Comparative Assessment of Scoring Functions on a Diverse Test Set

Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang*

State Key Laboratory of Bioorganic Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai, P. R. China

Received January 9, 2009

Scoring functions are widely applied to the evaluation of protein−ligand binding in structure-based drug design. We have conducted a comparative assessment of 16 popular scoring functions implemented in mainstream commercial software or released by academic research groups. A set of 195 diverse protein−ligand complexes with high-resolution crystal structures and reliable binding constants were selected through a systematic nonredundant sampling of the PDBbind database and used as the primary test set in our study. All scoring functions were evaluated in three aspects, that is, "docking power", "ranking power", and "scoring power", and all evaluations were independent from the context of molecular docking or virtual screening. As for "docking power", six scoring functions, including GOLD::ASP, DS::PLP1, DrugScore^{PDB}, GlideScore-SP, DS::LigScore, and GOLD::ChemScore, achieved success rates over 70% when the acceptance cutoff was root-mean-square deviation < 2.0 Å. Combining these scoring functions into consensus scoring schemes improved the success rates to 80% or even higher. As for "ranking power" and "scoring power", the top four scoring functions on the primary test set were X-Score, DrugScore^{CSD}, DS::PLP, and SYBYL::ChemScore. They were able to correctly rank the protein−ligand complexes containing the same type of protein with success rates around 50%. Correlation coefficients between the experimental binding constants and the binding scores computed by these scoring functions ranged from 0.545 to 0.644. Besides the primary test set, each scoring function was also tested on four additional test sets, each consisting of a certain number of protein−ligand complexes containing one particular type of protein. Our study serves as an updated benchmark for evaluating the general performance of today's scoring functions. Our results indicate that no single scoring function consistently outperforms others in all three aspects. Thus, it is important in practice to choose the appropriate scoring functions for different purposes.

INTRODUCTION

Structure-based drug design relies on some computational methods to tackle problems from lead identification to lead optimization and beyond.[1−9] Molecular docking is perhaps the most widely applied method in structure-based drug design, which predicts the preferred orientation of one molecule (*ligand*) to a second (*receptor*) when binding to each other to form a stable complex. As demonstrated in many previous studies,[10−12] today's molecular docking programs, such as DOCK,[13] AutoDock,[14−16] FlexX,[17] Surflex,[18,19] LigandFit,[20] GOLD,[21,22] and Glide,[12,23] can identify the correct binding pose of a flexible ligand to its receptor in seconds or minutes with reasonable accuracy. An essential component of these programs is a computational method evaluating the fitness between the ligand and receptor, which is normally referred to as scoring function. Such a scoring function guides the conformational and orientational search of ligand binding poses. Knowledge of the preferred binding pose in turn may be used to predict the strength of association between two molecules. If a whole library of molecules are docked onto a given molecular target, they can be ranked according to their predicted binding affinities, and only the most promising candidates are worth testing in subsequent experiments. This is the basic idea of virtual screening,[24−27] which is an established cost-effective approach to the discovery of novel lead compounds in structure-based drug design.

A variety of computational methods have been developed for computing receptor−ligand binding affinities, which are reviewed from time to time.[28,29] Free energy perturbation[30] and thermodynamics integration[31] conduct integration along the free energy pathway between two closely resembled systems. "End-point" approaches, such as MM-PB/SA[32] and linear interaction energy,[33,34] assume that the free energy change in a receptor−ligand binding process can be computed by only considering the difference between the unbound state and the bound state. All of the methods mentioned above rely on exhaustive conformational samplings, typically through molecular dynamics or Monte Carlo simulations, to compute ensemble averages. Unlike these methods, a scoring function typically considers only one low-energy snapshot of the receptor−ligand complex structure in computation. It does not compute ensemble averages or consider the unbound state of the two binding molecules explicitly. Therefore, they are fast enough for high-throughput applications in structure-based drug design, such as molecular docking and de novo design, and so on. In addition, scoring functions are often developed as generic models. They are in principle applicable to various protein−ligand binding systems without reparameterization, giving them another technical advantage in practice.

* Corresponding author phone: 86-21-54925128; e-mail: wangrx@mail.sioc.ac.cn.

10.1021/ci9000053 CCC: $40.75 © 2009 American Chemical Society
Published on Web 04/09/2009

Since the 1990s, several dozens of scoring functions have already been reported in the literature. New scoring functions are still emerging. Current scoring functions can be roughly classified as force-field-based methods,[13−16,21,22] empirical scoring functions,[12,17,23,35−44] and knowledge-based statistical potentials.[45−51] Force-field-based methods employ classical force fields to compute the direct noncovalent interactions between the protein and ligand, such as van der Waals and electrostatic energies. They are often augmented by a GB/SA or PB/SA term in order to compute solvation energies. Empirical scoring functions decompose the overall binding free energy into several energetic terms. Each term is computed with a somewhat intuitive algorithm, and the weight factors of all terms are derived from a regression analysis on a set of protein−ligand complexes with known binding affinities. Hence, empirical scoring functions are also referred to as regression-based methods. Knowledge-based scoring functions compute protein−ligand interactions as a sum of distance-dependent statistical potentials between the protein and ligand. A notable feature of them is that the deduction of such potentials only needs the knowledge of protein−ligand complex structures. Such knowledge is relatively rich and is still increasing rapidly due to the contributions from structural biologists.

Publicly available scoring functions are either implemented in commercial molecular modeling software or are released by researchers in academia. One certainly should not expect the performance of all scoring functions to be on the same level. Thus, an objective assessment of scoring functions has become a critical and intriguing subject. Such an assessment will help the users of scoring functions to choose the reliable ones in their studies. It is also desirable for the developers of scoring functions to improve their methods. A number of comparative studies of docking/scoring methods have already been reported in the literature.[10,11,52−67] A brief summary of such studies is given in the Supporting Information (part I). These studies basically took two approaches. The first approach tests scoring functions on some sets of protein−ligand complexes with known three-dimensional structures and binding affinity data.[52−55] Each scoring function is typically evaluated by its ability of reproducing the known binding poses and binding affinities of those protein−ligand complexes. The second approach evaluates scoring functions in combination with molecular docking programs.[10,11,56−67] The performance of each docking/scoring scheme is evaluated by its ability to reproduce the known binding poses of protein−ligand complexes. In addition, each docking/scoring scheme can also be evaluated by the enrichment factors observed in virtual screening trials against certain molecular targets, which reflect its ability to distinguish true active compounds from inactive ones. Normally, the inactive compounds considered in such virtual screening trials are randomly selected from some popular databases such as the Available Chemical Directory and the ZINC database.[68] Some special data sets, such as the Directory of Useful Decoys,[69] are also compiled to provide a more elaborate selection of decoys for virtual screening trials.

Our opinion is that the second approach mentioned above certainly has practical value since it can identify the optimal docking/scoring combinations on the molecular targets of interests. Nevertheless, it may not be appropriate for evaluating the intrinsic qualities of scoring functions because the final outcomes of a docking/scoring scheme are also affected by other factors. For example, the enrichment factor observed in a virtual screening trial is also dependent on the quality of the molecular docking program, the molecular target, and even the contents of the compound library considered in the screening. As indicated in the Supporting Information (part I), the results reported by this type of study are context-dependent and sometimes even conflicting. For example, Warren et al.[10] reported that GOLD produced higher enrichment factors than Glide in virtual screenings against Factor Xa, whereas Chen et al.[57] reported that Glide outperformed GOLD on the same target in similar virtual screening trials.

We thus choose the first approach for the purpose of assessing scoring functions. The key idea is to isolate the "scoring" step from the context of molecular docking or virtual screening so that the intrinsic qualities of scoring functions can be judged objectively while other factors have a minimal influence. This approach has been applied in one of our previous studies of scoring functions.[55] In that study, we used a set of 100 protein−ligand complexes to test a total of 11 scoring functions. All scoring functions were applied to the ensembles of possible ligand binding poses, which were generated previously, to see if they were able to identify the true binding poses among decoys. In this way, different scoring functions were compared objectively on the same grounds. This approach is well-accepted by the scientific community. In fact, the test set used in that study has since become a benchmark adopted in a number of studies on scoring methods.[70−75]

Here, we will report another comparative assessment of scoring functions following the same approach. This study has substantial improvements in several aspects as compared to our previous study.[55] It covered a larger collection of 16 scoring functions implemented in main-stream commercial software or available from academic research groups. Considering the possible roles of scoring functions in structure-based drug design, these scoring functions were evaluated in terms of three basic features, namely "docking power", "ranking power", and "scoring power". A series of tests regarding these features was conducted on a high-quality set of 195 diverse protein−ligand complexes and four additional sets of particular protein−ligand complexes. Our study represents an updated benchmark for evaluating the general performance of today's scoring functions, which provides useful guidance for both the users and the developers of scoring functions.

MATERIALS AND METHODS

Scoring Functions under Assessment. A total of 14 scoring functions implemented in several main-stream molecular modeling software programs were assessed in our study, including five scoring functions (LigScore,[35] PLP,[38,39] PMF,[48−51] Jain,[42] and LUDI[43,44]) in the Discovery Studio software (version 2.0),[76] five scoring functions (D-Score, PMF-Score, G-Score, ChemScore, and F-Score) in the SYBYL software (version 7.2),[77] GlideScore[12,23] in the Schrödinger software (version 8.0),[78] and three scoring functions (GoldScore, ChemScore,[40,41] and ASP[45]) in the GOLD software (version 3.2).[21,22] In addition, two stand-alone scoring functions released by academic groups, that is, DrugScore[46,47] and X-Score (version 1.2),[36] were also

assessed. Compared to the 11 scoring functions assessed in our previous study,[55] five new scoring functions were added in this study. This panel provides a fairly broad coverage of the scoring functions available to the public today. These scoring functions can be classified into three categories: (*i*) force-field-based methods, including GOLD::GoldScore and SYBYL::G-Score/D-Score; (*ii*) empirical scoring functions, including GOLD::ChemScore, DS::LigScore/PLP/Jain/LUDI, SYBYL::F-Score, GlideScore, and X-Score; and (*iii*) knowledge-based statistical potentials, that is, potentials of mean forces (PMF), including GOLD::ASP, DS::PMF, SYBYL::PMF-Score, and DrugScore. Brief descriptions of all 16 scoring functions along with the key parameters used in our computation are given in the Supporting Information (part II).

Several scoring functions in our test have different versions or provide multiple options, including LigScore (LigScore1 and LigScore2), PLP (PLP1 and PLP2), and LUDI (LUDI1, LUDI2, and LUDI3) in Discovery Studio; GlideScore (GlideScore-SP and GlideScore-XP); DrugScore (DrugScore^{PDB} and DrugScore^{CSD}); and X-Score (HPScore, HMScore, and HSScore). All of these variations have been assessed in our study. If all of these variations are treated as different scoring functions, the total number of scoring functions considered in our study was actually 29. For the sake of convenience, in each test, only the results produced by the best version/option of a certain scoring function are reported in this manuscript. The complete set of results can be found in the Supporting Information (parts VI and VII). In addition, the scoring functions in our test produce binding scores in different units and signs. In our study, the signs of the binding scores produced by some scoring functions, including GlideScore, DrugScore, and the five scoring functions in SYBYL, were reversed so that more positive binding scores always indicated higher binding affinities.

Compilation of Test Sets. We chose to use a total of 195 diverse protein−ligand complexes as the primary test set in our study. This test set was selected through a systematic mining of the PDBbind database.[79,80] The PDBbind database provides a collection of the experimentally determined binding data of the protein−ligand complexes deposited in the Protein Data Bank (PDB).[81] This database is now maintained through collaboration between our group and Prof. Shaomeng Wang's group at the University of Michigan. The 2007 version of the PDBbind database was considered in our study, which consisted of binding data of over 3100 protein−ligand complexes. Not all of them, however, are "healthy" enough for the purpose of validating scoring functions. Therefore, we applied a number of filters in order to select among them the qualified ones. These filters can be summarized briefly as follows:

(1) Concerns about the quality of structures. Only the protein−ligand complexes whose structures are determined through crystal diffraction were considered. Each qualified complex structure must have an overall resolution better than or equal to 2.5 Å. In addition, both the protein and the ligand need to be complete in the crystal structure.

(2) Concerns about the quality of binding data. Only the protein−ligand complexes with known dissociation constants (K_d) or inhibition constants (K_i) were considered. In addition, both the protein and the ligand used in the binding assay have to match exactly the ones used in structure determination.

Figure 1. Selection of the primary test set. Each pawn represents a protein−ligand complex, whose height indicates its binding affinity. Complexes formed by different types of proteins are in different colors.

(3) Concerns about the components of complexes. Only noncovalently bound protein−ligand complexes were considered. Each qualified complex should be formed by one protein molecule and one ligand molecule in a binary manner. In other words, there should not be multiple ligands bound in close vicinity at a common binding site. The ligand molecule must not contain any uncommon elements, such as Be, B, Si, and metal atoms. In addition, its molecular weight shall not exceed 1000. Oligo-peptides (up to nine residues) and oligo-nucleotides (up to three residues) are also considered as valid small-molecule ligands.

The outcome of the above selection is the so-called "refined set" of the PDBbind database. The refined set of PDBbind (version 2007) consists of 1300 protein−ligand complexes. This set of complexes could not be adopted directly as the primary test set for our study since it has a considerable level of redundancy in its contents. Nevertheless, it provided a good starting point for selecting the test set for our study. For this purpose, the refined set was grouped into clusters by sequence similarity computed by BLAST. A similarity cutoff of 90% was applied in clustering. Each resulting cluster typically consisted of complexes formed by a particular type of protein. A total of 65 clusters in the refined set were found to contain at least four protein−ligand complexes. For each cluster, the one with the highest binding affinity, the one with the lowest binding affinity, and the one with a binding affinity close to the mean value were selected as the representatives of this cluster. As a result, a total of $65 \times 3 = 195$ protein−ligand complexes were selected, which is termed by us as the "core set" of the PDBbind database. A graphical illustration of the above procedure is given in Figure 1. A full list of the protein−ligand complexes included in our primary test set is given in the Supporting Information (part III).

In addition to the primary test set, four test sets were also used in our study, each containing a certain number of protein−ligand complexes formed by one particular type of protein. The first set contained 112 HIV protease complexes, the second contained 73 trypsin complexes, the third

contained 44 carbonic anhydrase complexes, and the last contained 38 thrombin complexes (Supporting Information, part III). All of these complexes were also selected from the refined set of PDBbind (version 2007), and thus they had the same level of quality in terms of structure and binding data as the primary test set. These four proteins were chosen since they were the four most populated proteins in this data set. These four additional test sets overlapped with the primary test set because the latter also contained complexes formed by these four types of proteins. The overlapping parts, however, could be safely ignored since typically only three protein−ligand complexes were relevant in each case.

Preparation of the Complex Structures. Coordinates of the complexes in all test sets were downloaded from the Protein Data Bank.[81] The original structural files from the PDB were processed so that they could be readily utilized by the software used in our study. Basically, each complex was split into a complete "biological unit" of the protein molecule and the ligand molecule. Atomic types and bond types of the ligand molecule were automatically assigned by the I-interpret program.[82] They were then visually inspected and corrected if necessary. Hydrogen atoms were added to the protein molecule and the ligand molecule by using the SYBYL software. For the sake of convenience, both the protein and the ligand were set according to a simple protonation scheme under neutral pH: all carboxylic acid and phosphonate groups were deprotonated, while all aliphatic amine and guanidino/amidino groups were protonated. In order to apply some force-field-based scoring functions, the protein molecule was assigned the AMBER FF99 charges, while the ligand was assigned the MMFF94 charges. We did not attempt to explore the influence of other charge sets on our assessment results because, among all 16 scoring functions in our test, only the force-field-based SYBYL:: D-Score considers atomic partial charges. All water molecules included in the crystal structure were removed since no scoring function under our assessment was really able to consider them. The protein was saved in the PDB format, while the ligand was saved in the Mol2 format and the SD format. Metal ions, if residing inside the binding pocket and coordinately bound to the ligand and the protein, were saved with the protein molecule. No structural optimization was performed at this step on either the protein or the ligand in order to retain their coordinates exactly the same as those in the original PDB file.

Besides the original structure of each protein−ligand complex in the primary test set, a decoy set of the ligand binding poses was needed in order to evaluate the "docking power" of a scoring function, which will be explained later in this manuscript. The decoy set of each protein−ligand complex used in our study was generated through a multistep process, which is illustrated conceptually in Figure 2. First, four molecular docking programs, including LigandFit in Discovery Studio, Surflex and FlexX in SYBYL, and GOLD, were applied to generate an initial ensemble of the possible binding poses of the given ligand. Relevant parameters used by each molecular docking program were carefully controlled to produce diverse binding poses rather than some converged ones. Detailed descriptions of this step are given in the Supporting Information (part IV). The outputs from all four programs were combined, which typically resulted in an ensemble of ~2000 binding poses for each protein−ligand complex. Binding poses whose root-mean-square deviations (rmsd's) from the true binding pose, that is, the one observed in the crystal structure, were greater than 10 Å were identified and discarded, since they were typically well off the desired binding site. Second, all remaining binding poses were grouped into 10 bins with an interval of 1 Å according to their rmsd values with respect to the true binding pose. The binding poses in each bin were further clustered into 10 clusters according to their internal similarities (also measured quantitatively by rmsd values) using the "*rms_analysis*" tool in the GOLD software. The binding pose with the lowest noncovalent interaction energy with the protein in each cluster was selected as the representative of that cluster. The noncovalent interactions between each ligand binding pose and the protein, including van der Waals and electrostatic components, were computed with the Tripos force field by using the SYBYL software. The final outcome of the above process was a total of $10 \times 10 = 100$ nonredundant, low-energy binding poses of the ligand for each given protein−ligand complex.

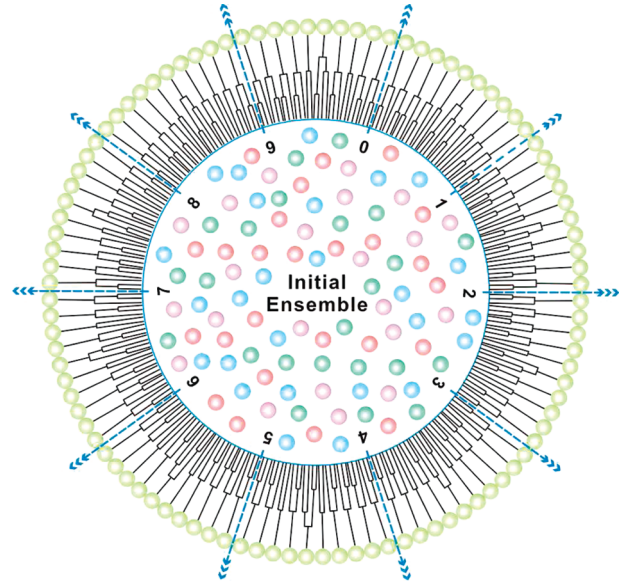

Figure 2. Preparation of the decoy set of each protein−ligand complex. Each small circle represents a certain binding pose of the ligand. An initial ensemble of binding poses was generated by four molecular docking programs. They were grouped into 10 bins according to their rmsd values with respect to the true binding pose. The binding poses in each bin were then clustered to select 10 representative low-energy binding poses. The final outcomes were a total of $10 \times 10 = 100$ binding poses evenly distributed between rmsd = 0 and 10 Å, that is, the green circles on the outer shell.

Evaluation Methods. We propose that a scoring function can be evaluated in three essential aspects, namely, "docking power", "ranking power", and "scoring power". These three features correspond to the three possible roles of scoring functions in structure-based drug design.

i. "Docking Power". This refers to the ability to identify the true ligand binding pose among computer-generated decoys. Ideally, the true binding pose should be identified as the one with the best binding score. This ability is essential for a scoring function used in a molecular docking program ("docking function") to make reasonable predictions. In our study, each scoring function under assessment was applied to score the decoy set of each protein−ligand complex in the primary test set. In each case, the rmsd between the best-scored binding pose and the true binding pose of the ligand

was computed regarding all heavy atoms. If this rmsd value fell within a predefined range, for example, 2.0 Å, it was recorded as a success. After this judgment was completed on the entire test set, an overall success rate was obtained for the given scoring function as a quantitative measurement of its docking power. Another set of tests after including the true ligand binding pose in the decoy set of each protein−ligand complex was also conducted. Both sets of results will be reported in this manuscript.

ii. "Ranking Power". This refers to the ability to correctly rank different ligands bound to the same protein according to their binding affinities when the correct binding poses of these ligands are known. This ability is obviously desired in virtual screening since an ideal virtual screening is expected to rank the compounds with higher binding affinities to the top. As described earlier, our primary test set contained 65 families of protein−ligand complexes. Each family consisted of three complexes formed between a high-affinity, a medium-affinity, and a low-affinity ligand and a common protein. Selection of these three complexes aimed at maximizing the binding affinity range in each family as much as possible. In our study, each scoring function under assessment was applied to compute a binding score for each protein−ligand complex in the primary test set. We then examined whether the rank of the three-member complexes in each family by the computed binding scores was in accordance with their known binding affinities. If so, a successful case was recorded for the given scoring function. Note that, in theory, there are six possible ways to rank three samples, but only one of them is correct. An overall success rate of the given scoring function was obtained after this examination was completed on this test set.

The ranking power of each scoring function was also assessed on the four additional test sets. In each case, the ranking power was measured by the Spearman correlation coefficient (R_s) as well as the Pearson correlation coefficient (R_p) between the known binding constants and the binding scores produced by the given scoring function.

iii. "Scoring Power". This refers to the ability of producing binding scores that are correlated, preferably in a linear manner, with experimentally measured binding affinities when protein−ligand complex structures are known. The ranking power defined above is evaluated on different ligands bound to a common target protein. In contrast, scoring power emphasizes the performance of a scoring function across different types of protein−ligand complexes. It measures the general ability of a scoring function in binding affinity prediction, which is probably the most challenging aspect among all three evaluated in our study. Each scoring function in our test was applied to compute the binding scores of the 195 complexes in the primary test set. The scoring power of each scoring function on this test set was measured by the Pearson correlation coefficient (R_p) between its binding scores and the known binding constants. Note that, if a scoring function produces a negative (unfavorable) binding score or fails to compute a certain protein−ligand complex due to miscellaneous reasons, this case will not be considered in the computation of the above two statistical properties.

It is possible that some steric clashes still exist between the protein and ligand even in high-resolution crystal structures. Some scoring functions are sensitive to such clashes and will not produce meaningful binding scores in such cases. To address this problem, we repeated our computation on optimized protein−ligand complex structures. In each case, the protein structure was kept fixed. The observed ligand binding pose was relaxed within its binding site using the built-in protocols in SYBYL, Discovery Studio, GOLD, and Schrödinger, respectively, in order to apply the scoring functions implemented in these software suites (Supporting Information, part II). Here, we did not attempt to derive a "standard" set of optimized ligand binding poses to evaluate all scoring functions. Instead, each scoring function was evaluated on the structures prepared by the same software in which it was implemented, an approach most likely to be adopted in practice. As for the two stand-alone scoring functions, that is, DrugScore and X-Score, they were evaluated on the optimized ligand binding poses produced by the Discovery Studio software. Both sets of results, which were obtained on original and optimized complex structures separately, will be reported in this manuscript.

Classification of Subsets of Protein−Ligand Complexes. We also examined each scoring function on some particular subsets of protein−ligand complexes sharing common properties in an attempt to obtain more subtle judgments on its performance. For this purpose, we computed three properties relevant to protein−ligand binding for each complex in the primary test set. The first was the buried percentage of the solvent-accessible surface area of the ligand upon binding. The second was the buried percentage of the molecular volume of the ligand inside the binding pocket. These two properties measure how much a ligand molecule is buried inside the binding site. The third was a "hydrophobic index" of the binding pocket. It was computed by summing up the fragmental log *D* value of each amino acid residue that was in direct contact with the bound ligand. Detailed methods for computing these properties are given in the Supporting Information (part V).

For each given property, the Z-Score of the *i*th protein−ligand complex in the primary test set was computed as

$$\text{Z-Score}_i = \frac{f_i - \mu}{\sigma}$$

Here, f_i is the value of a certain property of this complex, while μ and σ are the mean value and the standard deviation of this property observed on the entire test set, respectively. The entire test set was then divided into three subsets with Z-Scores falling in the ranges of $(-\infty, -1)$, $[-1, +1]$, and $(+1, +\infty)$, separately. Conceptually, these three subsets consisted of samples which were considerably lower than the average, around the average, and considerably higher than the average with respect to a given property. The docking power and scoring power of each scoring function under our assessment were recalculated on all three sets of subsets. Since the three members in each complex family did not necessarily fall into the same subset, the ranking power of each scoring function was not assessed on these subsets.

RESULTS AND DISCUSSION

Selection of the Primary Test Set. The two cornerstones of our assessment of scoring functions are the test sets and the evaluation methods. Our study aims at assessing the general performance of scoring functions. A desired test set for this purpose thus should have the following features: (1) Each protein−ligand complex in this test set must have a

high-resolution structure as well as reliable binding data in order to validate scoring functions. (2) It must contain a variety of protein−ligand complexes rather than a congeneric set of protein−ligand complexes. (3) It must contain a substantial number of samples to achieve statistical significance. At the same time, it should not contain too many samples to be computationally acceptable.

The primary test set used in our study is based on the PDBbind refined set, which itself is an assembly of protein−ligand complexes with both high-resolution structures and reliable binding data selected by a set of stringent criteria. Thus, it meets the first requirement mentioned above. This test set consists of a total of 195 protein−ligand complexes formed by 65 different species of proteins. Binding constants of these complexes range from 1.40 to 13.96 (in log units), spanning over 12 orders of magnitude. Molecular weights of the ligands in these complexes range from 103 to 974, while numbers of the rotatable bonds in these ligand molecules range from 0 to 32. This test set meets the second requirement in terms of the diversity at the protein side as well as the ligand side. The total number of samples included in this test set, that is, 195, is also adequate for achieving statistical significance. Although the size of the test set is not the primary goal of our efforts, our test set is in fact already larger than most test sets used in other comparative studies of docking/scoring (see the Supporting Information, part I). The computational cost of a test set of this size is also very acceptable to us. Thus, this test set meets all of the requirements mentioned above.

This test set is a major improvement as compared to our previous study of scoring functions.[55] In that study, a test set of 100 diverse protein−ligand complexes was used for scoring function assessment. Although those complexes were also selected by a set of criteria to ensure the quality of complex structures and binding data, the starting pool for selection was actually a random collection of 230 protein−ligand complexes in the PDB. Therefore, only limited diversity was presented in that test set. In addition, there was also a certain level of redundancy among the final 100 selected protein−ligand complexes. For example, there were 10 complexes formed by trypsin, accounting for 10% of the total population, whereas there was only one complex formed by HIV-1 protease. Consequently, one would expect that the assessment results based on this test set would be biased more toward trypsin than HIV-1 protease. In fact, the above drawbacks were also shared more or less by the test sets used in other comparative studies on docking/scoring methods (see the Supporting Information, part I), such as the study by Ferrara et al. (189 complexes),[52] the study by Marsden et al. (205 complexes),[53] the study by Chen et al. (164 complexes),[57] and the study by Perola et al. (150 complexes).[62] In contrast, the test set used in this study was compiled through a systematic sampling of the PDBbind database. Since the PDBbind database is based on the entire PDB, this test set can also be considered as the outcome of a systematic mining of the entire PDB. Moreover, our test set is compiled with a strict control on redundancy and an emphasis on maximal diversity. A test set like this will certainly provide a solid ground for assessing scoring functions. To the best of our knowledge, the only set of protein−ligand complexes employed in the development/validation of docking/scoring methods with quality comparable to ours is the Astex diverse set.[83] This data set, a total of 85 high-quality protein−ligand complexes, was also derived through a systematic mining of the entire PDB with an emphasis on diversity.

Docking Power on the Primary Test Set. Molecular docking is basically a process of conformational sampling, aimed at determining the most favorable binding pose of a given ligand to its target receptor. Most molecular docking programs rely on scoring functions to evaluate the fitness between the ligand and receptor. Thus, the "docking power" of a scoring function is of vital importance for this purpose. As described in the Materials and Methods section, our assessment of the docking power was based on the decoy sets prepared for each protein−ligand complex in the test set. This approach has been successfully applied in our previous study[55] as well as some other studies on scoring functions.[52] Obviously, such a decoy set would better sample the possible binding poses of a given ligand as completely as possible. In principle, it can be generated by using a molecular docking program or through molecular dynamics simulation. We prefer the former approach since the outcomes are relatively easier to control. In our previous study,[55] the AutoDock program was employed for this purpose. Relying on one particular molecular docking program may not be the best solution since each molecular docking program has its own bias in conformational sampling. In order to overcome this potential pitfall, we have employed four molecular docking programs in this study, including LigandFit, GOLD, Surflex, and FlexX. These four programs adopt a shape-directed algorithm, a genetic algorithm, a molecular similarity-based algorithm, and an incremental construction algorithm as the conformational search engine, respectively. Combining the outcomes of these four programs is likely to cover the possible binding poses of a given ligand more thoroughly. In addition, the final 100 low-energy binding poses were selected through systematic clustering. Little human interference was needed during the entire process. Compared to the one applied in our previous study, our new method for preparing the decoy set of each complex is apparently more reasonable. The docking power of each scoring function is expected to be reflected more objectively on these new decoy sets.

The very basic approach in our study for evaluating the docking power of a given scoring function was to examine whether the best-scored ligand binding pose selected by this scoring function resembles the one observed in the crystal structure closely enough. Success rates of all 16 scoring functions under three different cutoffs (rmsd < 1.0, 2.0, and 3.0 Å) are shown in Figure 3. One can see that these scoring functions perform quite differently in this test: three scoring functions produce success rates over 60% when the acceptance cutoff is rmsd < 1.0 Å, including GOLD::ASP, DS::PLP1, and DrugScore^{PDB}, while several scoring functions produce success rates below 30%. It is not surprising to observe that the success rates of all scoring functions increase under lower standards. For example, GOLD::ASP achieves a high success rate close to 90% when a low-resolution docking (rmsd < 3.0 Å) is acceptable.

Figure 4 compares the success rates of all scoring functions if the true ligand binding pose is not included in the decoy set for each protein−ligand complex. After this treatment, the success rates of all scoring functions decrease by 0∼5%,

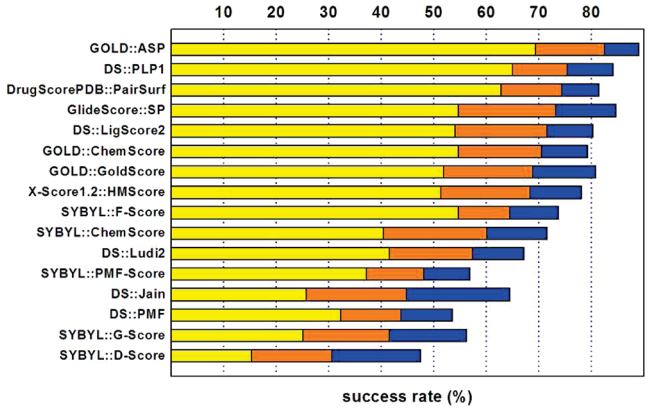

Figure 3. Comparison of the success rates of 16 scoring functions on the primary test set when the cutoff is rmsd < 1.0 Å (yellow bars), < 2.0 Å (orange bars), or < 3.0 Å (blue bars), respectively. The true ligand binding poses were included in the decoy sets in this test. Scoring functions are ranked by the success rates when the acceptance cutoff is rmsd < 2.0 Å.

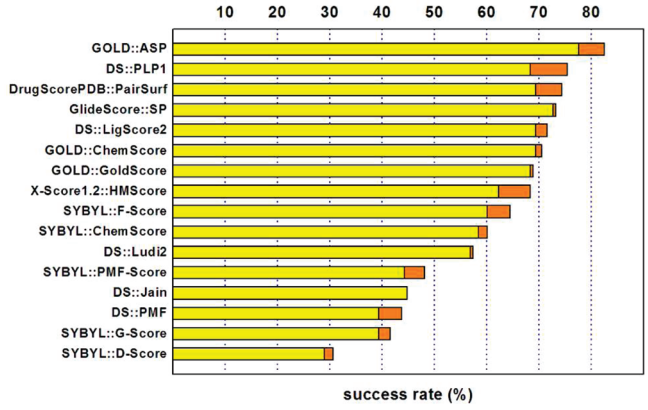

Figure 4. Comparison of the success rates of 16 scoring functions on the primary test set when the true binding pose is included (orange bars) or not (yellow bars) in the decoy set of each protein−ligand complex. The acceptance cutoff was rmsd < 2.0 Å in this test. Scoring functions are ranked by the success rates when the true binding poses are considered.

which is not significant. This indicates that the decoy sets of most protein−ligand complexes in our test set include some binding poses that are close enough to the true binding poses, and thus, whether the true binding poses are included in the decoy sets or not does not have a significant impact on the outcomes of our test. This should be attributed to the new method for preparing the decoy sets developed in this study. Adding the true ligand binding pose to the decoy set will certainly lead to an even more complete sampling of ligand binding poses. For the sake of convenience, we will only report and discuss the results when the true ligand binding poses are included in the decoy sets in the remaining parts of this manuscript unless specified.

The above analyses are all based on the results when only the best-scored ligand binding pose is considered in each case. In practice, it is normally possible to let a molecular docking program output multiple binding poses of a given ligand for further selection. Figure 5 compares the success rates of all 16 scoring functions if one, two, or three top-ranked ligand binding poses are considered. One can see that success rates of all scoring functions increase considerably if more top-ranked binding poses are considered. In particular, the three selected scoring functions, that is, GOLD::ASP,

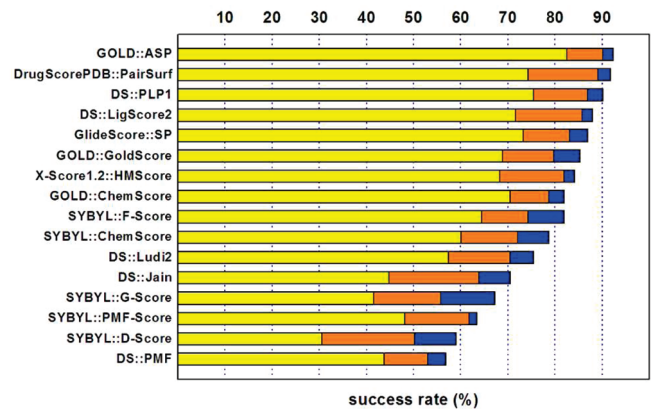

Figure 5. Comparison of the success rates of 16 scoring functions on the primary test set when a ligand binding pose is found within rmsd < 2.0 Å from the true one if the top one (yellow bars), the top two (orange bars), or the top three (blue bars) best-scored binding poses are considered. Scoring functions are ranked by the success rates when the top three binding poses are considered.

DS::PLP1, and DrugScore^{PDB}, achieve high success rates over 90% by considering three top-ranked binding poses. Our results indicate that it is a good practice in molecular docking to analyze a few representative top-ranked binding poses of a given ligand molecule. If the correct binding pose is missed as the best-scored one, it is probably among the next few if a good scoring function is employed.

It should be noted that the ranks of all scoring functions by their success rates does not get altered much in our tests. Three scoring functions, that is, GOLD::ASP, DS::PLP1, and DrugScore^{PDB}, outperform others under different criteria. DS::PLP1 and DrugScore^{PDB} were also identified among the best ones in terms of docking power in our previous study,[55] while GOLD::ASP is a newly developed scoring function since then. It is also interesting to notice that all three of these scoring functions are based on computing pairwise interactions between the protein and ligand, suggesting that this type of scoring function may be the preferred choice of "docking functions".

Multiple scoring functions can be further combined into consensus scoring schemes[84] in order to obtain improved performances. Consensus scoring has become a popular practice, especially in structure-based virtual screening studies, although its rationale is still a subject of study.[85−87] In this study, we have also tested all possible combinations of the three selected scoring functions mentioned above plus GlideScore-SP. Since the binding scores calculated by different scoring functions are typically given in different units, technically it is not possible to compute the consensus scores simply by summing up the binding scores given by all individual scoring functions. Therefore, we adopt the "rank-by-rank" strategy in consensus scoring[87] to combine the results of multiple scoring functions; that is, the consensus score of each binding pose in the decoy set is the average rank given by all individual scoring functions in a given consensus scoring scheme. For example, if a certain binding pose is ranked as number 1 by scoring function A and as number 5 by scoring function B, its final consensus score by this double-scoring scheme is $(1 + 5)/2 = 3$. The best-scored binding pose is then compared with the true binding pose, and the success rate of each consensus scoring scheme is derived accordingly. The results are summarized in Table 1.

Table 1. Success Rates of Some Consensus Scoring Schemes in "Docking Power" Evaluation[a]

scoring function	success rate (%)	double scoring	success rate (%)	triple scoring	success rate (%)	quadruple scoring	success rate (%)
A	82.5	A + B	85.8	A + B + C	83.1	A + B + C + D	88.0
B	75.4	A + C	88.0	A + B + D	86.3		
C	74.3	A + D	86.3	A + C + D	86.3		
D	73.2	B + C	80.3	B + C + D	83.1		
		B + D	82.5				
		C + D	80.9				

[a] A = GOLD::ASP; B = DS::PLP1; C = DrugScore^{PDB}; D = GlideScore-SP.

Table 2. Correlations between the Experimentally Measured Binding Constants and the Binding Scores Computed by 16 Scoring Functions on the Primary Test Set

	on original complex structures				on optimized complex structures			
scoring function[a]	N[b]	R_p[c]	SD[d]	R_s[e]	N	R_p	SD	R_s
X-Score::HMScore	195	0.644	1.83	0.705	195	0.649	1.82	0.701
DrugScore^{CSD}	195	0.569	1.96	0.627	195	0.589	1.93	0.649
SYBYL::ChemScore	195	0.555	1.98	0.585	194	0.622	1.87	0.668
DS::PLP1	195	0.545	2.00	0.588	194	0.529	2.03	0.569
GOLD::ASP	193	0.534	2.02	0.577	194	0.518	2.04	0.558
SYBYL::G-Score	195	0.492	2.08	0.536	195	0.522	2.03	0.579
DS::LUDI3	195	0.487	2.09	0.478	194	0.477	2.10	0.478
DS::LigScore2	193	0.464	2.12	0.507	194	0.479	2.10	0.505
GlideScore-XP	178	0.457	2.14	0.435	187	0.555	2.01	0.556
DS::PMF	193	0.445	2.14	0.448	194	0.471	2.11	0.482
GOLD::ChemScore	178	0.441	2.15	0.452	186	0.528	2.05	0.553
by NHA[f]	195	0.431	2.15	0.517	195	0.431	2.15	0.517
SYBYL::D-Score	195	0.392	2.19	0.447	195	0.388	2.20	0.443
DS::Jain	189	0.316	2.24	0.346	190	0.339	2.26	0.362
GOLD::GoldScore	169	0.295	2.29	0.322	188	0.329	2.26	0.386
SYBYL::PMF-Score	190	0.268	2.29	0.273	180	0.235	2.31	0.235
SYBYL::F-Score	185	0.216	2.35	0.243	181	0.238	2.31	0.208

[a] Scoring functions are ranked by the Pearson correlation coefficients obtained on the original complex structures. [b] Number of complexes receiving positive (favorable) binding scores by this scoring function. [c] Pearson correlation coefficients. [d] Standard deviations in linear correlation (in log K_d units). [e] Spearman correlation coefficients. [f] Using the number of heavy atoms on each ligand as the only variable in correlation.

As one can see in Table 1, all consensus scoring schemes tested in our study outperform any individual scoring function. The improvements are not trivial (>10%) in some cases. Thus, our results support the idea that consensus scoring is also an effective strategy for identifying correct ligand binding poses. However, it is largely unpredictable which combinations of scoring functions will produce the optimal results. Therefore, one may want to test all possible combinations of scoring functions on appropriate samples in practice. It should be emphasized though that it is not reasonable to include relatively poor scoring functions in a consensus scoring scheme. In addition, the advantage of triple-scoring or even quadruple-scoring schemes over double-scoring schemes seems to be unclear in our test. Double-scoring may be reliable enough for practical uses.

Scoring Power on the Primary Test Set. Predicting the correct binding mode of a given ligand and its binding affinity are two related but different aims for scoring functions. Our definition of the "scoring power" of a scoring function emphasizes the ability to produce binding scores correlated to experimentally measured binding affinities across diverse protein−ligand complexes. The statistical data produced by all 16 scoring functions are summarized in Table 2. It has been repeatedly observed that, in many test sets used for evaluating binding affinity predictions, there is a strong correlation between the size of the ligands and their binding affinities. Therefore, we also investigated this issue in our study. We chose to use the number of heavy atoms (NHA) on each ligand to quantify the size of each ligand. Unlike some other properties related to molecular size, such as volume or surface area, this property does not need any special algorithm or parameter to compute and thus is readily reproducible by other researchers. The correlation coefficient between NHAs and experimental binding constants of the primary test set was computed to be 0.431. Embarrassingly, one can see that the results produced by nearly half of the scoring functions in our test are not better than this (Table 2). The top three scoring functions in this test are X-Score, DrugScore^{CSD}, and SYBYL::ChemScore, which produced correlation coefficients between 0.55 and 0.64 and standard deviations below 2.00 log K_d units (corresponding to ~2.6 kcal/mol in terms of standard binding free energy at room temperature) in original crystal structures. DS::PLP is arguably in fourth place with a slightly inferior performance. The scoring power of these scoring functions is apparently superior to the NHA-based approach. Scatter plots of the experimental binding constants and the computed binding scores produced by these four scoring functions are given in Figure 6.

It is also interesting to investigate the intercorrelations between the scoring functions mentioned above. The correlations between the binding scores computed by the four scoring functions selected above are summarized in Table 3. One can see that at least a moderate correlation can be

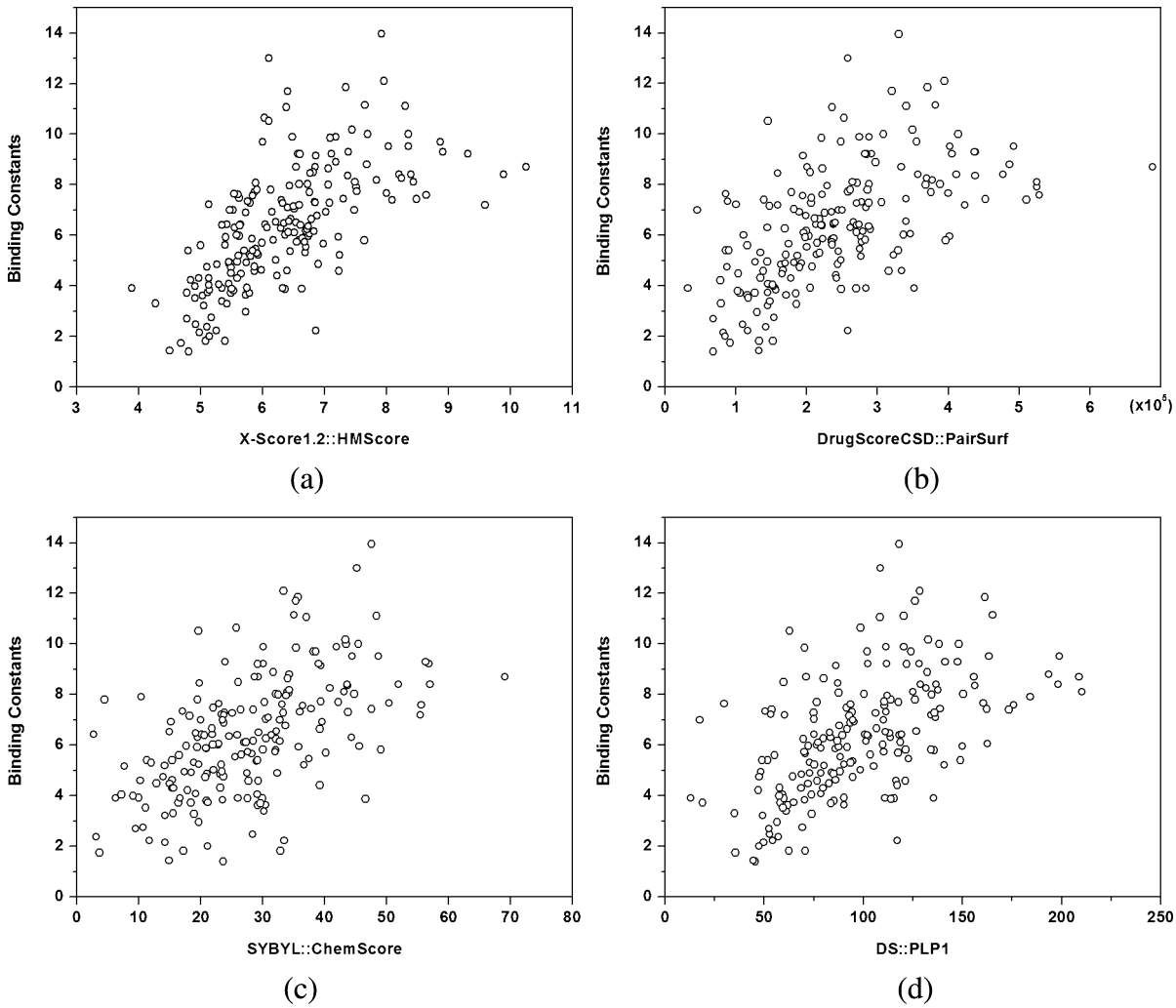

Figure 6. Correlations between the experimentally measured binding constants (in $-\log K_d$ units) of the 195 protein−ligand complexes in the primary test set and the binding scores computed by (a) X-Score::HMScore ($R = 0.644$), (b) DrugScore^{CSD}::PairSurf ($R = 0.569$), (c) SYBYL::ChemScore ($R = 0.555$), and (d) DS::PLP1 ($R = 0.545$).

Table 3. Intercorrelations between the Binding Scores Computed by Four Selected Scoring Functions on the Primary Test Set ($N = 195$)

correlation coefficient (R)	DrugScore^{CSD}:: PairSurf	SYBYL:: ChemScore	DS::PLP1
X-Score::HMScore	0.854	0.769	0.783
DrugScore^{CSD}::PairSurf		0.640	0.925
SYBYL::ChemScore			0.594

observed between any two of them. This is understandable since virtually all scoring functions are designed to reflect some basic features in protein−ligand interactions, such as hydrogen bonds and hydrophobic contacts. Moreover, the binding scores computed by these scoring functions are all correlated to the known binding constants to some extent so that some intercorrelations among themselves are natural. For some reason, DrugScore^{CSD} and DS:PLP1 exhibited a very high correlation ($R = 0.925$) in the primary test set, whereas the lowest correlation ($R = 0.594$) is observed between SYBYL::ChemScore and DS:PLP1. It is reasonable to expect that an effective consensus scoring scheme would better combine complementary scoring functions rather than highly correlated ones. As indicated in Table 1, the consensus scoring schemes containing both DrugScore and PLP1 indeed performed less successfully as compared to the other schemes at the same level of complexity.

The four scoring functions mentioned above do not have a problem in computing all of the protein−ligand complexes in our primary test set using the original crystal structures. In contrast, a few scoring functions in our test, including GlideScore, GOLD::ChemScore, and GOLD::GoldScore, fail to produce meaningful binding scores, that is, positive binding scores, in a certain number of protein−ligand complexes in this scenario (Table 2). All three of these scoring functions have a term accounting for the repulsions between the protein and ligand and therefore are sensitive to the remaining clashes in crystal structures. Somewhat improved results are obtained for them when optimized ligand binding poses are used in computation instead. This indicates that some sort of optimization of ligand binding poses is desired prior to the application of these scoring functions. Interestingly, the statistics of X-Score and DrugScore^{CSD} do not get altered much either in original crystal structures or optimized structures. Such scoring functions may be more welcome in practice because their results are not sensitive to minor changes in ligand binding poses.

As indicated in Table 2, the scoring powers exhibited by today's scoring functions are apparently not at the same level

as their docking powers. Even the best scoring functions in our test produce only moderate correlations between their binding scores and experimental binding constants. This level of performance can be understood considering the remarkable difficulties in binding affinity prediction. In fact, it is quite common in molecular modeling that structures are relatively easier to predict than energetic properties. On the other hand, our primary test set is specially designed to present a maximal diversity on the protein side as well as the ligand side. Some protein−ligand complexes with extremely high or extremely low binding affinities are included, which are certainly very difficult to interpret and thus have significant negative impacts on the statistical data (*R* and SD values) of scoring functions (see Figure 6 and the Supporting Information, part VI). It is actually encouraging to observe that some scoring functions still produce acceptable correlations in such a diverse test set. As far as we know, no other methods for binding affinity calculation aside from scoring functions have been tested extensively in this manner. In addition, statistical data, such as *R* and SD values, are strictly dependent on the contents of the data set under study. As discussed later in this paper, some scoring functions are certainly able to produce nice correlations in particular sets of protein−ligand complexes. It is seen too often in the literature that scoring functions, and sometimes other QSAR models, are compared simply by statistical data regardless of context. It is our opinion that a fair comparison of different scoring functions has to be made on a common benchmark.

Our results also suggest that regression-based empirical scoring functions, if they are well-developed, seem to be more capable in terms of scoring power, although knowledge-based scoring functions may produce acceptable results as well. The true predictive power of a regression-based model outside its training set could be a matter of concern. We tested the regression-based scoring functions, such as X-Score, SYBYL::ChemScore, DS::PLP1, and GlideScore, in their "native" forms available to us in order to obtain results which are reproducible by other researchers. The original training sets of these scoring functions all have certain overlaps with the primary test set used in our study. The impact of such overlaps is actually a subtle issue, since such overlaps may or may not have positive contributions to the performance of these scoring functions. In order to investigate this issue, we recalibrated X-Score using the remaining 1105 protein−ligand complexes in the PDBbind refined set after removing the 195 protein−ligand complexes in the primary test set (1300 − 195 = 1105) and named the outcome X-Score version 1.3. This special version of X-Score was also subjected to the same set of tests as other scoring functions, and the results are summarized in the Supporting Information (Tables S5−S8 and S10−S16). Indeed, some marginal difference between the results of X-Score v.1.2 and v.1.3 was observed, but the difference was somewhat in an unexpected manner. The readers may refer to the Supporting Information (part II) for more discussion. It is desirable that every scoring function in our study should be evaluated on a separate test set independent from its training set. However, this approach is technically not practical because the training sets used by those scoring functions are not always clearly documented. Even if this approach were practical, it would bring another concern in regard to fair comparison, as discussed above, since those scoring functions were evaluated on an array of test sets with different contents. Due to these reasons, we believe that it is still more appropriate to base our assessment of scoring functions on a common test set. One however should interpret with caution the results of these empirical scoring functions reported in our study.

Table 4. Success Rates of 16 Scoring Functions in "Ranking Power" Evaluation on the Primary Test Set

scoring function[a]	success rates (%) on original complex structures	success rates (%) on optimized complex structures
X-Score::HSScore	58.5	52.3
DS::PLP2	53.8	46.2
DrugScore^{CSD}	52.3	49.2
SYBYL::ChemScore	47.7	52.3
SYBYL::D-Score	46.2	46.2
SYBYL::G-Score	46.2	36.9
GOLD::ASP	43.1	49.2
DS::LUDI3	43.1	43.1
DS::Jain	41.5	35.4
DS::PMF	41.5	35.4
SYBYL::PMF-Score	38.5	33.8
GOLD::ChemScore	36.9	41.5
DS::LigScore2	35.4	47.7
GlideScore-XP	33.8	35.4
by NHA[b]	32.3	32.3
SYBYL::F-Score	29.2	36.9
GOLD::GoldScore	23.1	38.5

[a] Scoring functions are ranked by their success rates based on original complex structures. [b] Ranking by the number of heavy atoms on each ligand.

Ranking Power on the Primary and Additional Test Sets. Another new feature of our study is the introduction of the ranking power of a scoring function, which is most relevant to virtual screening studies. Virtual screening aims at distinguishing promising molecules from others so that subsequent experimental efforts can be focused on them. It is thus essential for a virtual screening approach to rank given molecules preferably in the order of their binding affinities to the desired molecular target. As described in the Materials and Methods section, we have defined the "ranking power" of a scoring function as the ability to correctly rank the known ligands bound to a common molecular target by their binding affinities when their true binding modes are known. Our primary test set consists of 65 families of protein−ligand complexes, each family featuring a high-affinity ligand, a medium-affinity ligand, and a low-affinity ligand bound to a common type of protein. If a given scoring function ranks the three complexes in a family correctly, it gets a point. An overall success rate can be derived once this is repeated throughout the entire test set. This test can be considered as 65 miniature virtual screening trials on a wide spectrum of proteins.

The success rates of all 16 scoring functions in this test are summarized in Table 4. The top four scoring functions in this test are X-Score, DS::PLP2, DrugScore^{CSD}, and SYBYL::ChemScore, which achieve success rates over 50% on either original or optimized complex structures. This level of performance is approximately the same as what we have observed in the test of scoring powers. Remember that ranking a number of protein−ligand complexes correctly does not require the computed binding scores to correlate with experimental binding constants in a linear manner. Thus,

Table 5. Ranking Power of Selected Scoring Functions on the Four Additional Test Sets[a]

HIV protease ($N = 112$)				trypsin ($N = 73$)			
scoring functions	R_s[b]	R_p[c]	SD[d]	scoring functions	R_s	R_p	SD
by NHA[e]	0.140	0.172	1.62	by NHA[e]	0.603	0.655	1.28
A: X-Score::HPScore	0.339	0.341	1.54	A: X-Score::HSScore	0.824	0.817	0.97
B: SYBYL::ChemScore	0.228	0.276	1.58	B: DS::Ludi2	0.791	0.823	0.96
C: DS::PMF04	0.200	0.183	1.61	C: DS::PLP2	0.774	0.797	1.02
D: DrugScore^{PDB}::PairSurf	0.170	0.225	1.60	D: SYBYL::ChemScore	0.773	0.829	0.95
A + B	0.304			A + B	0.845		
A + C	0.291			A + C	0.814		
A + D	0.266			A + D	0.818		
B + C	0.225			B + C	0.831		
B + D	0.205			B + D	0.808		
C + D	0.194			C + D	0.812		
carbonic anhydrase ($N = 44$)				thrombin ($N = 38$)			
scoring functions	R_s	R_p	SD	scoring functions	R_s	R_p	SD
by NHA[e]	0.273	0.443	1.25	by NHA[e]	0.555	0.622	1.66
A: DS::PLP2	0.772	0.800	0.84	A: DS::PLP1	0.672	0.692	1.53
B: SYBYL::G-Score	0.646	0.706	0.99	B: SYBYL::G-Score	0.626	0.667	1.58
C: SYBYL::ChemScore	0.631	0.699	1.00	C: DrugScore^{CSD}::Pair	0.622	0.651	1.61
D: SYBYL::PMF-Score	0.618	0.627	1.09	D: X-Score::HSScore	0.586	0.666	1.58
A + B	0.780			A + B	0.699		
A + C	0.757			A + C	0.653		
A + D	0.763			A + D	0.666		
B + C	0.686			B + C	0.641		
B + D	0.713			B + D	0.601		
C + D	0.735			C + D	0.644		

[a] All results in this table are obtained on the basis of the original crystal structures of the protein−ligand complexes in these test sets. The results obtained on the optimized complex structures can be found in the Supporting Information (part VII). [b] Spearman correlation coefficients. [c] Pearson correlation coefficients. [d] Standard deviations in linear correlation (in log K_d units). [e] Using the number of heavy atoms on each ligand as the only variable in correlation.

this observation surprises us since we have expected that a good ranking power is a less challenging goal as compared to scoring power. It is interesting to notice that the top four scoring functions in this test are exactly the same as the ones in the test of scoring powers (Table 2). This is logical since a scoring function capable of "scoring" is automatically good at "ranking" as well. There are also exceptions. For example, SYBYL::D-Score, a simple force-field-based scoring function proposed in early years, is among the worst in terms of scoring power. Nevertheless, its performance in this test is almost comparable to the top four mentioned above. This observation perhaps explains the many successful applications of the DOCK program in virtual screening. Most of today's scoring functions are developed to reproduce either experimental binding data, such as empirical scoring functions, or experimental structures, such as knowledge-based scoring functions. This observation prompts the idea that a "ranking function" could be different from a "scoring function" or a "docking function". It is possible that a scoring function can be developed with an emphasis on its ranking power from the very beginning.

The ranking power of all 16 scoring functions was further assessed on the four additional test sets, each consisting of some protein−ligand complexes formed by one particular type of protein. The statistical data of the four top scoring functions on each test set are summarized in Table 5. The complete set of results can be found in the Supporting Information (part VII). Apparently, the performance of a scoring function is case-dependent, and the intrinsic characteristics of each target protein may explain it. An interesting observation is that, when binding constants have an obvious correlation with the size of the ligands, such as in the cases of trypsin and thrombin complexes, scoring functions tend to work reasonably well ($R_s = 0.59\sim0.82$). When such a correlation is not obvious, scoring functions may still rank the given protein−ligand complexes correctly, such as in the case of carbonic anhydrase complexes ($R_s = 0.62\sim0.77$). Nevertheless, it is also possible that scoring functions totally fail in such cases. For example, a very low correlation is observed between the binding constants of the 112 HIV protease complexes considered in our study and the size of the ligands in these complexes ($R_s = 0.14$). In fact, none of the scoring functions in our study is able to correctly rank these complexes to a meaningful extent ($R_s < 0.34$). It is well-known that the binding process between HIV protease and a ligand molecule involves considerable conformational changes. The enthalpic as well as entropic factors in such a process are certainly difficult for scoring functions to capture. We suspect that this is the primary reason why HIV protease complexes are particularly difficult for correct ranking by scoring functions. Nevertheless, this hypothesis remains to be explored in the future.

Although no scoring function consistently outperformed others in these four test sets, one can notice that the top scoring functions selected in each of these four test sets are mostly among the top ones selected on a diverse test set, that is, the primary test set used in our study. The scoring functions listed at the lower part of Table 4 do not have much chance to appear in Table 5. This indicates that a comparative assessment conducted on a diverse test set like ours has practical value for end-users: it narrows possible choices down to a few promising candidates to start with. The top scoring functions selected in a diverse test set are of course more consistent, if the test set represents an

Table 6. Classification of the Primary Test Set by Three Properties

classification criterion	symbol of subset	number of complexes	Z-Score
buried percentage of the solvent-accessible surface area of the ligand	A1	30	(−2.87, −1.00)
	A2	129	[−1.00, +1.00]
	A3	36	(+1.00, +2.54)
buried percentage of the molecular volume of the ligand	B1	40	(−2.37, −1.00)
	B2	124	[−1.00, +1.00]
	B3	31	(+1.00, +1.43)
hydrophobic index of the binding pocket	C1	31	(−3.05, −1.00)
	C2	126	[−1.00, +1.00]
	C3	38	(+1.00, 2.89)

adequate level of diversity. It is thus an apparent technical advantage of using a diverse test set rather than some particular protein−ligand complexes for assessing scoring functions.

We also tested the ranking power of consensus scoring on these four additional test sets. The results produced by all six possible double-scoring schemes combining the four top scoring functions on each test set are also listed in Table 5. The Spearman correlation coefficient of each double-scoring scheme was computed through the "rank-by-rank" strategy for consensus scoring,[87] that is, the final rank of each sample is the average rank given by two individual scoring functions. One can see that the results of these double-scoring schemes are constantly better than those of either individual scoring function. Nevertheless, the improvements are basically marginal. The determining factor is still the quality of the individual scoring functions. This observation is in accordance with what we have concluded regarding the docking power of consensus scoring schemes (Table 1).

Performance on Individual Subsets. As described in the Materials and Methods section, the primary test set was divided into subsets by three essential features of protein−ligand complexes, including the buried percentage of solvent-accessible surface area of the ligand, the buried percentage of the molecular volume of the ligand, and a hydrophobic index of the binding pocket on protein. These structural properties were computed independently from any scoring function in our test. The detailed information of each set of subsets is summarized in Table 6. Scoring functions can be assessed in greater detail by examining their performance on these subsets.

The docking powers and scoring powers of all 16 scoring functions in our assessment of these subsets are summarized in Figure 7. Complete results can be found in the Supporting Information (part VI). It should be mentioned that the performance of each scoring function in this series of tests may be discussed through an in-depth analysis of its algorithm. This type of discussion, however, may not be in the interest of general readers. Some scoring functions, especially the ones implemented in commercial software, lack detailed documentation so that they are available to us as black boxes, which makes an in-depth analysis of their algorithms impossible. Therefore, our discussion below will focus on the overall trends revealed in this series of tests.

As for docking power, the performance of each scoring function varies considerably among different subsets (Figure 7). For example, the success rates given by GOLD::ASP on subsets A1, A2, and A3 are 53.3%, 82.2%, and 94.4%, respectively. In fact, most scoring functions in our assessment tend to achieve higher success rates when the ligand is buried to a larger extent upon binding. This is true no matter whether the buried percentage of solvent-accessible surface area (subsets A1−A3) or the buried percentage of molecular volume (subsets B1−B3) of the ligand is used as the criterion in subset classification. This trend can be understood since, when a ligand molecule is more constrained inside the binding pocket, it will be easier for a scoring function to identify its correct binding pose. On the contrary, if binding of the ligand molecule occurs on a relatively flat surface, it will be more difficult to distinguish the true binding pose from decoys. The same trend is also observed in terms of scoring power for most scoring functions in our test, although the scoring power of a scoring function is generally lower than its docking power (Figure 7).

Quite different trends in docking power and scoring power are observed when the hydrophobic index of binding pocket is the criterion for subset classification. Obviously, most scoring functions demonstrate better docking powers when binding pockets are hydrophilic (subset C1). In such cases, one would expect that some polar interactions, such as hydrogen bonds, are dominant in protein−ligand binding. Misplacement of the ligand molecule will not retain such interactions at a maximum. It is thus relatively easy to distinguish the true ligand binding pose from decoys. In contrast, hydrophobic interactions are less specific and directional in nature. Therefore, when the binding pocket is hydrophobic (subset C3), the true ligand binding pose is not so distinctive from that of decoys, which is indicated by the relatively poor docking powers of scoring functions in this subset.

Nevertheless, it is clear that most scoring functions demonstrate better scoring powers on subset C3 in which hydrophobic interactions are likely to be the dominant factor in protein−ligand binding (Figure 7). It seems that hydrophobic interactions are relatively easier to quantify with simple models, such as algorithms based on solvent-accessible surface areas. In contrast, hydrogen bonds are more complicated to model. Formation of a hydrogen bond in protein−ligand binding is inevitably accompanied by desolvation of the donor and the acceptor. Thus, the net contribution of a hydrogen bond to protein−ligand binding free energy is typically a small number, which is certainly difficult to compute accurately. In addition, hydrogen bonds formed between the protein and ligand are often seen in a network rather than isolated. The assumption of additivity may not be valid at all for bifurcate or multifurcate hydrogen bonds. All of these factors account for the poor scoring powers observed in subset C1. Developing better algorithms for modeling polar interactions should be a major aim for future scoring functions.

Implications to Further Development of Scoring Functions. Our study has revealed that today's scoring functions are generally more capable of identifying the correct ligand binding poses. Several scoring functions under our assessment produced very encouraging results in this regard. Improving the performance of scoring functions in binding affinity prediction, that is, scoring power and ranking power, seems to be a more urgent goal for further developments. This requires continuous efforts in designing better algorithms for polar interactions, solvation/desolvation energies, and the elusive configurational entropies. At the same

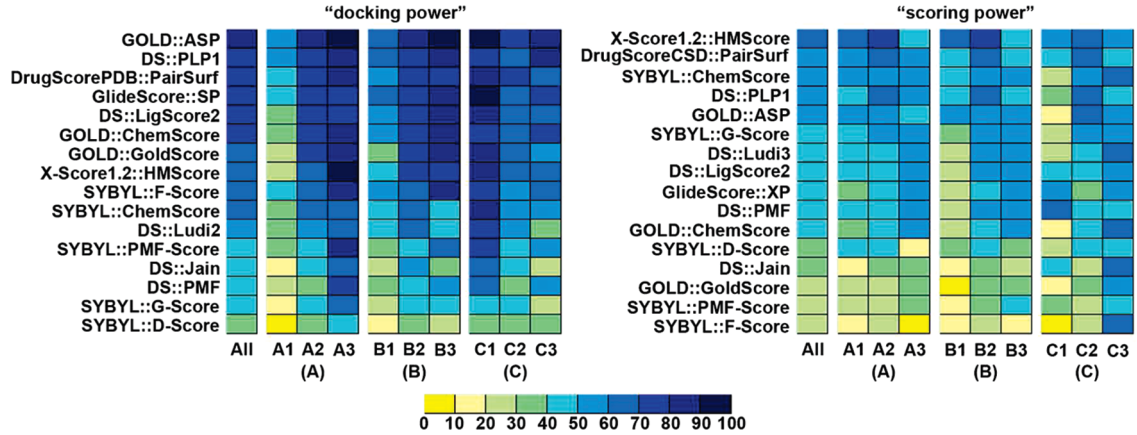

Figure 7. "Docking power" and "scoring power" of all 16 scoring functions on the subsets in the primary test set. Three sets of subsets were classified by (A) buried percentage of the solvent-accessible surface area of the ligand, (B) buried percentage of the molecular volume of the ligand, and (C) the hydrophobic index of the binding pocket. Here, scoring functions are ranked by their performance on the entire primary test set.

time, scoring functions still need to remain in relatively simple forms to keep their efficiency in high-throughput jobs, which makes this goal even more challenging. Considering these difficulties, our opinion is that the goal of developing a good generic scoring function, which is expected to perform reasonably well in all applicable systems, might be too ambitious. In fact, all of the 16 scoring functions assessed in our study were developed as generic scoring functions. Their performance in the four additional test sets clearly demonstrated that none of them could perform consistently well in all cases (Table 5).

In order to further improve scoring power and ranking power, it may be a practical strategy to develop customized scoring functions. Some target-specific scoring methods have already been reported in the literature.[45,88] Another option is to develop scoring functions applicable to a group of molecular targets sharing common structural features. For example, a scoring function strong at characterizing hydrophobic interactions is expected to perform better when hydrophobic interactions are the dominant factor in protein−ligand binding. Developing customized scoring functions is more practical than before since publicly available structures and binding constants of various protein−ligand complexes are increasing in availability constantly, which has provided abundant raw material for this purpose. If still relying on existing scoring functions, one may want to find out the appropriate "applicable space" for each of them through extensive tests. Then, one can apply the right scoring functions to the right problem through a divide-and-conquer strategy.

CONCLUSIONS

Our study aims at setting up a new benchmark for scoring function assessment, which has substantial improvements over previous similar studies in several respects. Our study covered a total of 16 popular scoring functions implemented in main-stream commercial software or released by academic groups. All scoring functions were evaluated on a high-quality set of 195 diverse protein−ligand complexes and four additional sets of particular protein−ligand complexes. As for "docking power", the best scoring function found in our test was GOLD::ASP, which was able to identify the correct ligand binding pose out of computer-generated decoys with a high success rate of 82.5% (when the acceptance cutoff was rmsd < 2.0 Å). Another five scoring functions, including DS::PLP1, DrugScore^{PDB}, GlideScore-SP, DS::LigScore, and GOLD::ChemScore achieved success rates over 70% under the same criterion. The success rates could be improved to 80% or even higher when these scoring functions were combined into consensus scoring schemes. As for "ranking power" and "scoring power", the top four scoring functions found in our primary test set were X-Score, DrugScore^{CSD}, DS::PLP, and SYBYL::ChemScore. They were able to correctly rank the protein−ligand complexes formed by a common type of protein in about 50% of the cases. Correlation coefficients between the experimental binding constants and the binding scores produced by these scoring functions on this test set ranged from 0.545 to 0.644. Thus, today's scoring functions are generally more capable of predicting binding modes than binding affinities. Generally speaking, no single scoring function consistently outperforms the others in all three aspects. Scoring functions based on summing up pairwise protein−ligand interaction potentials seem to be more capable in terms of docking power, while regression-based empirical scoring functions seem to have certain advantages in terms of ranking/scoring power. It is thus important to choose the appropriate scoring functions for different purposes. Our results obtained on four additional sets of protein−ligand complexes indicate that the scoring functions relatively more successful in the primary test set tend to perform better in these sets as well. Thus, a comparative assessment conducted on a diverse test set is helpful for end-users in a way that it narrows the possible choices down to a few promising candidates with which to start.

ACKNOWLEDGMENT

The authors are grateful for the financial support from the Chinese National Natural Science Foundation (grant nos. 20772149 and 90813006), the Chinese Ministry of Science and Technology (the 863 high-tech project, grant no. 2006AA02Z337), and the Science and Technology Commission of Shanghai Municipality (grant no. 074319113).

Supporting Information Available: Descriptions of the 16 scoring functions in our test, detailed information of all

test sets, methods for decoy set generation and subset classification, and the complete statistical results produced by all scoring functions in all tests. This material is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES AND NOTES

(1) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813–1818.
(2) Klebe, G. Recent Developments in Structure-Based Drug Design. *J. Mol. Med.* **2000**, *78*, 269–281.
(3) Gane, P. J.; Dean, P. M. Recent Advances in Structure-Based Rational Drug Design. *Curr. Opin. Struct. Biol.* **2000**, *10*, 401–404.
(4) Amzel, L. M. Structure-Based Drug Design. *Curr. Opin. Biotechnol.* **1998**, *9*, 366–369.
(5) Marrone, T. J.; Briggs, J. M.; McCammon, J. A. Structure-Based Drug Design: Computational Advances. *Annu. Rev. Pharmacol. Toxicol.* **1997**, *37*, 71–90.
(6) Blundell, T. L. Structure-Based Drug Design. *Nature (London)* **1996**, *384*, 23–26.
(7) Verlinde, C. L.; Hol, W. G. Structure-Based Drug Design: Progress, Results and Challenges. *Structure* **1994**, *2*, 577–587.
(8) Robertus, J. Structure-Based Drug Design Ten Years On. *Nat. Struct. Biol.* **1994**, *1*, 352–354.
(9) Colman, P. M. Structure-Based Drug Design. *Curr. Opin. Struct. Biol.* **1994**, *4*, 868–874.
(10) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(11) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57*, 225–242.
(12) 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.
(13) 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.
(14) 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.
(15) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed Automated Docking of Flexible Ligands to Proteins: Parallel Applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293–304.
(16) Goodsell, D. S.; Olson, A. J. Automated Docking of Substrates to Proteins by Simulated Annealing. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 195–202.
(17) 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.
(18) Jain, A. Surflex-Dock 2.1: Robust Performance from Ligand Energetic Modeling, Ring Flexibility, and Knowledge-Based Search. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 281–306.
(19) Jain, A. N. Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-Based Search Engine. *J. Med. Chem.* **2003**, *46*, 499–511.
(20) Venkatachalam, C. M.; Jiang, X.; Oldfield, T.; Waldman, M. LigandFit: A Novel Method for the Shape-Directed Rapid Docking of Ligands to Protein Active Sites. *J. Mol. Graphics Modell.* **2003**, *21*, 289–307.
(21) 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.
(22) 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.
(23) Friesner, R. A.; Murphy, R. B.; Repasky, M. P.; Frye, L. L.; Greenwood, J. R.; Halgren, T. A.; Sanschagrin, P. C.; Mainz, D. T. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.
(24) McInnes, C. Virtual Screening Strategies in Drug Discovery. *Curr. Opin. Chem. Biol.* **2007**, *11*, 494–502.
(25) Shoichet, B. K. Virtual Screening of Chemical Libraries. *Nature (London)* **2004**, *432*, 862–865.
(26) Lyne, P. D. Structure-Based Virtual Screening: An Overview. *Drug Discovery Today* **2002**, *7*, 1047–1055.
(27) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening - an Overview. *Drug Discovery Today* **1998**, *3*, 160–178.
(28) Gilson, M. K.; Zhou, H.-X. Calculation of Protein-Ligand Binding Affinities. *Annu. Rev. Biophys. Biomol. Struct.* **2007**, *36*, 21–42.
(29) Gohlke, H.; Klebe, G. Approaches to the Description and Prediction of the Binding Affinity of Small-Molecule Ligands to Macromolecular Receptors. *Angew. Chem., Int. Ed.* **2002**, *41*, 2644–2676.
(30) Kollman, P. Free Energy Calculations: Applications to Chemical and Biochemical Phenomena. *Chem. Rev.* **1993**, *93*, 2395–2417.
(31) Jorgensen, W. L. Free Energy Calculations: A Breakthrough for Modeling Organic Chemistry in Solution. *Adv. Drug Delivery Rev.* **1989**, *22*, 184–189.
(32) Massova, I.; Kollman, P. Combined Molecular Mechanical and Continuum Solvent Approach (MM-PBSA/GBSA) to Predict Ligand Binding. *Perspect. Drug Discovery Des.* **2000**, *18*, 113–135.
(33) Carlson, H. A.; Jorgensen, W. L. An Extended Linear Response Method for Determining Free Energies of Hydration. *J. Phys. Chem.* **1995**, *99*, 10667–10673.
(34) Aqvist, J.; Medina, C.; Samuelsson, J.-E. A New Method for Predicting Binding Affinity in Computer-Aided Drug Design. *Protein Eng.* **1994**, *7*, 385–391.
(35) Krammer, A.; Kirchhoff, P. D.; Jiang, X.; Venkatachalam, C. M.; Waldman, M. LigScore: A Novel Scoring Function for Predicting Binding Affinities. *J. Mol. Graphics Modell.* **2005**, *23*, 395–407.
(36) 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.
(37) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A New Empirical Method for Estimating the Binding Affinity of a Protein-Ligand Complex. *J. Mol. Model.* **1998**, *4*, 379–394.
(38) Gehlhaar, D. K.; Bouzida, D.; Rejto, P. A. *Rational Drug Design: Novel Methodology and Practical Applications*; American Chemical Society: Washington, DC, 1999; pp 292−311.
(39) Gehlhaar, D. K.; Verkhivker, G. M.; Rejto, P. A.; Sherman, C. J.; Fogel, D. R.; 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.
(40) Baxter, C. A.; Murray, C. W.; Clark, D. E.; Westhead, D. R.; Eldridge, M. D. Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. *Proteins: Struct., Funct., Genet.* **1998**, *33*, 367–382.
(41) 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.
(42) 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.
(43) 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.
(44) 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.
(45) Mooij, W. T. M.; Verdonk, M. L. General and Targeted Statistical Potentials for Protein-Ligand Interactions. *Proteins: Struct., Funct., Bioinf.* **2005**, *61*, 272–287.
(46) Velec, H. F. G.; Gohlke, H.; Klebe, G. DrugScoreCSD - Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of Near-Native Ligand Poses and Better Affinity Prediction. *J. Med. Chem.* **2005**, *48*, 6296–6303.
(47) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.
(48) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.
(49) Muegge, I. Effect of Ligand Volume Correction on PMF Scoring. *J. Comput. Chem.* **2001**, *22*, 418–425.
(50) Muegge, I. A Knowledge-Based Scoring Function for Protein-Ligand Interactions: Probing the Reference State. *Perspect. Drug Discovery Des.* **2000**, *20*, 99–114.
(51) 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.
(52) 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.

(53) Marsden, P. M.; Puvanendrampillai, D.; Mitchell, J. B. O.; Glen, R. C. Predicting Protein-Ligand Binding Affinities: A Low Scoring Game. *Org. Biomol. Chem.* **2004**, *2*, 3267–3273.
(54) Wang, R.; Lu, Y.; Fang, X.; Wang, S. An Extensive Test of 14 Scoring Functions Using the PDBbind Refined Set of 800 Protein-Ligand Complexes. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2114–2125.
(55) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46*, 2287–2303.
(56) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47*, 1599–1608.
(57) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model.* **2006**, *46*, 401–415.
(58) Evers, A.; Klabunde, T. Structure-Based Drug Discovery Using GPCR Homology Modeling: Successful Virtual Screening for Antagonists of the Alpha1a Adrenergic Receptor. *J. Med. Chem.* **2005**, *48*, 1088–1097.
(59) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of Automated Docking Programs as Virtual Screening Tools. *J. Med. Chem.* **2005**, *48*, 962–976.
(60) Kontoyianni, M.; Sokol, G. S.; McClellan, L. M. Evaluation of Library Ranking Efficacy in Virtual Screening. *J. Comput. Chem.* **2005**, *26*, 11–22.
(61) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47*, 558–565.
(62) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56*, 235–249.
(63) Hu, X.; Balaz, S.; Shelver, W. H. A Practical Approach to Docking of Zinc Metalloproteinase Inhibitors. *J. Mol. Graphics Modell.* **2004**, *22*, 293–307.
(64) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and Application of Multiple Scoring Functions for a Virtual Screening Experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333–344.
(65) Bursulaya, B.; Totrov, M.; Abagyan, R.; Brooks, C. Comparative Study of Several Algorithms for Flexible Ligand Docking. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 755–763.
(66) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.
(67) 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.
(68) Irwin, J. J.; Shoichet, B. K. ZINC - a Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(69) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.
(70) Lee, J.; Seok, C. A Statistical Rescoring Scheme for Protein-Ligand Docking: Consideration of Entropic Effect. *Proteins: Struct., Funct., Bioinf.* **2008**, *70*, 1074–1083.
(71) Ruvinsky, A. M. Calculations of Protein-Ligand Binding Entropy of Relative and Overall Molecular Motions. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 361–370.
(72) Ruvinsky, A. M. Role of Binding Entropy in the Refinement of Protein-Ligand Docking Predictions: Analysis Based on the Use of 11 Scoring Functions. *J. Comput. Chem.* **2007**, *28*, 1364–1372.
(73) Huang, S. Y.; Zou, X. An Iterative Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions: II. *J. Comput. Chem.* **2006**, *27*, 1876–1882.
(74) Zhang, C.; Liu, S.; Zhu, Q.; Zhou, Y. A Knowledge-Based Energy Function for Protein-Ligand, Protein-Protein, and Protein-DNA Complexes. *J. Med. Chem.* **2005**, *48*, 2325–2335.
(75) Raha, K.; Merz, K. M., Jr. Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein-Ligand Complexes. *J. Med. Chem.* **2005**, *48*, 4558–4575.
(76) *The Discovery Studio Software*, version 2.0; Accelrys Software Inc.: San Diego, CA, 2001.
(77) *The Sybyl Software*, version 7.2; Tripos Inc.: St. Louis, MO, 2006.
(78) *The Schrödinger Software*, version 8.0; Schrödinger, LLC: New York, 2005.
(79) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind Database: Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
(80) 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.
(81) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(82) Zhao, Y.; Cheng, T.; Wang, R. Automatic Perception of Organic Molecules Based on Essential Structural Information. *J. Chem. Inf. Model.* **2007**, *47*, 1379–1385.
(83) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance. *J. Med. Chem.* **2007**, *50*, 726–741.
(84) 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.
(85) Yang, J. M.; Chen, Y. F.; Shen, T. W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.
(86) 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.
(87) 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.
(88) Jansen, J. M.; Martin, E. J. Target-Biased Scoring Approaches and Expert Systems in Structure-Based Virtual Screening. *Curr. Opin. Chem. Biol.* **2004**, *8*, 359–364.

CI9000053