Targeted scoring functions for virtual screening

Markus H.J. Seifert

4SC AG, Am Klopferspitz 19A, D-82152 Planegg-Martinsried, Germany

The benefit offered by virtual screening methods during the early drug discovery process is directly related to the predictivity of scoring functions that assess protein–ligand binding affinity. The scoring of protein–ligand complexes, however, is still a challenge: despite great efforts, a universal and accurate scoring method has not been developed up to now. Targeted scoring functions, in contrast, enhance virtual screening performance significantly. This review analyzes recent developments and future directions in the area of targeted scoring functions.

Introduction

Protein–ligand binding affinity is the key determinant of biological activity and therefore the major objective in the early drug discovery process. At that stage, experimental and virtual screening of chemical libraries is used to identify compounds with sufficient affinity to allow for starting and promoting a medicinal chemistry program. The benefit offered by virtual screening methods [1,2] is directly related to the predictivity of scoring functions that estimate protein–ligand binding affinity, whereby a better predictivity reduces the number of compounds that have to be screened physically. Currently, scoring functions derived from empirical data dominate in practical applications, mainly for three reasons: firstly, they are easy to evaluate and allow for screening millions of compounds within hours. Secondly, massive amounts of empirical data are available that can be used to train scoring functions [3–6]. Thirdly, *ab initio* simulations of molecular binding events are too time-consuming, especially for high-throughput applications. Empirical scoring of protein–ligand complexes, however, is still a challenge. Despite a long history of research, the final goal of developing a universal, fast and accurate scoring method has not been achieved up to now [7]. Therefore, efforts to devise generic scoring functions with increased predictivity are ongoing and various methodologies are explored, for example, force-field approaches [8], potentials of mean force [9] and consensus scoring [10]. Target-specific compound collections [11] allow improvement of the results of virtual screening, because their more restricted composition avoids some of the problems associated with current scoring functions. Recently, large-scale crystal structure data were used to derive various empirical scoring functions, which achieved impressive, internally validated, correlation coefficients between experimental and predicted binding affinities of $R^2 = 0.57–0.72$ and standard errors around one logarithmic unit [12]. External validation, which provides a more realistic impression of predictivity, revealed correlation coefficients up to $R^2 = 0.38$. This correlation corresponds to the following percentages of correctly classified compounds: 39–51%, 53–78% and 35–83% for low ($\mathrm{p}K_\mathrm{i} < 5.0$), medium ($5.0 < \mathrm{p}K_\mathrm{i} < 8.0$) and high affinity compounds ($\mathrm{p}K_\mathrm{i} > 8.0$), respectively. In general, these results are at the upper limit of what has ever been achieved with conventional approaches [13].

What does this predictivity mean, however, for virtual screening? Assuming a 'best case' scenario, the best values cited above generalize to a real world setting, that is binding modes are generated by protein–ligand docking software and completely different chemotypes are encountered during screening. Under this very optimistic assumption, a database consisting of 1 million low affinity compounds and 100 high affinity compounds will be classified into $(1–0.51) \times 1{,}000{,}000 = 490{,}000$ false positives and only $0.83 \times 100 = 83$ true positives. Although this is an artificial example, it clearly illustrates the main challenge: screening highly biased databases requires accurate scoring functions, with a particular focus on the correct identification of inactive or low affinity compounds. A calculation of the probability for identifying true actives by virtual screening (Box 1) highlights the level of

E-mail address: mhj.seifert@gmx.de. *E-mail address:* markus.seifert@4sc.com.

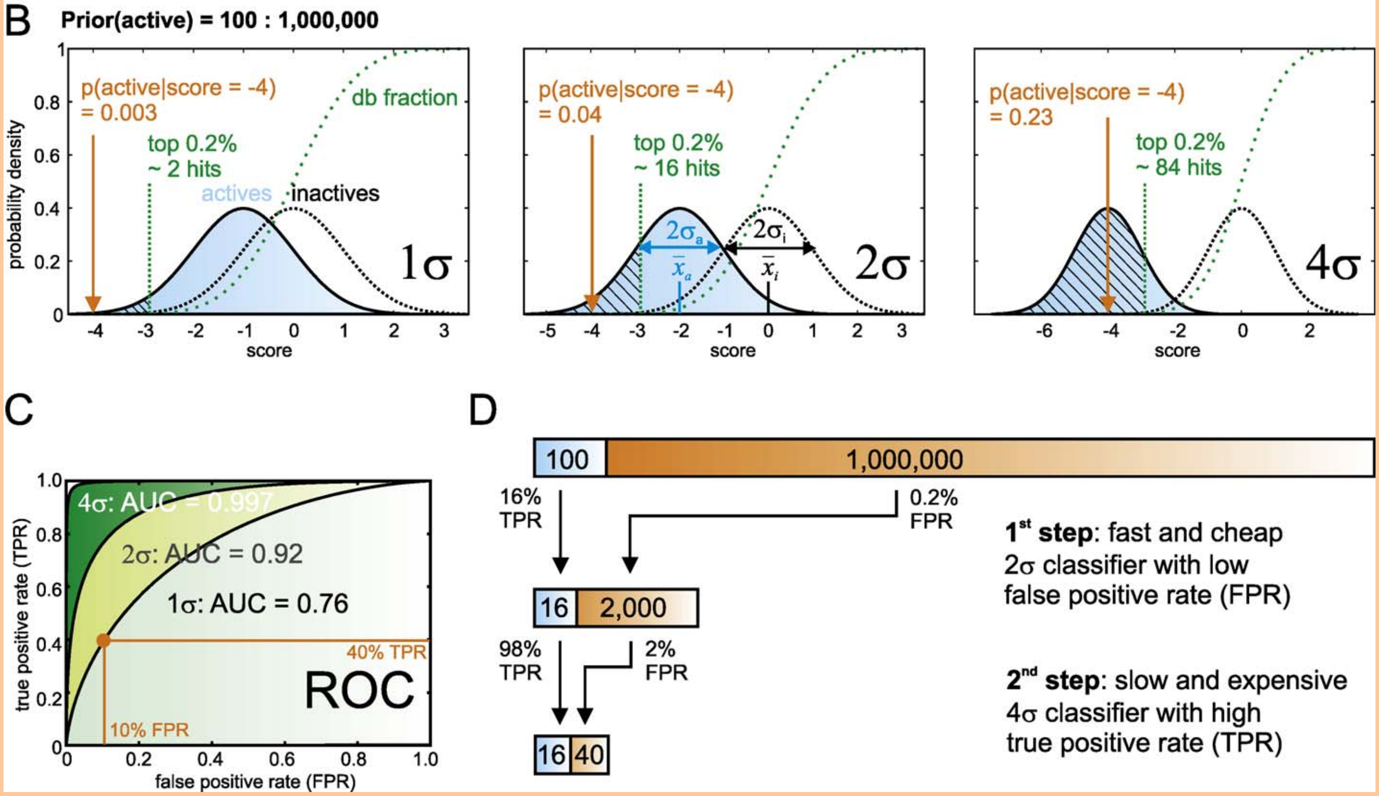
1359-6446/06/$ - see front matter © 2009 Elsevier Ltd. All rights reserved. doi:10.1016/j.drudis.2009.03.013

BOX 1

Probability of finding active molecules by virtual screening

A

$$\log \frac{p(a\,|\,x)}{1-p(a\,|\,x)} = \log \frac{p(a)}{1-p(a)} + \log \frac{\sigma_i}{\sigma_a} + \frac{1}{2}\left[\left(\frac{x-\bar{x}_i}{\sigma_i}\right)^2 - \left(\frac{x-\bar{x}_a}{\sigma_a}\right)^2\right]$$



A typical task for virtual screening aims at identifying, for example 100 active compounds out of a database of 1,000,000 compounds by examining only the 2000 top-scoring compounds (0.2%) of a score sorted ranking list. The prior probability of activity is $p(a) = 100/1{,}000{,}000 = 0.01\%$. Focusing on the classification of active and inactive compounds, a good virtual screening performance means that a scoring function is able to separate the score distributions for active and inactive compounds. In practice, various sources of error – including inadequate models, imperfect training data and so on – will broaden the distributions which causes overlap and classification noise. Assuming normal distributions $N(x_a, \sigma_a)$ and $N(x_i, \sigma_i)$ for the score distributions of active and inactive compounds, respectively, an expression for the odds of finding an active compound at a particular score x can be derived directly from Bayes' formula [16]. Eq. (A) can be solved for $p(a|x)$, which stands for the probability that a single compound which has received a score x is indeed active. A separation of one standard deviation (1σ) between the score distributions of actives and inactives leads to the following situation (B): a compound with a score of −4 has a probability of only 0.003 for being active and on average 2 hits are found within the top 2000 molecules. Accordingly, there is considerable risk that a particular screening run ends without any hits. A separation of two standard deviations (2σ) improves the situation significantly. Now $p(a|-4)$ is ten times higher (0.04) and approximately 16 hits are found on average within the top 2000. More than 50 hits are found and probabilities >10% for a single compound with score −4 are achieved with separations of three (3σ) or four (4σ) standard deviations. This allows for selecting only a few compounds and nevertheless having a good chance to get true hits. 3σ or 4σ classifiers, however, are rarely found for real world molecular data. Realistically, a separation of roughly 2σ on large and diverse datasets seems to be appropriate for successful virtual screening. Choosing the number of top-scoring molecules for subsequent evaluation corresponds to setting the 'operating point' of a classification method, which refers to specific trade-off between true and false positive rates (see orange dot). All possible operating points of a classifier lie on the 'receiver operating characteristic' (ROC) curve (C). The area under the ROC curve (AUC) is an established measure for classification quality in many disciplines and gives the probability of correctly ranking a randomly chosen pair of an active and an inactive compound (for an instructive demo see http://www.anaesthetist.com/mnm/stats/roc/Findex.htm). Of course, different classifiers can be combined. For example, virtual screening in pharmaceutical industry is usually performed as a two-step process where the ranking list from virtual screening is reviewed by a human expert. This is roughly equivalent to a combination of a fast and inexpensive 2σ classifier with a slow and expensive 4σ classifier (D). Obviously, this setup allows for extracting true hits even from a highly biased database.

predictivity that is necessary for successful virtual screening. Generic scoring functions, which are currently applied for virtual screening, seemingly lack this accuracy of affinity prediction [14].

Nevertheless, virtual screening is able to deliver true hits in many real world applications [15]. This is because of two important facts: on one hand, a two-step process consisting of virtual screening and subsequent postprocessing, for example by human experts, is able to alleviate the impact of this requirement (Box 1C). On the other hand, the calculation in Box 1 is based on the implicit assumption that the score is the only information about the compound, which is considered during virtual screening. In reality, experienced users will gather all available information about ligand(s), target(s) and so on, before screening. This additional information will inevitably influence the setup of virtual screening and subsequent compound selection. As a result, virtual screening usually performs better than anticipated solely from scoring function predictivity.

For industrial applications, however, it is desirable to account for this additional information in a more formal way by identifying workflows that systematically exploit prior information for increasing efficiency and reducing the risk of failure. The design and optimization of targeted scoring functions is a viable route for achieving this goal. Targeted scoring functions aim at incorporating the prior knowledge directly into the scoring function, instead of applying it as prefilter or postfilter for virtual screening. This review provides an overview of current methodologies and applications of targeted scoring functions.

Current approaches

The concept of developing targeted scoring functions is similar to modern machine learning procedures [16]: a training dataset is used to learn a statistical model, which is subsequently applied to predict new entities. In general, statistical learning methods can be broken down to the minimization of an objective function, which usually consists of a term for model quality and a penalty for model complexity. This allows for classifying the approaches for targeted scoring functions according to the objective function and the algorithm used for minimizing the objective function (Table 1).

Protein–ligand docking, however, exhibits one important difference to conventional machine learning techniques: the multiple instance problem [17]. To minimize the objective function, scoring function parameters are changed, which, in turn, lead to different top-scoring binding modes and thereby to a modification of the objective function. This feedback loop is commonly addressed by iterative optimization. Table 1 additionally describes the type of scoring function, the type of modification and the molecular target(s). Because of interesting methodological similarities, a few approaches for generic scoring function optimization and fingerprint-based 'scoring functions' have been compiled in Table 1 as well.

Two standard measures exist for evaluating the predictivity of scoring functions: should quantitative activities of compounds, like IC_{50} or K_i values, be known, correlation coefficients (R^2) are used. For qualitative data, that is compounds labeled active or inactive, the 'receiver operating characteristic' (ROC) curve and its area under the curve (AUC) are applied. The AUC gives the probability of correctly ranking a randomly chosen pair of an active and an inactive compound (Box 1). ROC curves are similar to enrichment curves, but in contrast to them, they have a sound statistical foundation and are insensitive to the number of active and inactive molecules [18].

Extended scoring functions

Most commonly, targeted scoring functions are based on established scoring functions that are either extended, recalibrated, or accompanied by special filters. AutoShim [19,20], for example, does not modify the scoring function directly, but instead adds new pharmacophore points, so-called 'shims', to the active site of the docking target. The ligands interact with these pharmacophores which modify their score. The weighting of the individual pharmacophore points is calibrated by partial least squares (PLS) regression to experimental affinities. An iterative procedure is applied to allow for binding mode changes during the calibration process. Predictive $R^2{}_s$ in the range of 0.21–0.57 were achieved [19]. For five out of seven kinase targets AutoShim performed better than the best generic scoring function cited in [14] ($R^2 = 0.32$). The generic AIScore [21] adds new hydrogen-bonding terms from *ab initio* quantum chemical calculations to an established scoring function. Trained on 101 protein–ligand complexes ($R^2 = 0.72$–0.76) it achieved a $R^2 = 0.21$ for 799 complexes from the PDBbind, in comparison to only $R^2 = 0.03$ for the original scoring function. Ligand-specific scoring functions focused on carbohydrates [22] and peptides ($R^2 = 0.80$–0.86) [23] have been developed using 2D quantitative structure–activity relationship (QSAR) methods. 3D QSAR models were also used to score protein–ligand complexes and achieved an internally validated correlation coefficient of $R^2 = 0.61$ for 79 thrombin inhibitors [24]. The addition of a term for the buriedness of receptor interactions to a generic scoring function improved the ranking of binding modes [25].

Recalibration using binary data

Using binary data for scoring function optimization has recently gained particular popularity. Initially this was proposed by Smith *et al.* [26] for optimizing the binding mode prediction of ligands in contrast to noise compounds (decoys). Provided that the binning procedure avoids ambiguous class assignments, binary data have several favorable properties: firstly, the classification of binding affinity into active and inactive reduces the impact of data inconsistencies caused by experimental factors. Secondly, it allows for incorporating data about inactive molecules, where no experimental binding affinities can be determined. For example, Pham and Jain [27,28] optimized parameters of a scoring function by sampling parameter space and observing the changes in the mean squared error (MSE) of pK_i prediction. They tweaked the standard MSE function to take into account decoys, for which no pK_i values exist, and added terms for scoring function constraints. The AUC clearly increased for two targets (polyA-ribose polymerase AUC = 0.89 → 0.99, HIV protease AUC = 0.91 → 0.96) and remained stable for most of the other targets. Notably, the default scoring function already provided a remarkable discrimination of actives and inactives. The assessment of statistical significance of the observed changes is difficult, however, because no values for the correlation of default and optimized scores are given [29]. The target-specific optimization (TOP) approach [30] focuses completely on ligand–decoy (LD) discrimination. The TOP objective

TABLE 1

Overview of recent reports on optimized and/or targeted scoring functions

Approach	Type	Method	Score	Target(s)	Objective and algorithm	Refs
AutoShim	T	Add	PP	Kinase	IC_{50} prediction optimized by iterative partial least squares regression	[19,20]
AIScore	G	Add	QSAR	Various	ΔG prediction optimized by iterative grid search with continuous downhill refinement	[21]
BALLDock/Slick	T	Add	QSAR	Carbohydrates	ΔG prediction optimized by multiple linear regression	[22]
Hetenyi *et al.*	T	Add	QSAR	Peptide ligands	ΔG prediction optimized by multiple linear regression	[23]
AFMoC^{con}	T	Adjust	QSAR	Thrombin	pK_i prediction optimized by partial least squares regression	[24]
O'Boyle *et al.*	G	Add	QSAR	Various	Average rank of ligand poses versus decoy poses optimized by brute-force algorithm	[25]
Smith *et al.*	G	Adjust	QSAR	Various	Average rank of native ligand pose versus decoy poses optimized by genetic algorithm	[26]
Pham *et al.*	T, G	Adjust	QSAR	Various	pK_d prediction (incl. ligands and decoys) optimized by random walk and line search	[27,28]
TOP	T	Adjust	QSAR	CDK2, ERα, COX2	Discrimination of ligand and decoy scores optimized by iterative taboo search	[30]
POEM	T	Adjust	QSAR	Kinase, ATPase	Binding mode prediction improved by iterative learning (neural networks) and optimization (genetic algorithm)	[32]
Andersson *et al.*	G	Adjust	Various	Various	Binding mode prediction optimized by partial least squares regression	[33]
SSM	T	Adjust	QSAR	Various	Binding mode (and ΔG) prediction optimized by Random Forest model	[34]
ITScore	G	Adjust	PMF	Various	Mean success rate of correct docking optimized by iterative Boltzmann-weighted averaging algorithm	[35,36]
DrugScore^{RNA}	T	Adjust	PMF	RNA	Binding mode prediction optimized by inverse Boltzmann approach	[37]
FLAP	T	Filter	FP	Kinase, TK, ERα, FXa	Matching of ligands and proteins evaluated by comparison of 3D pharmacophore fingerprints	[38]
IFP	T	Filter	FP	ADRB2	(Ant)agonist prediction by Tanimoto comparison of interaction fingerprints	[39]
IFS	T	Filter	FP	mGluR5	Antagonist prediction by Tanimoto comparison of interaction fingerprints	[40]
Kumar *et al.*	T	Filter	FP	$TMPK_{mt}$	Inhibitor identification by cluster analysis of interaction fingerprints	[41]
Gozalbes *et al.*	T	Filter	Various	Kinase	Score thresholds are learned from docking known ligands	[42]
TS-VS	T	Filter	Various	ERα	Reduction of conformational artifacts by a four-step scoring and filtering procedure	[43]

function is based on the analysis of variance (ANOVA), which is applied in classical statistics to assess the significance of differences between samples from two or more groups. Three different scoring functions for cyclin-dependent kinase 2 (CDK2), estrogen receptor α (ERα) and cyclooxygenase 2 (COX2) were trained on small samples from the respective 'directory of useful decoys' (DUD) LD datasets [31]. This resulted in statistically significant and practically relevant improvements in LD discrimination for CDK2 and COX2, and – for all targets – in ligand-random (LR) molecule discrimination. The strongest increase occurred for CDK2: $AUC_{LD} = 0.67 \rightarrow 0.83$ ($P < 0.0001$) and $AUC_{LR} = 0.60 \rightarrow 0.89$ ($P < 0.0001$). The true positive rate at 5% false positive rate ($TPR_{5\%}$) was enhanced as well: $TPR_{5\%,LD} = 0.21 \rightarrow 0.36$ ($P < 0.01$), and $TPR_{5\%,LR} = 0.09 \rightarrow 0.51$ ($P < 0.0001$). Considerable improvements were also found in an external validation using a large CDK2 dataset of 17,550 molecules. In general, it was observed that LR molecule discrimination profits more from the optimization than LD discrimination. This stems from the fact that random molecules typically differ much more strongly from the ligands than the property-matched decoys.

Optimization by statistical models

The approach of parameter optimization utilized by TOP and similar methods resembles what is known as 'design of experiments' in statistics. In fact, such methods have already been used for the optimization of protein–ligand docking software. For example, 'parameter optimization using ensemble methods' (POEM) has been proposed by Antes *et al.* [32] to find optimal parameters for two established scoring functions on kinases and ATPases. POEM is based on approximating the response to parameter changes by a statistical model, which is subsequently used to locate new, promising parameter combinations. In a similar fashion, Andersson *et al.* [33] used a PLS approach to refine the parameters of scoring functions for binding mode reproduction. The 'supervised scoring model' of Teramoto and Fukunishi [34] follows a slightly different approach: a rough linear correlation between binding free energy and root mean square deviation (RMSD) to the native ligand is utilized to improve predictions of binding affinity. This correlation was weak for the original scoring function, but was considerably improved by a Random Forest model (Spearman correlation coefficient $R^2_s = 0.69–0.90$) that was trained to predict binding mode RMSD from the original scoring function terms. Ranking ligands and decoys from the DUD dataset resulted in improved enrichments compared to the original scoring function.

Optimized potentials of mean force

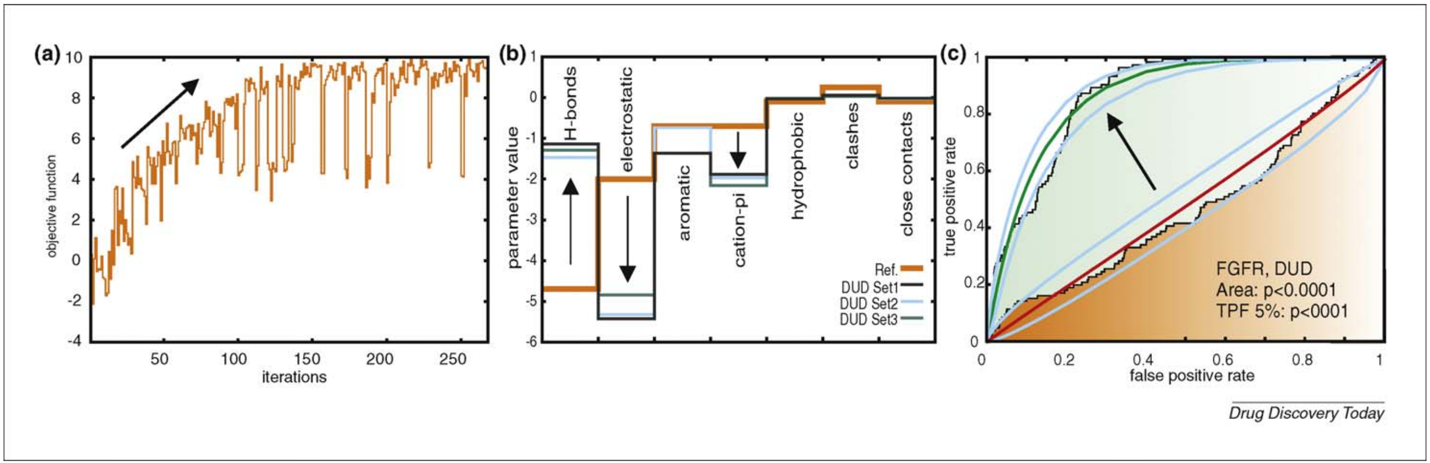
Another class of optimized scoring functions is represented by the generic ITScore [35,36] and the target-specific DrugScore^{RNA} [37]. Both are 'potentials of mean force', but the derivation is completely different: the distance-dependent DrugScore^{RNA} potentials were derived from crystal structures containing RNA using established methods and were able to predict binding affinities of 15 RNA–ligand complexes with a Spearman correlation coefficient of $R^2_s = 0.37$. ITScore, in contrast, applies an iterative strategy to improve the pair potentials until they correctly discriminate between experimental binding modes and decoy ligand poses. On a test set of 140 protein–ligand complexes a correlation coefficient for affinity prediction of $R^2 = 0.55$ was achieved.

Fingerprint scoring functions

Fingerprint-based scoring is another efficient method for target-specific virtual screening. For example, Baroni *et al.* [38] developed FLAP, which evaluates the complementarity of proteins and ligands by comparing 3D pharmacophore fingerprints. This method outperformed standard methods for ligand- and structure-design. de Graaf and Rognan [39] used a topological scoring function based on molecular interaction fingerprints (IFPs) for virtual screening of agonists of the β2 adrenergic receptor. Various combinations of established scoring functions with IFP were evaluated and the best results for agonists and antagonists were around AUC = 0.991 and 0.995, respectively. These values are based on a database of 13 known antagonists/inverse agonists, 13 known partial/full agonists and 980 drug-like, physico-chemically similar compounds randomly selected from commercial suppliers. Interestingly, conventional extended connectivity fingerprints for chemical similarity were as effective as the IFP approach: AUC = 0.998 and 0.990 for agonists and antagonists, respectively. A similar approach called 'interaction fingerprint scoring' (IFS) was evaluated by Radestock *et al.* [40]. They were able to demonstrate significantly higher enrichment rates of antagonists of metabotropic glutamate receptor 5 when using IFS compared to five established scoring functions. Another application of IFPs has been reported recently by Kumar *et al.* who were able to identify inhibitors of thymidine monophosphate kinase from *M. tuberculosis* [41].

Target-specific filtering

Target-specific virtual screening can be realized by appropriate filtering procedures as well. Gozalbes *et al.*, for example, learned score thresholds for 6 scoring functions from docking 123 kinase inhibitors into 3 kinase structures [42]. Virtual screening of a kinase-targeted library (1440 compounds) and a generic collection of drugs and drug-like compounds (2500 compounds) showed that compounds that pass all thresholds have a higher chance for being kinase inhibitors. The target-specific virtual screening (TS-VS) developed by Knox *et al.* used atom distance and hydrogen-bonding constraints to improve docking results [43]. TS-VS identified three novel ligands of estrogen receptor α by virtual screening of



Drug Discovery Today

FIGURE 1

Convergence, reproducibility and significance. The workflow of validating an approach to targeted scoring function for these criteria is illustrated using as an example the validation of D-TOP which is the multi-factorial version of TOP. **(a)** The global optimization algorithm [54,55] of D-TOP quickly converges to a stable set of parameters with an increased value for the objective function. **(b)** Like TOP, D-TOP uses a small sample of the training data for the optimization procedure. Here, the training data consisted of small samples (400 compounds in total) from DUD datasets of p38α, Src, EGFR and CDK2. Different samples lead only to minor effects on the final parameters of the scoring function. **(c)** After successful internal validation (31,756 compounds in total) an external validation was performed using DUD data (7055 compounds in total) for FGFR1 and VEGFR2. Significant improvements of the area under the ROC curve and the true positive fraction at 5% false positive rate ($TPF_{5\%}$) were found for FGFR1 (AUC = 0.46 → 0.86) and VEGFR2 (AUC = 0.66 → 0.77, data not shown). The ROC curves with initial (orange) and optimized (green) parameters and the fitted ROC curves (red and dark green) with their 95% confidence intervals (cyan) are depicted. These confidence intervals result from binormal models that take into account the correlation of the scores [56]. Because the same molecules are used during docking with initial and optimized parameters, it is important to consider this correlation for assessing the significance of the differences in those ROC curves.

~200,000 compounds. One of them had an IC_{50} of 53 nM, was selective against estrogen receptor β, and exhibited antiproliferative activity in the micromolar range.

Future directions

Evaluation of methods

As can be seen from the discussion above, targeted scoring functions exhibit better predictivities on average than their generic counterparts. It is, however, difficult to assess which approach is the most promising. There is still no gold standard for evaluating scoring functions, although this problem has become a subject of much discussion [44]. For example, a comparison of the AUC values achieved by different methods is tempting, but not necessarily valid. Some scoring functions [27,28,39] were trained and evaluated on backgrounds of various collections of random molecules, each of which contained approximately 1000 molecules. A

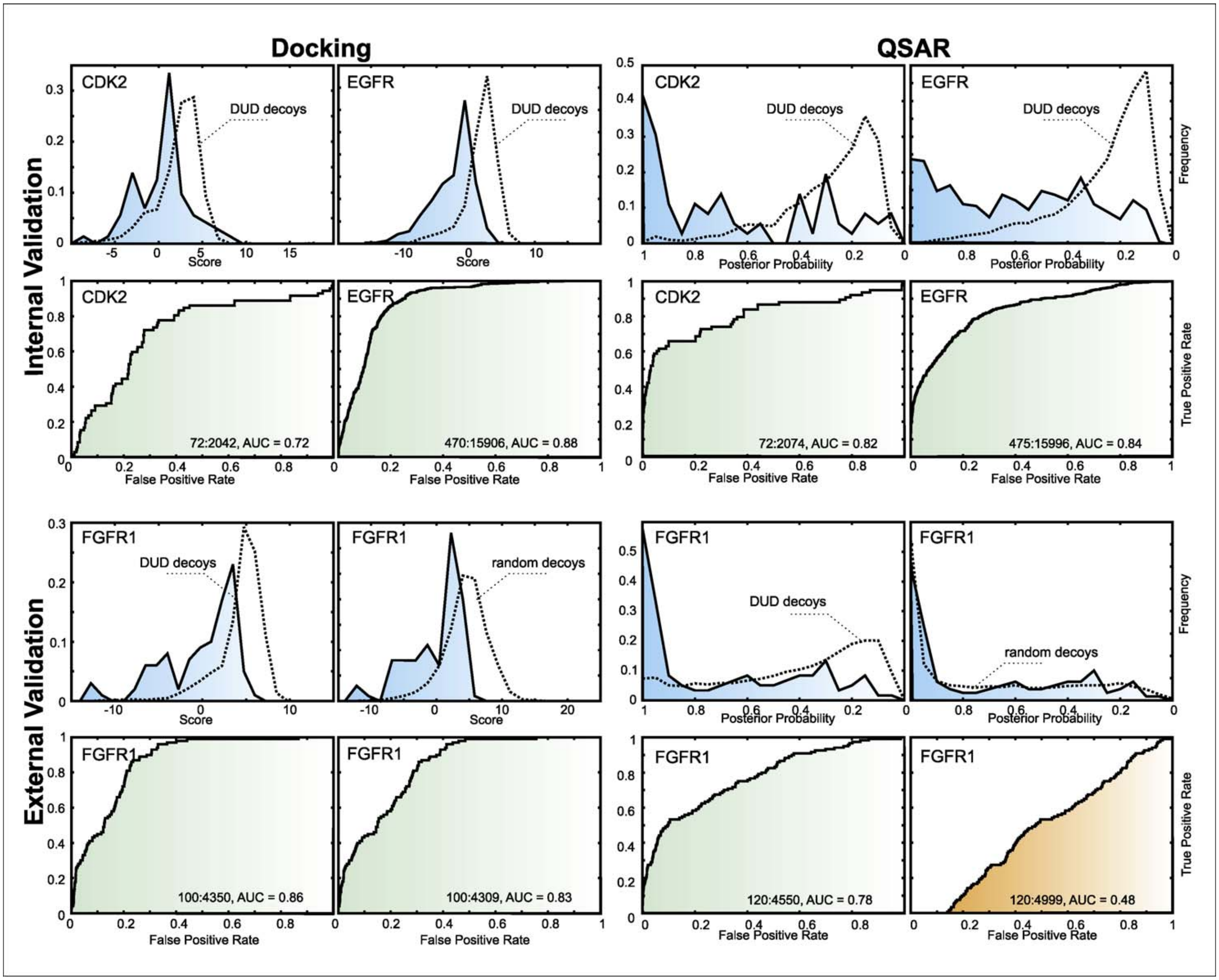


FIGURE 2

Comparison of a targeted scoring function with a binary QSAR model for kinase inhibitor scoring. D-TOP was used to optimize a scoring function for the docking of kinase inhibitors. Using small samples from the DUD datasets for CDK2, EGFR, p38α (data not shown) and Src (data not shown) as training data, new scoring function parameters were identified that lead to improved predictivity for ligands and decoys on the full DUD datasets for all four targets (top panel, left: score distributions – ligands in blue, decoys as dotted line – and ROC curves – filled with green – are shown). Additionally, they perform well in an external validation using the DUD data for FGFR1 (lower panel, left) and VEGFR2 (data not shown). Replacing the DUD decoys with random drug-like compounds from commercial suppliers does not have a significant impact on the performance of that scoring function. For comparison, a binary QSAR model was built using the exactly the same training and test dataset. The binary QSAR method first transforms the descriptor vectors into orthogonal eigenvectors, which serve as input for a naïve Bayes classifier [50,57]. For the sake of comparability, only seven independent eigenvectors were used for creating the model, similar to the seven free parameters of the docking scoring function. The ROC curves show that the model is highly predictive for the full dataset of ligands and decoys (internal validation, top panel, right). Subsequently, the same external validation was performed as for the docking method (lower panel, right). Obviously, DUD ligands can be discriminated from DUD decoys for both targets, FGFR1 (lower panel, right) and VEGFR2 (data not shown). DUD ligands, however, cannot be discriminated from random decoys at all using this model (lower panel, right, orange ROC curve). This indicates a strong dependence of the QSAR model's predictivity on dataset structure, an effect that is independent of the QSAR model complexity and target protein. Evidently, this stands in contrast to the results of the D-TOP optimized scoring functions, which do not suffer from this disadvantage.

relatively small number of molecules, however, leads to large estimation errors [45]. Although a random selection of molecules may represent a challenge for a scoring function because of its potentially large diversity, trivial property differences between active compounds and random decoys result in undeservedly good performance measures [45,46]. Therefore, it is advisable to evaluate methods on large, property-matched datasets, even if the new parameters are learned from small samples of training data. TOP [30], for example, used 50–100 actives and 100 inactives from the DUD LD dataset for training, but the test datasets contained up to 11,582 inactive molecules. This results in more conservative AUC values. It is clear from the introductory discussion that a realistic estimate of the AUC and the corresponding true and false positive rates is important. Because of the enormous size of chemical space, a large number of inactives in the test dataset are necessary to simulate the effects of highly biased databases that are typically used for virtual screening. The DUD datasets [31], for example, have ratios of active to inactive molecules of 1:36, which makes them particularly interesting, in addition to their public availability. Additionally augmenting the DUD datasets with a sufficient number of random molecules allows for getting a realistic and meaningful impression of virtual screening performance.

Scoring functions for target classes

Scoring functions for target classes provide a possibility to increase efficiency as targeting not only single proteins, but instead a full class of proteins brings a significant saving in computational costs. Particularly, targeting classes of proteins provides a suitable compromise between larger applicability domains compared to single target scoring functions and a better performance in comparison to generic scoring functions. Additionally, the diversity of the training data increases thereby improving robustness. For example, 'surrogate' AutoShim [20] uses the AutoShim [19] procedure in one crystal structure to reproduce affinity data for a different protein of the same target class. Another extension, 'ensemble surrogate' AutoShim refers to replacing the crystal structure with an ensemble of structure representing a universal receptor for the AutoShim procedure. Forty-one of 51 'ensemble surrogate' AutoShim models for kinases exceeded the $R^2 = 0.32$ criterion, which was the best performance for conventional docking [14].

A target-class approach, called D-TOP, was implemented within the framework of TOP [30] as well: the one-way ANOVA objective function of TOP was replaced in a straightforward manner with a function based on multi-factorial ANOVA [47] to maximize the discrimination of ligands and decoys for several targets of a target class and multiple protein conformations simultaneously. An initial evaluation study on protein kinases found significant and robust improvements for all kinases in internal and external validation (Seifert, unpublished). For example, the ranking of true ligands was significantly improved within the DUD datasets for FGFR1 (AUC = 0.46 → 0.86) and VEGFR2 (AUC = 0.66 → 0.77), although these targets were not part of the training data (Fig. 1). These results clearly underpin the potential of target-class approaches to scoring function design.

Technical issues

Convergence, reproducibility and significance are the further important parameters for the design and evaluation of targeted scoring functions (Fig. 1). Firstly, fast convergence of the optimization algorithm is important to limit the computational costs. Secondly, robustness with respect to different samples of training data is necessary to ensure reproducibility and reliability. Finally, the significance of the improvements in virtual screening performance has to be assessed in a rigorous manner. These technical issues can be addressed by methods that have been developed in computational engineering, operations research and statistics. For example, software packages are available that implement algorithms for parameter estimation, uncertainty quantification and sensitivity analysis (e.g. [48]). These software packages additionally include advanced optimization methods (e.g. [49]).

As mentioned above, computational costs play an important role. The complexity of protein–ligand docking algorithms exceeds by far that of conventional machine learning methods. In return for this complexity, protein–ligand docking has to provide additional advantages. For example, Fig. 2 shows a comparison of a kinase-targeted D-TOP-scoring function for protein–ligand docking with an established machine learning method, that is binary QSAR [50]. Obviously, on the same dataset and with same degrees of freedom, the QSAR approach gives better or at least comparable results for the internal validation when the dataset composition, DUD ligands and DUD decoys, is similar to the training data. However, when confronted with a different dataset structure, DUD ligands and random decoy molecules, the binary QSAR approach fails, in contrast to the targeted scoring function. Presumably, this interesting phenomenon is not unique to D-TOP and binary QSAR, and certainly justifies more research on the multiple prospects of targeted scoring functions for protein–ligand docking and virtual screening.

Conclusions

Targeted scoring functions are able to improve the predictivity of virtual screening, thereby lowering false positive rates and increasing the probability for identifying true hits. Thus efficiency is increased and risk of failure is reduced. The ultimate benefit of targeted scoring functions for drug discovery depends not only on the false positive rate itself, but also on the costs caused by false positives and false negatives [51]. Taking into account these costs, for example for postprocessing, secondary assays and so on, the determination of the required optimal statistical properties, that is true and false positive rate, of a scoring function used for a particular application is possible. In contrast to generic scoring functions, targeted scoring functions enable to meet these requirements more easily as the size of applicability domain can be traded off for a higher predictivity [52,53]. Therefore, targeted scoring functions are a particularly promising method for improving virtual screening.

Acknowledgements

The author is indebted to Daniel Vitt and Bernd Kramer for unwavering support, and to all colleagues in the Chem- and Bioinformatics Department for the excellent working climate. In particular, Thomas Herz and Ruth Seifert are acknowledged for their valuable comments on the manuscript.

References

1 Kontoyianni, M. *et al.* (2008) Theoretical and practical considerations in virtual screening: a beaten field? *Curr. Med. Chem.* 15, 107–116

2 Sperandio, O. *et al.* (2006) Receptor-based computational screening of compound databases: the main docking-scoring engines. *Curr. Protein Pept. Sci.* 7, 369–393

3 Liu, T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Res.* 35, 198–201

4 Wang, R. *et al.* (2004) The PDBbind database: collection of binding affinities for protein–ligand complexes with known three-dimensional structures. *J. Med. Chem.* 47, 2977–2980

5 Wishart, D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* 34, D668–D672

6 PubChem, National Center for Biotechnology Information (NCBI), National Library of Medicine, USA

7 Moitessier, N. *et al.* (2008) Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* 153 (Suppl. 1), S7–S26

8 Yin, S. *et al.* (2008) MedusaScore: an accurate force field-based scoring function for virtual drug screening. *J. Chem. Inf. Model.* 48, 1656–1662

9 Zhao, X. *et al.* (2008) An improved PMF scoring function for universally predicting the interactions of a ligand with protein, DNA, and RNA. *J. Chem. Inf. Model.* 48, 1438–1447

10 Feher, M. (2006) Consensus scoring for protein–ligand interactions. *Drug Discov. Today* 11, 421–428

11 Orry, A.J. *et al.* (2006) Structure-based development of target-specific compound libraries. *Drug Discov. Today* 11, 261–266

12 Sotriffer, C.A. *et al.* (2008) SFCscore: scoring functions for affinity prediction of protein–ligand complexes. *Proteins* 73, 395–419

13 Wang, R. *et al.* (2004) An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein–ligand complexes. *J. Chem. Inf. Comput. Sci.* 44, 2114–2125

14 Warren, G.L. *et al.* (2006) A critical assessment of docking programs and scoring functions. *J. Med. Chem.* 49, 5912–5931

15 Seifert, M.H. and Lang, M. (2008) Essential factors for successful virtual screening. *Mini Rev. Med. Chem.* 8, 63–72

16 Hastie, T. *et al.* (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer Science + Business Media p. 185

17 Dietterich, T.G. *et al.* (1997) Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* 89, 31–71

18 Triballeau, N. *et al.* (2005) Virtual screening workflow development guided by the ''receiver operating characteristic'' curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J. Med. Chem.* 48, 2534–2547

19 Martin, E.J. and Sullivan, D.C. (2008) AutoShim: empirically corrected scoring functions for quantitative docking with a crystal structure and IC50 training data. *J. Chem. Inf. Model.* 48, 861–872

20 Martin, E.J. and Sullivan, D.C. (2008) Surrogate AutoShim: predocking into a universal ensemble kinase receptor for three dimensional activity prediction, very quickly, without a crystal structure. *J. Chem. Inf. Model.* 48, 873–881

21 Raub, S. *et al.* (2008) AIScore: chemically diverse empirical scoring function employing quantum chemical binding energies of hydrogen-bonded complexes. *J. Chem. Inf. Model.* 48, 1492–1510

22 Kerzmann, A. *et al.* (2008) BALLDock/SLICK: a new method for protein–carbohydrate docking. *J. Chem. Inf. Model.* 48, 1616–1625

23 Hetényi, C. *et al.* (2006) Combination of a modified scoring function with two-dimensional descriptors for calculation of binding affinities of bulky, flexible ligands to proteins. *J. Am. Chem. Soc.* 128, 1233–1239

24 Breu, B. *et al.* (2007) Consensus adaptation of fields for molecular comparison (AFMoC) models incorporate ligand and receptor conformational variability into tailor-made scoring functions. *J. Chem. Inf. Model.* 47, 2383–2400

25 O'Boyle, N.M. *et al.* (2008) Using buriedness to improve discrimination between actives and inactives in docking. *J. Chem. Inf. Model.* 48, 1269–1278

26 Smith, R. *et al.* (2003) Analysis and optimization of structure-based virtual screening protocols. New methods and old problems in scoring function design. *J. Mol. Graph. Model.* 22, 41–53

27 Pham, T.A. and Jain, A.N. (2008) Customizing scoring functions for docking. *J. Comput. Aided Mol. Des.* 22, 269–286

28 Pham, T.A. and Jain, A.N. (2006) Parameter estimation for scoring protein–ligand interactions using negative training data. *J. Med. Chem.* 49, 5856–5868

29 Hanley, J.A. and McNeil, B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148, 839–843

30 Seifert, M.H. (2008) Optimizing the signal-to-noise ratio of scoring functions for protein–ligand docking. *J. Chem. Inf. Model.* 48, 602–612

31 Huang, N. *et al.* (2006) Benchmarking sets for molecular docking. *J. Med. Chem.* 49, 6789–6801

32 Antes, I. *et al.* (2005) POEM: parameter optimization using ensemble methods: application to target specific scoring functions. *J. Chem. Inf. Model.* 45, 1291–1302

33 Andersson, C.D. (2007) A multivariate approach to investigate docking parameters' effects on docking performance. *J. Chem. Inf. Model.* 47, 1673–1687

34 Teramoto, R. and Fukunishi, H. (2007) Supervised scoring models with docked ligand conformations for structure-based virtual screening. *J. Chem. Inf. Model.* 47, 1858–1867

35 Huang, S.Y. and Zou, X. (2006) An iterative knowledge-based scoring function to predict protein–ligand interactions: I. Derivation of interaction potentials. *J. Comput. Chem.* 27, 1866–1875

36 Huang, S.Y. and Zou, X. (2006) An iterative knowledge-based scoring function to predict protein–ligand interactions: II. Validation of the scoring function. *J. Comput. Chem.* 27, 1876–1882

37 Pfeffer, P. and Gohlke, H. (2007) DrugScore^{RNA} – knowledge-based scoring function to predict RNA–ligand interactions. *J. Chem. Inf. Model.* 47, 1868–1876

38 Baroni, M. *et al.* (2007) A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J. Chem. Inf. Model.* 47, 279–294

39 de Graaf, C. and Rognan, D. (2008) Selective structure-based virtual screening for full and partial agonists of the beta2 adrenergic receptor. *J. Med. Chem.* 51, 4978–4985

40 Radestock, S. *et al.* (2008) Homology model-based virtual screening for GPCR ligands using docking and target-biased scoring. *J. Chem. Inf. Model.* 48, 1104–1117

41 Kumar, A. *et al.* (2009) Knowledge based identification of potent antitubercular compounds using structure based virtual screening and structure interaction fingerprints. *J. Chem. Inf. Model.* 49, 35–42

42 Gozalbes, R. *et al.* (2008) Development and experimental validation of a docking strategy for the generation of kinase-targeted libraries. *J. Med. Chem.* 51, 3124–3132

43 Knox, A.J. *et al.* (2007) Target specific virtual screening: optimization of an estrogen receptor screening platform. *J. Med. Chem.* 50, 5301–5310

44 Stouch, T. (2008) Special issue on evaluation of computational methods. *J. Comput. Aided Mol. Des.* 22, 131 (and all articles therein)

45 Hawkins, P.C. *et al.* (2008) How to do an evaluation: pitfalls and traps. *J. Comput. Aided Mol. Des.* 22, 179–190

46 Verdonk, M.L. *et al.* (2004) Virtual screening using protein–ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* 44, 793–806

47 Bortz, J. (2005) *Statistik* (6th edn), Springer Medizin Verlag pp. 247–288

48 Eldred, M.S. *et al.* (2006) DAKOTA, a multilevel parallel object-oriented framework for design optimization, parameter estimation, uncertainty quantification, and sensitivity analysis: Version 4.0 Developers Manual, *Sandia Technical Report SAND2006-4056*. URL http://www.cs.sandia.gov/DAKOTA/index.html (accessed Sep 18, 2008)

49 Jones, D.R. *et al.* (2004) Efficient global optimization of expensive black-box functions. *J. Global Optim.* 13, 455–492

50 Labute, P. (1999) Binary QSAR: a new method for the determination of quantitative structure activity relationships. *Pac. Symp. Biocomput.* 444–455

51 Nicholls, A. (2008) What do we know and when do we know it? *J. Comput. Aided Mol. Des.* 22, 239–255

52 Wolpert, D.H. and Macready, W.G. (1997) No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* 1, 67

53 Schaffer, C. (1994) A conservation law for generalization performance. In *Proceedings of the 11th International Conference on Machine Learning* (Cohen, W.W. and Hirsh, H., eds), In pp. 259–265, Morgan Kaufmann

54 Jones, D.R. *et al.* (1993) Lipschitzian optimization without the lipschitz constant. *J. Optim. Theory Appl.* 79, 157–181

55 Gablonsky, J. and Kelley, C. (2001) A locally-biased form of the DIRECT algorithm. *J. Global Optim.* 21, 27–37

56 ROCKIT, Version 1.1b (2007) Kurt Rossmann Laboratories for Radiological Image Research, University of Chicago. http://www-radiology.uchicago.edu/krl/roc_soft6.html (accessed Sep 18, 2009)

57 Xia, X. *et al.* (2004) Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* 47, 4463–4470