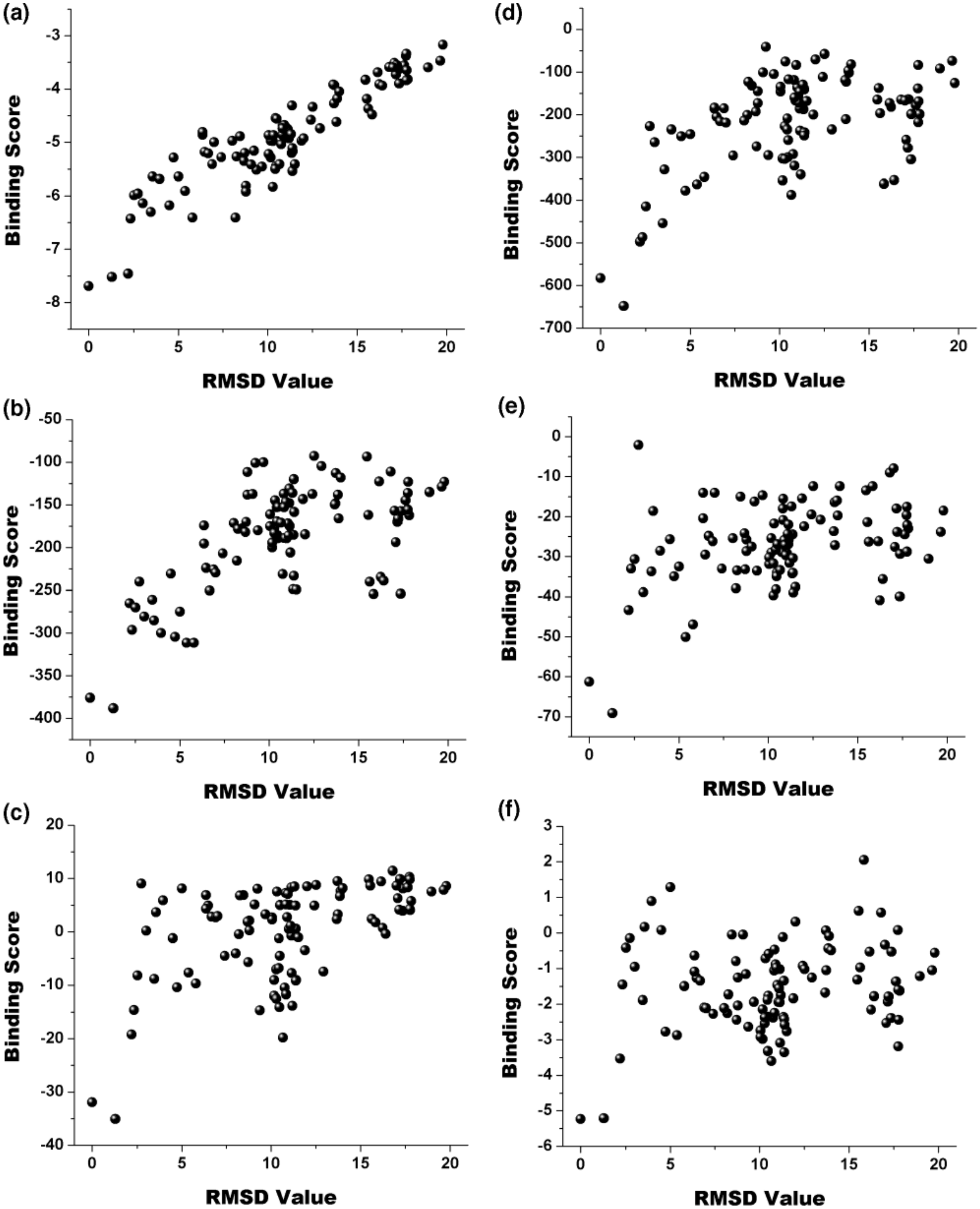

Figure 4. Correlations between rmsd values (Å) and binding scores of the 101 docked conformations of PDB entry 1CBX given by (a) X-Score ($R_s = 0.877$), (b) DrugScore ($R_s = 0.548$), (c) F-Score ($R_s = 0.478$), (d) LUDI ($R_s = 0.425$), (e) PLP ($R_s = 0.328$), and (f) LigScore ($R_s = 0.135$).

X-Score demonstrates a clear rmsd–score correlation throughout the entire conformational ensemble, which gives a picture of a wide smooth funnel on the protein–ligand complexation energy surface. DrugScore also tends to give such a correlation. For the other four scoring functions, i.e., F-Score, LUDI, PLP, and LigScore, such a correlation is generally not observed between 4 and 20 Å. Their scores will drop sharply only when the conformations are really close to the true one. In addition, their energy surfaces are more rugged because of the existence of some false minima, such as the one around 10 Å.

It is reasonable to expect that scoring functions that are able to give such an rmsd–score correlation will lead to a faster convergence to the global minimum when they are applied to conformational sampling. Or in other words, given the same amount of effort spent on conformational sampling, such scoring functions will have better chances to find the global minimum. In our test, those scoring functions showing better rmsd–score correlations do not necessarily achieve better success rates in identifying the correct docked conformations. This is because all the scoring functions are applied to pregenerated conformational ensembles, and thus, their efficiency, rather than their accuracy, in molecular docking is not explicitly revealed. This is similar to the fact that even an F-1 professional will not make too much difference if racing only on straight lanes. We expect that implementing a scoring function like X-Score or DrugScore into a molecular docking program will make conformational sampling smarter and thus lead to more efficient molecular dockings.

Analysis of the Outliers. There are 7 complexes in our test set for which none of the 11 scoring functions is able to pick out the correct conformation within an rmsd threshold of 2.0 Å. They are PDB entries 1CLA, 3CLA, 4CLA, 1RGL, 1TET, 1THA, and 1TLP. An analysis of these protein–ligand complexes may help to reveal the shortcomings embedded in today's scoring functions.

Among these outliers, 1CLA, 3CLA, and 4CLA are complexes formed between chloramphenicol and type III chloramphenicol acetyltransferases (3CLA is the wild type of chloramphenicol acetyltransferase; 1CLA is a S148A mutant; while 4CLA is a L160F mutant). In these three complex structures, one remarkable feature is that an entire layer of water molecules exist on the protein–ligand binding interface (see Figure 5a). None of the H-bonding groups on the ligand, i.e., two hydroxyl groups, one amide group, and one nitro group, is in direct contact with the protein. Instead, their interactions with the protein are mediated by some water molecules. The positions of those water molecules are conserved in all of the three complex structures. As described in Methods, in our study all of the water molecules are removed from complex structures because none of the 11 scoring functions can really handle such water-mediated protein–ligand interactions. This explains their failures in the cases of 1CLA, 3CLA, and 4CLA: after the removal of those water molecules, the experimentally observed conformation is not likely to be favored because it is somewhat suspended in the binding pocket. Instead, those scoring functions tend to find other locations for the ligand molecule where it can form direct interactions with the protein. For example, the best-scored conformation predicted by F-Score is shown in Figure 5b. This conformation is not quite native-like because it is not even bound in a cavity. The best-scored conformation predicted by DrugScore, LigScore, and PLP is shown in Figure 5c. This one is interesting in the sense that the ligand is placed inside a small hole. However, as revealed in the crystal complex structure, that hole is filled with water molecules and is not an alternative binding pocket. The inability to consider the water-mediated protein–ligand interactions is a major defect in today's scoring functions because these interactions are frequently observed in protein–ligand complexation. Rarey et al. have attempted to refine FlexX by placing discrete water molecules inside the binding pocket while docking the ligand molecule.[29] Although they have not observed significantly improved results by doing so, we believe this is a correct direction to pursue.

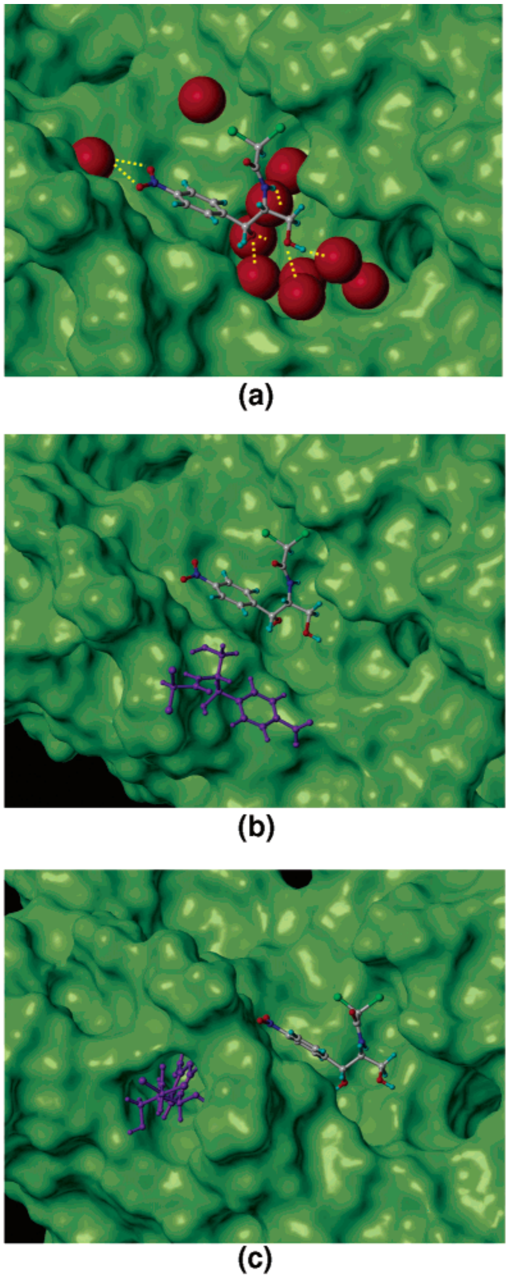

Figure 5. Type III chloramphenicol acetyltransferase in complex with chloramphenicol (PDB entry 3CLA). Chloramphenicol is shown in CPK color with ball-and-stick model. (a) Water molecules on protein–ligand binding interface are shown in red with space-filling model. Dashed yellow lines represent possible H-bonds. (b) Predicted bound conformation by F-Score (in violet, rmsd = 11.1 Å). (c) Predictied bound conformation by DrugScore, LigScore, and PLP (in violet, rmsd = 12.7 Å).

Complex 1THA, i.e., transthyretin in complex with 3,3′-diiodo-L-thyronine, reveals another typical tough situation. In this case, the ligand resides in a shallow groove instead of a well-defined pocket (see Figure 6). Not many specific interactions exist between the ligand and the protein, which may explain the moderate protein–ligand binding affinity ($-\log K_d = 5.35$). The bound conformation predicted by DrugScore is shown

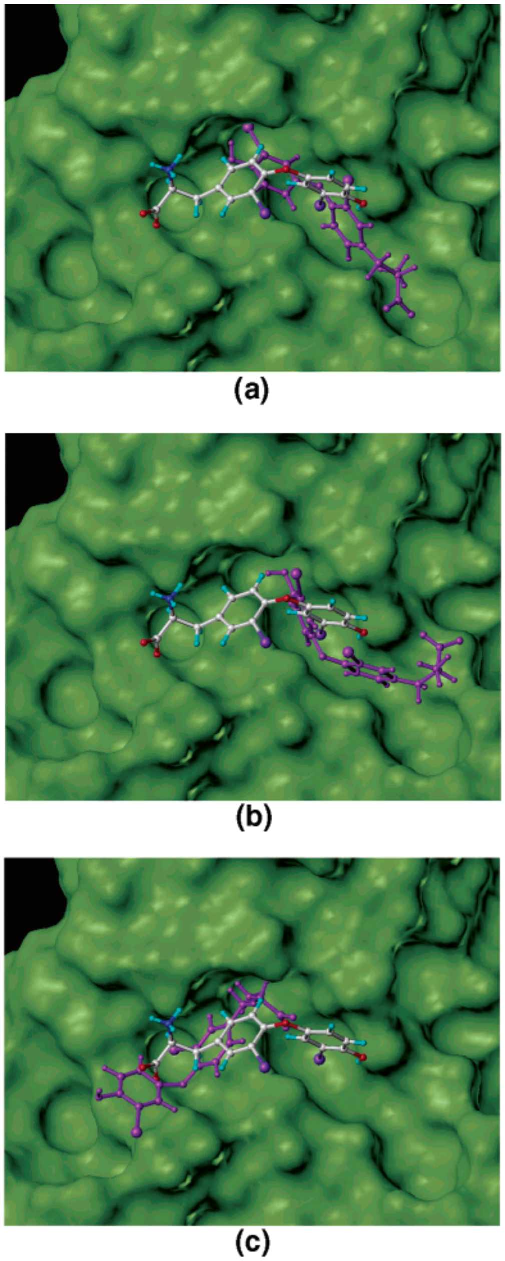

Figure 6. Transthyretin in complex with 3,3′-diiodo-L-thyronine (PDB entry 1THA). 3,3′-Diiodo-L-thyronine is shown in CPK color with ball-and-stick model; (a) predicted bound conformation by DrugScore (in violet, rmsd = 9.1 Å); (b) predicted bound conformation by F-Score and LigScore (in violet, rmsd = 9.9 Å); (c) predicted bound conformation by PLP (in violet, rmsd = 8.8 Å).

in Figure 6a; the one predicted by F-Score and LigScore is shown in Figure 6b; while the one predicted by PLP is shown in Figure 6c. All of these predicted bound conformations are located inside the same groove and partially overlap the experimentally observed conformation. But they all shift to the left side or the right side. In addition, the orientation of the ligand molecule is totally wrong by all of these four scoring functions. It appears that on a relatively flat surface it is simply more difficult to identify a specific binding area. The failures in the case of 1RGL and 1TET are also due to the same reason. It is clear that today's scoring functions still need to be refined to handle such protein−ligand complexes correctly.

Conclusions

We have tested 11 popular scoring functions on 100 protein−ligand complexes and have evaluated several aspects of their performance. Unlike previous studies of docking/scoring methods, in which scoring functions are always tested in the context of some molecular docking programs, we separate the docking procedure and the scoring procedure by performing a fairly complete conformational sampling first and then applying the scoring functions. In this way, all of the scoring functions are evaluated on the same ground, and the results are least affected by any particular docking program.

Among all the scoring functions we have tested, F-Score, LigScore, PLP, LUDI, DrugScore, and X-Score exhibit better docking accuracy than the energy function implemented in the AutoDock program. These scoring functions are able to identify the experimentally observed conformation among a large number of computer-generated decoys for 66−76% of the complexes in the test set. Considering the remarkable diversity presented in the test set, this level of success rate is impressive. Moreover, combining any two or three of these six scoring functions into a consensus scoring scheme further improves the success rate to nearly 80% or even higher. These results suggest that, given an adequate conformational sampling, the performance of today's best scoring functions is totally acceptable for molecular docking tasks. Thus, one may want to reexamine the notion that scoring function is the primary problem in molecular docking. Docking program developers should pay more attention to refining the conformational sampling methods. This request becomes even more imperative for virtual database screening because each molecule has to be processed in a very short time in such applications. As we have pointed out, conformational sampling can be more efficient if it is guided by a scoring function that is able to construct a funnel-shaped energy surface for protein−ligand complexation, such as X-Score and DrugScore. Such scoring functions will most likely lead to a faster convergence to the global minimum in conformational sampling.

However, our study by no means suggests that scoring functions do not need any further improvement. Our tests reveal that binding affinity prediction remains a serious problem. For the 100 complexes in the test set, only X-Score, DrugScore, PLP, and G-Score give moderate correlations between their binding scores and experimentally determined protein−ligand binding affinities. Unable to predict binding affinities accurately will be a major problem for virtual database screening because true hits may still be missed even when they are correctly docked. An ideal scoring function for molecular docking tasks should be good at both "docking" and "scoring". It is encouraging to see that some of the scoring functions under our test, such as X-Score, DrugScore, and PLP, have demonstrated a reasonable balance between these two aspects. We expect that more scoring functions of this kind will appear in the future.

Acknowledgment. The authors are grateful to Cap CURE Foundation (2001 Young Investigator Award to Dr. Renxiao Wang) and the Department of Defense

(Grant No. DOD DAMD17-93-V-3018 to Dr. Shaomeng Wang) for their financial support.

Appendix

The 11 scoring functions tested in our study are briefly described below, including the scoring function implemented in the AutoDock program, four scoring functions from the LigFit module in Cerius2, version 4.6 (LigScore, PLP, PMF, and LUDI), four scoring functions from the CScore module in SYBYL, version 6.8 (F-Score, G-Score, D-Score, and ChemScore), and another two stand-alone scoring functions (DrugScore, version 1.2, and X-Score, version 1.0). More details of these scoring functions can be found in the cited references.

(1) AutoDock. In AutoDock, the overall docking energy of a given ligand molecule is expressed as the sum of intermolecular interactions between the complex and the internal steric energy of the ligand.

$$\begin{aligned} E_{\text{dock}} &= E_{\text{vdw}} + E_{\text{H-bond}} + E_{\text{electrostatic}} + E_{\text{internal}} \\ &= \sum_{\text{protein}} \sum_{\text{ligand}} \left(\frac{A_{ij}}{{d_{ij}}^{12}} - \frac{B_{ij}}{{d_{ij}}^{6}} \right) + \sum_{\text{protein}} \sum_{\text{ligand}} E(t) \times \\ &\quad \left(\frac{C_{ij}}{{d_{ij}}^{12}} - \frac{D_{ij}}{{d_{ij}}^{10}} \right) + \sum_{\text{protein}} \sum_{\text{ligand}} 332.0 \frac{q_i q_j}{\epsilon(d_{ij})\, d_{ij}} + \\ &\quad \left\{ \sum_{\text{ligand}} \left(\frac{A_{ij}}{{d_{ij}}^{12}} - \frac{B_{ij}}{{d_{ij}}^{6}} \right) + \sum_{\text{ligand}} E(t) \left(\frac{C_{ij}}{{d_{ij}}^{12}} - \frac{D_{ij}}{{d_{ij}}^{10}} \right) + \right. \\ &\quad \left. \sum_{\text{ligand}} 332.0 \frac{q_i q_j}{4 d_{ij}\, d_{ij}} \right\} \end{aligned}$$

Here, the first three terms are in vacuo force field energies for intermolecular interactions: a Lennard-Jones 12-6 dispersion−repulsion term, a directional 12-10 hydrogen bonding term, where $E(t)$ is an angular weight factor, and a Coulombic electrostatic potential. The fourth term accounts for the internal steric energy of the ligand molecule, which also consists of these three elements. All these terms are taken from the early version of AMBER force field,[30] but the parameters used in these terms are especially tailored to yield the best results for molecular docking tasks.[2]

(2) Cerius2/LigScore. This scoring function is implemented in the LigFit module of the Cerius2 software. To the best of our knowledge, it is not published anywhere. According to the description in the Cerius2 user manual, it is a sum of three terms,

$$\mathrm{p}K_i = A - (B)(\mathrm{vdW}) + (C)(C_{+\mathrm{pol}}) - (D)(\mathrm{Totpol}^2)$$

where vdW is a softened Lennard-Jones 6-9 potential, $C_{+\mathrm{pol}}$ is a count of the buried polar surface area between the complex involving attractive protein−ligand interactions, Totpol^2 is the square of the buried polar surface area between the complex involving both attractive and repulsive protein−ligand interactions. In our study, we choose the CFF force field parameters with the "exact pairwise" option to calculate the vdW term. The Cerius2 user manual mentions that this scoring function was calibrated by fitting to known protein−ligand binding affinities. Thus, it falls into the empirical scoring function category.

(3) Cerius2/PLP. This empirical scoring function[6,7] is a sum of pairwise linear potentials between ligand and protein heavy atoms with parameters dependent on interaction type. It can be expressed conceptually as

$$E_{\text{total}} = E_{\text{H-bond}} + E_{\text{repulsion}} + E_{\text{contact}}$$

Ligand and protein heavy atoms are classified as hydrogen bond donors, acceptors, donor/acceptors, or nonpolar. Each pair of interacting atoms is then assigned one of the three interaction types: hydrogen bonding between donors and acceptors, repulsive donor−donor and acceptor−acceptor contacts, and generic dispersion of other contacts. Both the hydrogen bonding and repulsive terms are modulated by a scaling factor that imparts a crude distance and angular dependence. Small (fluorine and metal ion), medium (carbon, oxygen, and nitrogen), and large (sulfur, phosphorus, chlorine, and bromine) atoms are assigned atomic radii of 1.4, 1.8, and 2.2 Å, respectively. These parameters are derived from interatomic distances observed from a large number of high-quality crystal structures.

In Cerius2, there are two versions of this scoring function, namely, PLP1 and PLP2. They use slightly different algorithms and parameters sets. In our study, we have tested both of them and found that they give comparable results in a statistical sense. Therefore, we only report the results given by PLP1 and use it to represent PLP throughout this paper.

(4) Cerius2/PMF. This potential of mean force (PMF) scoring function is based on the work of Muegge et al.,[12−14] who analyzed 697 protein−ligand complex structures from the Protein Data Bank and derived a set of distance-dependent interaction potentials for various atom pairs. Both enthalpic and entropic effects are assumed to be included implicitly in this potential. The protein−ligand interaction energy is then defined as a sum of potentials over all heavy atom pairs between the complex:

$$\mathrm{PMF} = \sum_{\text{protein}} \sum_{\text{ligand}} A_{ij}(d_{ij})$$

According to the authors' description in their paper, a distance cutoff of 6 Å for carbon−carbon interactions and a cutoff of 9 Å for all the other interactions are used in our study. It should be mentioned that the same approach is also implemented in SYBYL. To avoid duplication and confusion, in this paper we only report the results of the Cerius2 version of this scoring function.

(5) Cerius2/LUDI. This empirical scoring function is developed by Böhm[8,9] and is one of the pioneering empirical scoring functions. It dissects protein−ligand binding free energy as

$$\begin{aligned}\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{H-bond}} \sum_{\text{H-bond}} f(\Delta R,\Delta\alpha) + \\ \Delta G_{\mathrm{ionic}} \sum_{\text{ionic}} f(\Delta R,\Delta\alpha) + \\ \Delta G_{\mathrm{hydrophobic}} \sum_{\text{hydrophobic}} |A_{\mathrm{hydrophobic}}| + \\ \Delta G_{\mathrm{rotor}} N_{\mathrm{rotor}} + \Delta G_0\end{aligned}$$

The first two terms account for the hydrogen bonds formed between the complex, where “neutral” and “ionic” hydrogen bonds are treated separately. The contribution of each hydrogen bond is scaled by a distance- and angle-dependent function in order to penalize the deviations from an ideal geometry. The third term accounts for the hydrophobic effect, which calculates the buried hydrophobic molecular surface. The fourth term counts all the rotatable single bonds (rotors) in the ligand, which is supposed to be related to the torsional entropy loss of the ligand upon protein−ligand complexation. The last term is a regression constant. This scoring function was calibrated by fitting known dissociation constants of 87 protein−ligand complexes.

There are three different versions of this scoring function in Cerius2, namely, LUDI1, LUDI2, and LUDI3. According to the Cerius2 user manual, only LUDI2 has its weight factors before each term derived by fitting to experimentally determined binding affinities. In fact, we have tested all of the three versions in our study and found that LUDI2 indeed outperforms the other two versions. Therefore, we only report the results given by LUDI2 and use it to represent LUDI throughout this paper.

(6) SYBYL/F-Score. This empirical scoring function is based on the one implemented in the molecular docking program FlexX.[3] It is actually a twist of the LUDI scoring function:

$$\begin{aligned}\Delta G_{\mathrm{bind}} = \Delta G_{\mathrm{H-bond}} \sum_{\text{H-bond}} f(\Delta R,\Delta\alpha) + \\ \Delta G_{\mathrm{ionic}} \sum_{\text{ionic}} f(\Delta R,\Delta\alpha) + \\ \Delta G_{\mathrm{aromatic}} \sum_{\text{aromatic}} f(\Delta R,\Delta\alpha) + \\ \Delta G_{\mathrm{contact}} \sum_{\text{contact}} f(\Delta R) + \Delta G_{\mathrm{rotor}} N_{\mathrm{rotor}} + \Delta G_0\end{aligned}$$

Just like LUDI, the first two terms account for neutral and ionic hydrogen bonds. The third term calculates the interactions between aromatic groups on both sides, which are scaled by a distance- and angle-dependent function. The fourth term is a general distance-dependent potential for protein−ligand atom contacts. The fifth term is the standard rotor term accounting for torsional entropy loss. The last term is a regression constant. This scoring function was originally calibrated by reproducing the three-dimensional structures of 19 protein−ligand complexes.

(7) SYBYL/G-Score. This force field scoring function is based on the one implemented in the molecular docking program GOLD.[4] It is the sum of a protein−ligand complexation term, a hydrogen bonding term, and an internal energy term.

$$\begin{aligned}E_{\mathrm{total}} &= E_{\mathrm{complex}} + E_{\mathrm{H-bond}} + E_{\mathrm{internal}} \\ &= \sum_{\text{protein}} \sum_{\text{ligand}} \left(\frac{A_{ij}}{d_{ij}^{\,8}} - \frac{B_{ij}}{d_{ij}^{\,4}} \right) + \\ &\quad \sum_{\text{protein}} \sum_{\text{ligand}} [(E_{\mathrm{da}} + E_{\mathrm{ww}}) - (E_{\mathrm{dw}} + E_{\mathrm{aw}})] + \\ &\quad \left\{ \sum_{\text{ligand}} \left(\frac{C_{ij}}{d_{ij}^{\,12}} - \frac{D_{ij}}{d_{ij}^{\,6}} \right) + \sum_{\text{ligand}} \frac{1}{2} V \left[1 + \frac{n}{|n|} \cos(|n|\omega) \right] \right\}\end{aligned}$$

Here, the complexation term is calculated with a reparametrized Lennard-Jones 8-4 potential. The hydrogen bonding term is a sum of the individual energies from all the donor−acceptor pairs between the complex. The energy of each hydrogen bond is calculated with a complicated function considering the type and the geometry of the donor−acceptor pair. The internal energy of the ligand includes a dispersion−repulsion energy and a torsional energy, both of which are calculated according to the Tripos force field. This scoring function was originally calibrated by reproducing the three-dimensional structures of 100 protein−ligand complexes.

(8) SYBYL/D-Score. This scoring function is drawn from the molecular docking program DOCK.[1] It is a classical force field energy function, which sums van der Waals and electrostatic interactions between the complex:

$$E_{\mathrm{interaction}} = \sum_{\text{protein}} \sum_{\text{ligand}} \left(\frac{A_{ij}}{d_{ij}^{\,12}} - \frac{B_{ij}}{d_{ij}^{\,6}} + 332.0 \frac{q_i q_j}{\epsilon(d_{ij})\, d_{ij}} \right)$$

Here, the van der Waals energy is calculated with a Lennard-Jones 12-6 potential and the electrostatic energy is calculated with the Coulombic equation. The distance-dependent dielectric constant is used in our computation.

(9) SYBYL/ChemScore. This empirical scoring function is based on the work of Eldridge et al.,[5] which also has an equation similar to LUDI:

$$\begin{aligned}\Delta G = \Delta G_{\mathrm{H-bond}} \sum_{\text{H-bond}} f(\Delta R,\Delta\alpha) + \\ \Delta G_{\mathrm{metal}} \sum_{\text{metal}} f(\Delta R,\Delta\alpha) + \Delta G_{\mathrm{lipo}} \sum_{\text{lipo}} f(\Delta R) + \\ \Delta G_{\mathrm{rotor}} \sum_{\text{rotor}} f(P_{\mathrm{nl}},P'_{\mathrm{nl}}) + \Delta G_0\end{aligned}$$

The first term accounts for protein−ligand hydrogen bonding. Unlike LUDI, neutral and ionic hydrogen bonds are not differentiated here. The second term accounts for the coordinate bonding between the ligand and the metal ions residing inside the protein binding pocket. The third term accounts for the hydrophobic effect, which is calculated by summing a distance-dependent potential of all the hydrophobic atom pairs formed between the complex. The fourth term also counts rotors, but the contribution of each rotor is scaled by a complicated function to reflect the chemical nature of its environment. The fifth term is a regression constant. This scoring function was originally calibrated by reproducing the measured dissociation constants of 82 protein−ligand complexes.

(10) DrugScore. This potential of mean force approach is developed by Gohlke et al.[15] This scoring function combines distance-dependent pairwise potentials and solvent-accessible surface (SAS) dependent singlet potentials for protein and ligand atoms:

$$\Delta W = \gamma \sum_{\text{protein}} \sum_{\text{ligand}} \Delta W_{i,j}(r) + (1-\gamma) \times \left[\sum_{\text{ligand}} \Delta W_i(\text{SAS},\text{SAS}_0) + \sum_{\text{protein}} \Delta W_j(\text{SAS},\text{SAS}_0)\right]$$

Here, γ is an adjustable weight factor, normally set to 0.5. A set of 17 atom types are defined for both the protein and ligand atoms. The distance-dependent and surface-dependent potentials of each atom type are derived from 1374 protein−ligand complex structures.

The DrugScore program (version 1.2) used in our study is obtained directly from its authors. It provides three options for calculating protein−ligand interactions, which can be based on pure pairwise potentials, pure buried SAS potentials, or a combination of both. All of these three options have been tested in our study. We found that the pair−surface combination gives the best results, which is also consistent with the authors' original descriptions. Thus, we only report the results calculated by choosing this option and use it to represent DrugScore throughout this paper.

(11) X-Score. This empirical scoring function is recently developed in our group, which was formerly known as X-CScore.[11] Although it is originally designed for binding affinity estimation, we find that it also performs reasonably well for molecular docking tasks in our preliminary studies. It is worthwhile to investigate how its performance is compared to other scoring functions.

The X-Score program (version 1.0) is used in our study. Three individual scoring functions, HSScore, HPScore, and HMScore, are implemented in this version, which all include a van der Waals interaction term, a hydrogen bonding term, a hydrophobic effect term, a torsional entropy penalty, and a regression constant:

$$\text{HSScore} = (C_{\text{VDW},1})(\text{VDW}) + (C_{\text{H-bond},1})(\text{HB}) + (C_{\text{hydrophobic},1})(\text{HS}) + (C_{\text{rotor},1})(\text{RT}) + C_{0,1}$$

$$\text{HPScore} = (C_{\text{VDW},2})(\text{VDW}) + (C_{\text{H-bond},2})(\text{HB}) + (C_{\text{hydrophobic},2})(\text{HP}) + (C_{\text{rotor},2})(\text{RT}) + C_{0,2}$$

$$\text{HMScore} = (C_{\text{VDW},3})(\text{VDW}) + (C_{\text{H-bond},3})(\text{HB}) + (C_{\text{hydrophobic},3})(\text{HM}) + (C_{\text{rotor},3})(\text{RT}) + C_{0,3}$$

The van der Waals term (VDW) is calculated by a softened Lennard-Jones 8-4 potential. The hydrogen binding term (HB) calculates all of the hydrogen bonds between the complex with geometry-dependent functions. The rotor term (RT) calculates the number of "effective" rotors in the ligand molecule. These three terms are the same in all three scoring functions. As for the hydrophobic effect term, HSScore calculates the buried hydrophobic molecular surface of the ligand (HS), HPScore calculates pairwise hydrophobic atom contact potential (HP), while HMScore calculates the microscopic match of hydrophobic ligand atoms to the binding pocket (HM). All three scoring functions were calibrated by reproducing the known binding affinities of 200 protein−ligand complexes.

All of these three scoring functions have been tested in our study, and we found HSScore gives slightly better results than the other two. Thus, we use the results of HSScore to represent X-Score throughout this paper.

References

(1) Ewing, T. J. A.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(2) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(3) 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.

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

(5) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(6) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. B.; Fogel, L. J.; Freer, S. T. Molecular recognition of the inhibitor AG-1343 by HIV-1 Protease: Conformationally flexible docking by evolutionary programming. *Chem. Biol.* **1995**, *2*, 317−324.

(7) Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. In *Rational Drug Design: Novel Methodology and Practical Applications*; Parrill, L., Reddy, M. R., Ed.; American Chemical Society: Washington, DC, 1999; pp 292−311.

(8) Böhm, H. J. The development of a simple empirical scoring function to estimate the binding constant for a protein−ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 243−256.

(9) Böhm, H. J. Prediction of binding constants of protein ligands: a fast method for the polarization of hits obtained from de novo design or 3D database search programs. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 309−323.

(10) Wang, R.; Gao, Y.; Lai, L. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.

(11) Wang, R.; Lai, L.; Wang, S. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.* **2002**, *16* (2), 11−26.

(12) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(13) Muegge, I. A knowledge-based scoring function for protein−ligand interactions: Probing the reference state. *Perspect. Drug Discovery Des.* **2000**, *20*, 99−114.

(14) Muegge, I. Effect of ligand volume correction on PMF scoring. *J. Comput. Chem.* **2001**, *22*, 418−425.

(15) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(16) Ishchenko, A. V.; Shakhnovich, E. I. Small molecule growth 2001 (SMoG2001): An improved knowledge-based scoring function for protein−ligand interactions. *J. Med. Chem.* **2002**, *45*, 2770−2780.

(17) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

(18) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.

(19) Terp, G. E.; Johansen, B. N.; Christensen, I. T.; Jorgensen, F. S. A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein−ligand binding affinities. *J. Med. Chem.* **2001**, *44*, 2333−2343.

(20) Paul, N.; Rognan, D. ConsDock: A new program for the consensus analysis of protein−ligand interactions. *Proteins: Struct., Funct., Genet.* **2002**, *47*, 521−533.

(21) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, I. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235−242. http://www.rcsb.org/pdb/.

(22) *SYBYL*, version 6.8; Tripos Inc.; http://www.tripos.com/.

(23) *Cerius2*, version 4.6; Accelrys Inc.; http://www.accelrys.com/.

(24) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

(25) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281−295.

(26) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.

(27) Lane, D. M. *HyperStat Online Textbook*; http://davidmlane.com/hyperstat/index.html.

(28) Gruebele, M. Protein folding: The free energy surface. *Curr. Opin. Struct. Biol.* **2002**, *12*, 161−168 and the references therein.

(29) Rarey, M.; Kramer, B.; Lengauer, T. The particle concept: placing discrete water molecules during protein−ligand docking predictions. *Proteins: Struct., Funct., Genet.* **1999**, *4*, 17−28.

(30) Weiner, S. J.; Kollman, P. A.; Case, D. A. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765−784.

JM0203783