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

Sheng-You Huang, Sam Z. Grinter and Xiaoqin Zou*

Received 7th April 2010, Accepted 3rd August 2010
DOI: 10.1039/c0cp00151a

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 that are used for protein–ligand docking. The commonly-used assessment criteria and publicly available protein–ligand databases for performance evaluation of the scoring functions have also been presented and discussed. We end with a discussion of the challenges faced by existing scoring functions and possible future directions for developing improved scoring functions.

1. Introduction

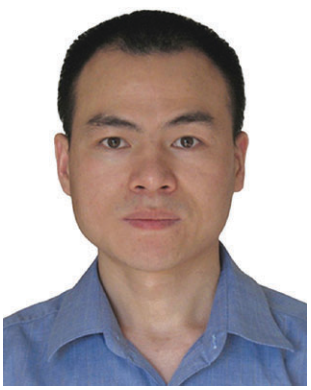
As the number of three-dimensional protein structures determined by experimental techniques grows, computational tools such as molecular docking have played an increasing role in the functional study of proteins and structure-based drug design.[1–6] In all the computational methodologies, one important 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 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 affinities 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

Department of Physics and Astronomy, Department of Biochemistry, Dalton Cardiovascular Research Center, and Informatics Institute, University of Missouri, Columbia, MO 65211, USA.
E-mail: zoux@missouri.edu; Fax: 573-884-4232; Tel: 573-882-6045



Sheng-You Huang

Dr Sheng-You Huang received his PhD in Physics from Wuhan University in 2003. He then became a postdoctoral fellow and is currently a research associate at the University of Missouri in Columbia. His research interests are molecular docking, scoring functions for protein–ligand and protein–protein interactions, and other molecular modeling including computer-aided drug design.

Sam Z. Grinter

Sam Grinter received his BSc in Physics and Mathematics from the University of Missouri in Columbia in 2007. He is now pursuing a PhD in bioinformatics. His primary research interest is developing algorithms for protein–ligand docking and applying these methods in virtual compound screening for structure-based drug design.

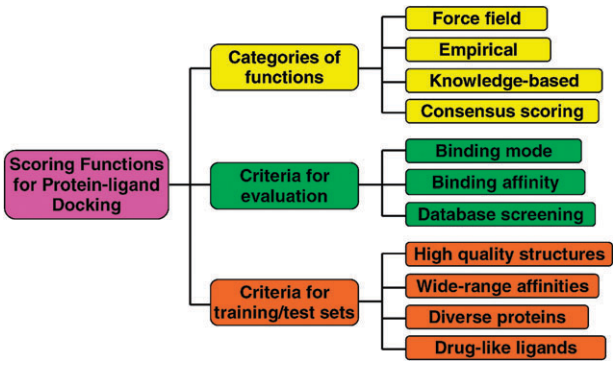


Fig. 1 An illustration of the categories and evaluations of the scoring functions for protein–ligand docking.

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 Fig. 1. Specifically, we will first briefly review different categories of scoring functions in section 2. We will then describe the commonly-used criteria for performance assessment of a scoring function in section 3. We will also review the publicly available protein–ligand databases for developing and validating scoring functions in section 4. Finally, challenges and future directions for scoring function development will be discussed in the Conclusion and Discussions.

Xiaoqin Zou

Xiaoqin Zou received her PhD in Physics from the University of California, San Diego in 1993. She did postdoctoral research first at the Case Western Reserve University and then at the University of California, San Francisco. She is currently an assistant professor at the University of Missouri in Columbia. Her research interests are physical modeling of protein–ligand and protein–protein interactions, structure-function relationship of membrane proteins, and structure-based drug design.

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

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

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 way to account for the solvent effect is to treat water molecules explicitly. Techniques such as free energy perturbation (FEP) and thermodynamic integration (TI) use explicit water representation (see ref. 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 water molecules are theoretically ideal/rigorous, the accuracy of the methods is also limited by the sampling issue and by the accuracy of the force field. This in turn depends on both the mathematical form and the parameterization thereof. To reduce the computational expense, some accelerated force field models have been

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]

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^{-1}), 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 course 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

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

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.

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.[83–85] The principle behind knowledge-based scoring functions is simple: Pairwise potentials are directly obtained from the occurrence frequency of atom pairs in a database using the inverse Boltzmann relation.[86–89] For protein–ligand studies, the potentials are calculated by

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

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.[89] 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 in these systems.[89,90] Moreover, the protein system is quite different from the simple fluid system due to the effects of atomic connectivity, excluded volume, composition, *etc*.[88] Therefore, $w(r)$ are not the true mean-force potentials, either.[90] Despite these limitations, the inverse Boltzmann method provides a simple and effective way to convert a histogram of atom–atom distances into a suitable function with the dimension of energy for complicated protein systems.[91] Since the pioneering work by Tanaka and Scheraga,[83] a large number of knowledge-based scoring functions have been developed and widely applied to protein structure prediction and protein–ligand studies (see ref. 92 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[88] 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*.[88] 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.[93] Recently, an improved version (DrugScore^{CSD})[42] was also developed based on the Cambridge Structural Database (CSD) of small molecules,[94] which contain low-molecular-weight structures with higher resolution than huge-molecular-weight structures in the Protein Data Bank (PDB).[95] 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 $R^2 = 0.28$.[96] 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[97] and inhibitors for matrix metalloprotease MMP-3.[98] 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 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.[36,37,99–101] 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 that can discriminate the native structures from decoys.[102–105] During the iteration procedure, the improvement for the potentials is guided through the difference between the predicted and experimentally observed pair distribution functions rather than through accurate calculation of the aforementioned reference state. Here, 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. Therefore, the iterative method circumvents the reference state problem faced by traditional knowledge-based scoring functions. Another advantage of the iterative method is its consideration of 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 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]

Inspired by the knowledge-based scoring functions, a knowledge-based quantitative structure–activity relationship (QSAR) approach has recently been introduced for scoring protein–ligand interactions.[106,107] In this type of QSAR approaches, the distance-dependent atom pair occurrences are used as descriptors for QSAR analysis by using a machine-learning method to fit the binding affinities of a training set. They differ from traditional knowledge-based scoring functions in using machine learning rather than reverse Boltzmann relationship, and in using binding affinities as well as structural data. One of the advantages of the machine-learning method is its ability to fit the binding affinities of a very large training set. For example, in a very recent study by Ballester and Mitchell,[107] the scoring function (RF-Score) derived from the machine-learning method yielded a high correlation (R = 0.953) for a large training set of 1105 protein–ligand complexes.

2.4 Consensus scoring

Despite a good number of scoring functions that have been developed, none of them is perfect in terms of accuracy and general applicability. Every scoring function has its advantages and limitations. To take the advantages and balance the deficiencies of different scoring functions, the consensus scoring technique has been introduced to improve the probability of finding correct solutions by combining the scores from multiple scoring functions.[108] The critical step in consensus scoring is the design of an appropriate consensus scoring strategy of individual scores so that the true modes/binders can be discriminated from others accordingly.[109,110] Commonly used consensus scoring strategies include vote-by-number, number-by-number, rank-by-number, average rank, linear combination, *etc*.[111] Examples of consensus scoring are 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 $\leq$2.0 Å, the prediction is considered successful. Because of its simplicity and ease of implementation, the rmsd criterion for binding mode prediction has been widely used in the field. However, this criterion could be problematic in some cases. For example, small or nearly symmetrical ligands are likely to obtain good rmsd values even 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, unimportant group may hide the correctness in prediction 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, may serve as a better index for ranking the complexes in order:

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

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 kth complex.

Compared to binding mode prediction, binding affinity prediction is more challenging to be assessed. 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. A higher enrichment at a fixed percentage of the ranked database can be taken to indicate a better scoring function. Another measurement for virtual database screening is AUC, the area under the receiver operating characteristic (*i.e.*, ROC) curve.[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, Wang *et al.*[122] showed that SYTYL/F-Score yields a good success rate (74%) in binding mode prediction with a test set of 100 protein–ligand complexes (Table 2), but performs poorly with a correlation coefficient of R = 0.30 in binding affinity prediction with the same test set (Table 3). Similar examples were also found in the comparative assessment of 16 scoring functions on a larger test set of 195 protein–ligand complexes by Cheng *et al.*[123] Success 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

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 $\leq$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

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 (Fig. 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 Wang *et al.*'s set in binding affinity prediction (Fig. 2 and Table 3). For the PMF validation set, all of the tested scoring functions perform better on the serine protease than the others (Fig. 2). Therefore, to fully evaluate the performance of a scoring function, the above three criteria should be examined with multiple test sets.

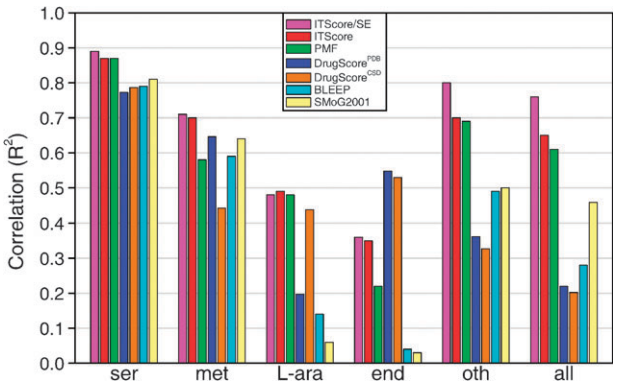


Fig. 2 Correlations of binding affinity predictions for 7 knowledge-based scoring functions with the PMF validation set of 77 protein–ligand complexes (all) that consists of five classes: 16 serine protease (ser), 15 metalloprotease (met), 18 L-arabinose binding protein (L-ara), 11 endothiapepsin (end), 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/).

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 exhaustive) criteria for constructing an appropriate training/test set include the following properties: 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 binding affinities and diverse protein types (except for developing a special scoring function for a specific protein family); The ligands should be drug-like and bind non-covalently to the protein. Overlaps between the training set and the test sets should be carefully avoided. Examples of the protein–ligand complex databases that can be used to construct a training or test set are listed as follows:

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 discussion

We have reviewed the scoring functions currently used for protein–ligand interactions in molecular docking. We have also described the commonly-used criteria/methods for scoring assessment in three different applications: binding mode prediction, binding affinity prediction, and database screening. Finally, we have briefly depicted the criteria for constructing an appropriate training/test set and the publicly available protein–ligand databases for such purposes.

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 effects may easily render useless the hard effort on accurate electrostatic calculations in force field scoring. Transition metal ions such as zinc

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]

Scoring function	Function type	Enrichment at the top 5% (%)			
		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 the active and inactive compounds.

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.

Acknowledgements

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 R21GM088517, 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 N. Brooijmans and I. D. Kuntz, *Annu. Rev. Biophys. Biomol. Struct.*, 2003, **32**, 335–373.
2 H. J. Böhm and M. Stahl, *Rev. Comput. Chem.*, 2002, **18**, 41–87.
3 W. Wang, O. Donini, C. M. Reyes and P. A. Kollman, *Annu. Rev. Biophys. Biomol. Struct.*, 2001, **30**, 211–243.
4 B. K. Shoichet, S. L. McGovern, B. Wei and J. J. Irwin, *Curr. Opin. Chem. Biol.*, 2002, **6**, 439–446.
5 M. R. Reddy and M. D. Erion, *Free Energy Calculations in Rational Drug Design*, Kluwer Academic, New York, 2001.
6 M. H. J. Seifert, J. Kraus and B. Kramer, *Curr. Opin. Drug Discov. Devel.*, 2007, **10**, 298–307.
7 A. N. Jain, *Curr. Protein Pept. Sci.*, 2006, **7**, 407–420.
8 T. Schulz-Gasch and M. Stahl, *Drug Discovery Today: Technol.*, 2004, **1**, 231–239.
9 R. Rajamani and A. C. Good, *Curr. Opin. Drug. Discov. Devel.*, 2007, **10**, 308–315.
10 H. Gohlke and G. Klebe, *Curr. Opin. Struct. Biol.*, 2001, **11**, 231–235.
11 M. K. Gilson and H. X. Zhou, *Annu. Rev. Biophys. Biomol. Struct.*, 2007, **36**, 21–42.
12 E. C. Meng, B. K. Shoichet and I. D. Kuntz, *J. Comput. Chem.*, 1992, **13**, 505–524.
13 B. K. Shoichet, A. R. Leach and I. D. Kuntz ID, *Proteins: Struct., Funct., Genet.*, 1999, **34**, 4–16.
14 B. Q. Wei, W. A. Baase, L. H. Weaver, B. W. Matthews and B. K. Shoichet, *J. Mol. Biol.*, 2002, **322**, 339–355.
15 X. Zou, Y. Sun and I. D. Kuntz, *J. Am. Chem. Soc.*, 1999, **121**, 8033–8043.
16 H.-Y. Liu, I. D. Kuntz and X. Zou, *J. Phys. Chem. B*, 2004, **108**, 5453–5462.
17 H.-Y. Liu and X. Zou, *J. Phys. Chem. B*, 2006, **110**, 9304–9313.
18 G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew and A. J. Olson, *J. Comput. Chem.*, 1998, **19**, 1639–1662.
19 R. Huey, G. M. Morris, A. J. Olson and D. S. Goodsell, *J. Comput. Chem.*, 2007, **28**, 1145–1152.
20 G. Jones, P. Willett, R. C. Glen, A. R. Leach and R. Talor, *J. Mol. Biol.*, 1997, **267**, 727–748.
21 M. Rarey, B. Kramer, T. Lengauer and G. Klebe, *J. Mol. Biol.*, 1996, **261**, 470–489.
22 R. A. Friesner, J. L. Banks, R. B. Murphy and T. A. Halgren, *J. Med. Chem.*, 2004, **47**, 1739–1749.
23 R. Abagyan, M. Totrov and D. Kuznetzov, *J. Comput. Chem.*, 1994, **15**, 488–506.
24 H. J. Böhm, *J. Comput.-Aided Mol. Des.*, 1994, **8**, 243–256.
25 H. J. Böhm, *J. Comput.-Aided Mol. Des.*, 1998, **12**, 309–323.
26 D. K. Gehlhaar, G. M. Verkhivker, P. A. Rejto, C. J. Sherman, D. B. Fogel and S. T. Freer, *Chem. Biol.*, 1995, **2**, 317–324.
27 D. K. Gehlhaar, D. Bouzida and P. A. Rejto, in *Rational Drug Design: Novel Methodology and Practical Applications*, ed. L. Parrill and M. R. Reddy, American Chemical Society, Washington, DC, 1999, pp. 292–311.
28 M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee, *J. Comput.-Aided Mol. Des.*, 1997, **11**, 425–445.
29 R. Wang, L. Liu, L. Lai and Y. Tang, *J. Mol. Model.*, 1998, **4**, 379–394.
30 R. Wang, L. Lai and S. Wang, *J. Comput.-Aided Mol. Des.*, 2002, **16**, 11–26.
31 A. N. Jain, *J. Med. Chem.*, 2003, **46**, 499–511.
32 *Cerius2*, version 4.6; Accelrys Inc.; http://www.accelrys.com/.
33 S. Yin, L. Biedermannova, J. Vondrasek and N. V. Dokholyan, *J. Chem. Inf. Model.*, 2008, **48**, 1656–1662.
34 S. Raub, A. Steffen, A. Kämper and C. M. Marian, *J. Chem. Inf. Model.*, 2008, **48**, 1492–1510.
35 C. A. Sotriffer, P. Sanschagrin, H. Matter and G. Klebe, *Proteins: Struct., Funct., Bioinf.*, 2008, **73**, 395–419.
36 S.-Y. Huang and X. Zou, *J. Comput. Chem.*, 2006, **27**, 1865–1875.
37 S.-Y. Huang and X. Zou, *J. Comput. Chem.*, 2006, **27**, 1876–1882.
38 S.-Y. Huang and X. Zou, *J. Chem. Inf. Model.*, 2010, **50**, 262–273.
39 I. Muegge and Y. C. Martin, *J. Med. Chem.*, 1999, **42**, 791–804.
40 I. Muegge, *J. Med. Chem.*, 2006, **49**, 5895–5902.

41 H. Gohlke, M. Hendlich and G. Klebe, *J. Mol. Biol.*, 2000, **295**, 337–356.
42 H. F. G. Velec, H. Gohlke and G. Klebe, *J. Med. Chem.*, 2005, **48**, 6296–6303.
43 C. Zhang, S. Liu, Q. Zhu and Y. Zhou, *J. Med. Chem.*, 2005, **48**, 2325–2335.
44 R. S. DeWitte and E. I. Shakhnovich, *J. Am. Chem. Soc.*, 1996, **118**, 11733–11744.
45 A. V. Ishchenko and E. I. Shakhnovich, *J. Med. Chem.*, 2002, **45**, 2770–2780.
46 J. B. O. Mitchell, R. A. Laskowski, A. Alex and J. M. Thornton, *J. Comput. Chem.*, 1999, **20**, 1165–1176.
47 J. B. O. Mitchell, R. A. Laskowski, A. Alex, M. J. Forster and J. M. Thornton, *J. Comput. Chem.*, 1999, **20**, 1177–1185.
48 C.-Y. Yang, R. Wang and S. Wang, *J. Med. Chem.*, 2006, **49**, 5903–5911.
49 W. T. Mooij and M. L. Verdonk, *Proteins: Struct., Funct., Bioinf.*, 2005, **61**, 272–287.
50 X. Zhao, X. Liu, Y. Wang, Z. Chen, L. Kang, H. Zhang, X. Luo, W. Zhu, K. Chen, H. Li, X. Wang and H. Jiang, *J. Chem. Inf. Model.*, 2008, **48**, 1438.
51 N. Huang, C. Kalyanaraman, J. J. Irwin and M. P. Jacobson, *J. Chem. Inf. Model.*, 2006, **46**, 243–253.
52 S. J. Weiner, P. A. Kollman and D. A. Case, *J. Am. Chem. Soc.*, 1984, **106**, 765–784.
53 S. J. Weiner, P. A. Kollman, D. T. Nguyen and D. A. Case, *J. Comput. Chem.*, 1986, **7**, 230–252.
54 S. A. Adcock and J. A. McCammon, *Chem. Rev.*, 2006, **106**, 1589–1615.
55 W. Rocchia, S. Sridharan, A. Nicholls, E. Alexov, A. Chiabrera and B. Honig, *J. Comput. Chem.*, 2002, **23**, 128–137.
56 J. A. Grant, B. T. Pickup and A. Nicholls, *J. Comput. Chem.*, 2001, **22**, 608–640.
57 N. A. Baker, D. Sept, S. Joseph, M. J. Holst and J. A. McCammon, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 10037–10041.
58 W. C. Still, A. Tempczyk, R. C. Hawley and T. Hendrickson, *J. Am. Chem. Soc.*, 1990, **112**, 6127–6129.
59 G. D. Hawkins, C. J. Cramer and D. G. Truhlar, *Chem. Phys. Lett.*, 1995, **246**, 122–129.
60 D. Qiu, P. S. Shenkin, F. P. Hollinger and W. C. Still, *J. Phys. Chem. A*, 1997, **101**, 3005–3014.
61 J. Gasteiger and M. Marsili, *Tetrahedron*, 1980, **36**, 3219–3228.
62 J. B. Li, T. H. Zhu, C. J. Cramer and D. G. Truhlar, *J. Phys. Chem. A*, 1998, **102**, 1820–1831.
63 J. Wang, P. Morin, W. Wang and P. A. Kollman, *J. Am. Chem. Soc.*, 2001, **123**, 5221–5230.
64 B. Kuhn, P. Gerber, T. Schulz-Gasch and M. Stahl, *J. Med. Chem.*, 2005, **48**, 4040–4048.
65 B. Kuhn and P. A. Kollman, *J. Med. Chem.*, 2000, **43**, 3786–3791.
66 D. A. Pearlman, *J. Med. Chem.*, 2005, **48**, 7796–7807.
67 P. A. Sims, C. F. Wong and J. A. McCammon, *J. Med. Chem.*, 2003, **46**, 3314–3325.
68 D. Huang and A. Caflisch, *J. Med. Chem.*, 2004, **47**, 5791–5797.
69 D. C. Thompson, C. Humblet and D. Joseph-McCarthy, *J. Chem. Inf. Model.*, 2008, **48**, 1081–1091.
70 H.-Y. Liu, S. Z. Grinter and X. Zou, *J. Phys. Chem. B*, 2009, **113**, 11793–11799.
71 N. Majeux, M. Scarsi, J. Apostolakis, C. Ehrhardt and A. Caflisch, *Proteins: Struct., Funct., Genet.*, 1999, **37**, 88–105.
72 M. Cecchini, P. Kolb, N. Majeux and A. Caflisch, *J. Comput. Chem.*, 2004, **25**, 412–422.
73 D. Huang, U. Luthi, P. Kolb, K. Edler, M. Cecchini, S. Audetat, A. Barberis and A. Caflisch, *J. Med. Chem.*, 2005, **48**, 5108–5111.
74 A. E. Cho, J. A. Wendel, N. Vaidehi, P. M. Kekenes-Huskey, W. B. Floriano, P. K. Maiti and W. A. Goddard, III, *J. Comput. Chem.*, 2005, **26**, 48–71.
75 A. Ghosh, C. S. Rapp and R. A Friesner, *J. Phys. Chem. B*, 1998, **102**, 10983–10990.
76 P. D. Lyne, M. L. Lamb and J. C. Saeh, *J. Med. Chem.*, 2006, **49**, 4805–4808.
77 C. R. W. Guimarães and M. Cardozo, *J. Chem. Inf. Model.*, 2008, **48**, 958–970.
78 T. J. A. Ewing, S. Makino, A. G. Skillman and I. D. Kuntz, *J. Comput.-Aided Mol. Des.*, 2001, **15**, 411–428.
79 D. T. Moustakas, P. T. Lang, S. Pegg, E. Pettersen, I. D. Kuntz, N. Brooijmans and R. C. Rizzo, *J. Comput.-Aided Mol. Des.*, 2006, **20**, 601–619.
80 K. A. Dill, *J. Biol. Chem.*, 1997, **272**, 701–704.
81 A. N. Jain, *J. Comput.-Aided Mol. Des.*, 1996, **10**, 427–440.
82 R. D. Head, M. L. Smythe, T. I. Oprea, C. L. Waller, S. M. Green and G. R. Marshall, *J. Am. Chem. Soc.*, 1996, **118**, 3959–3969.
83 S. Tanaka and H. A. Scheraga, *Macromolecules*, 1976, **9**, 945–950.
84 S. Miyazawa and R. L. Jernigan, *Macromolecules*, 1985, **18**, 534–552.
85 M. J. Sippl, *J. Mol. Biol.*, 1990, **213**, 859–883.
86 P. D. Thomas and K. A. Dill, *Proc. Natl. Acad. Sci. U. S. A.*, 1996, **93**, 11628–11633.
87 W. A. Koppensteiner and M. J. Sippl, *Biochemistry (Moscow)*, 1998, **63**, 247–252.
88 P. D. Thomas and K. A. Dill, *J. Mol. Biol.*, 1996, **257**, 457–469.
89 D. A. McQuarrie, *Statistical Mechanics*, Harper Collins Publishers, New York, 1976.
90 S.-Y. Huang and X. Zou, *Annu. Rep. Comput. Chem.*, 2010, **6**, 281–296.
91 C. K. Kirtay, J. B. O. Mitchell and J. A. Lumley, *QSAR Comb. Sci.*, 2005, **24**, 527–536.
92 X. Li and J. Liang, *Knowledge-based energy functions for computational studies of proteins* in *Computational Methods for Protein Structure Prediction and Modeling*, ed. Y. Xu, D. Xu and J. Liang, Springer, New York, 2006, **vol. 1**, pp. 71–124.
93 C. A. Sotriffer, H. Gohlke and G. Klebe, *J. Med. Chem.*, 2002, **45**, 1967–1970.
94 F. H. Allen, *Acta Crystallogr., Sect. B: Struct. Sci.*, 2002, **58**, 380–388.
95 H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235–242.
96 I. Nobeli, J. B. O. Mitchell, A. Alex and J. M. Thornton, *J. Comput. Chem.*, 2001, **22**, 673–688.
97 I. Muegge, Y. C. Martin, P. J. Hajduk and S. W. Fesik, *J. Med. Chem.*, 1999, **42**, 2498–2503.
98 S. Ha, R. Andreani, A. Robbins and I. Muegge, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 435–448.
99 S.-Y. Huang and X. Zou, *Proteins: Struct., Funct., Bioinf.*, 2007, **66**, 399–421.
100 S.-Y. Huang and X. Zou, *Protein Sci.*, 2006, **16**, 43–51.
101 S.-Y. Huang and X. Zou, *Proteins: Struct., Funct., Bioinf.*, 2008, **72**, 557–579.
102 P. Seetharamulu and G. M. Crippen, *J. Math. Chem.*, 1991, **6**, 91–110.
103 L. A. Mirny and E. I. Shakhnovich, *J. Mol. Biol.*, 1996, **264**, 1164–1179.
104 T. Huber and A. E. Torda, *Protein Sci.*, 2008, **7**, 142–149.
105 K. K. Koretke, Z. Luthey-Schulten and P. G. Wolynes, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**, 2932–2937.
106 W. Deng, C. Breneman and M. J. Embrechts, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 699–703.
107 P. J. Ballester and J. B. Mitchell, *Bioinformatics*, 2010, **26**, 1169–1175.
108 P. S. Charifson, J. J. Corkery, M. A. Murcko and W. P. Walters, *J. Med. Chem.*, 1999, **42**, 5100–5109.
109 R. Wang and S. Wang, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1422–1426.
110 R. D. Clark, A. Strizhev, J. M. Leonard, J. F. Blake and J. B. Matthew, *J. Mol. Graphics Modell.*, 2002, **20**, 281–295.
111 A. Oda, K. Tsuchida, T. Takakura, N. Yamaotsu and S. Hirono, *J. Chem. Inf. Model.*, 2006, **46**, 380–391.
112 G. E. Terp, B. E. Johansen, I. T. Christensen and F. S. Jorgensen, *J. Med. Chem.*, 2001, **44**, 2333–2343.
113 S. Betzi, K. Suhre, B. Chétrit, F. Guerlesquin and X. Morelli, *J. Chem. Inf. Model.*, 2006, **46**, 1704–1712.
114 R. Teramoto and H. Fukunishi, *J. Chem. Inf. Model.*, 2007, **47**, 526–534.
115 S. Bar-Haim, A. Aharon, T. Ben-Moshe, Y. Marantz and H. Senderowitz, *J. Chem. Inf. Model.*, 2009, **49**, 623–633.
116 R. A. Abagyan and M. M. Totrov, *J. Mol. Biol.*, 1997, **268**, 678–685.
117 R. T. Kroemer, A. Vulpetti, J. J. McDonald, D. C. Rohrer, J.-Y. Trosset, F. Giordanetto, S. Cotesta, C. McMartin,

M. Kihlen and P. F. W. Stouten, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 871–881.
118 D. Yusuf, A. M. Davis, G. J. Kleywegt and S. Schmitt, *J. Chem. Inf. Model.*, 2008, **48**, 1411–1422.
119 J. C. Baber, D. C. Thompson, J. B. Cross and C. Humblet, *J. Chem. Inf. Model.*, 2009, **49**, 1889–1900.
120 J. P. Egan, *Signal Detection Theory and ROC Analysis*, Academic Press, New York, 1975.
121 A. N. Jain, *J. Comput.-Aided Mol. Des.*, 2000, **14**, 199–213.
122 R. Wang, Y. Lu and S. Wang, *J. Med. Chem.*, 2003, **46**, 2287–2303.
123 T. Cheng, X. Li, Y. Li, Z. Liu and R. Wang, *J. Chem. Inf. Model.*, 2009, **49**, 1079–1093.
124 M. Stahl and M. Rarey, *J. Med. Chem.*, 2001, **44**, 1035–1042.
125 O. Roche, R. Kiyama and C. L. Brooks, *J. Med. Chem.*, 2001, **44**, 3592–3598.
126 D. Puvanendrampillai and J. B. Mitchell, *Bioinformatics*, 2003, **19**, 1856–1857.
127 T. Liu, Y. Lin, X. Wen, R. N. Jorrisen and M. K. Gilson, *Nucleic Acids Res.*, 2007, **35**, D198–D201.
128 R. Wang, X. Fang, Y. Lu and S. Wang, *J. Med. Chem.*, 2004, **47**, 2977–2980.
129 R. Wang, X. Fang, Y. Lu, C.-Y. Yang and S. Wang, *J. Med. Chem.*, 2005, **48**, 4111–4119.
130 M. L. Benson, R. D. Smith, N. A. Khazanov, B. Dimcheff, J. Beaver, P. Dresslar, J. Nerothin and H. A. Carlson, *Nucleic Acids Res.*, 2007, **36**, D674–D678.
131 P. Block, C. A. Sotriffer, I. Dramburg and G. Klebe, *Nucleic Acids Res.*, 2006, **34**, D522–D526.
132 A. N. Jain and A. Nicholls, *J. Comput.-Aided Mol. Des.*, 2008, **22**, 133–139.
133 P. Ferrara, H. Gohlke, D. J. Price, G. Klebe and C. L. Brooks, III, *J. Med. Chem.*, 2004, **47**, 3032–3047.
134 C. Bissantz, G. Folkers and D. Rognan, *J. Med. Chem.*, 2000, **43**, 4759–4767.
135 E. Perola, W. P. Walters and P. S. Charifson, *Proteins: Struct., Funct., Bioinf.*, 2004, **56**, 235–249.
136 G. L. Warren, C. W. Andrews, A.-M. Capelli, B. Clarke, J. LaLonde, M. H. Lambert, M. Lindvall, N. Nevins, S. F. Semus, S. Senger, G. Tedesco, I. D. Wall, J. M. Woolven, C. E. Peishoff and M. S. Head, *J. Med. Chem.*, 2006, **49**, 5912–5931.
137 M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson and C. W. Murray, *J. Med. Chem.*, 2007, **50**, 726–741.
138 N. Huang, B. K. Shoichet and J. J. Irwin, *J. Med. Chem.*, 2006, **49**, 6789–6801.

