Testing Assumptions and Hypotheses for Rescoring Success in Protein−Ligand Docking

Noel M. O'Boyle,[†] John W. Liebeschuetz, and Jason C. Cole*

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

Received May 6, 2009

In protein−ligand docking, the scoring function is responsible for identifying the correct pose of a particular ligand as well as separating ligands from nonligands. Recently there has been considerable interest in schemes that combine results from several scoring functions in an effort to achieve improved performance in virtual screens. One such scheme is consensus scoring, which involves combining the results from several rescoring experiments. Although there have been a number of studies that have investigated factors affecting success in consensus scoring, these studies have not addressed the question of why a rescoring strategy works in the first place. Here we propose and test two alternative hypotheses for why rescoring has the potential to improve results, using GOLD 4.0. The "consensus" hypothesis is that rescoring is a way of combining results from two scoring functions such that only true positives are likely to score highly. The "complementary" hypothesis is that the two scoring functions used in rescoring have complementary strengths; one is better at ranking actives with respect to inactives while the other is better at ranking poses of actives. We find that in general it is this hypothesis that explains success in a rescoring experiment. We also test an assumption of any rescoring method, which is that the scores obtained are representative of the fitness of the docked pose. We find that although rescored poses tended to have slightly higher clash values than their docked equivalents, in general the scores were representative.

INTRODUCTION

Despite more than a decade of research into improved scoring functions, a scoring function that can accurately predict binding affinities remains an elusive goal.[1] Even the simpler problem of identifying ligands from a data set of inactive molecules is a challenge for modern scoring functions, although for a given protein a particular scoring function may work very well. While there is certainly a need for the development of improved scoring functions with better performance over a wider range of protein families, it is also important to make the maximal use of currently available scoring functions. One of the ways to do this is to combine existing scoring functions in a so-called rescoring experiment.

In a rescoring experiment poses are generated by docking with one scoring function but are evaluated using a different scoring function. A common application of rescoring is consensus scoring, introduced by Charifson et al.,[2] where the docked pose is rescored several times with different scoring functions whose results are combined in some way. Several different strategies have been proposed for combining scores such as voting strategies[2,3] (thresholds are set for each scoring function), ranking by rank[4] (ranking based on the mean or sum of ranks of a particular molecule over all scoring functions), and regression schemes[3] (weighted combinations of scoring functions or their components). Consensus scoring can be regarded as a type of data fusion,[5−8] an area whose broad definition involves the combination of data and information from multiple sources. Feher, in a review of consensus scoring in protein−ligand docking,[9] presents multiple examples where consensus scoring methods have performed better than using the individual rescore values in terms of enrichments, pose prediction, and prediction of binding affinities.

Only a small number of studies have investigated how consensus scoring can improve results compared to rescoring using a single scoring function alone. Wang and Wang[10] used computer simulations on a virtual library generated by assuming a Gaussian distribution of activities in a data set of 5000 compounds. The scores for each of 10 supposed scoring functions were generated by assuming a normal distribution of errors around the actual activity. They found that the more scoring functions were used to calculate a consensus score, the smaller the error in prediction. This was attributed to the fact that the mean of repeated samplings tends toward the true value. More recent studies have questioned some of the assumptions of the simulations particularly in light of the fact that consensus scoring can lead to poorer results in some cases (for example, ref 11). In particular, as noted by Baber et al.[12] the ranks from different scoring functions are not necessarily independent; first, all good scoring functions score actives highly, and second, different scoring functions can share the same deficiencies and thus make correlated errors.

Consensus docking, as distinct from consensus scoring, is an alternative way of combining results from several scoring functions, although the terms are sometimes confused in the literature. Consensus docking involves combining in some way the scores or ranks obtained for poses docked with two or more different scoring functions; in consensus scoring, the poses are docked only once. In a study of consensus docking with five scoring functions and four protein targets,

* Corresponding author e-mail: cole@ccdc.cam.ac.uk.

[†] Current address: Analytical and Biological Chemistry Research Facility, University College Cork, Cork, Ireland.

10.1021/ci900164f CCC: $40.75 © 2009 American Chemical Society
Published on Web 07/31/2009

Yang et al.[13] found that the result did not always perform better than the best individual scoring function although it always did better than the average of the individual results. They concluded that consensus docking works best where the scoring functions involved have high individual performance and have different scoring characteristics. Baber et al.[12] found that, in the case of consensus methods for ligand-based virtual screening, different methods tend to agree more on the ranking of actives than of inactives, thus leading to few false positives. This is also likely to be the case here.

Here we address the question of how docking with one scoring function and rescoring with another can improve performance in a virtual screen. This is an underlying factor in the success of consensus scoring methods. While both Verdonk et al.[14] and Gohlke et al.[15] have touched on this question in the context of improved pose prediction, there have not been any studies which have specifically addressed the question of how rescoring with a single scoring function can improve performance in a virtual screen compared to docking alone.

We propose two hypotheses to explain how performance can be improved through rescoring. The 'consensus hypothesis' is that rescoring works for the same reasons that consensus methods work. Namely, by combining the results for two scoring functions you correct for false positives in one or the other. Overall, this hypothesis implies that for an active the highly scored docked poses found using one scoring function will still score highly when assessed with another scoring function.

A second hypothesis, the 'complementary hypothesis', is that rescoring works because the two scoring functions have complementary strengths which, when combined, lead to improved results. Specifically, one scoring function is better at predicting poses, but the other is better at scoring those poses. The implication of this hypothesis is that while the first scoring function is capable of correctly identifying likely poses, it is either poorer at ranking the binding affinities of different molecules in the same pose or at ranking different poses of the same molecule. Clearly this hypothesis is valid where the rescoring procedure involves a more accurate though computationally expensive calculation such as MM-GBSA (as in ref 16 for example). Here we focus on rescoring using scoring functions of similar computational cost compared to those used for docking.

In this study we test these hypotheses using the GOLD protein−ligand docking software[14,17] and the Astex Diverse Data Set of 85 protein−ligand complexes.[18] GOLD 4.0 has two empirical scoring functions, ChemScore[19] and GoldScore,[20] as well as a knowledge-based scoring function, ASP,[21] all of which can be used for both scoring and rescoring. We also test an assumption of any rescoring method, which is that the rescore values obtained are valid measures of a pose's fitness according to the scoring function used for rescoring.

METHOD

Data Set. Protein structures and actives were taken from the Astex Diverse Set.[18] This is a set of structures of 85 pharmaceutically relevant proteins and their actives. The protein crystal structures are high quality, and the active poses have been verified against the experimental density. Our data set also contained 99 inactive molecules for each protein. The set of inactives was taken from a previous study.[22] Briefly, the inactives for a particular protein were selected from an in-house data set of compounds available for purchase but were chosen to be physicochemically similar to the active but topologically distinct.

Docking and Rescoring. GOLD 4.0 was used for all docking and rescoring experiments. Docking experiments were carried out with a setting of 30000 genetic algorithm operations. The diverse solutions option was used to generate ten diverse poses for each molecule (maximum cluster size of 3, minimum intercluster rmsd of 2.5 Å). GOLD 4.0 provides three different scoring functions which can be used for docking or rescoring: GoldScore, ChemScore, and ASP (Astex Statistical Potential). A fourth scoring function, ChemScoreRDS,[22] has been developed for rescoring. The docking procedure uses the same search algorithm (a genetic algorithm) for all scoring functions which means that search space coverage is consistent across the three scoring functions. In the remainder of the paper, GS will be used as an abbreviation for GoldScore and CS for ChemScore.

In a typical docking experiment, each active along with its corresponding 99 inactives was docked to its corresponding protein. The rank of each active with respect to its 99 inactives was calculated, and the mean value across the 85 proteins was reported. In addition, the pose prediction accuracy was calculated as the number of active poses within 2.0 Å of the crystal pose; note that for each active only its top-ranked pose was considered. Since the docking procedure is stochastic, each docking experiment was repeated 25 times to obtain mean and standard deviation values for the mean rank of actives and pose prediction accuracy.

In a rescoring experiment, the poses obtained by docking with one scoring function were re-evaluated using a second scoring function. During this process, the structure of the ligand is optimized using a simplex procedure to the local minimum. This is the same simplex procedure that is used by GOLD as the last step in a docking experiment (similar to the rigid-body minimization described by Meng et al.[23]). This is a necessary step, particularly for rescoring, as an artifactually poor score would otherwise be obtained calculated score particularly where the scoring function does not have a smooth response function.[24] In general, all ten poses for each molecule were rescored, and the largest score was used to calculate the rank.

The following notation will be used to indicate a rescoring experiment: CS/rASP. This example indicates that CS docked poses were rescored using ASP.

Consensus Scoring. Here the molecules were docked with one scoring function, while the other two scoring functions were used to rescore the docked poses. The rank-by-rank strategy was used to combine the rescoring results into a consensus rank. Since 10 poses were generated for each molecule, as an initial step the pose with the best consensus rank was chosen. This involved calculating the rank of each pose with respect to Scoring Function A and then with Scoring Function B and finally identifying the pose with the lowest mean rank.

Using these best poses, the consensus rank for an active of a particular protein was calculated as follows:

(1) The rank of each active and each inactive in terms of Scoring Function A was calculated.

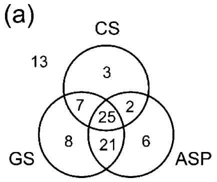

Figure 1. Venn diagrams contrasting the performance of the three scoring functions for different proteins. The results are taken from a single repetition (the first repetition) of the docking experiments. (a) The number of actives placed in the top-ranked position. For example there are 25 actives (out of 85) which all three scoring functions correctly place in the top-ranked position. (b) Poses correctly predicted; that is, where the top-ranked pose is within 2.0 Å rmsd of the crystal structure. For example there are 53 proteins where all three scoring functions correctly predict the active pose.

(2) The rank of each active and each inactive in terms of Scoring Function B was calculated.

(3) For each active and inactive, the mean of (1) and (2) is found.

(4) The rank of the active in terms of the mean values in (3) is found.

The mean rank of the 85 actives was then calculated as well as the standard deviation. The following notation will be used to indicate a consensus scoring experiment: GS/(rASP + rCS). This example indicates that consensus scoring was used to combine ASP and CS rescores of GS docked poses.

Consensus Docking. Consensus docking involves combining the scores from docking using two or more different scoring functions. The same procedure was used as for consensus scoring, namely steps (1) to (4) described above. The only difference is that the poses used were the highest scoring poses from separate docking experiments.

Comparison of Rescore Values and Docked Values. Rescore values were obtained from the first repetition of the 25 experiments described above. For each protein we identified the GS pose, *P*, of the active with the highest score, *R*, when rescored with CS.

For comparison, docked values were obtained as follows: for each protein in the Astex Diverse Set, 100 poses of the active molecule were generated by docking with CS using the diverse solutions option (maximum cluster size of 3, minimum intercluster rmsd of 0.1 Å). For each of these docked poses a ΔChemScore value was calculated by subtracting *R* and an rmsd value was calculated with respect to *P*.

RESULTS AND DISCUSSION

Testing Hypotheses for Rescoring Success. The data set was docked using each of ChemScore, GoldScore, and ASP in turn. Each docking experiment was repeated 25 times. Figure 1(a) compares the ability of each scoring function to rank the active in the top position, while Figure 1(b) compares their ability at pose prediction. It is clear that while there is a great deal of overlap between the abilities of the three scoring functions, each scoring function has its own strengths: for example, to take the case of CS, there are 3 actives which only CS places in the top-ranked position, and similarly there are 5 actives for which only CS correctly predicts the poses.

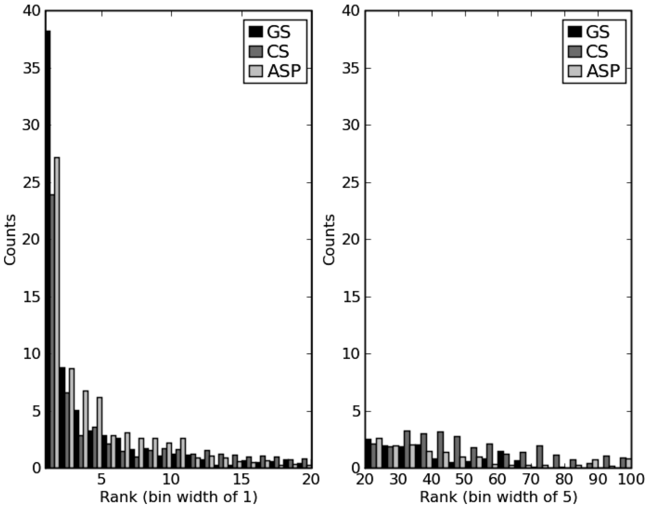

Figure 2. Histogram showing the distribution of the ranks of the 85 actives for GoldScore, ChemScore, and ASP. Results are an average from 25 repetitions of each docking experiment.

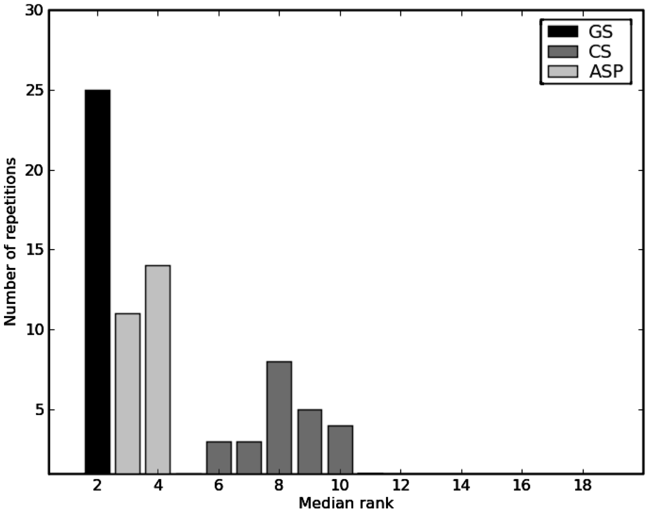

Figure 3. Histogram showing the distribution of the median rank of the actives across 25 repetitions of each docking experiment.

Figure 2 shows that the resulting distribution of ranks of the actives is highly skewed with a peak at rank 1 which declines exponentially. The active is found in one of the top five positions in 52 out of 85 cases for GS, 58 for ASP, and 39 for CS. The distribution of the median rank for each of the 25 repetitions is shown in Figure 3. For a skewed distribution the median is a better estimator of the average than the mean. However for the purposes of this study, we will focus on changes in the mean, as the median is insensitive to improvements in the ranks of actives anywhere except in the extreme left of the distribution. The mean rank of actives and pose prediction accuracy are shown in Table 1 as the mean and standard deviation from 25 repetitions of each docking experiment.

The results for the docking and rescoring experiments shown in Table 1 illustrate the potential of rescoring over docking with a single scoring function alone. Although GS poses cannot on average be improved by rescoring, CS poses are greatly improved by rescoring with either of the other two scoring functions, while ASP poses are improved by rescoring with GS. An interesting result is that in every case the mean rank obtained for a scoring function on its own is the same as or poorer than when that scoring function is

Table 1. Pose Prediction Accuracy and Scoring Performance from 25 Repetitions of the Same Docking Experiments[a]

docking function	rescoring function	mean rank of actives	number of poses correctly predicted (top-ranked pose only)	number of poses correctly predicted (any of the 10 poses)
GS	-	8.9 (0.4)	67.1 (1.3)	80.7 (1.4)
GS	CS	15.8 (0.7)		
GS	ASP	9.5 (0.6)		
CS	-	20.5 (0.7)	68.2 (2.0)	79.7 (1.7)
CS	ASP	11.0 (0.8)		
CS	GS	7.2 (0.8)		
CS	CSrds	12.6 (0.6)		
ASP	-	11.0 (0.5)	65.4 (2.0)	80.3 (1.6)
ASP	GS	7.9 (0.5)		
ASP	CS	19.3 (0.9)		

[a] The standard deviation is shown in parentheses. Each docking experiment produced 10 diverse poses (see Methods).

used to rescore poses from another scoring function. For example, when used to score its own poses GS gives a mean rank of 8.9, whereas when used to score CS and ASP poses it gives 7.2 and 7.9, respectively. Where does this improvement originate?

In a rescoring experiment, one or several of the top-ranked poses may be rescored. For the purposes of investigating the proposed hypotheses we have used all ten poses generated. Although we investigated the effect of choosing different numbers of poses to rescore (results not shown), the standard deviation of the results from the 25 repetitions was too large to allow any definite conclusions. In order to generate multiple binding poses during the docking, the diverse solutions option was used with maximum cluster size of 3 and intercluster rmsd of 2.5 Å. When the docking results are analyzed for CS, ASP, and GS, the average number of correct poses in each case is 2.9 (rising to 3.1 if the 4 proteins where a correct pose is not found are excluded). It is also worth noting that the number of poses with scores within 10% of the top-scoring pose was approximately 5 but varied widely with a standard deviation of approximately 2.5.

In the following sections we investigate two hypotheses that have been proposed to explain why rescoring docked poses with a different scoring function often leads to an improvement: the consensus hypothesis and the complementary hypothesis.

The starting point for testing the proposed hypotheses is the assumption that rescoring can improve results compared to docking alone, and furthermore that this can be explained in terms of the performance of the individual scoring functions. While it is true that just by chance a particular scoring function or rescoring combination can lead to improved results for a particular protein and series of actives, by combining the results for 85 docking experiments it is possible to get an overview of what exactly is necessary in order to achieve improved results through rescoring.

The Consensus Hypothesis. The first question we wished to answer was whether rescoring success is due to an averaging effect between the scoring functions which serves to eliminate false positives, the 'consensus hypothesis'. Since the consensus hypothesis relies on the same arguments used to explain success in both consensus docking and consensus scoring, it is useful to investigate how those consensus methods perform on this data set. The results for consensus docking are shown in Table 2. When the ranks of the docked poses from CS and GS are combined the resulting mean rank is 11.0. This is, perhaps not unsurprisingly, intermediate between the rank of CS on its own (20.5) and GS on its own (8.9), although much closer to the better result. On the other hand, if GS is combined with ASP, a scoring function which appears to have a similar ability to score actives highly (mean rank of 11.0), the consensus rank is better than either on its own at 7.0.

Table 2. Consensus Docking Results

docking functions	consensus docking
CS and ASP	11.6 (0.4)
ASP and GS	7.0 (0.3)
GS and CS	11.0 (0.4)
CS and ASP and GS	8.1 (0.2)

Combining the results of different predictors is known to work best where the predictors make uncorrelated errors.[25] If they make the same errors, for example if both GS and CS are poor at predicting activity in lipophilic pockets, their combination is not likely to do any better. Table 3 shows that the mean Spearman correlations of the ranks of the actives from 25 repetitions are 0.39 for CS vs GS, 0.43 for GS vs ASP, and 0.21 for CS vs ASP. Although the low correlation for CS vs ASP means that the consensus score at 11.6 is almost as good as ASP on its own (11.0), it is difficult to improve a good predictor by combining its results with a poorer predictor.

The results for consensus scoring are shown in Table 4. The trend is similar to that obtained for consensus docking in that the best score is obtained when ASP and GS are combined. There is an additional confounding factor, of course, in that the three consensus scores are calculated using different poses. It is worth noting that the scores for consensus scoring are in every case as good as, or better than, those obtained for the corresponding consensus docking, although the latter actually requires more computation. This can be attributed to the fact that the rescore values on which the consensus scores are based are in every case as good as, or better than, the scores obtained from the original docking (on which the consensus dockings are based).

In relation to rescoring, if the consensus hypothesis is true, the effect of swapping the order of the scoring and rescoring function should be small as we are still combining the same functions. However, this is clearly not the case. For example, when combining CS and GS, CS/rGS has a rank of 7.2, while GS/rCS has a rank of 15.8. In fact, in every case the final rank is close to that found for the rescoring function on its own.

However, an alternative way of calculating the rank after rescoring would be to use the consensus of the docked poses and the rescored poses. This is the same procedure as described in the Methods for consensus scoring except that instead of combining the results from two rescoring experiments, one rescore is combined with the results from the original docking. Table 5 shows the resulting mean rank of the actives. Now the effect of swapping the order of scoring and rescoring is much smaller: GS + GS/rCS is 7.9 while *vice versa* (CS + CS/rGS) gives 10.1; CS + CS/rASP is 12.0 and *v.v.* is 11.0; ASP + ASP/rGS is 7.2 and *v.v.* is 6.0.

Overall our results show that for a regular rescoring experiment, where the score from the rescoring function is

Table 3. Spearman Correlation between Ranks Obtained from the Docking and Rescoring Experiments Shown in Table 1[a]

	CS	GS/rCS	ASP/rCS	ASP	CS/rASP	GS/rASP	GS	ASP/rGS	CS/rGS
CS	-								
GS/rCS	**0.80**	-							
ASP/rCS	**0.84**	**0.75**	-						
ASP	0.21	0.32	0.31	-					
CS/rASP	0.23	0.32	0.27	**0.81**	-				
GS/rASP	0.18	0.42	0.21	**0.75**	**0.72**	-			
GS	0.39	0.49	0.40	0.43	0.41	0.46	-		
ASP/rGS	0.21	0.39	0.37	0.37	0.48	0.48	**0.71**	-	
CS/rGS	0.34	0.44	0.33	0.49	0.56	0.52	**0.72**	**0.62**	-

[a] The values shown are the mean values from 25 repetitions of the experiments. The mean standard deviation of these values is 0.04. Correlations greater than 0.60 are highlighted in bold.

Table 4. Consensus Scoring Results

docking function	consensus scoring functions	mean rank of actives
GS	CS+ASP	9.8 (0.9)
CS	ASP+GS	7.4 (0.8)
ASP	GS+CS	10.4 (1.0)

Table 5. Consensus of Score from Docking and Rescoring

docking function	rescoring functions	consensus mean rank of actives
GS	ASP	6.0 (0.5)
	CS	7.9 (0.6)
CS	GS	10.1 (0.8)
	ASP	12.0 (0.6)
ASP	CS	11.0 (0.6)
	GS	7.2 (0.6)

used to rank the molecules in the data set, the consensus hypothesis does not hold. The scores from the initial scoring function serve only to filter out all but the top ten poses. For a pose to score highly in the end, it must score highly according to the rescoring function. The correlations shown in Table 3 support this point. All of the correlations above 0.60 are associated with pairs of experiments that involve the same function used for the final scoring. However if, rather than using the ranks from the rescoring function, the ranks from the scoring and rescoring functions are combined in a consensus score, the hypothesis holds true. In that case, if scoring functions that perform well on their own are combined, the result is better than either on its own (Table 5).

The Complementary Hypothesis. Up until now, we have focused on the ranks of the actives versus the inactives. On this basis, CS appears to perform much poorer in general than either of GS or ASP. For a particular protein of course, CS may do better (for 21 proteins, CS ranks the active higher than GS and for 10 of these the difference is greater than 5 ranks), but here we are interested in the general trend. It may come as a surprise then that, when we look at the pose prediction accuracy of the top-ranked poses, CS performs as well as GS and ASP, with 68.2 poses within 2.0 Å of the crystal structure compared to 67.1 for GS and 65.4 for ASP (Table 1).

How can these results be reconciled? One possibility is that the correct pose was not found in several instances by ASP and GS. This would indicate a docking failure rather than a scoring problem. However, for all three scoring functions the top 10 poses contain the correct pose for on average 80 proteins (Table 1). An alternative hypothesis is that the ability of a scoring function to correctly rank different poses of the same molecule (pose prediction) is independent, at least to some degree, of its ability to separate actives from inactives (virtual screening). This would explain how CS could be as good as ASP and GS at pose prediction yet have quite a different performance in ranking the active.

Further support for this idea is provided by the ChemScoreRDS rescoring function (CSrds), a modification of CS that scales hydrogen bond interactions based on their burial depth.[25] Crucially, the terms in the scoring function were optimized for rescoring performance of CS docked poses. As a result it is an excellent example of a scoring function that performs well at rescoring (see Table 1) but is wholly unsuitable for docking. Initial experiments investigating the use of CSrds for docking showed that an extensive reweighting of CS terms would be required as deeply buried hydrogens bonds were formed at the expense of protein−ligand clashes and unreasonable geometries. This distinction between the performance of a scoring function in rescoring versus docking has previously been noted.[4]

This leads directly to the complementary hypothesis: that an improvement can be obtained if one scoring function (the one used for docking) is better at scoring different poses of the same molecule, while the other scoring function (the one used for rescoring) is better at relative scoring of different molecules. On this basis, CS could be used for docking, so long as either ASP or GS is used for rescoring. When the CS poses are rescored with GS, a mean rank of 7.2 (0.8) is obtained (better than GS on its own), while with ASP a value of 11.0 (0.8) is found (as good as ASP on its own). Earlier we saw that the consensus docking of GS and ASP gave results better than either on its own; here, the consensus score CS/r(GS + ASP) gives 7.4 (0.8) (Table 4), which is equivalent to the CS/rGS value. However, all of these values are markedly better than those obtained using the CS scores themselves: 20.5 (0.7). The converse should also hold; if CS is used for rescoring the results should always get worse, which is true.

An alternative way to exploit this complementarity would be to dock with the scoring function that is better at ranking actives versus inactives and then to identify the best pose by rescoring with the scoring function that is better at pose prediction. Once the best pose is identified, the rank should be calculated with the original score from the docking. It should be expected that this method would not perform as well as the first, as we are not docking with the scoring function that is better at identifying the correct pose. For comparison with the earlier results, if CS is used to identify the best GS-docked poses the mean rank improves from 8.9 to 7.6, while the ASP ranks is unchanged (from 11.0 to 10.7). Since ASP and GS are just as good as CS at pose prediction we would not expect to see much change in any case, but the contrast with the results obtained if the CS rescore values were used is quite clear: GS/rCS is 15.8 while ASP/rCS is 19.3 (Table 1).

In a typical rescoring experiment, all of the docked poses are rescored and the one with the highest rescore value used. This is the method used here as there is no reason to prefer the top-ranked CS pose over the top-ranked ASP or GS poses. However if one scoring function clearly outperforms the others at pose prediction, the optimal result should be obtained by just rescoring the pose that is top-ranked according to the docking scoring function. It should also be noted that if too many poses are included the results will tend toward those obtained if docking with the rescoring function.

One clear conclusion from this section is that if a particular scoring function does not perform well in ranking actives versus inactives, the scores from that scoring function should not be used in determining the final ranking either as part of a consensus scoring scheme nor as a rescoring function. However, if that scoring function performs well at pose prediction, it may be used to generate poses for rescoring with an alternate scoring function or for identifying (but not scoring) the correct pose in a rescoring experiment.

Are the Rescore Values Valid Measures of a Pose's Fitness? In GOLD the final step before calculating the score of a docked pose involves a local optimization using a simplex procedure. This means that the docked pose will be at a local maximum with respect to the scoring function used to dock it. Given the size of the conformational and translational space available to a putative ligand, the incorporation of a local optimization procedure means that the search algorithm used in docking (a genetic algorithm in the case of GOLD) just needs to identify a pose in the correct 'valley' of the scoring function, a much easier task.

However, the docked pose is unlikely to also be at a local maximum with respect to the scoring function used to rescore the pose. As a result, a local optimization is also necessary before a valid rescore value can be calculated (this is the default in GOLD although in published rescoring studies such an optimization is not always carried out). However, such an optimization is constrained to the valley in which the pose finds itself, and there may be another deeper valley nearby which is equally valid. This could lead to artifactually poor rescore values.

To test this, the ChemScore values obtained by rescoring GoldScore poses were compared to those for very similar poses generated in a ChemScore docking experiment to identify whether the ChemScore scoring function had a greater local maximum in an adjacent region of configuration space. To do this, an exhaustive docking was carried out with ChemScore to generate 100 poses for each active with the goal of saturating the binding site in the most highly scoring region. The options used for the docking ensured that although 100 distinct poses were generated, that they could differ by as little as 0.1 Å rmsd. These poses were compared to those from the GS/rCS experiment by calculating the rmsd with respect to the original GS pose and by calculating ΔChemScore, the increase in the ChemScore value compared to that obtained in the rescoring experiment.

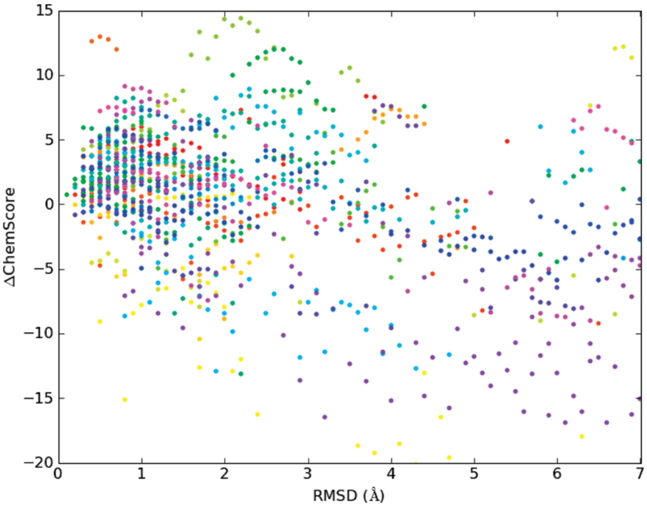

Figure 4. The rmsd plotted against ΔChemScore. rmsd values are rounded to the nearest 0.1 Å. For clarity only the highest scoring pose for a particular rmsd and protein is shown. A different color is used for each protein, but these should only be used as guides as the shade of the color may differ only slightly from protein to protein.

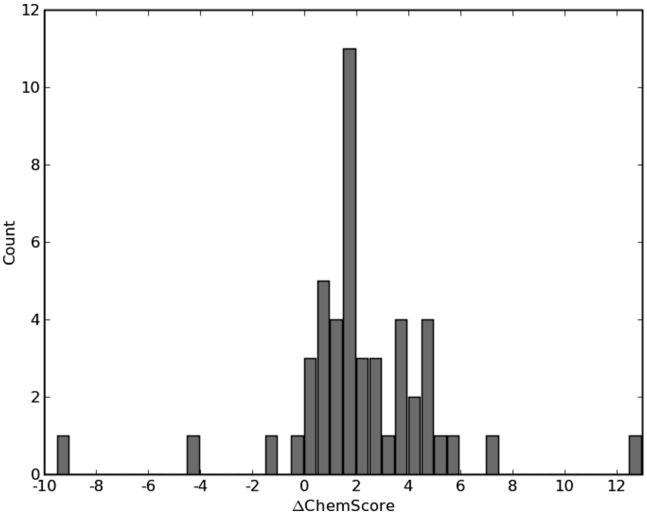

Figure 5. For all CS docked poses within 0.5 Å of the GS docked pose, the histogram shows the improvement in ChemScore values compared to a CS rescore of the GS docked pose. For each of the 48 proteins where such a pose was found, only the pose with the largest improvement is shown. The bin size is 0.5 CS units.

A plot of rmsd versus ΔChemScore is shown in Figure 4 and as a histogram in Figure 5 for those poses within 0.5 Å of a GoldScore docked pose. A docking experiment explores the entire binding site and so is likely to find a pose somewhere with a higher ChemScore value than that obtained by rescoring a GoldScore pose. For 48 proteins it was possible to find similar (within 0.5 Å rmsd) docked poses. In general these scored better than the rescored poses (see Figure 5), but the median improvement was only around 1.9 ChemScore units, about two-thirds the score for a perfect hydrogen bond (about 3.3 ChemScore units). If this value is used as a lower-bound for ΔChemScore, 14 proteins have a pose that has ΔChemScore > 3.3 and rmsd ≤ 0.5 Å. Of these, only four have ΔChemScore > 5.0: 1hww (ΔChemScore

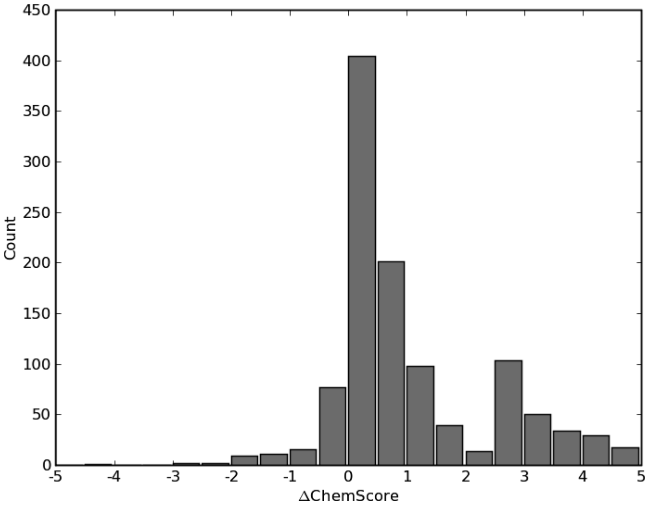

Figure 6. For all CS docked poses within 0.5 Å of the GS docked pose, the histogram shows the decrease in the clash contribution to the ChemScore compared to a CS rescore of the GS docked pose. The bin size is 0.5 CS units.

12.6, 0.4 Å rmsd), 1q41 (5.9, 0.3 Å), 1v4s (5.2, 0.5 Å), and 2bsm (7.2, 0.5 Å).

In the crystal structure of 1hww, the ligand is bound to a zinc atom in the active site through two hydroxyl groups with an O−Zn distance of about 2.3 Å. The docked ChemScore pose is quite similar (distances around 2.5 Å). However, in the docked GoldScore pose one of the hydroxyl groups is closer to the Zn (2.1 Å) and the other further away (3.1 Å). The simplexed ChemScore structure is almost identical, and only a single metal−ligand bond receives a score. This accounts for 6 ChemScore units of the 12.6. The remainder is mainly accounted for by a high value for atom−atom clashes in the rescored structure. A similar effect occurs with 1q41 where the initial GoldScore pose is missing a hydrogen bond to a carbonyl on the protein backbone (O−O distance is 3.4 Å) compared to a nearby docked ChemScore pose (O−O distance is 3.2 Å). Again, the other main contribution to the difference in scores is a difference in clash (2.9 CS units less for the docked pose). For 1v4s, the difference in scores is a mixture of slightly poorer hydrogen bonds and larger penalties for clashes and internal energy. For 2bsm, an amide group of the docked ChemScore pose makes two hydrogen bonds in opposing directions, while the rescored GoldScore pose sits is shifted slightly so that one of the hydrogen bonds is shortened, while the other is missed. There is also a contribution of 2 CS units from clashes.

For the four poses whose scores increased by more than 5.0 CS units, the main effect was due to an additional hydrogen bond. These cases do not appear to be a failure of the rescore procedure to find the local minimum; rather it seems that even using an rmsd cutoff of 0.5 Å can result in poses which differ in their hydrogen or metal bonding interactions.

Given that clash values appeared to be overestimated for these four poses, we looked at the increase in clash values across all CS-docked poses within 0.5 Å of the GS-docked pose. The results (Figure 6) show a median decrease of only 0.6 CS units for the rescored poses indicating that the clash terms used by GoldScore and ChemScore are largely in agreement.

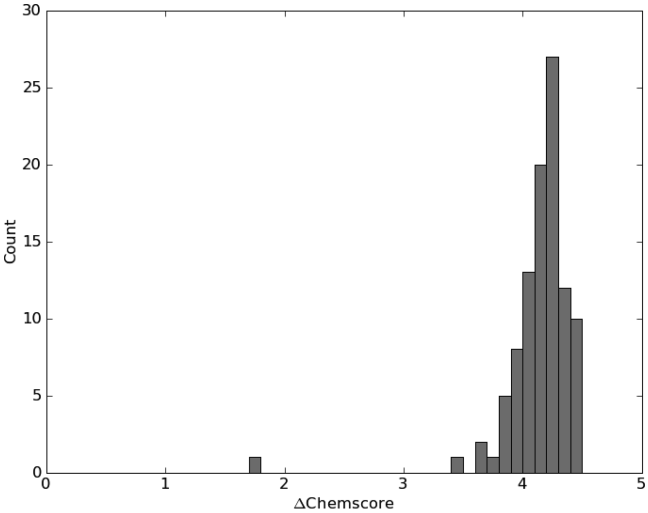

Figure 7. Histogram showing the improvement of the ChemScore values for 100 repetitions of a simplex optimization of the crystal structure of 1gkc. Bin size is 0.1 CS units.

It should also be noted that the simplex procedure itself is not deterministic, and the result can vary depending on the initial guess used in the procedure. This is illustrated by the histogram in Figure 7 which shows the improvement in scores (ChemScore) obtained in 100 repetitions of the simplexing procedure for the crystal structure of 1gkc. If the single outlier is ignored, the standard deviation of the scores is around 0.4, which corresponds to about 1/10 of the score contributed by an ideal hydrogen bond. The existence of the outlier, at about 2.5 units less than the mean, shows however that there is a small but finite probability of the algorithm failing to converge to the local maximum.

There are also other more subtle effects that can influence rescored values. For example, both GoldScore and ChemScore have a built-in concept of a perfect hydrogen bond, with deviations from this in terms of angles or distances receiving a lower score. However, the exact terms used by each scoring function are not identical. This means that a perfect hydrogen bond according to one scoring function might be suboptimal according to the other. This has consequences for rescoring since the original docked pose will have the protein rotatable hydrogens optimized for making the best hydrogen bond to the ligand as defined by the scoring function used in docking. Although the rescoring step includes simplex optimization of the ligand, the protein rotatable hydrogens will remain fixed, and it is possible that poorer hydrogen bonds will result.

On a final note, although we have highlighted a number of factors that can affect the accuracy of rescore values, it is important to remember that so long as a virtual screening experiment is a level playing field - that is, so long as the inactive molecules are rescored in the same way as the active molecules - if the scoring function works well, it should still be able to score the actives highly relative to the inactives.

CONCLUSIONS

This study is a first step in understanding the factors underlying success in a rescoring experiment. Although the complete picture is not yet clear, several aspects have been resolved which we hope can provide guidance to practitioners in the field.

The consensus hypothesis states that rescoring success is due to an averaging effect across the combined scoring functions such that only true positives score highly. However this hypothesis does not hold true except where a consensus score between the scoring and rescoring functions is calculated explicitly.

Rather, success appears to be explained by the complementary hypothesis, which states that success occurs where complementary scoring functions are combined such that the scoring function used for docking is better at ranking poses of the same molecule, while the rescoring function is better at ranking actives with respect to inactives. A scoring function that does not perform well when used on its own in a virtual screen will not yield good results if used to rescore. However as illustrated by the case of CS/rGS, such a scoring function may still yield very good results if its poses are rescored.

While we have focused on a broad overview in order to draw general conclusions regarding the underlying basis of rescoring, where the performance of individual scoring functions on a particular target or family of targets can be established this can guide the choice of a suitable scoring or rescoring strategy. The Astex Diverse Set, for example, has four families of proteins multiply represented. The best performing protocols are GS for the 5 serine proteases (average rank 2.1) and for the 11 kinases (6.1), ASP for the 3 phosphodiesterases (6.1), and CS/rGS for the 9 nuclear receptors (4.5). Work is currently ongoing using larger virtual screening data sets to establish the most appropriate docking and rescoring protocols for a range of pharmaceutically important target types.

REFERENCES AND NOTES

(1) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(2) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100–5109.

(3) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035–1042.

(4) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graphics Modell.* **2002**, *20*, 281–295.

(5) Willett, P. Enhancing the effectiveness of ligand-based virtual screening using data fusion. *QSAR Comb. Sci.* **2006**, *25*, 1143–1152.

(6) Hsu, D. F.; Chung, Y.-S.; Kristal, B. S. Combinatorial Fusion Analysis: Methods and Practices of Combining Multiple Scoring Systems. In *Advanced Data Mining Technologies in Bioinformatics*; Hsu, H.-H., Ed.; Idea Group Inc.: Hershey, PA, 2006; pp 32−62.

(7) Hsu, D. F.; Taksa, I. Comparing rank and score combination methods for data fusion in information retrieval. *Inf. Retrieval* **2005**, *8*, 449–480.

(8) Ng, K. B.; Kantor, P. B. Predicting the effectiveness of naïve data fusion on the basis of system characteristics. *J. Am. Soc. Inf. Sci.* **2000**, *51*, 1177–1189.

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

(10) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422–1426.

(11) Xing, L.; Hodgkin, E.; Liu, Q.; Sedlock, D. Evaluation and application of multiple scoring functions for a virtual screening experiment. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 333–344.

(12) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46*, 277–288.

(13) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 1134–1146.

(14) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved protein-ligand docking using GOLD. *Proteins* **2003**, *52*, 609–623.

(15) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337–356.

(16) Lyne, P. D.; Lamb, M. L.; Saeh, J. C. Accurate prediction of the relative potencies of members of a series of kinase inhibitors using molecular docking and MM-GBSA scoring. *J. Med. Chem.* **2006**, *49*, 4805–4808.

(17) Jones, G.; Willett, P.; Glen, R. C. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **1995**, *245*, 43–53.

(18) Hartshorn, M. J.; Verdonk, M. L.; Chessari, G.; Brewerton, S. C.; Mooij, W. T. M.; Mortenson, P. N.; Murray, C. W. Diverse, high-quality test set for the validation of protein-ligand docking performance. *J. Med. Chem.* **2007**, *50*, 726–741.

(19) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.

(20) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.

(21) Mooij, W. T. M.; Verdonk, M. L. General and targeted statistical potentials for protein-ligand interactions. *Proteins* **2005**, *61*, 272–287.

(22) O'Boyle, N. M.; Brewerton, S. C.; Taylor, R. Using buriedness to improve discrimination between actives and inactives in docking. *J. Chem. Inf. Model.* **2008**, *48*, 1269–1278.

(23) Meng, E. C.; Gschwend, D. A.; Blaney, J. M.; Kuntz, I. D. Orientational sampling and rigid-body minimization in molecular docking. *Proteins* **1993**, *17*, 266–278.

(24) Cole, J. C.; Murray, C. W.; Nissink, J. W.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins* **2005**, *60*, 325–332.

(25) Dietterich, T. G. Machine-learning research: four current directions. *AI Mag.* **1998**, *18*, 97–136.

CI900164F