J Comput Aided Mol Des (2009) 23:289–299
DOI 10.1007/s10822-008-9258-8

Scoring confidence index: statistical evaluation of ligand binding mode predictions

Maria I. Zavodszky · Andrew W. Stumpff-Kane · David J. Lee · Michael Feig

Received: 22 August 2008 / Accepted: 29 December 2008 / Published online: 20 January 2009

Abstract Protein-ligand docking programs can generate a large number of possible binding orientations for each ligand candidate. The challenge is to identify the orientations closest to the native binding mode using a scoring method. Many different scoring functions have been developed for protein-ligand scoring, but their performance on binding mode prediction is often target-dependent. In this study, a statistical approach was employed to provide a confidence measure of scoring performance in finding close to the correct docked ligand orientations. It exploits the fact that the scores provided by an adequately performing scoring function generally improve as the ligand binding modes get closer to the correct native orientation. For such cases, the correlation coefficient of scores versus distances is expected to be highest when the most native-like orientation is used as a reference. This correlation coefficient, called the correlation-based score (CBScore), was used as an indicator of how far the docked pose was from the native orientation. The correlation between the original scores and CBScores as well as the range of CBScores were found to be good measures of scoring performance. They were combined into a single quantity, called the scoring confidence index. High values of the scoring confidence index were indicative of pronounced and relatively smooth binding energy landscapes with easily discernable global minima, resulting in reliable binding mode predictions. Low values of this index reflected rugged energy landscapes making the prediction of the correct binding mode very difficult and often unreliable. The diagnostic ability of the scoring confidence index was tested on a non-redundant set of 50 protein-ligand complexes scored with three commonly employed scoring functions: AffiScore, DrugScore and X-Score. Binding mode predictions were found to be three times more reliable for complexes with scoring confidence indices in the upper half than for cases with values in the lower half of the resulting range of 0–1.6. This new confidence measure of scoring performance is expected to be a valuable tool for virtual screening applications.

Keywords Binding orientation · Correlation-based score · Energy landscape · Protein-ligand docking · Scoring function

Abbreviations

CBScore	Correlation-based score
PDB	Protein Data Bank
RMSD	Root-mean-square deviation
PSR	Pseudo-RMSD
SCI	Scoring confidence index

Electronic supplementary material The online version of this article (doi:10.1007/s10822-008-9258-8) contains supplementary material, which is available to authorized users.

M. I. Zavodszky (✉) · A. W. Stumpff-Kane · M. Feig
Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824-1319, USA
e-mail: zavodszk@msu.edu

M. I. Zavodszky · M. Feig
Quantitative Biology Initiative, Michigan State University, East Lansing, MI 48824, USA

D. J. Lee
Lyman Briggs College, Michigan State University, East Lansing, MI 48824, USA

M. Feig
Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA

Introduction

Structure-based computational methods are widely used in the screening of large databases of small organic compounds for the identification of potent inhibitors that can be developed into new drugs [1–6]. The successful selection of actual inhibitors typically requires both the prediction of the correct ligand binding mode and the accurate estimation of relative binding affinities based on that conformation. The key challenges in finding the correct binding mode for a given ligand revolve around efficient conformational sampling and scoring. During sampling, a large number of possible ligand conformations are generated in the protein binding site, at least some of which have to be sufficiently close (<2.0 Å) to the correct binding mode for the docking protocol to be considered successful. A suitable scoring function then has to be able to select the correct binding conformation from a possibly large number of correct and incorrect decoys as the top-scoring configuration. After screening the entire database, the best scoring pose for each docked molecule along with its estimated binding affinity is used to generate a ranked hit list. Top scoring inhibitor candidates are then selected for experimental testing and further structure-based refinement.

Commonly, the same scoring function is used both to identify the best binding orientation of a ligand and to predict its binding affinity. The two tasks are somewhat different, though, and good performance of a given scoring function with ranking ligand poses does not guarantee similarly good performance when estimating binding affinities [6–8]. Numerous studies have compared the performance of docking protocols using one or multiple scoring functions on many different targets [5–7, 9]. Most of these studies are retrospective in nature with the testing done on known complexes. Conclusions as to which scoring functions and/or docking protocols perform best are implicitly assumed to be transferable to new, untested cases, even though it is well known that scoring performance is often target dependent [3, 6, 7, 9–12]. Therefore, a method that can predict the reliability of a given scoring function in finding the correct binding pose for a particular target *without knowledge of the correct answer* would be of great value. Such a scoring function assessment would provide a confidence estimate for docking results and could allow the use of alternate scoring functions and/or additional sampling in cases where an initially chosen scoring method did not perform reliably on a given set of docked ligand orientations.

Here, a statistical approach for providing such a scoring function assessment is introduced extending a previously published correlation-based formalism for enhancing the performance of scoring methods in protein structure prediction [13]. The method is based on the assumption that a reliable scoring or energy function awards gradually better scores to the docked ligand poses as they approach the correct binding orientation [14]. For such cases, with the correct binding mode corresponding to the global minimum, there is a significant correlation between the scores and the distances of the docked poses from the correct binding mode. This also implies that if a docked conformation close to the correct binding mode is used as the reference instead of the correct pose, the correlation between score and distance from that reference remains similarly high. This idea can be exploited by calculating a new score based on the correlation coefficients of score versus distance from each of the docked poses. The result is an enhanced correlation between the new score and the distance from the correct conformation. The improved correlation in turn was shown to be successful in aiding in the identification of near-native decoys in protein structure prediction applications [13]. In scoring protein-ligand complexes, the correlation of score versus distance also improved with the correlation-based score but we had limited success in actually improving the selection of good binding poses over simply applying the original scoring function. However, further analysis revealed ways to detect whether a given scoring function was likely to find correctly docked conformations from the top-scoring decoys and to attach a confidence value to docking predictions. The resulting scoring confidence index did not improve the selection of good binding poses for any one scoring function but allowed the identification of cases for which an alternative scoring functions and/or the generation of additional samples could improve the identification of correct binding poses. The new methodology was tested with three different scoring functions on a non-redundant set of 50 protein-ligand complexes. The results are described and discussed in more detail in the following after a short overview of the computational methodology.

Materials and methods

Dataset

About 50 protein-ligand complexes were selected from three different databases to ensure a wide range of systems with good resolutions: 20 structures were taken from the Ligand-Protein Database (http://lpdb.scripps.edu) [15], 12 from the training database of X-Score [16], and 18 structures were added from the Protein Data Bank [17]. All but four structures have a resolution of 2.5 Å or better (Table 1, Supplementary material). The dataset is non-redundant with each protein and ligand represented only once.

Docking and scoring

Ligands extracted from Protein Data Bank files were protonated and transformed into mol2 format using InsightII (Accelrys, Inc., San Diego, CA). To ensure thorough conformational sampling of each ligand, low-energy conformations were generated for each molecule using Omega version 2 (OpenEye Scientific Software, Santa Fe, NM). SLIDE [18–20] version 3.1 was used to dock the conformers into the binding sites of their corresponding target proteins from the complex structures. Protein side chains and ligands were treated flexibly during the docking. The number of docked ligand orientations obtained with this protocol was between 71 and 1,000, with an average of 622 dockings per complex. By default, each docked pose was scored with SLIDE's recently updated scoring function AffiScore, which is a weighted sum of hydrophobic contacts, hydrogen bonds, salt bridges, metal interactions, unsatisfied and repulsive polar interactions (Zavodszky MI, Tonero ME, He L, Arora S, Namilikonda S, and Kuhn LA, unpublished data). This was followed by rescoring all poses using the original implementation of DrugScore [21] and X-Score [16]. When target side-chains were rotated during ligand docking, the changed protein conformations were used when scoring the docked ligands with all three scoring functions.

To avoid the problem of artificially high RMSD values caused by automorphs, ligands having large chemical groups with chemically equivalent atoms in symmetry-related positions were excluded from the dataset. The remaining groups that could have caused such problems were rather small (phenyl groups or carboxylic acids) which were not expected to lead to very high RMSDs. To verify this, RMSD calculations were performed by taking into account the symmetry of such groups for one of the smallest ligands (1cbx) where this effect could be the largest. The maximum false increase in RMSD detected due to ignoring the symmetry of a COO^- group was 0.11 Å, while the same effect due to the phenyl group was 0.27 Å. For larger ligands this effect would be even smaller. In addition, the most suspicious outliers with high RMSDs and favorable scores were inspected using molecular graphics tools. No cases explainable by automorphs were found.

Correlation-based scoring enhancement

The correlation-based scoring enhancement method has been described previously [13]. It is summarized briefly in the following: taking one complex at a time, Pearson correlation coefficients are calculated between score and root mean square deviation (RMSD) for each docked pose used as a reference state. The resulting correlation coefficient is then assigned to the reference pose as the new correlation-based score (CBScore). The calculation of the Pearson correlation coefficient assumes a linear relationship between score and RMSD, yet empirical observation confirms a non-linear relationship (Fig. 1). A comparison of the correlation coefficients of scores versus RSMD with correlation coefficients of scores versus lnRMSD confirmed that scores correlated better with the logarithm of the RSMD (Fig. 2). Consequently, the original method was modified to calculate this score as the correlation between the scores and the natural logarithm of the RMSD in this study. The formula for calculating this modified CBScore for pose i is:

$$\mathrm{CBScore}_i = \frac{N\sum_{j\neq i}^{N} s_j \ln d_{ij} - \sum_{j\neq i}^{N} s_j \sum_{j\neq i}^{N} \ln d_{ij}}{\sqrt{\left(N\sum_{j\neq i}^{N} s_j^2 - \left(\sum_{j\neq i}^{N} s_j\right)^2\right)\left(N\sum_{j\neq i}^{N} \ln d_{ij}^2 - \left(\sum_{j\neq i}^{N} \ln d_{ij}\right)^2\right)}} \quad (1)$$

where d_{ij} is the RMSD between poses i and j, s_j is the original score of pose j, and N is the number of docked poses. It is worth noting that for a given pose i, the correlation-based score $\mathrm{CBScore}_i$ does not depend on the original score s_i of that pose.

Results

Application of CBScore as a scoring function for protein-ligand complexes

The CBScore was calculated for the non-redundant test set of 50 protein-ligand complexes scored with each of the three scoring functions as described in "Methods". As expected, the correlation-based method resulted in increased correlation between CBScore and RMSD compared to the correlation of the original score versus RMSD across all three scoring functions. The average correlation coefficients with the original score and CBScore were 0.46 and 0.62 for AffiScore, 0.51 and 0.60 for DrugScore, 0.57 and 0.65 for X-score, respectively. The correlation coefficients improved in 80% of the complexes for AffiScore (by an average of 0.23), in 64% of the complexes for DrugScore (by an average of 0.19) and in 68% of the complexes for X-Score (by an average of 0.18) (Fig. 1, Supplementary material). Test cases with no significant improvement in the correlations mostly involved targets where the initial correlation of the score versus RMSD (relative to X-ray structure) was low. In cases where the initial correlation between score and RMSD was more significant, the correlation-based method almost always improved the correlation coefficient by a significant amount. However, in

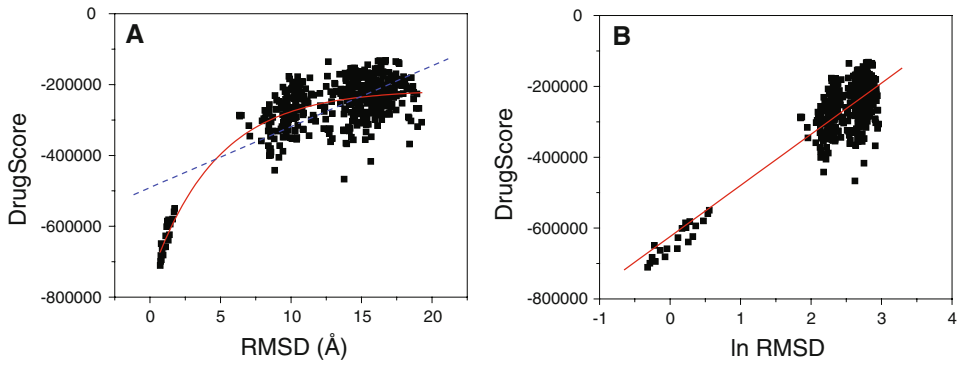

Fig. 1 An exponential function of the type $y = P1 + P2*(1-\exp(-x/P3))$ (*solid line*) provides a better fit to the DrugScore versus RMSD data than a linear function (*dashed line*) as illustrated for the complex 1eed (**a**). As a consequence, a linear correlation exists between the scores and the natural logarithm of the RMSD (**b**)

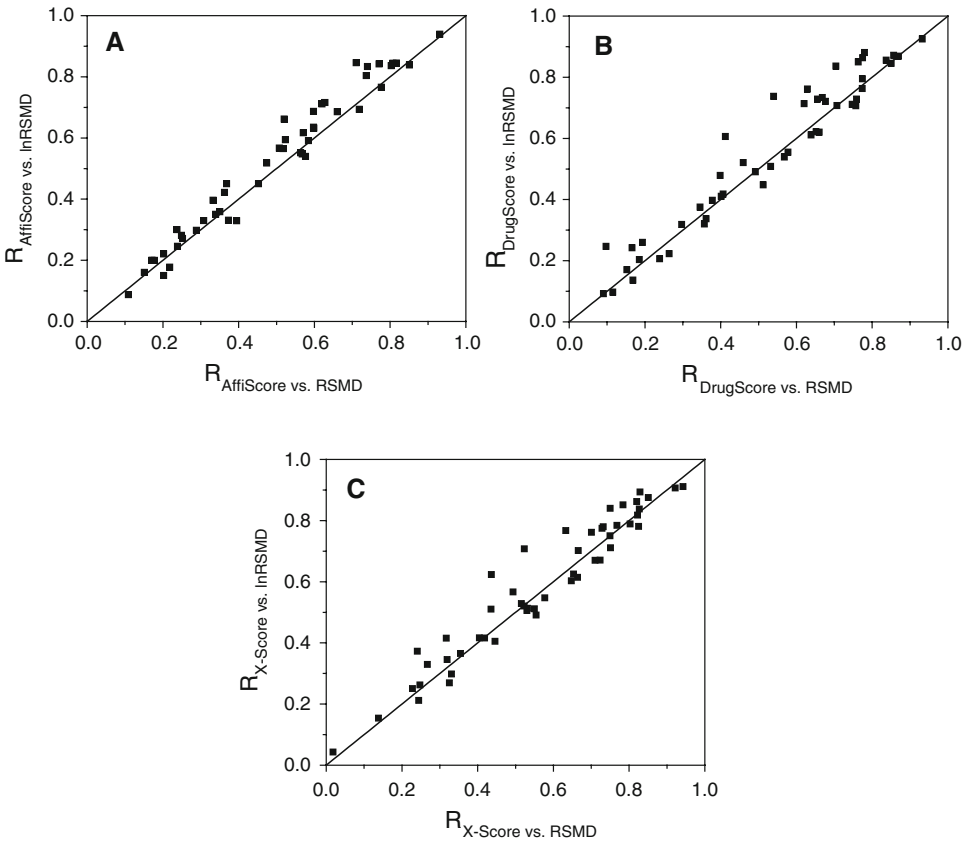

Fig. 2 Correlation coefficients between score and RSMD versus correlation coeffcients between score and lnRSMD for AffiScore (**a**), DrugScore (**b**), and X-Score (**c**). Data points above the *diagonal* represent cases in which scores correlate better with the logarithm of the RSMD than with RMSD

spite of increases in overall correlation between CBScores and RMSD in most cases, the selection of correct binding poses did not improve consistently when using the CBScore over the original score. In particular, the application of CBScore often resulted in the selection of worse conformations in cases where the original score already picked a very good binding pose as the top-scoring conformation (Fig. 2, Supplementary material).

Application of CBScore as pseudo-RMSD in the assessment of scoring functions

In most of the test cases, the CBScores were highly correlated with RMSD and thus were useful as an approximate metric for RMSD. Indeed, plots of the original score versus the derived CBScore, or more precisely versus the maximum CBScore for a given target minus the CBScores for individual poses, were strikingly similar to plots of the original scores versus RMSD (Fig. 3). Therefore, the correlation between the original score and CBScore could be used as an indicator of how the original score was correlated with the RMSD from the correct binding mode and, consequently, how reliable a given scoring function was in identifying a good binding pose.

The potential for using the correlation coefficient between score and the CBScore as a scoring assessment metric was examined with the 50 protein-ligand complexes and three scoring functions. Instead of taking the CBScore

directly, the logarithm of a new quantity called pseudo-RMSD (PSR) was used to obtain positive correlation coefficients and better capture the non-linear relationship between score and distance. The PSR of pose i of complex j was calculated as:

$$\mathrm{PSR}_{ij} = \mathrm{CBScore}_{\mathrm{MAX}_j} - \mathrm{CBScore}_{ij} + 0.01 \quad (2)$$

with $\mathrm{CBScore}_{\mathrm{MAX}_j}$ corresponding to the maximum correlation-based score for complex j and $\mathrm{CBScore}_{ij}$ corresponding to the correlation-based score of pose i in the case of complex j. The addition of 0.01 was necessary to maintain positive arguments of the logarithmic function. A high, near 0.8, correlation was found between the correlation coefficients of score versus lnRMSD and the correlation coefficients of score versus lnPSR for all three scoring functions for the set of complexes tested (Table 1).

Application of CBScore ranges in the assessment of scoring functions

Closer inspection of the CBScores revealed that its range (difference between maximum and minimum values) for each system was also a good indicator of how well the original score correlated with RMSD. A wider range of CBScores for a particular system corresponded to a better correlation between the original score and RMSD while a narrow range indicated poor correlation. Application to the 50 protein-ligand complexes resulted in correlations above 0.8 between the CBScore ranges and the correlation coefficients of score versus lnRMSD for all three scoring functions (Table 1).

Table 1 Correlation between correlation coefficients of score versus lnRMSD and correlation coefficients of score versus lnPSR, between correlation coefficients of score versus lnRMSD and the ranges of the CBScore, and between correlation coefficients of score versus lnRMSD and the scoring confidence index from 50 protein-ligand complexes based on AffiScore, DrugScore and X-score scoring functions

	$R_{\mathrm{AffiScore\ vs.\ RMSD}}$	$R_{\mathrm{DrugScore\ vs.\ RMSD}}$	$R_{\mathrm{X\text{-}Score\ vs.\ RMSD}}$
$R_{\mathrm{Score\ vs.\ lnPSR}}$	0.80	0.80	0.76
CBScore range	0.83	0.82	0.81
SCI	0.87	0.87	0.86

Application of the scoring confidence index in the assessment of scoring functions

Although the CBScore ranges and the PSR-based correlation coefficient provided similar information, they exhibited some degree of complementarity. Through empirical testing we found that the product of the CBScore range and the correlation between the original scores and lnPSR was a better scoring function quality assessment than each of the individual metrics alone. The product was named scoring confidence index (SCI), and was calculated as follows for complex j:

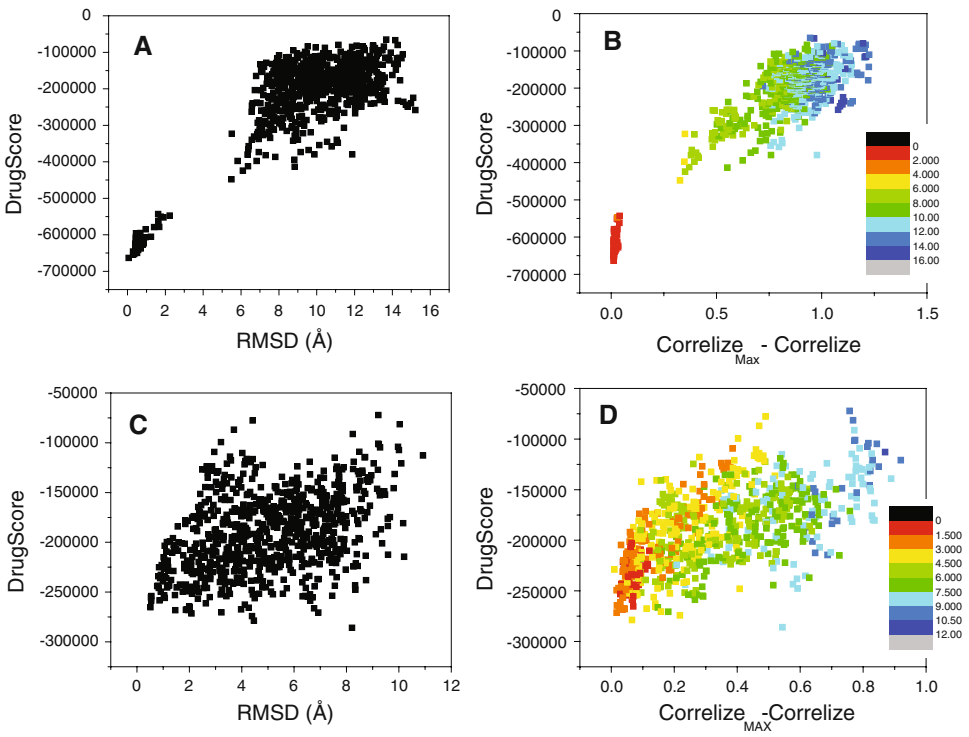

Fig. 3 DrugScore versus RMSD and DrugScore versus $\mathrm{CBScore}_{\mathrm{MAX}}-\mathrm{CBScore}$ for 1c2t (**a**, **b**) and 1bid (**c**, **d**). The data points from **b** and **d** are colored by RMSD. There is striking similarity between plots representing the score as a function of RMSD and score versus CBScore

$$\mathrm{SCI}_j = R_{Cj} \times \left(\mathrm{CBScore}_{\mathrm{MAX}_j} - \mathrm{CBScore}_{\mathrm{MIN}_j}\right) \quad (3)$$

where R_{Cj} is the correlation coefficient between score and lnPSR, $\mathrm{CBScore}_{\mathrm{MAX}_j}$ and $\mathrm{CBScore}_{\mathrm{MIN}_j}$ are the maximum and minimum values of the correlation-based scores for the protein-ligand complex j. The correlation coefficients between SCI and the correlation coefficients of score versus RMSD reached 0.86–0.87 for the three scoring functions (Table 1). Thus, it was expected that the SCI metric could predict the cases in which the original score was highly (or poorly) correlated with RMSD and therefore the scoring function was likely (or unlikely) to identify a good binding pose as the best-scoring conformation.

Figure 4 illustrates how the SCI performed for the individual structures in the test set. The use of SCI, which *does not* require knowledge of the correct native conformation, is contrasted with the use of the correlation coefficient between score and RMSD, which *does* require knowledge of the native state. It can be seen that both lead to a very similar correlation with the RMSD of the best prediction.

In the practical context of applying docking methods for ligand screening, a sufficiently accurate binding pose would typically correspond to an RMSD of less than 2.0 Å from a crystallographically observed conformation for a given protein-ligand complex. It is therefore of high practical relevance to predict whether a given scoring function is likely to identify a conformation below 2.0 Å as the best scoring prediction. The use of the correlation coefficient of score versus RMSD as a metric for such an assessment minimized false positives and false negatives when a cutoff of about 0.5 was chosen. This corresponded to a cutoff of 0.8 for SCI when considering its larger range. Indeed, it was found that complexes with SCI values above 0.8 had almost all of the top scoring poses close to the correct binding mode with an RMSD of less than 2.0 Å. (Fig. 4b, d, f). In contrast, complexes with SCI values below 0.8 had a large fraction of top-scoring poses far from the correct binding mode.

The accuracy of the confidence value provided by the SCI can be scrutinized further by analyzing the occurrence of true and false positive and negative predictions. SCI and RMSD values were each divided into two regimes: SCI below and above 0.8, and RMSD of the best-scoring docking conformation from the correct binding mode below and above 2.0 Å. In the plot of SCI versus RMSD, this gave rise to four quadrants as indicated in Fig. 4b, d, f. Complexes in the first quadrant (SCI $>$ 0.8, RMSD $<$ 2.0 Å) are the true positives. As summarized in Table 2, they account for almost half of the structures for DrugScore and X-Score and about a third of the cases for AffiScore. The false positives are in the upper right quadrant. For these cases, the top scoring poses were incorrect in spite of the high SCI values. The false positive rate was quite low (4%) for AffiScore and DrugScore, and higher for X-Score (18%). The complexes in the lower left and right quadrants are the false and true negatives. They were of similar magnitude. Taken together, the true positives and true negatives were the correct predictions and accounted for 58–74% of the cases (Table 2).

False prediction rates calculated separately for cases above and below the SCI cutoff of 0.8 were about 15% for cases above the SCI cutoff and about 45% for cases below the SCI cutoff (Table 3). This meant that a high SCI score reliably indicated that a given scoring function was able to identify a good docking pose below 2.0 Å for a given decoy set, while a low SCI score warned that the prediction of the best-scoring conformation was unreliable.

In practice, the following protocol provides one possible way to use the SCI scores to increase the likelihood that the top scoring pose will be correct and identify those cases where that cannot be achieved with relatively high certainty. Since Affiscore is the native scoring function of SLIDE (the scores are computed for each docked orientation and saved in the output file) and the fastest of the three scoring functions used, first we ranked the Affiscore SCI values. 17 out of the 50 cases studied had SCI values above 0.8, which meant relatively high confidence in the accuracy of the poses predicted by Affiscore for these 17 cases. Indeed, there were only 2 false positive predictions among them, providing a 12% false prediction rate. The next step was to score the remaining 33 cases with DrugScore, a scoring function that required no additional preparation, and which proved to have a low false positive pose prediction rate in general. Out of the 33 cases, 10 had SCI values above 0.8, out of which one was a false positive prediction. With this, the number of cases for which the best binding mode can be predicted with high confidence increased to 27 with an overall false prediction rate of 11% (3 out of 27). Using DrugScore SCI alone would have predicted high confidence for 26 cases with two false positives. One of these false positives had an RMSD of 10.16 Å relative to the 2.61 Å false positive prediction of AffiScore. Even though this is a modest improvement in prediction accuracy, this stepwise protocol has the potential to reduce the computation time significantly since AffiScore is approximately 10 times faster than DrugScore. Applying the third scoring function, X-Score, to the remaining 23 cases did not bring about further improvements: although there were 7 more cases with X-Score SCI values above 0.8, five of them were false positives. These are the problem cases for X-Score, when it seemed to fall into the trap of local minima. In summary, the combination of AffiScore and DrugScore with the help of the scoring confidence calculations provided a binding mode prediction for 27 cases with a false pose prediction rate of 11%.

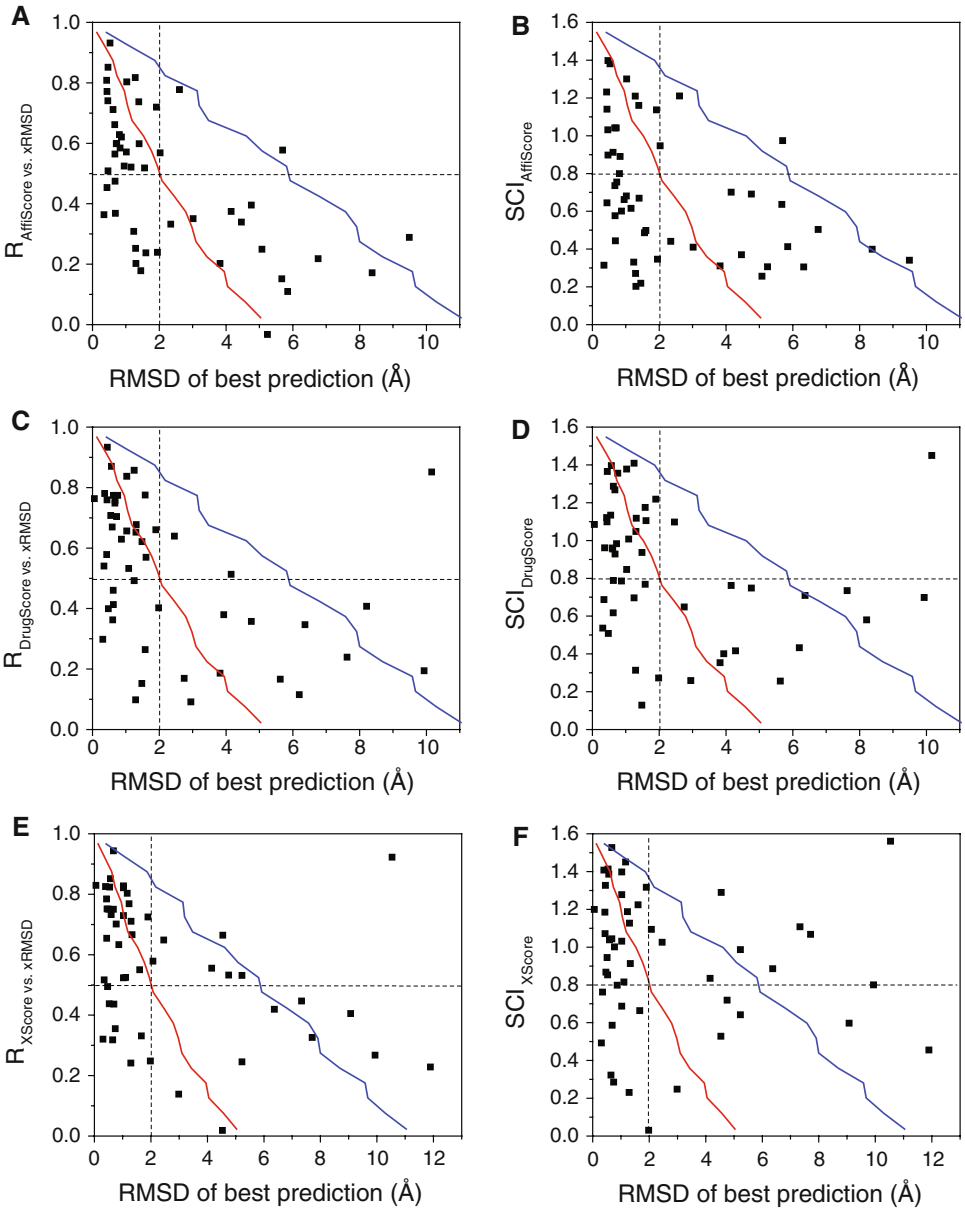

Fig. 4 Correlation between score and RMSD versus RMSD of best prediction from the decoy set according to the score (**a**, **c**, **e**) and scoring confidence index versus RMSD of best prediction (**b**, **d**, **f**). The plots are divided into four regions according to an RMSD cutoff of 2.0 Å and cutoffs of 0.5 and 0.8 for score/RMSD correlation coefficients and SCI, respectively (see text). In the context of scoring function assessment the regions correspond to true positives (*upper left*), false positives (*upper right*), false negatives (*lower left*), and true negatives (*lower right*). *Red* and *blue curves* correspond to theoretical estimates of mean and mean plus two SD for the correlation coefficient between score and RMSD versus the RMSD of the best prediction, respectively (Fig. 5)

Table 2 True/false positive/negatives in predicting the best-scoring RMSD from the correct binding mode based on the SCI score with different scoring functions

	SCI	RMSD (Å)	AffiScore (%)	DrugScore (%)	X-Score (%)
True positive	>0.8	≤2.0	30	48	48
False positive	>0.8	>2.0	4	4	18
True negative	≤0.8	>2.0	28	26	14
False negative	≤0.8	≤2.0	38	22	20

Using either scoring function alone without the scoring confidence estimates would lead to 30–34% false prediction rates for the set of 50 cases without any information about which of these predictions were likely to be correct.

Discussion

The correct prediction of ligand binding poses is a critical part of ligand screening protocols as the prerequisite for accurate estimates of ligand binding affinities. Correct

Table 3 False prediction rates for protein-ligand complexes with low and high confidence values

	False prediction rate (%)			
	AffiScore	DrugScore	X-Score	Average
High confidence set SCI > 0.8	11.8	7.7	27.3	15.6
Low confidence set SCI ≤ 0.8	42.4	54.2	41.2	45.9

binding conformations (with an RMSD of less than 2.0 Å from an experimentally determined complex) can be predicted only when the sampling protocol generates such conformations and when the scoring function is able to select the best docking conformation from a set of predicted poses. In this study, test sets were generated with sufficient sampling to ensure that at least one good pose (at or below 2.0 Å RMSD) was generated for each complex so that a perfect scoring function could identify a ligand binding conformation in all of the test cases. The only exception was the complex 21gs, where the best docked orientation had an RMSD of 2.19 Å. Although standard docking protocols often generate sampling sets that include at least some fraction of docked poses close to the correct binding mode, it is not obvious how to ensure that good conformations are actually present. While this issue will need to be addressed in the future, we are focusing here on the assessment of scoring functions in cases where good conformations are present.

In case of a perfect scoring function, the scores should gradually decrease as the ligand approaches its correct binding mode. This is expected to be reflected by a high correlation between the scores and RMSD values. However, even for such cases, the maximum achievable correlation is limited by local roughness in the underlying energy surface in addition to inherent deficiencies in a given scoring function. As a result, the best-scoring solutions were essentially incorrect poses in approximately 30% of the complexes examined, regardless of the scoring function used. This performance could be considered quite respectable when compared to other scoring methods [5, 6, 8, 11]. However, this means that in an actual screening application there is still a 1 in 3 chance that predicted docking conformations are incorrect and that subsequently calculated binding affinities are not meaningful. While further improvements in the available scoring functions may provide better success rates, this study focuses on the development of a confidence measure that a given scoring function is able to select a good docking conformation from a given set of ligand poses. Such information allows the use of alternate scoring functions and/or additional sampling in low-confidence cases to improve the success rate in identifying good binding poses. The SCI score introduced in this study provides such a confidence measure.

To assure that SCI is not merely a reflection of simple properties of the ligands that might influence docking success, we computed the molecular weights, number of polar (nitrogen and oxygen) and nonpolar (carbon and sulfur) atoms, and the number or non-terminal rotatable bonds and correlated them with the SCI scores. None of these simple measures showed strong correlation with SCI. The maximum correlation of 0.36 was observed between the number of nonpolar atoms and the DrugScore SCI (Table 2, Supplementary material). Among these simple properties, the best predictor of scoring success was the size, but only for the largest ligands (with molecular weight over 630 Da). All of these high molecular weight ligands had top ranking poses under 2.0 Å (Fig. 3, Supplementary material), but they accounted for only 16% of the cases (8 out of 50). For smaller, drug-like ligands, none of the computed properties could predict scoring performance. Therefore, the possibility that the observed correlation between SCI and scoring success was itself a function of more fundamental measures appears not to be the case.

There are two concepts underlying the development of the SCI: The first one is the generally accepted idea that the correct binding mode corresponds to the lowest energy state of the protein-ligand complex and a relatively smooth path should lead to this state in order for the ligand to find it efficiently. This is intuitively called a binding funnel [14], although the legitimacy of this concept remains to be fully validated. The other concept is that the correlation between score and distance from the correct solution is indicative of the likelihood that the top scoring solution will be a good docking. The SCI metric implemented these ideas by using the recently introduced CBScore which did not require knowledge of the correct binding conformation. It was found that a cutoff of 0.8 could be established for the SCI metric above which the best prediction according to a given scoring function was most likely within 2.0 Å from the correct conformation.

In order to better understand the data in Fig. 4 and, in particular, to aid the interpretation of false positives, a theoretical analysis with a synthetic scoring function was carried out as follows: 10,000 sets of data each with 1,000 "conformations" were generated with one-dimensional distances ranging from 0 to 10. For each conformation, a score was calculated based on the distance with added Gaussian noise. The width of the Gaussian noise was varied for each of the data sets so that correlation coefficients between the artificial scores and distance varied between 0 and 1. The correlation coefficients between scores and distances for each set were calculated and plotted against the distances of the best scoring points from

the origin (Fig. 5). Data points were then binned by the correlation coefficient (bin size 0.05) to obtain the mean and standard deviation as a function of the distance as shown in Fig. 5. This analysis highlighted a clear correlation between the correlation coefficient and distance of the best-scoring conformation under the assumption of a linearly correlated scoring function with added noise. Furthermore, even at high correlation coefficients, there were some outliers, that would have been considered false positives given cutoffs for the distance and correlation coefficient.

Interestingly, we have found that an average correlation coefficient between scores and distances of 0.5 corresponded to a distance of 2 in the theoretical data given a maximum distance of 10. This compared well with the previously established cutoff of 0.5 for the correlation of score versus RMSD used to reliably identify predictions below 2.0 Å, and it suggested a more general prescription for determining score cutoffs to achieve a certain level of accuracy. Moreover, the false prediction rate on the theoretical data (percent of top scoring points with RSMD > 2.0) was on average 15% in the high confidence interval and 60% in the low confidence interval. These values established theoretical limits for an ideal, but noisy scoring function. The empirically obtained values for the protein-ligand complex test set were actually slightly better for AffiScore and DrugScore and worse for X-Score. The poorer performance with X-Score was most likely due to a

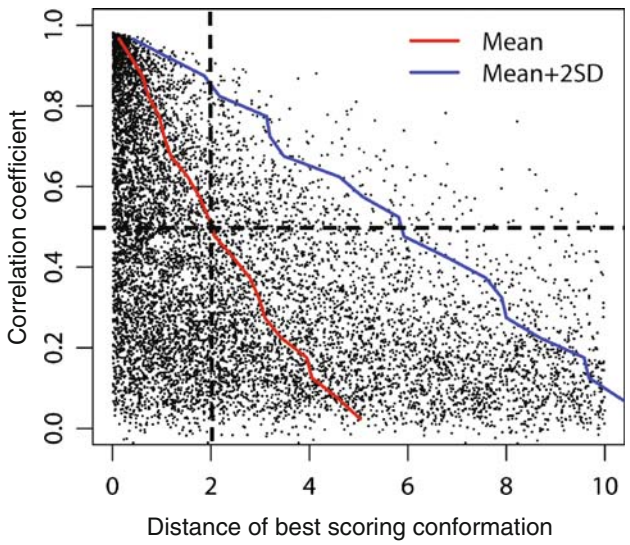

Fig. 5 Correlation coefficients between scores and distances versus distances of the best scoring conformations from the origin in arbitrary units from theoretical data based on a synthetic scoring function (see text). The *red curve* corresponds to the mean correlation coefficients as a function of distance; the *blue curve* is the mean plus two SD

number of cases where the scoring function identified non-native minima.

To further analyze the occurrence of false positives, the mean and the mean plus two standard deviations (SD) from the theoretical analysis were overlaid on the data from the test sets in Fig. 4. Points below the mean + 2SD line were expected due to noise. Indeed, almost all of the data points fell below that line for AffiScore and DrugScore, but a larger fraction of outliers above the line were present for X-score. While this analysis indicated that false positives were unavoidable, it also suggested that false positives were most likely to occur with low RMSD values just above the 2.0 Å cutoff. Furthermore, the theoretical analysis implied that a high rate of false negatives should be expected. While this meant that some good predictions could be missed, these might be recovered through the use of a different scoring function or through rescoring of an enlarged set of decoys.

It is instructive to examine some of the false positive and false negative cases in more detail. Most of the false positive cases were comprised of complexes for which the overall correlation between original scores and RMSD was relatively high (correlation coefficients above 0.5), yet the best scoring predictions were incorrect poses. These cases fell into two major categories: On one hand there were the complexes with good correlation between score and RSMD with only a few outliers with very favorable scores (Fig. 6a). In the second category there were complexes for which the global minimum corresponded to the wrong orientation (Fig. 6b). To some extent, these cases might have resulted from pathologies of the underlying scoring function, such as ignoring the ligand and protein internal conformational strains and the dependence of hydrogen bond strength on geometry. Small van der Waals overlaps between the protein and the docked ligand are also difficult to handle correctly. The three scoring functions used in this study scored a few well buried poses (especially in the case of retinoic acid from the complex 1cbs) with small overlaps very favorably. In reality, these overlaps might not be resolvable by induced fit. Another factor to be considered is that the RMSD might be an imperfect measure of correctness of a given pose relative to the correct conformation and there might be concerns that the native conformation is in fact not as well defined as suggested by crystallographic data.

For most of the false negative cases, the overall correlation between the original scores and RMSD was quite low. There were essentially two scenarios. In the first one, the overall correlation was poor and good predictions were found by very small differences in score, probably to some extent by chance (Fig. 6). In these cases, the SCI score rightly indicated a poor correlation and predicted a low level of confidence in the predicted binding mode. The other possibility of obtaining false negatives was due to

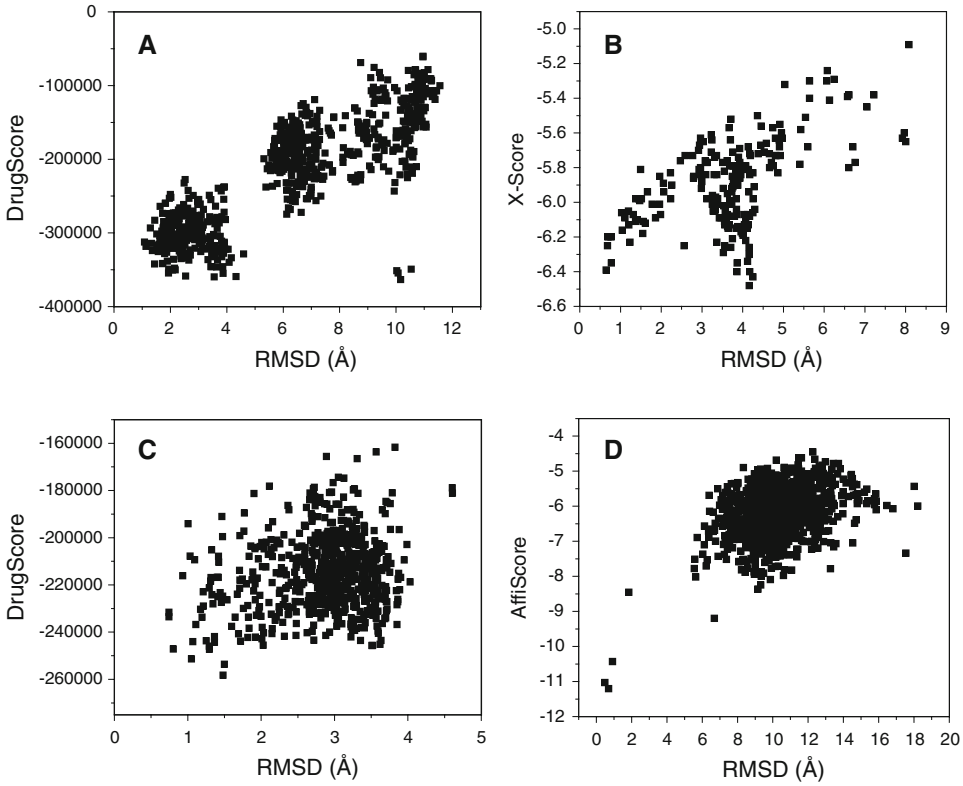

Fig. 6 Score versus RMSD from the correct conformation in selected false positive: (1cbs, **a** and 5cna, **b**) and false negative cases (1abe, **c** and 1csc, **d**). The corresponding SCI values and the score versus RMSD correlations are: 1.449 and 0.851 (**a**), 0.834 and 0.555 (**b**), 0.128 and 0.152 (**c**), 0.443 and 0.368 (**d**)

non-uniform sampling, as shown in Fig. 6d. In this case, few structures close to the correct binding mode had favorable scores, while all of the other conformations were far from the correct conformation with poor correlation between score and RMSD. Such cases could be diagnosed by clustering the poses and calculating the distances between the cluster centers. A large gap separating a few top scoring poses from every other pose is indicative of insufficient sampling. Sampling non-uniformity can be remedied by generating additional samples in regions of space that are poorly sampled otherwise. This is expected to lead to better overall correlation and a larger SCI score so that these cases could become true positives.

The quality of the crystal structure can also affect the performance of the scoring confidence index. Although an attempt was made to include only complexes with good overall resolution, there were four cases for which the average B-factors for the ligands were more than double of the target average B-factors: 1avn, 1bid, 1eed, and 21gs (Table 1, Supplementary material). It is noteworthy, that in three out of these four cases (1avn, 1bid, 21gs) the scoring functions usually failed to identify the correct binding mode, and the SCI values were low. We might speculate that low SCI values are consequences of incorrect X-ray poses or perhaps of the existence of multiple binding modes in these cases.

The scoring confidence index was derived independently of any given primary scoring function and is therefore expected to be scoring-function independent. In fact, we have found that the SCI metric was useful with all three scoring functions tested here, although the false prediction rates varied between AffiScore and DrugScore on one hand and X-score on the other. This seemed to be due to the fact that X-Score identified the global energy minimum in the wrong place for more cases than the other scoring functions. On average, X-Score had higher correlations with RMSD than AffiScore and DrugScore, but this resulted in misleadingly high SCI values for complexes with wrong global energy minima. The SCI value could indicate whether a gradual descent toward the global energy minimum existed or not, but could not detect an entirely incorrect global energy minimum. Therefore, the success of the SCI metric depended to some extent on the ability of the underlying scoring function to avoid false global minima at non-native conformations.

Conclusions

A statistically derived scoring confidence index (SCI) was introduced to provide a simple yet promising novel diagnostic tool for assessing scoring performance in protein-ligand docking without knowledge of the correct native binding pose. In a virtual screening or inhibitor design project, the application of SCI-based confidence criteria would allow the selection of the optimal scoring methods

for different systems and diagnose cases where additional sampling may lead to better prediction of correct binding modes. This is in contrast to previous approaches where scoring functions have been tailored to work well with a given target [22]. Because of its generality, the scoring confidence index is expected to be also applicable in the context of protein–protein docking and structure prediction. Future studies will test how the SCI metric can be used in a protocol that employs alternate scoring functions and allows for additional generation of decoys in cases of predicted poor performance of a given scoring function.

Acknowledgments The authors thank Gerhard Klebe (University of Marburg) and Holger Gohlke (University of Frankfurt) for providing DrugScore, Shaomeng Wang (University of Michigan) for providing X-Score, and OpenEye Scientific Software (Santa Fe, NM) for providing Omega to us. This work was supported in part through National Science Foundation CAREER grant MCB 0447799 (to MF), National Institute of Health grant GM 084953 (to MF), and an Alfred P. Sloan Foundation fellowship (to MF).

References

1. Davies JW, Glick M, Jenkins JL (2006) Curr Opin Chem Biol 10:343. doi:10.1016/j.cbpa.2006.06.022
2. Klebe G (2006) Drug Discov Today 11:580. doi:10.1016/j.drudis.2006.05.012
3. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49:5851. doi:10.1021/jm060999m
4. Stahl M, Rarey M (2001) J Med Chem 44:1035. doi:10.1021/jm0003992
5. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) J Chem Inf Model 46:401. doi:10.1021/ci0503255
6. Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912. doi:10.1021/jm050362n
7. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CLIII (2004) J Med Chem 47:3032. doi:10.1021/jm030489h
8. Wang R, Lu Y, Wang S (2003) J Med Chem 46:2287. doi:10.1021/jm0203783
9. Wang R, Lu Y, Fang X, Wang S (2004) J Chem Inf Comput Sci 44:2114. doi:10.1021/ci049733j
10. Kontoyianni M, McClellan LM, Sokol GS (2004) J Med Chem 47:558. doi:10.1021/jm0302997
11. Perola E, Walters WP, Charifson PS (2004) Proteins 56:235. doi:10.1002/prot.20088
12. Schulz-Gasch T, Stahl M (2004) Drug Discov Today Technol 1:231
13. Stumpff-Kane AW, Feig M (2006) Proteins 63:155. doi:10.1002/prot.20853
14. Tsai CJ, Kumar S, Ma B, Nussinov R (1999) Protein Sci 8:1181
15. Roche O, Kiyama R, Brooks CLIII (2001) J Med Chem 44:3592. doi:10.1021/jm000467k
16. Wang R, Lai L, Wang S (2002) J Comput Aided Mol Des 16:11. doi:10.1023/A:1016357811882
17. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235. doi:10.1093/nar/28.1.235
18. Schnecke V, Kuhn LA (2000) Perspect Drug Discov Des 20:171. doi:10.1023/A:1008737207775
19. Zavodszky MI, Kuhn LA (2005) Protein Sci 14:1104. doi:10.1110/ps.041153605
20. Zavodszky MI, Sanschagrin PC, Korde RS, Kuhn LA (2002) J Comput Aided Mol Des 16:883. doi:10.1023/A:1023866311551
21. Gohlke H, Hendlich M, Klebe G (2000) J Mol Biol 295:337. doi:10.1006/jmbi.1999.3371
22. Gohlke H, Klebe G (2002) J Med Chem 45:4153. doi:10.1021/jm020808p