Published in final edited form as:

Phys Chem Chem Phys. 2010 October 28; 12(40): 12899–12908. doi:10.1039/c0cp00151a.

Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions

Sheng-You Huang,

Sam Z. Grinter,

Xiaoqin Zou*

Department of Physics and Astronomy, Department of Biochemistry, Dalton Cardiovascular Research Center, and Informatics Institute University of Missouri Columbia, MO 65211

Abstract

The scoring function is one of the most important components in structure-based drug design. Despite considerable success, accurate and rapid prediction of protein-ligand interactions is still a challenge in molecular docking. In this perspective, we have reviewed three basic types of scoring functions (force-field, empirical, and knowledge-based) and the consensus scoring technique used in protein-ligand docking. The commonly-used criteria/methods and publicly available protein-ligand databases that are used to evaluate the performance of scoring functions are also depicted and discussed. We end with a discussion of the challenges faced by existing scoring functions and possible future directions for developing improved scoring functions.

Keywords

scoring functions; molecular docking; protein-ligand interactions; criteria for evaluations; protein-ligand databases

1 Introduction

As the number of three-dimensional protein structures determined by experimental techniques grows, computational tools such as molecular docking have played an increasing role in the functional study of proteins and structure-based drug design.[1–6] In all the computational methodologies, one important problem is the development of an energy scoring function that can rapidly and accurately describe the interaction between protein and ligand. Several reviews on scoring are available in the literature.[7–11]

There are three important applications of scoring functions in molecular docking. The first of these is the determination of the binding mode and site of a ligand on a protein.[9] Given a protein target, molecular docking generates hundreds of thousands of putative ligand binding orientations/conformations at the active site around the protein. A scoring function is used to rank these ligand orientations/conformations by evaluating the binding tightness of each of the putative complexes. An ideal scoring function would rank the experimentally determined

*Corresponding author. zoux@missouri.edu, 573-884-4232 (fax).

binding mode most highly. Given the determined binding mode of a ligand, scientists would be able to gain a deep understanding of the molecular mechanism of ligand binding and to further design an efficient drug by modifying the protein or ligand.[9]

The second application of a scoring function, which is related to the first application, is to predict the absolute binding affinity between protein and ligand. This is particularly important in lead optimization.[4] Lead optimization refers to the process to improve the tightness of binding for low-affinity hits or lead compounds that have been identified. During this process, an accurate scoring function can greatly increase the optimization efficiency and save costs by computationally predicting the binding affinity between the protein and modified ligands before the much more expensive step of ligand synthesis and experimental testing.

The third application, perhaps the most important one in structure-based drug design, is to identify the potential drug hits/leads for a given protein target by searching a large ligand database, i.e. virtual database screening.[6] A reliable scoring function should be able to rank known binders most highly according to their binding scores during database screening. Given the expensive cost of experimental screening and sometimes unavailability of high-throughput assays, virtual database screening has played an increasingly important role in drug discovery.

All of these three applications, ligand binding mode identification, binding affinity prediction, and virtual database screening, are related to each other. Presumably, an accurate scoring function would perform equally well on each of them. Despite over a decade of development, scoring is still an open question. Many existing scoring functions perform well only on one or two of the three applications. Roughly, the scoring functions can be grouped into three basic types according to how they are derived: force field-based, empirical, and knowledge-based. In this perspective, we have reviewed several important aspects of scoring functions for protein-ligand docking, as outlined in Figure 1. Specifically, we will first briefly review different categories of scoring functions in Section 2. We will then describe the commonly used criteria that are used to evaluate the performance of a scoring function in Section 3. We also review the publicly available protein-ligand databases for developing and validating a scoring function in Section 4. Finally, challenges and future directions for scoring function development will be discussed in the Conclusion and Discussions.

2 Brief review of scoring functions

Over the years, different scoring functions have been developed that exhibit different accuracies and computational efficiencies. In this section, we will briefly review the scoring functions in literature developed for protein-ligand interactions in molecular docking. Some of the commonly-used scoring functions are summarized in Table 1 and grouped into three broad categories.

2.1 Force field scoring function

Force field (FF) scoring functions are developed based on physical atomic interactions,[51] including van der Waals (VDW) interactions, electrostatic interactions, and bond stretching/

bending/torsional forces. Force field functions and parameters are usually derived from both experimental data and *ab initio* quantum mechanical calculations according to the principles of physics. Despite its lucid physical meaning, a major challenge in the force field scoring functions is how to treat the solvent in ligand binding.

One typical force field scoring function in molecular docking is the scoring function of DOCK whose energy parameters are taken from the Amber force fields.[12,52,53] The scoring function is composed of two energy components of Lennard-Jones VDW and an electrostatic term

$$E = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \right) \quad (1)$$

where r_{ij} stands for the distance between protein atom i and ligand atom j, A_{ij} and B_{ij} are the VDW parameters, and q_i and q_j are the atomic charges. Here, the effect of solvent is implicitly considered by introducing a simple distance-dependent dielectric constant $\varepsilon(r_{ij})$ in the Coulombic term. Despite the computational efficiency of the force field scoring function of DOCK, the distance-dependent dielectric factor cannot account for the desolvation effect, an important solvent effect that charged groups favor aqueous environments whereas non-polar groups tend to stay in non-aqueous environments. The desolvation energy is a many-body interaction term and depends on specific geometric and chemical surrounding environments of the considered solute atoms. If the desolvation effect is ignored, a scoring function would be biased on Coulombic electrostatic interactions and therefore would tend to select highly charged ligands.

A rigorous method to account for the solvent effect such as free energy perturbation (FEP) and thermodynamic integration (TI) is to treat water molecules explicitly (see refs 3 and 54 for review). However, these methods, together with their simplified approaches such as LIE, PROFEC and OWFEG (see ref 3 and references therein) are too computationally expensive to be used in virtual database screening. In addition, while simulations with explicit waters is theoretically ideal/rigorous, the accuracy of the results may be limited by the high computational cost of sampling due to the inclusion of water molecules in real applications. To reduce the computational expense, some accelerated force field models have been developed for scoring use in molecular docking by treating water as a continuum dielectric medium. Typical examples of such implicit solvent models include the Poisson-Boltzmann/surface area (PB/SA) model[55–57] and the generalized-Born/surface area (GB/SA) model,[58–60] that are often used in post-scoring of docking programs. Shoichet and colleagues applied a modified Born equation to calculate the electrostatic component of ligand solvation.[13,14] In their study, the electrostatic potential of the protein is calculated by using the finite-difference Poisson-Boltzmann (PB) method implemented in DelPhi,[12,55] and partial atomic charges are calculated with the Gasteiger algorithm[61] implemented in the program SYBYL (Tripos) or with the semi-empirical quantum mechanical approach implemented in the program AMSOL.[62] The desolvation energy penalty for the ligand was calculated by assuming full desolvation of each ligand atom or of the whole ligand. The

method was validated by screening the Available Chemicals Directory (ACD) against T4 lysozyme mutants and dihydrofolate reductase (DHFR).

The PB/SA[63–69] and GB/SA[15–17,70–77] approaches have been successfully used for relative potency studies and virtual screening tests. For example, Zou et al. accounted for the solvation effect in ligand binding free energy calculations using a GB/SA approach.[15,16] Specifically, the solvent-screened electrostatic interactions and the electrostatic desolvation costs are calculated with the GB model. The hydrophobic contributions for non-polar atoms are estimated using the solvent-accessible surface areas (SA) of the atoms. The van der Waals energies are calculated using Lennard-Jones potentials. Then the weights for the electrostatic, van der Waals, and hydrophobic contributions to the free energy of binding are optimized so that the predicted binding scores agree well with the experimental affinity data for known inhibitors and known inhibitors are distinguished from random molecules in database screening. The GB/SA formulation implemented in DOCK[78,79] as SDOCK was validated on three systems: dihydrofolate reductase (DHFR), trypsin, and a fatty acid-binding protein. To enhance the computational efficiency, a pairwise format of GB was parameterized for protein-ligand docking,[16,17] which takes only about 0.5s per orientation (with minimization) on a Silicon Graphics Octane R12000 workstation.

After thorough and systematic comparison between PB and GB on protein-ligand complexes with a wide range of electrostatic component of binding energies (from −5 to 25 kcal/mol), Zou and colleagues showed that being able to reproduce the solvation energy of a ligand or a protein calculated with PB is not necessarily suitable for ligand binding calculations. Additional quantities should be used for evaluation, particularly quantities such as the partial desolvation energy of the receptor.[17,70] To warrant the accuracy and efficiency, they proposed a multiscale GB approach for the use of virtual screening. In this approach atoms are divided into two groups: The few atoms in the first group are most likely to be critical to binding electrostatics; their contributions are calculated with accurate GB models at the sacrifice of speed. The rest atoms (second group) may be treated with a fast GB method.[70]

In addition to the challenge in rapidly and accurately accounting for the solvent effect in electrostatics, how to combine individual energy terms is also difficult. Usually, empirical weighting coefficients have to be introduced because each energy component is calculated from unrelated methods.[15,16,18,19] For example, the electrostatic component can be calculated with Coulombic, PB or GB approaches. The VDW energy component is commonly represented by Lennard-Jones potentials. The hydrophobic interaction term is often approximated as being proportional to the change of solvent-accessible surface area. These terms have quite different scales, and thereby cannot be added up without weighting factors. The weighting factors are obtained by fitting experimental binding data, etc. There may be more than one set of empirical weighting coefficients to achieve comparable answers.[15,16] Although it is possible to find appropriate weighting coefficients for a specific protein or protein family, it is difficult to obtain a universal set for diverse protein-ligand complexes. Furthermore, even accurate electrostatic energy calculations can be blown off by poor treatment of entropic contributions. Finally, it is well-known that individual free energy terms may not be additive.[80]

2.2 Empirical scoring function

A second kind of scoring functions are empirical scoring functions, which estimate the binding affinity of a complex on the basis of a set of weighted energy terms

$$\Delta G = \sum_i W_i \cdot \Delta G_i \quad (2)$$

where ΔG_i represents different energy terms such as VDW energy, electrostatics, hydrogen bond, desolvation, entropy, hydrophobicity, etc. The corresponding coefficients W_i are determined by fitting the binding affinity data of a training set of protein-ligand complexes with known three-dimensional structures.[24–30,32–35,81,82]. Compared to the force field scoring functions, the empirical scoring functions are much faster in binding score calculations due to their simple energy terms.

By calibrating with a dataset of 45 protein-ligand complexes, Böhm developed an empirical scoring function (SCORE1) consisting of four energy terms: hydrogen bonds, ionic interactions, the lipophilic protein-ligand contact surface, and the number of rotatable bonds in the ligand.[24] This empirical scoring function was further improved by expanding the dataset to 82 protein-ligand complexes with known 3D structures and binding constants and by considering the energy parameters for the following terms: the number and geometry of intermolecular hydrogen bonds and ionic interactions, the size of the lipophilic contact surface, the flexibility of the ligand, the electrostatic potential in the binding site, water molecules in the binding site, cavities along the protein-ligand interface, and specific interactions between aromatic rings.[25] Eldridge et al. presented an empirical scoring function referred to as ChemScore by taking into account hydrogen bonds, metal atoms, the lipophilic effects of atoms, and the effective number of rotatable bonds in the ligand.[28] The scoring function was calibrated using 82 ligand-receptor complexes with known binding affinities and was tested using two other sets of 20 and 10 protein-ligand complexes, respectively. Based on a larger set of 200 protein-ligand complexes, Wang et al. developed a new empirical scoring function, X-Score, consisting of four energy terms including VDW interactions, hydrogen bonds, hydrophobic effects and effective rotatable bonds.[30]

By including different empirical energy terms, empirical scoring functions have been used in many well-known protein-ligand docking programs such as FlexX[21] and Surflex[31]. How to avoid double-counting problems is a difficult issue for empirical scoring functions with many energy terms. Their general applicability may also depend on the training set due to their nature of fitting binding affinities of a small dataset. With the rapid increase in the number of protein-ligand complexes with known 3D structures and binding affinities, it is possible to develop a relatively general empirical scoring function by training with known binding constants of thousands of diverse protein-ligand complexes.

In addition to fitting a set of weighted energy terms to the binding affinities of a training set, recently, inspired by the knowledge-based scoring functions, a knowledge-based quantitative structure-activity relationship (QSAR) approach has been introduced for scoring protein-ligand interactions.[83] In the knowledge-based QSAR method, the atom pair occurrence and

distance-dependent atom pair features are used to generate the interaction potentials by using a machine-learning method to fit the binding affinities of a training set. One advantage of the machine-learning method is that it can fit the binding affinities of a very large training set due to the inclusion of more parameters through knowledge-based atom pair features. For, example, in a very recent study by Ballester and Mitchell,[84] the derived scoring function (RF-Score) by a machine-learning method yielded a high correlation ($R = 0.953$) on a large training set of 1105 protein-ligand complexes.

2.3 Knowledge-based scoring function

A third kind of scoring functions are knowledge-based scoring functions (also referred to as statistical-potential based scoring functions), which employ energy potentials that are derived from the structural information embedded in experimentally determined atomic structures.[85–87] The principle behind knowledge-based scoring functions is simple: Pairwise potentials are directly obtained from the the occurrence frequency of atom pairs in a database using the inverse Boltzmann relation.[88–91] For protein-ligand studies, the potentials are calculated by

$$w(r) = -k_B T \ln[g(r)], \quad g(r) = \rho(r)/\rho^*(r) \tag{3}$$

where k_B is the Boltzmann constant, T is the absolute temperature of the system, $\rho(r)$ is the number density of the protein-ligand atom pair at distance r, and $\rho^*(r)$ is the pair density in a reference state where the interatomic interactions are zero.

The idea of the inverse Boltzmann method for knowledge-based potentials comes from statistical mechanics in the physics field.[91] According to the analytical integral equation for the pair distribution function $g(r)$ in the simple fluid system, the interaction potentials by the inverse Boltzmann method are actually the mean-force potentials rather than the true potentials.[91,92] In addition, The protein system is also much different from the simple fluid system in statistical mechanics due to the effects of atomic connectivity, excluded volume, composition, etc.[90] Therefore, the knowledge-based potentials are also not the true mean-force potentials in the physics of the simple fluid system. Howerer, despite these limitations, the inverse Boltzmann method provides a simple and effective alternative method to derive the interaction scores from the structural information for complicated protein systems.[92,93] Since the pioneering work by Tanaka and Scheraga,[85] a large number of knowledge-based scoring functions have been developed and widely applied to protein structure prediction and protein-ligand studies (see ref 94 for review).

Compared to the force field and empirical scoring functions, the knowledge-based scoring functions offer a good balance between accuracy and speed. Because the potentials in eqn 3 are extracted from the structures rather than from attempting to reproduce the known affinities by fitting, and because the training structural database can be large and diverse, the knowledge-based scoring functions are quite robust and relatively insensitive to the training set.[36,37,39,40] Their pairwise characteristic also enables the scoring process to be as fast as empirical scoring functions.

However, there is a challenge in deriving knowledge-based scoring functions, which is the reference state (see eqn 3). As pointed out by Thomas and Dill[90] and other groups, an accurate reference state is not achievable. Therefore, how to calculate $\rho^*(r)$ of the reference state becomes a longstanding hurdle in deriving knowledge-based potentials. Below we will use the reference state treatment to classify various knowledge-based scoring functions.

Most of the current knowledge-based scoring functions approximate the reference state with an atom-randomized state by ignoring the effects of excluded volume, interatomic connectivity, etc.[90] Gohlke et al. developed a knowledge-based scoring function (DrugScore) based on 17 atom types and 1376 protein-ligand complex structures.[41] The scoring function consists of a distance-dependent pair-potential term and a surface-dependent singlet-potential term. It was validated by using two sets of protein-ligand complexes (91 and 68 complexes in each set). A further comparative evaluation of DrugScore and AutoDock shows that DrugScore yields slightly superior results in flexible docking.[95] Recently, an improved version (DrugScore^{CSD})[42] was also developed based on the Cambridge Structural Database (CSD) of small molecules,[96] which contain low-molecular-weight structures with higher resolution than huge-molecular-weight structures in the Protein Data Bank (PDB).[97] Mitchell et al. presented a statistical potential model, BLEEP, using 40 atom types.[46] This model was tested on nine serine protease-inhibitor complexes and obtained a correlation coefficient of 0.71 (or $R^2 = 0.50$) between the calculated energy scores and the experimental binding data. A further test on a set of 90 protein-ligand complexes shows a good correlation ($R = 0.74$ or $R^2 = 0.55$) in affinity predictions.[47] Application of BLEEP to the 77 complexes used by Muegge and Martin[39] yields a correlation of $R^2 = 0.28$.[98] Based on 725 protein-ligand complexes from the PDB, Ishchenko and Shakhnovich presented an improved version of SMoG[44] (referred to as SMoG2001).[45] SMoG2001 uses 13 atom types, two distance intervals, and a reference state determined by a self-consistent method. Applying SMoG2001 to Muegge and Martin's test set gives a correlation coefficient of 0.68 (or $R^2 = 0.46$).[45] Yang et al. presented a new knowledge-based scoring function M-Score to account for the mobility of protein atoms based upon 2331 protein-ligand complexes.[48] M-Score describes the location of each protein atom by a Gaussian distribution based upon the isotropic B-factors, which results in a smoothing effect on the pairwise distribution functions and thereby smoothen its knowledge-based potentials.

In addition to adopting the traditional atom-randomized reference state, researchers have also tried to improve the accuracy of the reference state by introducing some corrections or scalings. The potential model by Muegge and Martin, PMF (potential of mean force), was the first knowledge-based scoring function to be extensively tested for affinity predictions.[39] Based on 34 ligand atom types and 16 protein atom types, the distance-dependent pair potentials were derived using 697 protein-ligand structures in which a ligand volume factor is introduced to correct the reference state. The model was tested on a diverse set of 77 protein-ligand complexes with known binding affinities and outperformed LUDI[24] and SMoG[44], yielding a high correlation ($R^2 = 0.61$) between the calculated scores and the experimental binding constants. The PMF scoring function was also successfully applied to docking/scoring studies of weak ligands for the FK506 binding protein[99] and inhibitors for matrix metalloprotease MMP-3[100]. Recently, a newer version of PMF (PMF04) has been

developed using a much larger database of 7152 protein-ligand complexes from the PDB and received similar results.[40] Zhang et al. developed a knowledge-based statistical energy function for protein-ligand, protein-protein, and protein-DNA complexes by using 19 atom types and a distance-scale finite ideal-gas reference state (DFIRE).[43] The scoring function obtained a correlation coefficient of 0.63 on 100 protein-ligand complexes, 0.73 for 82 protein-protein complexes, and 0.83 for 45 protein-DNA complexes, respectively.

No matter whether one chooses to use an atom-randomized state or a more physical approximation, the accuracy of the reference state remains a problem in knowledge-based scoring functions. The problem is more prominent for binding mode predictions and virtual screening, as the pairwise potentials, which are derived from nicely-bound structures, are not sufficiently sensitive to different ligand positions and may give good scores even to bad/wrong modes. Attempting to solve this problem, Huang and Zou have recently developed a new kind of knowledge-based scoring function (referred to as ITScore) using an iterative method so as to circumvent the accurate calculation of the reference state.[36,37,101–103] The basic idea of the iterative method is to adjust the pair potentials $u_{ij}(r)$ by iteration until the interaction potentials reproduce the experimentally determined pair distribution function in the training set, yielding a set of potentials discriminating the native structures from decoys.[104–107] During the iteration procedure, the improvement for the potentials is guided through the difference between the predicted and experimentally observed pair distribution functions instead of through accurate calculation of the aforementioned reference state, where the predicted pair distribution function $g_{ij}(r)$ is calculated from the ensemble of the native structures and a set of well-sampled decoys according to the Boltzmann probability. In such a way, the iterative method circumvents the reference state problem faced by traditional knowledge-based scoring functions. Another major advantage of the iterative method is that it considers the full binding energy landscape of the complexes by including both the native structures and decoys during the calculation of $g_{ij}(r)$, instead of considering only the energy minima (i.e., native structures) like what conventional knowledge-based scoring functions do. Extensive evaluations on diverse test sets showed that ITScore yielded good performances on predictions of ligand binding modes and affinities and on virtual screening of compound databases.[36,37] Very recently, Huang and Zou have included the solvation effect and configurational entropy in ITScore. The new scoring function, referred to as ITScore/SE, has further improved the performance of ITScore.[38]

2.4 Consensus scoring

Despite a large and ever increasing number of scoring functions developed, none of them is perfect in terms of accuracy and general applicability. Every existing scoring function has its advantage and limitation. To make use of the advantages and balance out the errors from different scoring functions, the consensus scoring technique has been introduced to improve the probability in finding correct solutions by combing the scores from multiple scoring functions.[108] The most important/challenging task in consensus scoring is how to make an appropriate consensus scoring strategy of individual scores so that the true modes/binders can be discriminated from others according to the consensus strategy.[109,110] Commonly used consensus scoring strategies include vote-by-number, numberby-number,

rank-by-number, average rank, and linear combination, etc.[111] Examples of consensus scoring have MultiScore,[112] X-Cscore,[30] GFscore,[113] SCS,[114] and SeleX-CS,[115]

3 Criteria for evaluating scoring functions

In response to the three important applications of a scoring function as described in Introduction, three related but independent criteria are commonly used to evaluate the performance of a scoring function for its ability in binding mode identification, binding affinity prediction, and virtual database screening.

One of the essential measures for the performance of a scoring function is its ability to distinguish native binding modes from decoys. Namely, given a set of decoys for a protein-ligand complex, a reliable scoring function should be capable of ranking the native structure to the top by the calculated binding scores. In docking applications, successful prediction of a native binding mode is commonly defined by the rmsd value between the top ligand conformations and the experimentally observed (native) structure. If rmsd is ≤ 2.0 Å, the prediction is considered successful. Because it is simple and easy to implement, the rmsd criterion for binding mode prediction has been widely used and accepted in the field. However, this criterion could be problematic in some cases. For example, small or nearly symmetrical ligands always have good rmsd values when they are randomly placed in a small active site. On the contrary, for a large flexible ligand, the large rmsd value due to a solvent-exposed irrelevant group may hide the correctness of the overall binding mode. To overcome these limitations, several alternative methods have been presented for pose evaluations, such as relative displacement error (RDE),[116] interaction-based accuracy classification (IBAC),[117] real space R-factor (RSR),[118] and Generally Applicable Replacement for rmsD (GARD).[119]

A second important measure for a scoring function is its ability to predict the binding affinity of a complex, i.e. how tightly the ligand binds the protein. It is generally difficult to achieve a score scale similar to experimental binding data. (Certainly, one may scale the calculated scores to fit the normal affinity range.) Therefore, the commonly-used criterion for affinity prediction is the Peason correlation between the calculated scores and the experimental data, which is calculated as follows:

$$R=\frac{\sum_{k=1}^{N}(x_k-\langle x\rangle)(y_k-\langle y\rangle)}{\sqrt{\left[\sum_{k=1}^{N}(x_k-\langle x\rangle)^2\right]\left[\sum_{k=1}^{N}(y_k-\langle y\rangle)^2\right]}} \quad (4)$$

where N is the number of tested complexes. x_k and y_k are the experimentally determined binding energy and the calculated score for k-th complex, respectively. $\langle\cdots\rangle$ is an arithmetic average over all the complexes. Yet, the correlation between the predicted and experimental binding energies does not have to be linear for a scoring function. Therefore, the Spearman correlation coefficient, which calculates the correlation between two sets of rankings, is also often used in the binding affinity prediction evaluation as

$$R_s = 1 - \frac{6\sum_{k=1}^{N} d_k^2}{N(N^2 - 1)} \quad (5)$$

where the complexes in the test set are ranked by their known affinities and calculated scores, respectively, and d_k is the difference in two rankings for the k-th complex. Compared to binding mode prediction, binding affinity prediction is more challenging to be evaluated. One major reason is the uncertainties of the collected experimental affinity data that may come from different experimental conditions by different research groups or the inherent experimental error of an assay.

The third criterion for assessing a scoring function is its capability of selecting potential binders (hits) from a large database of compounds for a given protein target. The practical application is virtual screening in computer-based drug design, which is often used to identify lead compounds in drug discovery. Virtual database screening tests whether or not a scoring function is able to rank the known binders/inhibitors above many inactive compounds in a database. The enrichment test is a commonly-used criterion to quantify the performance of a scoring function in virtual database screening. The enrichment is defined as the accumulated rate of active inhibitors/binders found above a certain percentile of the ranked database that includes the active binders and inactive ligands. The higher enrichment corresponds to a better scoring function at a fixed percentage of the ranked database. Another measurement for virtual database screening is the receiver operating characteristic (ROC).[120,121] This method is normally more appropriate when the number of inactive ligands is comparable to the number of active binders.

Theoretically, an accurate scoring function should be able to perform equally well on all of the three criteria on any test set. However, due to the inherent limitations, most of the existing scoring functions usually perform well on only one or two of the criteria and fail on others. For example, as shown in a previous study[37] that SYTYL/F-Score yields a good success rate (74%) in binding mode prediction on the test set of 100 protein-ligand complexes constructed by Wang et al.[122] (Table 2), but the F-Score performs poor with a correlation coefficient of $R = 0.30$ in binding affinity prediction on the same set (Table 3). Similar examples can also be found in the comparative assessment of 16 scoring functions on a larger test set of 195 protein-ligand complexes by Cheng et al.[123] To be successful in virtual database screening usually requires good performance in both binding mode and affinity predictions. A scoring function that yields a good correlation in binding affinity prediction does not necessarily perform well in database ranking.[124] For example, PMF-Score yielded a high correlation ($R^2 = 0.61$) in binding affinity prediction on the PMF validation set of 77 complexes (Figure 2), but performed much less satisfactorily in virtual database screening and failed to identify any binder on two of four tested targets at the 5% of the ranked database (Table 4). In addition, the performances of scoring functions are test set-dependent. For example, ITScore and PMF-Score perform significantly better on the PMF validation set than on the Wang at al.'s set in binding affinity prediction (Figure 2 and Table 3). For the PMF validation set, all of the tested scoring functions perform better on the

serine protease than the others (Figure 2). Therefore, to fully evaluate the performance of a scoring function, all of the three criteria should be examined on multiple test sets.

4 Databases for evaluating scoring functions

In addition to the success criteria for evaluating scoring functions, another important issue in developing an efficient scoring function is the construction of an appropriate training/test set. Commonly-used (but not limited) criteria for constructing an appropriate training/test set include: The complexes in the set should be high quality structures with no atomic clashes (e.g. crystal structures with high resolutions); The set of complexes should cover a wide range of proteins and binding affinities; The ligands should be drug-like and non-covalent with the protein. Examples of the protein-ligand complex databases that can be used to construct the training/test set include:

1. LPDB (http://lpdb.chem.lsa.umich.edu/)[125]
2. PLD (http://chemistry.st-andrews.ac.uk/staff/jbom/group/PLD.xls)[126]
3. Binding DB (http://www.bindingdb.org/bind/)[127]
4. PDBbind (http://sw16.im.med.umich.edu/databases/pdbbind/)[128,129]
5. Binding MOAD (http://www.bindingmoad.org/)[130]
6. AffinDB (http://www.agklebe.de/affinity)[131]

5 Conclusion and Discussions

We have reviewed the scoring functions currently used for protein-ligand interactions in molecular docking and the commonly-used criteria/methods for evaluating the performance of scoring functions in three different applications: binding mode prediction, binding affinity prediction, and database screening. The criteria for constructing an appropriate training/test set and publicly available protein-ligand databases for evaluating a scoring function are also briefly depicted.

Despite considerable progress, current scoring functions are still far from being universally accurate, considering the test set-dependency of their performance and the fact that many of the scoring functions failed on one or two of the three widely-used criteria. To improve the universal applicability of the empirical scoring functions, a large training set of complexes with known affinity data are desired for parameter fitting. For force field and knowledge-based scoring functions, explicit and accurate inclusion of the desolvation and entropic effects is requisite to improve the accuracy. The categorization of atom types with a good balance of the statistics of the pair occurrences and the number of atom types is also important for knowledge-based scoring functions. Extension of the pairwise potentials to many-body potentials theoretically will help improve the accuracy of knowledge-based scoring functions but practically remains unknown because of the introduction of many more parameters to be determined. Lack of a universal set of weighting coefficients for different energy terms for diverse protein-ligand complexes is a challenge for force field scoring functions. What is even more challenging, neglect or inaccurate treatment of

entropic effect may easily render the hard efforts on accurate electrostatic calculations in force field scoring. Transition metal ions such as zinc impose great parameterization difficulty for all scoring functions. Another issue is how to evaluate the increasing number of scoring functions being developed.[132] Comparing different scoring functions is not always possible if they are tested on different sets. Although some comparison studies have been done by researchers,[122,124,133–136] publicly available benchmarks such as CCDC/Astex set,[137] CSAR (http://www.csardock.org/), and DUD (http://dud.docking.org/)[138] are invaluable for development of new and existing scoring functions.

Acknowledgments

Support to XZ from OpenEye Scientific Software Inc. (Santa Fe, NM) and Tripos, Inc. (St. Louis, MO) is gratefully acknowledged. XZ is supported by NIH grant GM088517, the Research Board Award of the University of Missouri RB-07-32 and Research Council Grant URC 09-004. The work is also supported by Federal Earmark NASA Funds for Bioinformatics Consortium Equipment and additional financial support from Dell, SGI, Sun Microsystems, TimeLogic, and Intel.

References

1. Brooijmans N and Kuntz ID., Annu. Rev. Biophys. Biomol. Struct, 2003, 32, 335–373. [PubMed: 12574069]
2. Böhm HJ and Stahl M, Rev. Comput. Chem, 2002, 18, 41–87.
3. Wang W, Donini O, Reyes CM and Kollman PA, Annu. Rev. Biophys. Biomol. Struct, 2001, 30, 211–243. [PubMed: 11340059]
4. Shoichet BK, McGovern SL, Wei B and Irwin JJ, Curr. Opin. Chem. Biol, 2002, 6, 439–446. [PubMed: 12133718]
5. Reddy MR and Erion MD, Free Energy Calculations in Rational Drug Design, Kluwer Academic: New York, 2001.
6. Seifert MHJ, Kraus J and Kramer B, Curr. Opin. Drug Discov. Devel, 2007, 10, 298–307
7. Jain AN, Curr. Protein Pept. Sci, 2006, 7, 407–420. [PubMed: 17073693]
8. Schulz-Gasch T and Stahl M, Drug Discov. Today: Tech, 2004, 1, 231–239.
9. Rajamani R and Good AC, Curr. Opin. Drug. Discov. Devel, 2007, 10, 308–315
10. Gohlke H and Klebe G, Curr. Opin. Struct. Biol, 2001, 11, 231–235. [PubMed: 11297933]
11. Gilson MK and Zhou HX, Annu. Rev. Biophys. Biomol. Struct 2007, 36, 21–42. [PubMed: 17201676]
12. Meng EC, Shoichet BK and Kuntz ID, J. Comput. Chem, 1992, 13, 505–524.
13. Shoichet BK, Leach AR and Kuntz ID ID, Proteins, 1999, 34, 4–16. [PubMed: 10336382]
14. Wei BQ, Baase WA, Weaver LH, Matthews BW and Shoichet BK, J. Mol. Biol, 2002, 322, 339–355. [PubMed: 12217695]
15. Zou X, Sun Y and Kuntz ID, J. Am. Chem. Soc, 1999, 121, 8033–8043.
16. Liu H-Y, Kuntz ID and Zou X, J. Phys. Chem. B, 2004, 108, 5453–5462.
17. Liu H-Y and Zou X, J. Phys. Chem. B, 2006, 110, 9304–9313. [PubMed: 16671749]
18. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK and Olson AJ, J. Comput. Chem, 1998, 19, 1639–1662.
19. Huey R, Morris GM, Olson AJ and Goodsell DS, J. Comput. Chem 2007, 28, 1145–1152. [PubMed: 17274016]
20. Jones G, Willett P, Glen RC, Leach AR and Talor R, J. Mol. Biol, 1997, 267, 727–748. [PubMed: 9126849]
21. Rarey M, Kramer B, Lengauer T and Klebe G, J. Mol. Biol, 1996, 261, 470–489. [PubMed: 8780787]

22. Friesner RA, Banks JL, Murphy RB and Halgren TA, J. Med. Chem, 2004, 47, 1739–1749. [PubMed: 15027865]
23. Abagyan R, Totrov M and Kuznetzov D, J. Comput. Chem, 1994, 15, 488–506.
24. Böhm HJ, J. Comput.-Aided Mol. Des, 1994, 8, 243–256. [PubMed: 7964925]
25. Böhm HJ, J. Comput.-Aided Mol. Des, 1998, 12, 309–323. [PubMed: 9777490]
26. Gehlhaar DK, Verkhivker GM, Rejto PA, Sherman CJ, Fogel DB and Freer ST, Chem. Biol, 1995, 2, 317–324. [PubMed: 9383433]
27. Gehlhaar DK, Bouzida D and Rejto PA, In Rational Drug Design: Novel Methodology and Practical Applications; Parrill L; Reddy MR; Ed.; American Chemical Society: Washington, DC, 1999, 292–311.
28. Eldridge MD, Murray CW, Auton TR, Paolini GV, and Mee RP, J. Comput.-Aided Mol. Des, 1997, 11, 425–445. [PubMed: 9385547]
29. Wang R, Liu L, Lai L and Tang Y, J. Mol. Model, 1998, 4, 379–394.
30. Wang R, Lai L and Wang S, J. Comput.-Aided Mol. Des, 2002, 16, 11–26. [PubMed: 12197663]
31. Jain AN, J. Med. Chem, 2003, 46, 499–511. [PubMed: 12570372]
32. Cerius2, version 4.6; Accelrys Inc.; http://www.accelrys.com/.
33. Yin S, Biedermannova L, Vondrasek J and Dokholyan NV, J. Chem. Inf. Model, 2008, 48, 1656–1662. [PubMed: 18672869]
34. Raub S, Steffen A, Kämper A and Marian CM, J. Chem. Inf. Model, 2008, 48, 1492–1510. [PubMed: 18597446]
35. Sotriffer CA, Sanschagrin P, Matter H and Klebe G, Proteins, 2008, 73, 395–419. [PubMed: 18442132]
36. Huang S-Y and Zou X, J. Comput. Chem, 2006, 27, 1865–1875.
37. Huang S-Y and Zou X, J. Comput. Chem, 2006, 27, 1876–1882. [PubMed: 16983671]
38. Huang S-Y and Zou X, J. Chem. Inf. Model, 2010, 50, 262–273. [PubMed: 20088605]
39. Muegge I and Martin YC, J. Med. Chem, 1999, 42, 791–804. [PubMed: 10072678]
40. Muegge I, J. Med. Chem 2006, 49, 5895–5902. [PubMed: 17004705]
41. Gohlke H, Hendlich M and Klebe G, J. Mol. Biol, 2000, 295, 337–356. [PubMed: 10623530]
42. Velec HFG, Gohlke H and Klebe G, J. Med. Chem, 2005, 48, 6296–6303. [PubMed: 16190756]
43. Zhang C, Liu S, Zhu Q and Zhou Y, J. Med. Chem, 2005, 48, 2325–2335. [PubMed: 15801826]
44. DeWitte RS and Shakhnovich EI, J. Am. Chem. Soc, 1996, 118, 11733–11744.
45. Ishchenko AV and Shakhnovich EI, J. Med. Chem, 2002, 45, 2770–2780. [PubMed: 12061879]
46. Mitchell JBO, Laskowski RA, Alex A and Thornton JM, J. Comput. Chem, 1999, 20, 1165–1176.
47. Mitchell JBO, Laskowski RA, Alex A, Forster MJ and Thornton JM, J. Comput. Chem, 1999, 20, 1177–1185.
48. Yang C-Y, Wang R and Wang S, J. Med. Chem, 2006, 49, 5903–5911 [PubMed: 17004706]
49. Mooij WT and Verdonk ML, Proteins, 2005, 61, 272–287. [PubMed: 16106379]
50. Zhao X, Liu X, Wang Y, Chen Z, Kang L, Zhang H, Luo X, Zhu W, Chen K, Li H, Wang X and Jiang H, J. Chem. Inf. Model, 2008, 48, 1438C1447 [PubMed: 18553962]
51. Huang N, Kalyanaraman C, Irwin JJ and Jacobson MP, J. Chem. Inf. Model, 2006, 46, 243–253. [PubMed: 16426060]
52. Weiner SJ, Kollman PA and Case DA, J. Am. Chem. Soc, 1984, 106, 765–784.
53. Weiner SJ, Kollman PA, Nguyen DT and Case DA, J. Comput. Chem, 1986, 7, 230–252. [PubMed: 29160584]
54. Adcock SA, and McCammon JA. Chem. Rev, 2006, 106, 1589–1615. [PubMed: 16683746]
55. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A and Honig B, J. Comput. Chem, 2002, 23, 128–137. [PubMed: 11913378]
56. Grant JA, Pickup BT and Nicholls A, J. Comput. Chem, 2001, 22, 608–640.
57. Baker NA, Sept D, Joseph S, Holst MJ and McCammon JA, Proc. Natl. Acad. Sci. USA, 2001, 98, 10037–10041. [PubMed: 11517324]
58. Still WC, Tempczyk A, Hawley RC and Hendrickson T, J. Am. Chem. Soc, 1990, 112, 6127–6129.

59. Hawkins GD, Cramer CJ and Truhlar DG, Chem. Phys. Lett, 1995, 246, 122–129.
60. Qiu D, Shenkin PS, Hollinger FP and Still WC, J. Phys. Chem. A, 1997, 101, 3005–3014.
61. Gasteiger J and Marsili M, Tetrahedron, 1980, 36, 3219–3228.
62. Li JB, Zhu TH, Cramer CJ, and Truhlar DG, J. Phys. Chem. A, 1998, 102, 1820–1831.
63. Wang J, Morin P, Wang W, and Kollman PA, J Am Chem Soc, 2001, 123, 5221–5230. [PubMed: 11457384]
64. Kuhn B, Gerber P, Schulz-Gasch T and Stahl M, J. Med. Chem 2005, 48, 4040–4048. [PubMed: 15943477]
65. Kuhn B and Kollman PA, J. Med. Chem, 2000, 43, 3786–3791. [PubMed: 11020294]
66. Pearlman DA, J. Med. Chem, 2005, 48, 7796–7807. [PubMed: 16302819]
67. Sims PA, Wong CF and McCammon JA, J. Med. Chem, 2003, 46, 3314–3325. [PubMed: 12852762]
68. Huang D and Caflisch A, J. Med. Chem, 2004, 47, 5791–5797. [PubMed: 15509178]
69. Thompson DC, Humblet C and Joseph-McCarthy D, J. Chem. Inf. Model, 2008, 48, 1081–1091. [PubMed: 18465849]
70. Y Liu H, Grinter SZ and Zou X, J. Phys. Chem. B, 2009, 113, 11793–11799. [PubMed: 19678651]
71. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C and Caflisch A, Proteins, 1999, 37, 88–105. [PubMed: 10451553]
72. Cecchini M, Kolb P, Majeux N and Caflisch A, Automated docking of highly flexible ligands by genetic algorithms: A critical assessment. J. Comput. Chem, 2004, 25, 412–422. [PubMed: 14696075]
73. Huang D, Luthi U, Kolb P, Edler K, Cecchini M, Audetat S, Barberis A and Caflisch A, J. Med. Chem, 2005, 48, 5108–5111. [PubMed: 16078830]
74. Cho AE, Wendel JA, Vaidehi N, Kekenes-Huskey PM, Floriano WB, Maiti PK and III Goddard WA, J. Comput. Chem, 2005, 26, 48–71. [PubMed: 15529328]
75. Ghosh A, Rapp CS and A Friesner R, J. Phys. Chem. B, 1998, 102, 10983–10990.
76. Lyne PD, Lamb ML and Saeh JC, J. Med. Chem 2006, 49, 4805–4808. [PubMed: 16884290]
77. Guimaraes CRW and Cardozo M, J. Chem. Inf. Model 2008, 48, 958–970. [PubMed: 18422307]
78. Ewing TJA, Makino S, Skillman AG and Kuntz ID, J. Comput.-Aided Mol. Des, 2001, 15, 411–428. [PubMed: 11394736]
79. Moustakas DT, Lang PT, Pegg S, Pettersen E, Kuntz ID, Brooijmans N and Rizzo RC, J. Comput.-Aided Mol. Des, 2006, 20, 601–619. [PubMed: 17149653]
80. Dill KA, J. Biol. Chem, 1997, 272, 701–704. [PubMed: 8995351]
81. Jain AN, J. Comp.-Aided Mol. Des, 1996, 10, 427–440.
82. Head RD, Smythe ML, Oprea TI, Waller CL, Green SM and Marshall GR, J. Am. Chem. Soc, 1996, 118, 3959–3969.
83. Deng W, Breneman C and Embrechts MJ, J. Chem. Inf. Comput. Sci, 2004, 44, 699–703. [PubMed: 15032552]
84. Ballester PJ and Mitchell JB, Bioinformatics, 2010, 26, 1169–1175. [PubMed: 20236947]
85. Tanaka S and Scheraga HA, Macromolecules, 1976, 9, 945–950. [PubMed: 1004017]
86. Miyazawa S and Jernigan RL, Macromolecules, 1985, 18, 534–552.
87. Sippl MJ, J. Mol. Biol, 1990, 213, 859–883. [PubMed: 2359125]
88. Thomas PD and Dill KA, Proc. Natl. Acad. Sci. USA, 1996, 93, 11628–11633. [PubMed: 8876187]
89. Koppensteiner WA and Sippl MJ, Biochemistry (Moscow), 1998, 63, 247–252. [PubMed: 9526121]
90. Thomas PD and Dill KA, J. Mol. Biol, 1996, 257, 457–469. [PubMed: 8609636]
91. McQuarrie DA, Statistical Mechanics; Harper Collins Publishers: New York, 1976.
92. Huang S-Y and Zou X, Annu. Rep. Comput. Chem, 2010, 6, (submitted).
93. Kirtay CK, Mitchell JBO and Lumley JA, QSAR & Combinatorial Sci,, 2005, 4, 527–536.

94. Li X X and Liang J, Knowledge-based energy functions for computational studies of proteins. In Computational Methods for Protein Structure Prediction and Modeling, Eds. Xu Y, Xu D, Liang J, 2006, 1, 71–124.
95. Sotriffer CA, Gohlke H and Klebe G, J. Med. Chem, 2002, 45, 1967–1970. [PubMed: 11985464]
96. Allen FH, Acta Crystallogr, 2002, B58, 380–388.
97. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE, Nucleic Acids Res, 2000, 28, 235–242. [PubMed: 10592235]
98. Nobeli I, Mitchell JBO, Alex A and Thornton JM, J. Comput. Chem, 2001, 22, 673–688.
99. Muegge I, Martin YC, Hajduk PJ and Fesik SW, J. Med. Chem, 1999, 42, 2498–2503. [PubMed: 10411471]
100. Ha S, Andreani R, Robbins A and Muegge I, J. Comput.-Aided Mol. Des, 2000, 14, 435–448. [PubMed: 10896316]
101. Huang S-Y and Zou X, Proteins, 2007, 66, 399–421. [PubMed: 17096427]
102. Huang S-Y and Zou X, Protein Sci, 2007, 16, 43–51. [PubMed: 17123961]
103. Huang S-Y and Zou X, Proteins, 2008, 72, 557–579. [PubMed: 18247354]
104. Seetharamulu P and Crippen GM, J. Math. Chem, 1991, 6, 91–110
105. Mimy LA and Shakhnovich EI, J. Mol. Biol, 1996, 264, 1164–1179. [PubMed: 9000638]
106. Huber T and Torda AE, Protein Sci, 1998, 7, 142–149 [PubMed: 9514269]
107. Koretke KK, Luthey-Schulten Z and Wolynes PG, Proc. Natl. Acad. Sci. USA, 1998, 95, 2932–2937. [PubMed: 9501193]
108. Charifson PS, Corkery JJ, Murcko MA and Walters WP, J. Med. Chem, 1999, 42, 5100–5109. [PubMed: 10602695]
109. Wang R and Wang S, J. Chem. Inf. Comput. Sci, 2001, 41, 1422–1426. [PubMed: 11604043]
110. Clark RD, Strizhev A, Leonard JM, Blake JF and Matthew JB, J. Mol. Graph. Model, 2002, 20, 281–295. [PubMed: 11858637]
111. Oda A, Tsuchida K, Takakura T, Yamaotsu N and Hirono S, J. Chem. Inf. Model, 2006, 46, 380–391. [PubMed: 16426072]
112. Terp GE, Johansen BE, Christensen IT and Jorgensen FS, J. Med. Chem, 2001, 44, 2333–2343. [PubMed: 11428927]
113. Betzi S, Suhre K, Chétrit B, Guerlesquin F and Morelli X, J. Chem. Inf. Model, 2006, 46, 1704–1712. [PubMed: 16859302]
114. Teramoto R and Fukunishi H, J. Chem. Inf. Model, 2007, 47, 526–534. [PubMed: 17295466]
115. Bar-Haim S, Aharon A, Ben-Moshe T, Marantz Y and Senderowitz H, J. Chem. Inf. Model, 2009, 49, 623–633. [PubMed: 19231809]
116. Abagyan RA and Totrov MM, J. Mol. Biol, 1997, 268, 678–685. [PubMed: 9171291]
117. Kroemer RT, Vulpetti A, McDonald JJ, Rohrer DC, Trosset J-Y, Giordanetto F, Cotesta S, McMartin C, Kihlen M and Stouten PFW, J. Chem. Inf. Comput. Sci, 2004, 44, 871–881. [PubMed: 15154752]
118. Yusuf D, Davis AM, Kleywegt GJ and Schmitt S, J. Chem. Inf. Model, 2008, 48, 1411–1422. [PubMed: 18598022]
119. Baber JC, Thompson DC, Cross JB and Humblet C, J. Chem. Inf. Model, 2009, 49, 1889–1900. [PubMed: 19618919]
120. Egan JP, Signal detection theory and ROC analysis, Academic Press, New York, 1975.
121. Jain AN, J. Comput.-Aided Mol. Des, 2000, 14, 199–213 [PubMed: 10721506]
122. Wang R, Lu Y and Wang S, J. Med. Chem, 2003, 46, 2287–2303. [PubMed: 12773034]
123. Cheng T, Li X, Li Y, Liu Z and Wang R, J. Chem. Inf. Model, 2009, 49, 1079–1093. [PubMed: 19358517]
124. Stahl M and Rarey M, J. Med. Chem, 2001, 44, 1035–1042. [PubMed: 11297450]
125. Roche O, Kiyama R and Brooks CL, J. Med. Chem, 2001, 44, 3592–3598. [PubMed: 11606123]
126. Puvanendrampillai D and Mitchell JB, Bioinformatics, 2003, 19, 1856–1857. [PubMed: 14512362]

127. Liu T, Lin Y, Wen X, Jorrisen RN and Gilson MK, Nucleic Acids Res, 2007, 35, D198–D201. [PubMed: 17145705]
128. Wang R, Fang X, Lu Y and Wang S, J. Med. Chem, 2004, 47, 2977–2980. [PubMed: 15163179]
129. Wang R, Fang X, Lu Y, Yang C-Y and Wang S, J. Med. Chem, 2005, 48, 4111–4119. [PubMed: 15943484]
130. Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J and Carlson HA, Nucleic Acids Res, 2008, 36, D674–D678. [PubMed: 18055497]
131. Block P, Sotriffer CA, Dramburg I and Klebe G, Nucleic Acids Res, 2006, 34, D522–D526. [PubMed: 16381925]
132. Jain AN and Nicholls A, J. Comput.-Aided Mol. Des, 2008, 22, 133–139. [PubMed: 18338228]
133. Ferrara P, Gohlke H, Price DJ, Klebe G and Brooks CL III, J. Med. Chem, 2004, 47, 3032–3047. [PubMed: 15163185]
134. Bissantz C, Folkers G and Rognan D, J. Med. Chem, 2000, 43, 4759–4767. [PubMed: 11123984]
135. Perola E, Walters WP and Charifson PS, Proteins, 2004, 56, 235–249. [PubMed: 15211508]
136. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE and Head MS, J. Med. Chem, 2006, 49, 5912–5931. [PubMed: 17004707]
137. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN and Murray CW, J. Med. Chem, 2007, 50, 726–741. [PubMed: 17300160]
138. Huang N, Shoichet BK and Irwin JJ, J. Med. Chem, 2006, 49, 6789–6801. [PubMed: 17154509]

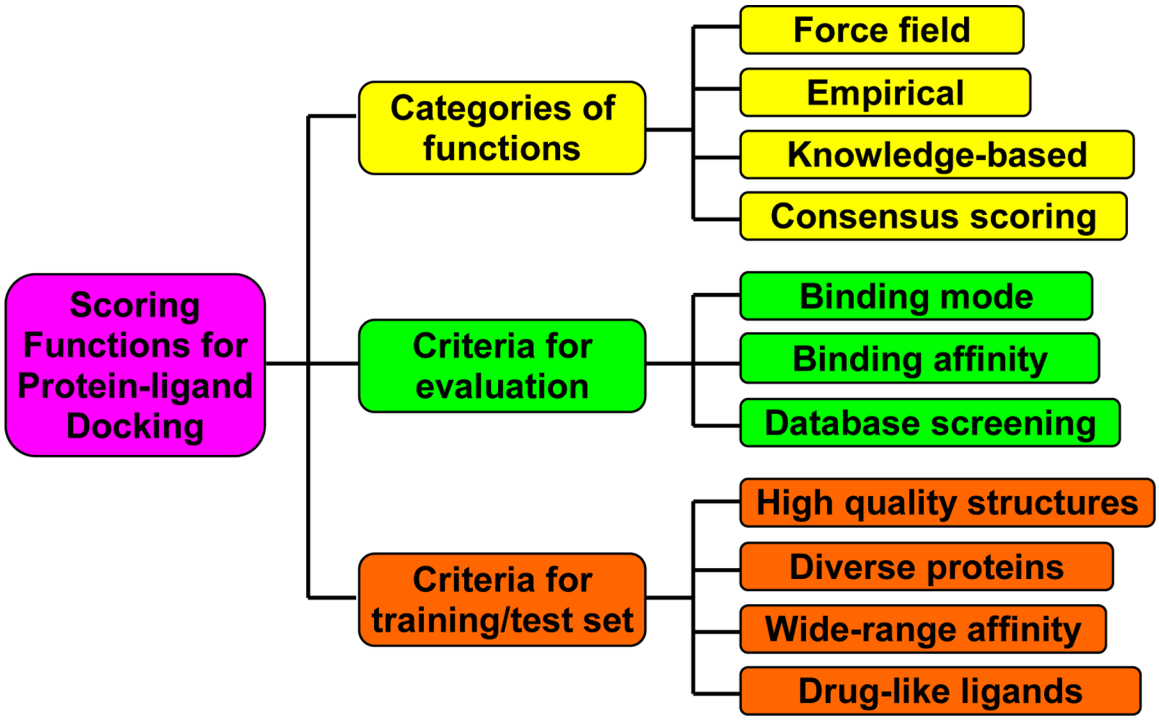

Figure 1:
An illustration of categories and evaluations for scoring functions in protein-ligand docking.

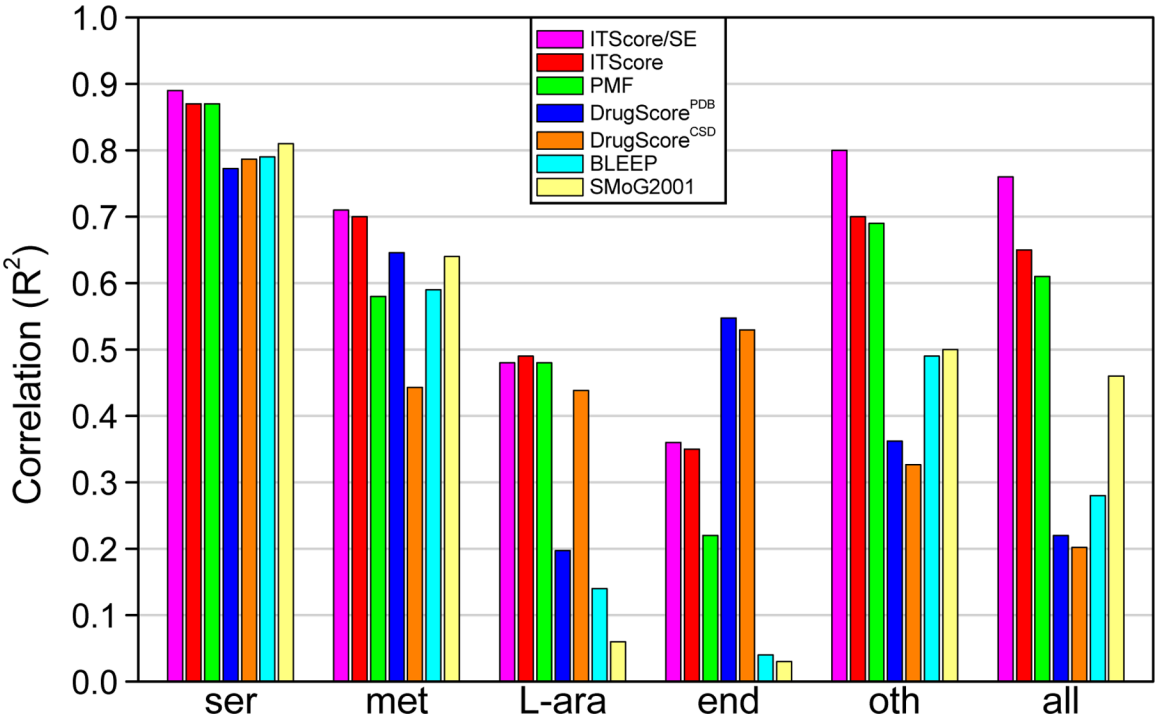

Figure 2:

Correlations of binding affinity predictions for 7 knowledge-based scoring functions with the PMF validation set of 77 protein-ligand complexes (all) that consists of five classes: 16 serine protease (ser), 15 metalloprotease (met), 18 L-arabinose binding protein (L-ara), 11 endothiapepsin (end), and 17 diverse protein-ligand complexes (oth).[39] The correlation parameter here is the square of correlation coefficient (R^2) rather than correlation coefficient itself (R) to maintain consistency with the original data. The correlation data for ITScore/SE, ITScore, BLEEP and SMoG2001 are taken from our previous study,[38] and those for DrugScore^{PDB} and DrugScore^{CSD} were calculated by the DrugScore^{ONLINE} server (http://pc1664.pharmazie.uni-marburg.de/drugscore/).

Table 1:

Types of scoring functions.

Type	Scoring function
Force field-based	DOCK,[12] DOCK3.5(PB/SA),[13,14] DOCK/GBSA(SDOCK),[15–17] AutoDock,[18,19] GOLD,[20] SYBYL/D-Score,[12] SYBYL/G-Score[20]
Empirical	FlexX,[21] Glide,[22] ICM,[23] LUDI,[24,25] PLP,[26,27] ChemScore,[28] SCORE,[29] X-Score,[30] Surflex,[31] SYBYL/F-Score,[21] LigScore,[32] MedusaScore,[33] AIScore,[34] SFCscore[35]
Knowledge-based	ITScore,[36–38] PMF,[39,40] DrugScore,[41,42] DFIRE,[43] SMoG,[44,45] BLEEP,[46,47] MScore,[48] GOLD/ASP,[49] KScore[50]

Table 2:

Success rates of 16 scoring functions for Wang et al.'s test set of 100 diverse protein-ligand complexes, using the criterion of rmsd ≤ 2.0 Å (from Huang and Zou, 2010).[38]

Scoring function	Type of scoring[a]	Success rate (%)
ITScore/SE[38]	K	91
DrugScore[CSD42]	K	87
ITScore[37]	K	82
Cerius2/PLP[26,27]	E	76
SYBYL/F-Score[21]	E	74
Cerius2/LigScore[32]	E	74
DrugScore[PDB41]	K	72
Cerius2/LUDI[24,25]	E	67
X-Score[30]	E	66
AutoDock[18]	F	62
DFIRE[43]	K	58
DOCK/FF[12]	F	58
Cerius2/PMF[39]	K	52
SYBYL/G-Score[20]	F	42
SYBYL/ChemScore[28]	E	35
SYBYL/D-Score[12]	F	26

[a]"K" stands for knowledge-based scoring functions, "E" for empirical scoring functions, and "F" for force field scoring functions, respectively.

Table 3:

Correlation coefficients between the experimentally determined binding energies and the calculated binding scores of 17 scoring functions for Wang et al.'s test set of 100 complexes (from Huang and Zou, 2010).[38]

Scoring function	Function type	Correlation (*R*)
ITScore/SE	K	0.65
ITScore	K	0.65
X-Score	E	0.64
DFIRE	K	0.63
$DrugScore^{CSD}$	K	0.62
$DrugScore^{PDB}$	K	0.60
Cerius2/PLP	E	0.56
SYBYL/G-Score	F	0.56
KScore	K	0.49
SYBYL/D-Score	F	0.48
SYBYL/ChemScore	E	0.47
Cerius2/PMF	K	0.40
DOCK/FF	F	0.40
Cerius2/LUDI	E	0.36
Cerius2/LigScore	E	0.35
SYBYL/F-Score	E	0.30
AutoDock	F	0.05

Table 4:

Enrichments of nine scoring functions at the top 5% of the ranked databases[a] on four targets of ERα, MMP3, fXa, and AChE (from Huang and Zou, 2006).[37]

		Enrichment at the top 5% (%)			
Scoring function	**Function type**	**ERα**	**MMP3**	**fXa**	**AChE**
ITScore	iterative/knowledge-based	19.2	68.3	34.9	37.0
DOCK/FF[12]	force-field-based	2.7	56.7	14.0	7.4
ICM-Score[23]	empirical	38.4	36.7	29.5	0.0
ICM-PMF[23]	knowledge-based	9.6	20.0	19.4	1.9
SYBYL/F-Score[21]	empirical	23.3	31.7	26.4	1.9
SYBYL/G-Score[20]	force-field-based	0.7	31.7	31.8	11.1
SYBYL/ChemScore[28]	empirical	0.0	73.3	23.3	9.3
SYBYL/PMF-Score[39]	knowledge-based	0.0	5.0	21.7	0.0
SYBYL/D-Score[12]	force-field-based	0.0	0.0	16.3	0.0
Maximum enrichments[b]		39.2	88.3	43.7	97.5

[a]For each protein target, the constructed database includes known inhibitors (146 for ERα, 60 for MMP3, 129 for fXa, and 54 for AChE) and 999 random, diverse drug-like molecules served as a set of inactive compounds.

[b]The last row lists the maximum theoretically possible enrichments at the top 5% of the ranked database, given the compositions of the databases including active and inactive compounds.