An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein−Ligand Complexes

Renxiao Wang, Yipin Lu, Xueliang Fang, 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 August 30, 2004

Fourteen popular scoring functions, i.e., X-Score, DrugScore, five scoring functions in the Sybyl software (D-Score, PMF-Score, G-Score, ChemScore, and F-Score), four scoring functions in the Cerius2 software (LigScore, PLP, PMF, and LUDI), two scoring functions in the GOLD program (GoldScore and ChemScore), and HINT, were tested on the refined set of the PDBbind database, a set of 800 diverse protein−ligand complexes with high-resolution crystal structures and experimentally determined K_i or K_d values. The focus of our study was to assess the ability of these scoring functions to predict binding affinities based on the experimentally determined high-resolution crystal structures of proteins in complex with their ligands. The quantitative correlation between the binding scores produced by each scoring function and the known binding constants of the 800 complexes was computed. X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP provided better correlations than the other scoring functions with standard deviations of 1.8−2.0 log units. These four scoring functions were also found to be robust enough to carry out computation directly on unaltered crystal structures. To examine how well scoring functions predict the binding affinities for ligands bound to the same target protein, the performance of these 14 scoring functions were evaluated on three subsets of protein−ligand complexes from the test set: HIV-1 protease complexes (82 entries), trypsin complexes (45 entries), and carbonic anhydrase II complexes (40 entries). Although the results for the HIV-1 protease subset are less than desirable, several scoring functions are able to satisfactorily predict the binding affinities for the trypsin and the carbonic anhydrase II subsets with standard deviation as low as 1.0 log unit (corresponding to 1.3−1.4 kcal/mol at room temperature). Our results demonstrate the strengths as well as the weaknesses of current scoring functions for binding affinity prediction.

INTRODUCTION

One of the key issues in structure-based drug discovery is prediction of the binding affinities of candidate ligand molecules to the target molecules. This is often referred to as the “scoring problem”. A whole spectrum of methods has been developed to solve this problem, and a group of approaches called “scoring functions” has gained popularity.[1−20] A scoring function computes the fitness score of a ligand molecule to its target protein based on a given complex structure. These empirical scoring functions do not require extensive conformational sampling and are very fast in binding affinity prediction, and some of them were also found to have reasonable accuracy. For these reasons, they have extensively been applied in high-throughput virtual library screening and detailed molecular docking studies. Not surprisingly, several recently developed molecular docking programs, such as FlexX,[21] GOLD,[22,23] and GLIDE,[24] have employed at least one scoring function as their internal scoring engine.

A number of scoring functions have been developed in the past decade, and a systematic and objective evaluation of their performance is clearly needed. Indeed, several recent studies[25−30] have offered comparative evaluations of a number of scoring functions in the context of molecular docking. For example, we recently tested 11 scoring functions on 100 selected protein−ligand complexes for their abilities to identify the experimentally determined binding mode in complex with their target proteins among an ensemble of computer-generated decoys.[31] Another study with a similar theme was also published very recently.[32] But docking the ligand molecule correctly onto the target protein is only part of the scoring problem. Ideally, the binding score computed for a ligand molecule to its protein should also reflect its experimentally measured binding affinity. In our previous study,[31] we also evaluated the correlations between their scores of those 11 scoring functions and the experimentally determined binding constants of the 100 protein−ligand complexes. It was found that X-Score, DrugScore, PLP in Cerius2, and G-Score in Sybyl provided better correlations than the other scoring functions in our test. However, that conclusion was obtained on a moderate-sized test set and should be treated with caution. A much larger test set is needed to arrive at a more objective conclusion in the performance of today’s scoring functions.

For many years, the availability of a large and high-quality set of high-resolution, experimentally determined three-dimensional (3D) structures of proteins in complex with ligands and their experimentally measured binding constants has been a bottleneck in scoring function development and

[*] Corresponding author phone: (734)615-0362; fax: (734)647-9647; e-mail: shaomeng@med.umich.edu.

10.1021/ci049733j CCC: $27.50 © 2004 American Chemical Society
Published on Web 10/26/2004

Table 1. Collections of Protein−Ligand Complexes with Known Binding Affinities in Prior to the PDBbind Database

approaches[a]	year of publish	total number of protein−ligand complexes cited	refs
Böhm (Score1)	1994	54	2
Jain	1996	34	3
Head et al. (VALIDATE)	1996	65	4
Eldridge et al. (ChemScore)	1997	112	5, 6
Böhm (Score2)	1998	94	7
Wang et al. (SCORE)	1998	181	8
Muegge et al. (PMF)	1999	225	9−11
Mitchell et al. (BLEEP)	1999	90	12, 13
Gohlke et al. (DrugScore)	2000	∼100	14, 15
Brooks et al. (LPDB)	2001	∼200	33
Ishchenko et al. (SMoG2001)	2002	119	16
Kellogg et al. (HINT)	2002	53	17
Wang et al. (X-Score)	2002	230	18
Mitchell et al. (PLD)	2003	∼ 270	34

[a] LPDB (Ligand−Protein Database) and PLD (Protein−Ligand Database) are Web-based databases of protein−ligand complexes; all of the other approaches are scoring functions.

validation. Table 1 provides a summary of the publicly available collections of such information reported since the 1990s. One can see clearly that the total number of such protein−ligand complexes has only increased to 200−300 since 1994. This bottleneck, however, has been overcome recently. Through a vigorous effort, we screened the entire Protein Data Bank (PDB)[36] to identify valid protein−ligand complexes and then systematically checked the original references of these protein−ligand complexes to collect the experimentally measured binding affinities. Our effort led to the creation of the PDBbind database,[35] which provided a compilation of binding affinities for nearly 1400 protein−ligand complexes in the PDB. After applying several additional criteria to filter out the entries that were not fully suitable for docking and scoring studies, such as covalently bound complexes, a set of 800 protein−ligand complexes were selected, which we called the “refined set”. This data set is much larger in size and also of better quality than any of the previous compilations of this kind.

In our present study, this “refined set” has been used to evaluate the performance of 14 popular scoring functions for their ability to predict the experimentally determined binding affinities. These include those 11 scoring functions evaluated in our previous study, i.e., X-Score,[18] DrugScore,[14,15] five scoring functions in the Sybyl software,[37] and four scoring functions in the Cerius2 software.[38] Another three scoring functions, i.e., GoldScore and ChemScore implemented in the GOLD program,[22,23] and HINT,[17] have also been included in our test. Our goal is to provide an objective assessment of the current scoring functions for their ability to predict the experimentally measured binding affinities for diverse ligand−protein complexes based upon high-resolution crystal structures.

METHODS

Construction of the Test Set. In this study, we used the “refined set” from the PDBbind database (version 2002)[35] as the test set to assess all of these 14 scoring functions. It consists of 800 diverse complexes formed between small organic molecules and over 200 different types of proteins. The detailed selection criteria of this set of protein−ligand complexes have been described in our recent work.[35] In brief, (i) all of the structures are determined by X-ray crystallography with resolution better or equal to 2.5 Å; (ii) all of the complexes are noncovalently bound complexes; (iii) all of the complexes are binary complexes; and (iv) all of the ligand molecules only consist of common organic elements, i.e., C, N, O, S, P, H and halogens, with molecular weights below 1000. These quality-control criteria were applied to eliminate those complexes that are not fully suitable for docking/scoring studies. All of the 800 complexes have experimentally measured K_d or K_i values. We use the negative logarithm of K_d or K_i value ($-\log K_d$ or $-\log K_i$) as the binding affinity scale in this paper and refer to it as “*binding constant*”. The binding constants of this set of protein−ligand complexes range from 0.60 to 13.96, spanning over 13 orders of magnitude, with a mean value of 6.46 and a standard deviation of 2.20.

Coordinates of all these complexes were downloaded from the Protein Data Bank.[36] For the convenience of computation, each complex was split into a protein molecule and a ligand molecule. All water molecules included in the crystal structure were removed since they are not considered by any of the scoring functions in our test. Metal ions, if residing inside the binding pocket and coordinately bound to the ligand and the protein, were saved as part of the protein. Hydrogen atoms were added to both of the proteins and the ligands. Atomic types and bond types of the ligand molecules were inspected and modified manually to ensure their correctness. The protein was assigned AMBER all-atom charges, and the ligand was assigned MMFF94 charges. No additional structural optimization was performed on either the protein or the ligand to keep their coordinates exactly the same as in the original PDB file. To meet the requirements of different scoring functions, the protein was saved in the PDB format and the Mol2 format, while the ligand was saved in the Mol2 format and the MDL SD (MACCS) format. All of the above work was done on an SGI Octane2 workstation using the Sybyl software.[37]

The entire test set, including the PDB codes, experimental binding constants, and the processed coordinate files of these 800 protein−ligand complexes, is available from the PDBbind Web set at http://www.pdbbind.org/.

Scoring Functions under Test. Fourteen popular scoring functions were chosen to be evaluated in this study, including two standalone scoring functions, X-Score (version 1.1) and DrugScore (version 1.2), five scoring functions (F-Score, G-Score, D-Score, PMF-Score, and ChemScore) implemented in the CScore module in the Sybyl software (version 6.9), four scoring functions (LigScore, PLP, PMF, and LUDI) implemented in the Ligand Scoring module in the Cerius2 software (version 4.6), two scoring functions (GoldScore and ChemScore) in the GOLD program (version 2.1), and the HINT scoring function implemented in the HINT program package (version 3.06). These scoring functions can be roughly classified into three groups: (i) empirical scoring functions, including X-Score, F-Score, ChemScore, LigScore, PLP, LUDI, and HINT, (ii) knowledge-based potentials, including DrugScore and PMF, and (iii) force-field based approaches, including D-Score and GoldScore. A brief review of these scoring functions is given in the Supporting Information. Special parameters or treatments applied in our work are also explained in the Supporting Information.

All of the 14 scoring functions were applied to compute the binding scores of the 800 protein−ligand complexes in the test set. For the scoring functions in Sybyl and Cerius2, the computation was automated using command scripts. For X-Score, DrugScore, and GOLD, the computation was automated through batch jobs. Since binding affinities are expressed in $-\log K_d$ units in this paper, we changed the sign of some scoring functions to ensure that a higher score always indicates a better binding affinity. It needs to be pointed out that in our study these 14 scoring functions were evaluated primarily based on the crystal structures of protein−ligand complexes. Even with a high resolution, some crystal structures may still have clashes between the proteins and the ligand molecules, which may be problematic for certain scoring functions to yield reasonable binding scores. An alternative approach would be to use minimized complex structures, derived from the crystal structures, in binding score computation. We applied this approach to the two scoring functions implemented in the GOLD program, but did not attempt it on the other scoring functions. Further discussion is given in the following Results and Discussion section.

Evaluation Methods. The performance of each scoring function set was measured by the linear correlation between its binding scores and the experimental binding constants of the protein−ligand complexes in the test set:

$$y = ax + b \tag{1}$$

In the above regression equation, y denotes the experimental binding constants (in $-\log K_i$ units), while x denotes the computed binding scores given by a scoring function. To provide quantitative measurements of the correlation, Pearson's correlation coefficient (R_p), standard deviation (SD), and unsigned mean error (ME) in regression were computed as

$$R_p = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{2}$$

$$SD = \sqrt{\sum_i [y_i - (ax_i + b)]^2/(N - 2)} \tag{3}$$

$$ME = \sum_i |y_i - (ax_i + b)|/N \tag{4}$$

The regression statistics of all of the 14 scoring functions are summarized in Table 2.

Besides computing the quantitative correlation between binding scores and experimental binding affinities, we also adopted a simple scheme to qualitatively assess a given scoring function for its ability to separate low-affinity complexes from high-affinity complexes. Given the fact that most of the samples in our test set have binding constants between 2 and 11, we divided this continuous binding affinity spectrum into three groups: the low-affinity group ($-\log K_d$ < 5.0), the medium-affinity group (5.0 ≤ $-\log K_d$ ≤ 8.0), and the high-affinity group ($-\log K_d$ > 8.0). The medium-affinity group, with K_i or K_d values from 10 μM to 10 nM, represents the binding affinity range of a typical lead compound in drug discovery. For each sample in the test set, we compared its experimentally determined binding constant and computed binding constant (the expectation value of y in eq 1): if the latter one fell in the same binding affinity group as the former one, a correct classification was counted. The success rate of each scoring function on each binding affinity group is listed in Table 3.

Table 2. Correlation Evaluation of 14 Scoring Functions on the Entire Test Set[a]

scoring function	N[b]	R_p	SD	ME	a	b
X-Score::HPScore	*800*	*0.514*	*1.89*	*1.47*	*0.71*	*2.03*
X-Score::HMScore	*800*	*0.566*	*1.82*	*1.42*	*0.92*	*1.18*
X-Score::HSScore	*800*	*0.506*	*1.90*	*1.48*	*0.93*	*1.24*
DrugScore::Pair	*800*	*0.473*	*1.94*	*1.51*	4.9×10^{-6}	*4.10*
DrugScore::Surf	*800*	*0.463*	*1.95*	*1.53*	7.2×10^{-5}	*4.48*
DrugScore::Pair/Surf	*800*	*0.476*	*1.94*	*1.50*	4.7×10^{-6}	*4.09*
Sybyl::D-Score	800	0.322	2.09	1.67	9.7×10^{-3}	5.00
Sybyl::PMF-Score	785	0.147	2.16	1.74	6.4×10^{-3}	5.92
Sybyl::G-Score	800	0.443	1.98	1.56	9.1×10^{-3}	4.34
Sybyl::ChemScore	*797*	*0.499*	*1.91*	*1.50*	9.1×10^{-2}	*3.90*
Sybyl::F-Score	732	0.141	2.19	1.77	2.1×10^{-2}	6.06
Cerius2::LigScore	717	0.406	2.00	1.57	0.79	4.63
Cerius2::PLP1	*800*	*0.458*	*1.96*	*1.52*	2.3×10^{-2}	*4.09*
Cerius2::PLP2	*800*	*0.455*	*1.96*	*1.53*	2.6×10^{-2}	*3.93*
Cerius2::PMF	795	0.253	2.13	1.71	1.1×10^{-2}	5.37
Cerius2::LUDI1	790	0.334	2.08	1.66	2.6×10^{-3}	4.88
Cerius2::LUDI2	799	0.379	2.04	1.62	4.2×10^{-3}	4.28
Cerius2::LUDI3	800	0.331	2.08	1.67	3.2×10^{-3}	4.68
GOLD::GoldScore	694	0.285	2.16	1.72	2.4×10^{-2}	5.33
GOLD::GoldScore_opt	772	0.365	2.06	1.63	3.0×10^{-2}	4.70
GOLD::ChemScore	741	0.423	2.00	1.56	8.5×10^{-2}	4.65
GOLD::ChemScore_opt	762	0.449	1.96	1.52	8.6×10^{-2}	4.41
HINT	800	0.330	2.08	1.65	0.20	6.36

[a] The relatively successful scoring functions in this test are in italics. [b] Number of protein−ligand complexes receiving positive binding scores from this scoring function. Only such complexes were included in regression analysis.

The above tests were performed on the entire test set, which consists of ligands with diverse chemical structures bound to more than 200 different proteins. Since it is also of great interest to examine how well a scoring function performs on ligand molecules bound to the same target protein, we evaluated these 14 scoring functions on the three most populated proteins in the test set, i.e., HIV-1 protease, trypsin, and carbonic anhydrase II (see Table 5). The regression statistics of each scoring function was recomputed on these three subsets of protein−ligand complexes. Besides the properties computed with eqs 2 to 4, we also computed the Spearman rank-order correlation coefficient (R_s) for each scoring function on these three subsets as an additional measure of their performance.

$$R_s = 1 - \frac{6 \times \sum_i (R_i - S_i)^2}{n^3 - n} \tag{5}$$

Spearman correlation coefficient provides a quantitative measurement of the correlation between two sets of ranks. In our case, R_i is the rank of complex i determined by its experimental binding constant, while S_i is the rank determined by a scoring function. The regression statistics of the 14 scoring functions on these three subsets of protein−ligand

Table 3. Qualitative Assessment of 14 Scoring Functions on Three Binding Affinity Groups

	success rate in classification[a]		
scoring function	low-affinity group ($-\log K_d < 5.0$)	medium-affinity group ($5.0 \leq -\log K_d \leq 8.0$)	high-affinity group ($-\log K_d > 8.0$)
X-Score::HPScore	33/205 = 16%	358/402 = 89%	48/193 = 25%
X-Score::HMScore	41/205 = 20%	348/402 = 87%	65/193 = 34%
X-Score::HSScore	29/205 = 14%	350/402 = 87%	53/193 = 27%
DrugScore::Pair	24/205 = 12%	359/402 = 89%	45/193 = 23%
DrugScore::Surf	11/205 = 5%	362/402 = 90%	45/193 = 23%
DrugScore::Pair/Surf	24/205 = 12%	358/402 = 89%	47/193 = 24%
Sybyl::D-Score	0/205 = 0%	384/402 = 96%	2/193 = 1%
Sybyl::PMF-Score	0/196 = 0%	395/396 = 99%	0/193 = 0%
Sybyl::G-Score	12/205 = 6%	359/402 = 89%	30/193 = 16%
Sybyl::ChemScore	38/204 = 19%	349/400 = 87%	40/193 = 21%
Sybyl::F-Score	0/182 = 0%	362/362 = 100%	0/188 = 0%
Cerius2::LigScore	11/186 = 6%	340/366 = 93%	16/165 = 10%
Cerius2::PLP1	24/205 = 12%	364/401 = 91%	35/193 = 18%
Cerius2::PLP2	30/205 = 15%	363/402 = 90%	32/193 = 17%
Cerius2::PMF	0/202 = 0%	390/400 = 97%	3/193 = 2%
Cerius2::LUDI1	1/203 = 0%	379/394 = 96%	9/193 = 5%
Cerius2::LUDI2	6/205 = 3%	378/401 = 94%	15/193 = 8%
Cerius2::LUDI3	1/205 = 0%	387/402 = 96%	9/193 = 5%
GOLD::GoldScore	0/178 = 0%	331/339 = 98%	4/177 = 2%
GOLD::GoldScore_opt	3/200 = 1%	366/385 = 95%	11/187 = 6%
GOLD::ChemScore	8/177 = 5%	345/376 = 92%	37/188 = 20%
GOLD::ChemScore_opt	20/187 = 11%	346/386 = 90%	38/189 = 20%
HINT	2/205 = 1%	388/402 = 97%	11/193 = 6%

[a] The denominator is the total number of samples that an individual scoring function could compute in a certain group; the numerator is the total number of correctly classified samples in this group by the given scoring function.

Table 4. Top Outliers in the Consensus of X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP

PDB code	$-\log K$ (exp.)[a]	mean error[b]	protein in the complex
7cpa	13.96	6.24	carboxypeptidase A
1ctu	11.92	5.95	cytidine deaminase
1swn	12.00	5.49	streptavidin
1qpb	1.36	5.46	pyruvate decarboxylase
1els	10.82	5.41	enolase
1swk	12.00	5.20	streptavidin
1duv	11.80	4.98	ornithine transcarbamoylase
1dqx	11.05	4.80	orotidine 5′-monophosphate decarboxylase
1lor	11.06	4.70	orotidine 5′-monophosphate decarboxylase
1zsb	0.60	4.57	carbonic anhydrase II
1if7	10.52	4.47	carbonic anhydrase II
1rbo	10.55	4.36	ribulose bisphosphate carboxylase/oxygenase
1xli	1.48	4.34	xylose isomerase
1n4k	10.05	4.33	inositol 1,4,5-trisphosphate receptor I
1b8o	10.64	4.29	purine nucleoside phosphorylase
1m0n	2.22	4.25	2,2-dialkylglycine decarboxylase
1bnn	10.00	4.20	carbonic anhydrase II
1m0o	2.31	4.09	2,2-dialkylglycine decarboxylase
5sga	2.85	3.98	proteinase A

[a] Experimentally measured K_d or K_i values in negative logarithm units. [b] The average discrepancy (in log units) between the experimental binding constant and the computed binding constants given by X-Score::HMScore, DrugScore::Pair, Sybyl::ChemScore, and Cerius2::PLP2.

Table 5. Three Subsets of Protein−Ligand Complexes Extracted from the Test Set

HIV-1 protease subset (82 complexes)

1a30 1a94 1a9m 1aaq 1ajv 1ajx 1b6j 1b6k 1b6l 1b6m
1b6n 1b6o 1b6p 1bdq 1bv7 1bv9 1bwa 1bwb 1c6y 1c70
1d4k 1d4l 1d4s 1d4y 1dmp 1g2k 1g35 1hbv 1hih 1hii
1hiv 1hos 1hpo 1hps 1hpv 1hpx 1hsg 1hsh 1htf 1htg
1hvh 1hvi 1hvj 1hvk 1hvl 1hvr 1hvs 1hwr 1hxw 1ivp
1izh 1izi 1k6c 1k6p 1k6t 1k6v 1kzk 1mes 1met 1meu
1mtr 1ody 1ohr 1pro 1qbr 1qbs 1qbt 1qbu 1sbg 1siv
1tcw 1tcx 2bpv 2bpy 3aid 4hvp 5hvp 5upj 6upj 7hvp
7upj 8hvp

trypsin subset (45 complexes)

1bra 1c1r 1c2d 1c5p 1c5q 1c5s 1c5t 1ce5 1f0t 1f0u
1g3b 1g3d 1g3e 1ghz 1gi2 1gi4 1gi5 1gi6 1gj6 1h4w
1j14 1j16 1j17 1k1i 1k1j 1k1l 1k1m 1k1n 1ppc 1pph
1qb1 1qb6 1qb9 1qbn 1qbo 1smf 1tng 1tnh 1tni 1tnj
1tnk 1tnl 1xuf 2bza 3ptb

carbonic anhydrase II subset (40 complexes)

1a42 1avn 1bcd 1bn1 1bn3 1bn4 1bnn 1bnq 1bnt 1bnu
1bnv 1bnw 1bzm 1cil 1cim 1cin 1cnw 1cnx 1cny 1g1d
1g45 1g46 1g48 1g4j 1g4o 1g52 1g53 1g54 1h4n 1i9n
1i9p 1if7 1if8 1okl 1okn 1yda 1ydb 1ydd 1zsb 2h4n

complexes are summarized in Table 6.

RESULTS AND DISCUSSION

Performance on the Entire Test Set. As shown in Table 2, four scoring functions, i.e., X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP, demonstrated moderate correlations ($R_p = 0.45-0.57$) between their binding scores and the experimentally determined binding constants for the entire 800 protein−ligand complexes. These four scoring functions reproduced the known binding constants of the entire test set with a standard deviation of 1.8−2.0 log units (corresponding to 2.5−2.7 kcal/mol in terms of binding free energy at room temperature). Scatter plots of experimental binding constants vs computed binding scores for these four scoring functions are shown in Figure 1. The standard deviations and average errors produced by these four scoring functions were generally 0.2−0.3 log units lower than the ones produced by the other scoring functions. It needs to point out that the scoring functions in our test actually demonstrated a continuous spectrum of accuracy level. The criterion of $R_p > 0.45$ was chosen in a more or less subjective

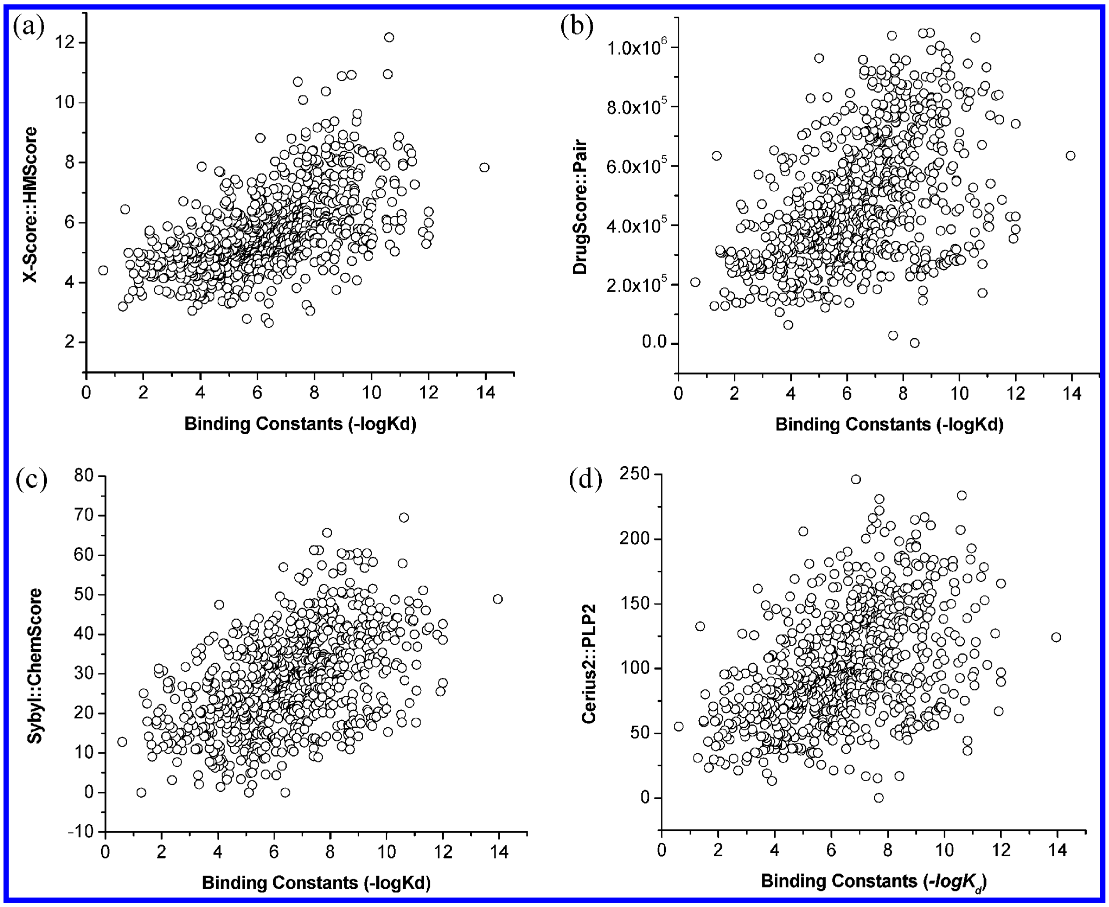

Figure 1. Binding constants versus the binding scores computed by (a) X-Score::HMScore, (b) DrugScore::Pair, (c) Sybyl::ChemScore, and (d) Cerius2::PLP2 on the entire test set ($N = 800$).

way. The accuracy of some other scoring functions were actually very close to this cutoff, such as Sybyl::G-Score. Some scoring functions, however, demonstrated basically no correlation between their scores and the experimental binding constants with very low R_p values.

Table 3 summarizes the performance of these 14 scoring functions when the test set was divided into three binding affinity groups. This test was designed to evaluate if a given scoring function can really differentiate low-affinity and high-affinity complexes from the others, which would be useful in database virtual screening projects. In this test, a number of scoring functions showed near-zero success rates for the low-affinity group and the high-affinity group. In contrast, the four scoring functions we selected above were able to identify some low-affinity and high-affinity complexes correctly, although their success rates for these two groups were only modest.

In our previous study,[31] we observed that X-Score, DrugScore, and Cerius2::PLP exhibited relatively better performance in binding affinity prediction on 100 selected protein−ligand complexes than other scoring functions. In this study, essentially the same "winners" were identified. In a study published very recently,[32] Ferrara et al. reported that ChemScore outperformed the other scoring functions implemented in the CScore module of Sybyl in binding affinity prediction on a total of 189 protein−ligand complexes. This is consistent with our observation in this study.

Another aspect that needs to be discussed is the robustness of a scoring function, which is as important as its accuracy. In a structure-based drug design project, one cannot expect that binding affinity prediction is always performed based on high-resolution, experimentally determined complex structures. It is therefore important for a scoring function to tolerate at least minor inaccuracies in a given complex structure and yet produce a good estimate of the binding affinity. In our study, we observed that some scoring functions failed to produce overall positive (favorable) binding scores for a certain number of protein−ligand complexes in the test set, e.g. Sybyl::F-Score failed in 68 cases, Cerius2::LigScore failed in 83 cases, GOLD::GoldScore failed in 106 cases, and GOLD::ChemScore failed in 59 cases. All of these scoring functions have a term for computing the dispersion/repulsion (van der Waals) interactions between the protein and the ligand. This term is helpful for docking the ligand molecule into its protein binding site since it penalizes the models having steric clashes between the protein and the ligand. However, it is not so unusual to observe bad contacts between protein and ligand molecule even in high-resolution crystal structures. When a scoring function of this type is applied to such a structure, the repulsion energy computed between the clashed atom pairs can overwhelm the other terms in the scoring function and lead to an overall negative binding score: a phenomenon we call "repulsion outbreak". The negative binding scores in such cases were often in an unrealistic range. As a consequence, inclusion of these cases in regression analysis would drive the statistics to be completely meaningless. To circumvent this problem, if a scoring function failed to yield overall positive binding scores for certain complexes in the test set, we simply excluded them from our consideration and performed the regression analysis using the rest of the test set. The total number of "valid" complexes each individual scoring function could compute is also given in Table 2.

As one can see in Table 2, the four relatively successful scoring functions we selected above were not only able to provide more accurate predictions but also robust enough to compute almost all of the complexes in the test set using the crystal structures directly without optimization (Sybyl::ChemScore failed in only three cases). We would like to mention that one method for tackling the "repulsion outbreak" problem is to set a ceiling to the repulsion energy computed between any pair of atoms so that the overall binding score will not be compromised. This simple yet effective method is indeed used in X-Score, Sybyl::D-Score, and Sybyl::G-Score in their van der Waals interaction terms. As revealed in Table 2, all of these three scoring functions had no problem in computing all of the 800 complexes in our test set.

It is also reasonable to expect that an optimization of the input complex structure to a local minimum prior to binding score computation, ideally driven by the same scoring function, will help to release the repulsion energies. In our study, we were able to perform such structural optimization for GOLD::GoldScore and GOLD::ChemScore using the facilities provided by the GOLD program. With this treatment, the failed cases for these two scoring functions were reduced considerably to 28 (GOLD::GoldScore_Opt) and 38 (GOLD::ChemScore_Opt), respectively. As a result of removing the repulsion energies in their binding scores, the accuracy of these two scoring functions was also improved. GOLD::ChemScore_Opt even demonstrated an accuracy level close to the four relatively successful scoring functions we selected above. In our study, however, we did not test other scoring functions on minimized structures for a number of reasons. First, as indicated above, some scoring functions did not need such treatment on the input structures and yet produced reasonable results. Second, minimizing a crystal structure is by no means a trivial job since a number of technical issues must be considered: which force field and what parameters should be chosen for this purpose? Should the complex structure be minimized in vacuum, with a continuum solvation model, or with the explicit water molecules observed in the crystal structure? How to control the minimization process so that the minimized structure will not deviate too much from the original crystal structure? Even if all these issues can be successfully addressed, it still remains debatable if a minimized structure is always more appropriate than an experimentally determined structure for binding affinity prediction. Nevertheless, based on what we observed for the two scoring functions in the GOLD program, it appears that a prior structural minimization, if performed properly, will be helpful especially for the scoring functions which may have the "repulsion outbreak" problem.

Some of the scoring functions in our test are different implementations of the same scoring function, for example, Sybyl::G-Score versus GOLD::GoldScore, Sybyl::ChemScore versus GOLD::ChemScore, and Sybyl::PMF-Score versus Cerius2::PMF. Our results showed that different implementations of the same scoring function could give notably different results (Tables 2 and 3). It is understandable since some implementations have their own modifications to the original scoring function. For example, the ChemScore implemented in GOLD differs from the original ChemScore by including additional terms. End-users of these scoring functions should be very cautious of this fact. It is interesting to observe that Sybyl::G-Score out-performed GOLD::GoldScore in our test even when a local structural optimization was performed, i.e., GOLD::GoldScore_opt. We believe this should be credited to the method in Sybyl::G-Score for handling the repulsion break problem.

Another related issue is that some of the scoring functions in our test, including X-Score, DrugScore, Cerius2::PLP, and Cerius2::LUDI, offer multiple variations for binding score computation. All of these variations have been tested in our study. As shown in Tables 2 and 3, although generally no particular variation significantly outperforms other available variations offered by the same scoring function, some interesting observations are noted. For X-Score, HMScore exhibited marginally but consistently better statistics than either HPScore or HSScore. For DrugScore, although the combination of pairwise and surface-based potentials ("Pair/Surf") was supposed to be more advanced, using the pairwise potentials alone ("Pair") was actually equally good. Since applying the "Pair" variation does not need the additional computation of surface areas, it may even be preferred in practice. For Cerius2::PLP, the difference between its two variations, PLP1 and PLP2, is not clear due to lack of documentation. Their performance was however on the same level. As for Cerius2::LUDI, the Cerius2 user manual mentions that LUDI2, unlike the other two variations, was actually calibrated with a set of protein−ligand complexes with known binding constants. In our test, LUDI2 indeed produced slightly better statistical results than LUDI1 or LUDI3.

Some of the scoring functions in our test, especially the empirical scoring functions such as X-Score, ChemScore, and LUDI, were calibrated using a training set of protein−ligand complexes from the PDB. Although the test set used in this study, i.e., the PDBbind refined set, was the outcome of a project totally independent to any scoring function, it does overlap with the training sets used by those empirical scoring functions. For example, there are 140 overlapping samples between the test set used in this study and the original training set of X-Score. Ideally, such overlapping samples should be removed from the test set to provide an unbiased assessment on the performance of these scoring functions. In our study, however, we did not attempt to do this since those overlapping samples only account for a small fraction of our test set. We expected that those overlapping samples would not have a significant impact on the overall statistics. Indeed, the regression analysis of X-Score after removing the 140 overlapping samples gave the following: for X-Score::HPScore, $N = 660$, $R_p = 0.493$, $SD = 1.91$, $ME = 1.49$; for X-Score::HMScore, $N = 660$, $R_p = 0.542$, $SD = 1.85$, $ME = 1.44$; for X-Score::HSScore, $N = 660$, $R_p = 0.484$, $SD = 1.92$, $ME = 1.50$. Compared to the results computed on the entire test set (see Table 2), no significant difference was observed. For other scoring functions, it should have an even less significant impact since their training sets are even smaller than that used by X-Score. Another technical problem preventing us from identifying

Table 6. Correlation Evaluation of 14 Scoring Functions on Three Subsets of Protein–Ligand Complexes[a]

	HIV-1 protease					trypsin					carbonic anhydrase II[b]				
scoring function	*N*	R_p	*SD*	*ME*	R_s	*N*	R_p	*SD*	*ME*	R_s	*N*	R_p	*SD*	*ME*	R_s
X-Score::HPScore	82	0.429	1.25	1.01	0.436	45	0.754	1.15	0.88	0.725	39	0.544	1.18	0.85	0.547
X-Score::HMScore	82	0.379	1.28	1.04	0.334	*45*	*0.823*	*0.99*	*0.75*	*0.824*	39	0.495	1.23	0.95	0.341
X-Score::HSScore	82	0.400	1.27	1.05	0.322	45	0.753	1.15	0.91	0.766	39	0.417	1.28	0.91	0.448
DrugScore::Pair	82	0.377	1.28	1.04	0.315	45	0.780	1.09	0.82	0.818	39	0.622	1.10	0.83	0.501
DrugScore::Surf	82	0.401	1.27	1.02	0.317	45	0.674	1.29	0.99	0.753	39	0.512	1.21	0.97	0.269
DrugScore::Pair/Surf	82	0.384	1.28	1.04	0.322	45	0.780	1.09	0.82	0.807	39	0.623	1.10	0.83	0.495
Sybyl::D-Score	82	0.342	1.30	1.03	0.305	45	0.617	1.37	0.98	0.736	39	0.584	1.14	0.86	0.441
Sybyl::PMF-Score	82	0.246	1.34	1.09	0.226	37	0.513	1.02	0.86	0.523	39	0.655	1.07	0.80	0.652
Sybyl::G-Score	82	0.350	1.30	1.05	0.335	45	0.580	1.42	1.06	0.728	39	0.643	1.08	0.79	0.649
Sybyl::ChemScore	82	0.376	1.28	1.05	0.350	45	0.761	1.13	0.91	0.749	39	0.609	1.12	0.76	0.663
Sybyl::F-Score	80	0.361	1.31	1.08	0.375	45	0.663	1.31	1.05	0.610	35	0.371	1.15	0.87	0.145
Cerius2::LigScore	81	0.528	1.18	0.99	0.496	40	0.392	1.59	1.27	0.467	18	0.154	1.78	1.34	−0.323
Cerius2::PLP1	82	0.458	1.23	1.02	0.395	45	0.729	1.19	0.88	0.785	39	0.718	0.98	0.76	0.606
Cerius2::PLP2	82	0.438	1.25	1.03	0.414	45	0.754	1.15	0.84	0.802	*39*	*0.735*	*0.96*	*0.67*	*0.781*
Cerius2::PMF	82	0.411	1.26	1.03	0.342	43	0.775	1.06	0.85	0.740	39	0.604	1.12	0.87	0.603
Cerius2::LUDI1	82	0.208	1.35	1.11	0.123	45	0.670	1.29	1.01	0.698	38	0.065	1.21	0.86	0.335
Cerius2::LUDI2	82	0.274	1.33	1.11	0.181	45	0.696	1.25	0.95	0.725	39	0.470	1.25	0.89	0.519
Cerius2::LUDI3	82	0.248	1.34	1.10	0.174	45	0.679	1.28	1.00	0.690	39	0.433	1.27	0.91	0.554
GOLD::GoldScore	69	0.386	1.25	1.00	0.391	36	0.029	1.65	1.32	−0.012	34	0.539	1.25	0.90	0.420
GOLD::GoldScore_opt	*78*	*0.555*	*1.13*	*0.92*	*0.579*	42	0.590	1.41	1.14	0.673	37	0.585	1.17	0.86	0.532
GOLD::ChemScore	78	0.404	1.19	0.98	0.386	44	0.388	1.61	1.33	0.348	39	0.498	1.22	0.89	0.307
GOLD::ChemScore_opt	80	0.429	1.24	1.02	0.393	44	0.520	1.49	1.21	0.565	39	0.639	1.08	0.80	0.454
HINT	82	0.313	1.32	1.04	0.264	45	0.135	1.73	1.37	0.251	39	0.599	1.13	0.78	0.689

[a] The best performing scoring function on each subset is in italics. [b] PDB entry 1ZSB was not included in evaluation for all of the scoring functions in this test.

and eliminating overlapping samples in our test set is that not every empirical scoring function has revealed its training set.

The last point we want to make in this section concerns outliers. As can be seen from Table 2, the overall performance even for those relatively successful scoring functions (R_p = 0.45–0.57, *SD* = 1.8–2.0 log units) was not very satisfactory. We found that these seemingly disappointing statistics were largely caused by a small number of significant outliers. We have identified the top 5% outliers for X-Score::HMScore, DrugScore::Pair, Sybyl::ChemScore, and Cerius2::PLP2, respectively, by checking the errors between the experimental and computed binding constants. The outliers in the consensus of these four scoring functions are listed in Table 4. We noticed that the average errors concerning these outliers were as large as 4–6 log units. These results showed that there is still much room left for the improvement of today's scoring functions.

The cause of these significant outliers could be two-fold. Because of the complexity of the protein–ligand binding process, binding affinity prediction in general remains a very difficult task. The remarkable structural diversity for both proteins and ligands in our test set and the very wide range of binding affinities makes it an extremely difficult task for any scoring function. Furthermore, some of the protein–ligand complexes in our test set do have extraordinary features. For example, there are two streptavidin–biotin complexes, 1SWN and 1SWK, among the 19 outliers listed in Table 4. Biotin is well-known to achieve exceptionally high affinity to streptavidin with relatively simple chemical structures. A scoring function trained to handle "normal" protein–ligand complexes therefore tends to fail in such cases.

The other possible reason may arise from experimentally determined binding affinity data. For example, the PDB entry 1ZSB, which is a complex formed between carbonic anhydrase II (E117Q mutant) and a transition state analogue acetazolamide, was reported to have an extremely low binding affinity with a K_d value of 250 mM ($-\log K_d$ = 0.60),[39] the lowest binding affinity in our test set. However, there are several other complexes formed between carbonic anhydrase II mutants and acetazolamide in our test set, all of which exhibit much higher binding affinities, e.g. 1YDA ($-\log K_d$ = 6.55), 1YDB ($-\log K_d$ = 8.24), 1YDD ($-\log K_d$ = 7.07), and 2H4N ($-\log K_d$ = 8.70). It is unclear why the ligand–protein complex in 1ZSB would have such low binding affinity. Ideally, such suspicious binding affinity data should be identified and confirmed with their original authors. At the present time, we were unable to carry out such a task due to the large size of our test set. Another issue is that binding assays are not performed under the same conditions. Difference in temperature, pH level, buffer, and other factors may lead to notable discrepancies in binding affinity measurement. One should keep all of these issues in mind when interpreting the evaluation results in this study.

Performance on Three Individual Subsets. We have described the performance of each individual scoring function for their ability to reproduce the experimentally determined binding affinities of 800 ligand–protein complexes consisting of more than 200 different proteins. However, one typically works with a particular target protein in a structure-based drug design project. Therefore, it is of great interest to evaluate the performance of today's scoring functions on a specific protein target. Accordingly, we have evaluated the performance of these 14 scoring functions on three subsets of protein–ligand complexes.

The HIV-1 Protease Subset. There are 82 HIV-1 protease complexes (including wide type and mutants) in our test set, with experimental binding constants ranging from 4.30 to 11.40 with a mean value of 8.49 and a standard deviation of 1.39 (all in $-\log K_d$ units). If using the Pearson coefficients (R_p) as the criterion, almost all of the scoring functions,

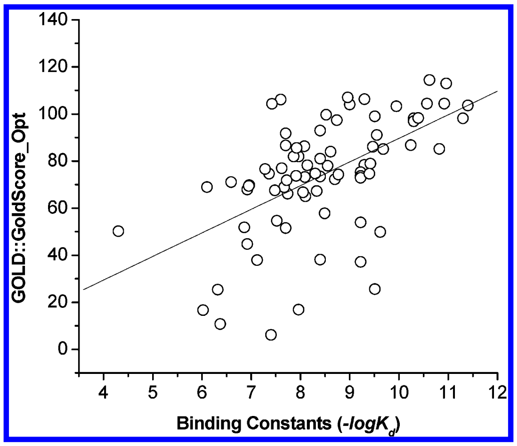

Figure 2. Binding constants versus the binding scores computed by GOLD::GoldScore_Opt for the HIV-1 protease subset (N = 82, R_p = 0.555, SD = 1.13, ME = 0.92).

including the four relatively successful scoring functions we selected in the foregoing test, demonstrated poorer performance on this subset than on the entire test set (see Table 6). The only exception was GOLD::GoldScore_Opt, which produced the best results on this subset among all of the scoring functions with a modest correlation coefficient of 0.555 and a standard deviation of 1.13 log units (Figure 2). This is surprising at the first glance since the GoldScore does not even have an explicit term for hydrophobic interaction, which is believed to be one major element in the binding of HIV-1 protease to its ligands. Nevertheless, as described in the Supporting Information, GoldScore applies an empirical weight factor of 1.375 to its van der Waals term to compensate for the hydrophobic contacts between the protein and the ligand.

The Trypsin Subset. There are 45 beta-trypsin (wide type and mutants) complexes in our test set, making it the second most populated protein class. The binding constants of these complexes range from 1.49 to 8.28 with a mean value of 5.45 and a standard deviation of 1.74 (all in $-\log K_d$ units). In contrast to the disappointing performance observed for the HIV-1 proteases, a number of scoring functions were quite successful with this subset: X-Score, DrugScore, Sybyl::ChemScore, Cerius2::PLP, and Cerius2::PMF all produced correlation coefficients (R_p and R_s) higher than 0.70 (see Table 6). X-Score::HMScore in particular yielded the best performance on this subset among all the tested scoring functions with a correlation coefficient of 0.823, a standard deviation of 0.99, and an unsigned mean error of 0.75 (corresponding to $\sim$1.0 kcal/mol in terms of binding free energy). It is encouraging to observe such a good correlation for a variety of chemical structures with binding constants spanning nearly 7 orders of magnitude (Figure 3).

The Carbonic Anhydrase II Subset. Human carbonic anhydrase type II (HCA II), with 40 complex structures in total, is the third most populated protein class in our test set. We decided to remove PDB entry 1ZSB from our regression analyses because of its suspicious binding affinity. Binding constants of the remaining 39 complexes range from 3.90 to 10.52 with a mean value of 8.50 and a standard deviation of 1.41 (all in $-\log K_d$ units). Note that this set of complexes may be challenging for scoring functions since (i) the average binding constant of this subset (8.50) is well above that of the entire test set (6.46) and (ii) binding constants of this subset have a narrow distribution as indicated by the small standard deviation (1.41). Despite these challenges, some scoring functions still produced correlation coefficients (R_p and R_s) higher than 0.6, including Sybyl::PMF-Score, Sybyl::G-Score, Sybyl::ChemScore, Cerius2::PLP, Cerius2::PMF, and HINT (see Table 6). These scoring functions typically yielded a standard deviation around 1.0 and an average error around 0.7. The best correlation was given by Cerius2::PLP2, which had a correlation coefficient of 0.735, a standard deviation of 0.96, and an unsigned mean error of only 0.67 ($<$1.0 kcal/mol in terms of binding free energy) for this subset (see Figure 4).

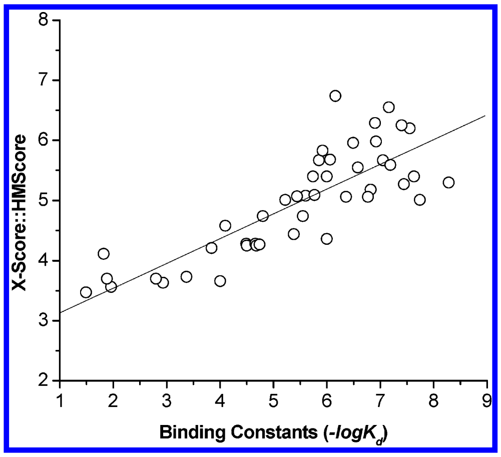

Figure 3. Binding constants versus the binding scores computed by X-Score::HMScore for the trypsin subset (N = 45, R_p = 0.823, SD = 0.99, ME = 0.75).

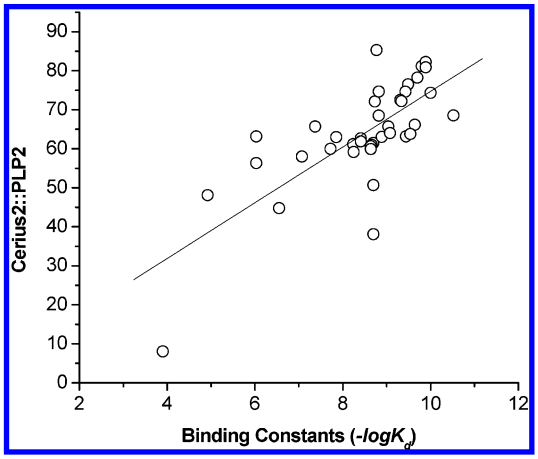

Figure 4. Binding constants versus the binding scores computed by Cerius2::PLP2 for the carbonic anhydrase II subset (N = 39, R_p = 0.735, SD = 0.96, ME = 0.67).

Our test has basically verified the expectation that scoring functions tend to work better on a specific set of protein−ligand complexes. For the trypsin subset and the HCA II subset we have investigated, a number of scoring functions produced quite satisfactory results. In some cases, such as X-Score on the trypsin subset, the performance demonstrated by a scoring function was quite impressive. Our results show that today's scoring functions are applicable to real structure-based drug design projects, despite their unsatisfactory performance on a nondiscriminative assembly of structurally diverse protein−ligand complexes. Furthermore, our study suggests that selecting the right scoring function through an objective test in prior is very important to ensure a reasonably successful prediction of the binding affinities for designed ligands, since none of these 14 scoring function performs consistently better than the others in our test.

Today's Scoring Function: Pros and Cons. With the development of more and more docking/scoring tools, objective comparison of the performance of these tools has become an important task. Our opinion is that, to make such a comparison, it is more informative to dissect the complicated docking/scoring procedure into some modular problems and conduct the assessment accordingly. For example, in our previous comparison of a number of scoring functions,[31] we focused on evaluating their abilities to identify the native binding pose out of an ensemble of computer-generated decoys. In this study, we focused on evaluating their abilities to predict the binding affinity when the correct protein−ligand complex structure is provided. Such information may serve as a general guidance when one needs to make a choice among several available scoring functions. Furthermore, testing these scoring functions on a large variety of protein−ligand complexes also helps us to achieve a better understanding of the strength and weakness of today's scoring functions. This also sheds light on the development of more accurate scoring functions.

In this study, we have included 14 scoring functions that are force-field-based, or PMF, or empirical in nature. A number of relatively successful scoring functions, i.e., X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP, reproduced the binding constants of the entire test set with a standard deviation of 1.8−2.0 log units (corresponding to 2.5−2.7 kcal/mol in binding free energy at room temperature). When applied to some specific sets of complexes, they were able to produce a standard deviation as low as 1.0 log units (~1.4 kcal/mol). It would be thus interesting to compare the accuracy of these scoring functions with "first-principle" based methods, such as free energy perturbation (FEP),[41] linear interaction energy approximation (LIE),[42,43] and MM-PBSA/GBSA.[44] However, to the best of our knowledge, those "first-principle" based methods have not been tested extensively. They were often applied to reproduce the binding affinities of a small and well-selected set of ligand molecules bound to a certain target protein. In such studies, the reported accuracy typically ranged from 1−2 kcal/mol.[41−44] Hence, these relatively successfully scoring functions evaluated in our current study and those "first-principle" based methods appear to have a similar accuracy in protein−ligand binding affinity prediction, but scoring functions only cost a small fraction of the computation time needed by those "first-principle" based methods.

Among these four relatively successful scoring functions, there are either empirical in nature such as X-Score, Sybyl::ChemScore, and Cerius2::PLP or knowledge-based PMF methods such DrugScore. Thus, it appears that both strategies are valid for developing scoring functions. The major advantage of the scoring functions in these two categories is that they can readily take advantage of the growing number of protein−ligand complexes to derive better formulated equations. Among these 14 scoring functions we have tested, there are also force field based scoring functions, i.e., Sybyl::D-Score, Sybyl::G-Score, and GOLD::GoldScore. Their performance in our test was generally worse than empirical scoring functions or knowledge-based PMF approaches. However, we have not investigated how well a classical force field can predict protein−ligand binding affinities if it is combined with a solvation model. We noted with a great interest that Ferrara et al. recently combined the CHARMm force field with a variety of solvation models, including the Poisson−Boltzmann model (PB), the generalized Born model (GB), and distance-dependent or constant dielectric functions for binding affinity prediction.[32] To optimize the performance of this approach, atomic partial charges and protonation states were carefully assigned on both of the protein and the ligand molecules. It was shown that this approach performs reasonably well in recognizing the native binding pose of decoys. Furthermore, it was shown to have a similar performance as the empirical ChemScore method in reproducing the experimentally determined binding affinities of 189 protein−ligand complexes.[32] Hence, force field-based approaches, if combined with appropriate solvent models, can also predict the binding affinities of protein−ligand complexes with a similar accuracy as current empirical and PMF-based scoring functions.

Table 7. Correlations between the Binding Scores Computed by Four Scoring Functions for the Entire Test Set

correlation coefficients (R_p) $N = 800$	DrugScore::Pair	Sybyl::ChemScore	Cerius2::PLP2
X-Score::HMScore	0.737	0.751	0.691
DrugScore::Pair		0.742	0.924
Sybyl::ChemScore			0.717

Another interesting observation in our study is that, regarding the binding scores they computed, today's scoring functions are actually quite similar to each other, even for those that belong to different categories. In Table 7, we have listed the correlations between the binding scores computed by X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP on the entire test set. Comparing the data in Tables 7 and 2, one can see that the intercorrelations between these scoring functions ($R_p = 0.69-0.92$) are in fact higher than the correlations between their scores and experimental binding constants ($R_p = 0.45-0.57$). In particular, DrugScore::Pair and Cerius2::PLP2 showed a remarkable correlation coefficient of 0.924. This observation is understandable since many scoring functions formulate their equations in a similar way. The good correlations between these scoring functions indicate that many of the current scoring functions share their strengths as well as their weaknesses. As indicated in Table 4, the most significant outliers identified in our test were actually common to X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP.

Given the fact that no scoring function performs consistently better than the others, one would wonder if the consensus scoring strategy would work better in our test. It has been shown that this strategy can reduce the false positives in virtual screening.[45] In our previous study,[31] we also observed that a combination of two or more scoring functions could lead to a better chance of identifying the native pose out of an ensemble of decoys. In our current study, we have tested all of the possible combinations of any two of the four relatively successful scoring functions on the entire test set. The results of all consensus scoring schemes we tested are summarized in Table 8. Note that since most scoring functions give binding scores in arbitrary units, caution must be taken when combining the results produced by two different scoring functions. In our consensus scoring experiments, the relative rank of each protein−ligand

Table 8. Consensus Scoring Schemes Tested on the Entire Test Set

scoring method	Pearson correlation coefficient (R_p)	Spearman correlation coefficient (R_s)
Single Scoring Function		
X-Score::HMScore	0.566	0.603
DrugScore::Pair	0.473	0.484
Sybyl::ChemScore	0.499	0.507
Cerius2::PLP2	0.455	0.478
Double Scoring Schemes		
X-Score::HMScore + DrugScore::Pair	0.573	0.586
X-Score::HMScore + Sybyl::ChemScore	0.586	0.597
X-Score::HMScore + Cerius2::PLP2	0.573	0.586
DrugScore::Pair + Sybyl::ChemScore	0.520	0.529
DrugScore::Pair + Cerius2::PLP2	0.476	0.488
Sybyl::ChemScore + Cerius2::PLP2	0.521	0.530

complex was used instead of its absolute binding score in correlation analysis, i.e., the final score of a complex computed by a consensus scoring scheme was the mean value of the ranks given by each component scoring function. Such scores were then correlated with the experimentally determined binding constants.

As shown in Table 8, once again we observed that double scoring schemes produced marginally but consistently better results than individual scoring functions. On the other hand, it is also clear that the accuracy level a consensus scoring scheme can reach is basically limited by the accuracy level of its component scoring functions. As we have attempted to demonstrate in a previous study,[46] consensus scoring only provides statistical advantages because of its multiple sampling nature. This is especially true when the component scoring functions in a consensus scoring scheme have fairly high intercorrelations between each other.

Today's scoring functions, even those relatively successful ones, have their significant drawbacks. One significant drawback we have noticed is that today's scoring functions are less capable of handling those complexes with either very low or very high affinities. This is clearly indicated in Table 3 and can also be seen in Figure 1. For example, X-Score::HMScore, the one showing the best statistics on the entire test set, systematically overscored the protein−ligand complexes with very low affinities ($-\log K_d < 3$) and underscored many protein−ligand complexes with very high affinities ($-\log K_d > 10$). We believe that this is caused in part by a lack of enough penalty terms in X-Score. In fact, the only penalty term in X-Score is a count of rotatable single bonds on the ligand molecule, while all of the other terms sum up favorable contributions from van der Waals interaction, hydrogen bonding, and hydrophobic effect (see the Supporting Information). Other scoring functions also share this drawback. Another possible reason is embedded in the way how a scoring function is developed. The quality of a scoring function is inevitably influenced by the contents of its training set. Considering that the training sets used for calibrating empirical scoring functions were primarily formed by protein−ligand complexes with medium-level binding affinities (which is also true for the test set we used in this study), it is not surprising that those scoring functions cannot handle the complexes with extremely low or extremely high binding affinities very well.

Another major drawback we know about today's scoring functions is that they do not have adequate coverage for some less frequently occurring factors in protein−ligand binding. Today's empirical scoring functions normally only consist of terms for computing some common types of interactions upon protein−ligand binding, such as van der Waals dispersion/repulsion, hydrogen bonding, and hydrophobic interactions. Knowledge-based PMF approaches do not use explicit terms, but they include such consideration implicitly with their atom typing schemes. Such a scoring function may be sufficient for handling "normal" protein−ligand complexes, such as those protein−ligand complexes in the subsets of trypsin and carbonic anhydrase II. However, if some unusual factors play an important role in the protein−ligand binding process, such a scoring function may fail. For example, we observed cation-π interactions in a number of HIV-1 protease complexes such as 1HVI, 1HVJ, 1HVK, and 1HVS. No scoring function in our test contemplates this type of interactions, although they are recognized as strong interactions in molecular recognition. There are also other factors that may be difficult to implement in a scoring function, such as the cooperative conformational rearrangement upon protein−ligand binding.

CONCLUSION

Fourteen popular scoring functions have been tested on a set of 800 protein−ligand complexes with high-resolution structures and experimentally determined binding affinities. Overall, only moderate correlations were found between the computed scores and the experimentally determined binding affinities of these protein−ligand complexes. Among them, X-Score, DrugScore, Sybyl::ChemScore, and Cerius2::PLP were found to produce better correlations between their scores and the experimentally determined binding constants than the other scoring functions. Notably, these four scoring functions were also robust enough to conduct their computation on the input crystal structures without any structural optimization.

Reevaluation of these 14 scoring functions on three subsets of protein−ligand complexes, i.e., HIV-1 protease complexes, trypsin complexes, and carbonic anhydrase II complexes, demonstrated that for the trypsin and carbonic anhydrase II subsets, the best current scoring functions were able to produce a standard deviation as low as 1.0 log units (~1.4 kcal/mol). Compared to the performance observed for the entire test set, we concluded that a scoring function may perform significantly better in the prediction of the binding affinities of ligand molecules bound to the same protein. Our test also showed that a scoring function's performance on a particular target protein is largely case-dependent. Therefore, in a practical structure-based drug design project, an objective evaluation of available scoring functions on the target protein of interest is still necessary in order to choose the most suitable scoring function for the project.

It is encouraging to observe that today's scoring functions are applicable to a wide range of protein−ligand systems, and in some cases they are even very successful. However, it is also clear that scoring functions, both empirical scoring functions and PMF-based approaches, still need to be improved significantly toward the ultimate goal to reliably predict the binding affinity of a protein−ligand complex even if an experimental structure is available.

ACKNOWLEDGMENT

The authors are grateful to the Cap CURE Foundation (2001 Young Investigator Award to Renxiao Wang) and the Department of Defense and the National Institutes of Health (to Shaomeng Wang) for their financial support. The authors also thank Dr. G. W. A. Milne for his critical reading of the manuscript. The excellent secretarial assistance of Ms. Karen Kreuzter is greatly appreciated.

Supporting Information Available: A brief review of the 14 scoring functions under test and the parameters and special treatments applied in this study. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Böhm, H. J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; Wiley-VCH Inc.: New Jersey, 2002; Vol. 18, pp 41−88.

(2) 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.

(3) Jain, A. N. Scoring noncovalent protein−ligand interactions: A continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 427−440.

(4) Head, R. D.; Smythe, M. L.; Oprea, T. I.; Waller, C. L.; Green, S. M.; Marshall, G. R. VALIDATE: A new method for the receptor-based prediction of binding affinities of novel ligands *J. Am. Chem. Soc.* **1996**, *118*, 3959−3969.

(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) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand−receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503−519.

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

(8) 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.

(9) 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.

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

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

(12) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP: potential of mean force describing protein−ligand interactions. I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

(13) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Forster, M. J.; Thornton, J. M. BLEEP: potential of mean force describing protein−ligand interactions. II. Calculation of binding energies and comparison with experimental data. *J. Comput. Chem.* **1999**, *20*, 1177−1185.

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

(15) Gohlke, H.; Hendlich, M.; Klebe, G. Predicting binding modes, binding affinities and "hot spots" for Protein−ligand complexes using a knowledge-based scoring function. *Perspect. Drug Discovery Des.* **2000**, *20*, 115−144.

(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) Cozzini, P.; Fornabaio, M.; Marabotti, A.; Abraham, D. J.; Kellogg, G. E.; Mozzarelli, A. Simple, Intuitive Calculations of Free Energy of Binding for Protein−Ligand Complexes. 1. Models without Explicit Constrained Water. *J. Med. Chem.* **2002**, *45*, 2469−2483.

(18) 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*, 11−26.

(19) Verkhivker, G.; Appelt, K.; Freer, S. T.; Villafranca, J. E. Empirical free energy calculations of ligand-protein crystallographic complexes. I. Knowledge-based ligand-protein interaction potentials applied to the prediction of human immunodeficiency virus 1 protease binding affinity. *Protein Eng.* **1995**, *8*, 677−691.

(20) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Rose, P. W. Deciphering common failures in molecular docking of ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 731−751.

(21) 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.

(22) 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.

(23) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43−53.

(24) 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.

(25) 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.

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

(27) 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.

(28) Perez, C.; Ortiz, A. R. Evaluation of Docking Functions for Protein−Ligand Docking. *J. Med. Chem.* **2001**, *44*, 3768−3785.

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

(30) Wilton, D.; Willett, P. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.

(31) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.

(32) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L. Assessing scoring functions for protein−ligand interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.

(33) Roche, O.; Kiyama, R.; Brooks, C. L., III Ligand-Protein DataBase: Linking Protein−Ligand Complex Structures to Binding Data. *J. Med. Chem.* **2001**, *44*, 3592−3598. http://lpdb.scripps.edu/.

(34) Puvanendrampillai, D.; Mitchell, J. B. O. Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein−ligand complexes. *Bioinformatics* **2003**, *19*, 1856−1857. http://www-mitchell.ch.cam.ac.uk/pld/.

(35) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977−2980. http://www.pdbbind.org/.

(36) 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/.

(37) The SYBYL software, version 6.9, Tripos Inc., http://www.tripos.com/.

(38) The Cerius2 software, version 4.6, Accelrys Inc., http://www.accelrys.com/.

(39) Huang, C. C.; Lesburg, C. A.; Kiefer, L. L.; Fierke, C. A.; Christianson, D. W. Reversal of the Hydrogen Bond to Zinc Ligand Histidine-119 Dramatically Diminishes Catalysis and Enhances Metal Equilibration Kinetics in Carbonic Anhydrase II. *Biochemistry* **1996**, *35*, 3439−3446.

(40) Babine, R. E.; Bender, S. L. Molecular Recognition of Protein−Ligand Complexes: applications to Drug Design. *Chem. Rev.* **1997**, *97*, 1359−1472, and the references therein.

(41) Kollman, P. Free energy calculations: Application to chemical and biological phenomena. *Chem. Rev.* **1993**, *7*, 2395−2417, and the references therein.

(42) Aqvist, J.; Medina, C.; Samuelsson, J. E. New method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385−391.

(43) Carlson, H. A.; Jorgensen, W. L. Extended linear response method for determining free energies of hydration. *J. Phys. Chem.* **1995**, *99*, 10667−10673.

(44) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discovery Des.* **2000**, *18*, 113−135.

(45) 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.

(46) 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.

(47) 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.

CI049733J