ELSEVIER

Consensus scoring for protein–ligand interactions

Miklos Feher

Campbell Family Institute for Breast Cancer Research, University Health Network, Toronto Medical Discovery Tower, 101 College Street, Suite 5-361, Toronto, Ontario, M5G 1L7, Canada

This article reviews the application of consensus scoring for cases when the target 3D structure is known. Comparing the performance of different methods is not a trivial task, and it appears that consensus scoring usually substantially improves virtual screening performance, contributing to better enrichments. It also seems to improve – albeit less dramatically – the prediction of bound conformations and poses. The prediction of binding energies is still rather inaccurate and although consensus scoring generally improves these predictions, more development is required before it can be used for this purpose in routine lead optimization.

Virtual screening has become an essential component of the drug discovery process [1]. Its principal aim is the *in silico* assay of different chemical structures for the purposes of establishing their binding affinities, to separate active and inactive molecules or establish their rank order of activity. Virtual screening approaches can use either the 3D structure of the target (target-based virtual screening or docking) or use active and inactive ligands (ligand-based approaches) to determine and rank those structures most likely to bind. This review is primarily concerned with target-based virtual screening. In target-based virtual screening, the 3D structure of the target has been previously determined experimentally (usually by X-ray crystallography or NMR) or comes from computational modeling (e.g. based on protein sequence homology to targets with known 3D structures).

The use of scoring functions in target-based virtual screening has been extensively reviewed recently [2–7]. It usually proceeds in two steps. First, possible conformations and poses of the ligand in the pocket are computationally generated (the docking step) and, second, these are ranked according to how well they fit the pocket (the scoring step). Additionally, results from the scoring step are often applied to compare and rank different chemical structures (called database enrichment or focusing). Scoring functions are applied in all of these stages. In many cases, the same scoring function is used in all steps, although separate functions have increasingly been introduced for these roles. Although it might appear quite natural that combining results from different methods would lead to some improvement in performance, in reality this concept has only recently been applied in drug discovery. In fact, the first applications of consensus scoring were in ligand-oriented environments, namely QSAR [8–10] and molecular similarity [11], although the applicability of the approach in virtual screening was recognized relatively soon thereafter [12–15]. Despite its short existence, the application of consensus scores in drug discovery has now become common and, in addition to published methods, users can choose between commercial (e.g. CScore by Tripos, DS LigandScore by Accelrys, and Model-Composer, MOE by Chemical Computing Group) and publicly available (X-score: scoring function to predict protein–ligand binding affinities, available at http://sw16.im.med.umich.edu/software/xtool) approaches. In the majority of cases, it is employed as a postprocessing step after docking runs. At the very simplest, hit lists from multiple docking programs can be compared and those molecules that pass all, or the majority, of virtual screens can be selected, leading to a marked improvement in enrichment [12]. Many complex combinations of screening scores have been tested, a glossary of the most common approaches is provided in Box 1. The objective of this article is to review some of the most important advances in this relatively young field.

How can we compare different virtual screening results?

To judge the performance of any virtual screening approach, and therefore consensus scoring, it is important to consider the various

Corresponding author: Feher, M. (mfeher@uhnres.utoronto.ca)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2006.03.009

BOX 1

A glossary of popular consensus scoring approaches

Voting (intersection). Pass–fail criteria are established for each method, overall decision is based on how many passes a molecule has. These pass–fail criteria are often arbitrary. Good overall enrichment can be achieved but often at the cost of low recovery rates. The intersection approach is a special case of voting for methods with binary output.

Coarse quantiles (positions in ranges) voting. Each scoring function casts a vote if the score falls in the top quantile of the range of values obtained for that scoring function across the dataset of interest. The consensus score is the total number of votes received. When using a single criterion (e.g. best half or best third) there is a large number of tied scores, which can be overcome by also referencing another criterion, for example, obtaining ranks by the best third criterion and breaking ties based on the best half.

Rank voting. Each method has a predefined number of votes for activity and the top ranking compounds using each scoring function are assigned those votes. Performance depends on how many votes each method receives.

Simple sum ranks (rank-sum, also related to average rank, rank-by-rank). Entries are ranked using each scoring function and the ranks are added up (in case of average ranks, this sum is divided by the number of properties).

Deprecated sum-ranks. Entries are ranked based on each scoring function, the worst rank for each entry is dropped and the sum of the ranks is calculated from the remaining ranks.

Worst-best ranks. Compounds are ranked based on each scoring function and each entry is assigned the second worst rank from these. In case of a higher number of scoring functions, the third or fourth worst rank might be used.

Weighted sum ranks. Ranks from each method are weighted, with the weight reflecting the importance of the given method. The sum of the weighted ranks is then used as the final score.

Regression schemes (multilinear regression, non-linear regression). The total score is expressed as the linear (or nonlinear) combination of the individual scores. The coefficients are fitted, so that the performance (enrichment, binding energy, etc.) of this equation is optimal.

Multivariate methods. See textbooks (e.g. Ref. [43]) for a detailed discussion of principal component analysis (PCA), projection to latent structures (PLS), discriminant analysis and other related methods.

BOX 2

Factors to consider in the performance comparison of different target-based screening or prediction approaches

(i) Characteristics of the studied receptors (e.g. size, shape, polarity, or hydrophobicity, of the active site), the number and diversity of the receptors considered, as well as the quality of the 3D receptor structure (e.g. resolution).

(ii) Preparation of the active site (e.g. protonation and minimization) and whether this is appropriate for the given docking program. Also, consider whether multiple possibilities for certain residues have been considered (e.g. residue ionization or tautomer possibilities for His, Asp and Glu residues).

(iii) In training- and test-set composition, how different are the actives and inactives? Can they be easily separated using only 1D measures (e.g. size or lipophilicity)?

(iv) Is it ensured that the ligand is docked in the 'correct' pose before scoring and how are multiple conformers and poses from the docking program handled (e.g. score all binding modes and choose best scoring one, or choose a single mode, assuming that it is the best-predicted pose).

(v) The type of docking method and scoring functions used. Systematic errors can be introduced if badly docked ligands are rescored with several scoring functions. Rescoring ligands with a scoring function different from the 'native' one can produce different results owing to slight differences in ligand placement and differences in predocking preparation of the receptor for which the native scoring function had been optimized. However, in case of certain combinations this approach seems to work well, for example, combining a scoring function that finds poses correctly with one that ranks them well. Differences also arise depending on whether and how the ligand is relaxed before the second scoring function is applied.

(vi) How were the results evaluated? Which solutions were accepted? Was the computationally top-scoring solution taken (most rigorous but hardest test), a selection from among the top scoring solution (e.g. the one closest to the experimental data among the ten top solutions) or simply the solution closest to the experimental data (easiest and least rigorous approach)?

parameters that can influence their operation. Although the important factors can differ depending on the application, there are some commonly recognized issues (these are summarized in Box 2). These generally relate to the preparation of the receptor, the composition of the training and test sets and the evaluation of the raw results. Differences in treatment of these factors can lead to big variations in performance and might make it difficult to evaluate the results. For example, possibly the most often applied approach in the literature for testing enrichments is using a large decoy set seeded with a few known actives. It has been argued recently that this process often leads to artificially high enrichment scores if the decoy molecules have different 1D properties – such as molecular weight or lipophilicity – from the actives [16]. Thus, for example, if the dataset contains inactives that are much larger in size than the actives and the target can only accommodate small ligands, the docking program might simply act as a size filter and any observed high performance might not reflect the ability of the scoring function to recognize the important ligand features responsible for activity. Although validation-set composition might depend on the particular application, it is generally more meaningful to measure performance by including only those inactives that are 'reasonably similar' to the actives acting on the given receptor. Also, to achieve high enrichments, it is generally accepted that molecules must be docked correctly, although some believe this to be less important [17].

Many scoring functions have been developed and optimized to work with specific docking algorithms and applying them to score results from other docking programs can lead to inaccuracies and errors. This arises because distances between some ligand and receptor atoms can vary when using different docking programs,

TABLE 1

Selected examples to illustrate the performance improvement when using consensus scoring[a]

Measure	Approach	Single methods	Consensus method	Ref.
(a) Enrichments				
Hit rates (%)	Intersection using three scoring functions	3	18	[12]
Hit rates (%)	Intersection using three scoring functions	10	65–70	[17]
Top compounds containing all actives (%)	Voting using three scoring functions	20	8.4	[18]
(b) Poses				
Ligands with top docked pose within 2Å of the crystal structure (%)	ConsDock	39–56	60	[15]
Ligands with top docked pose within 2Å of the crystal structure (%)	Average rank using three functions	66–76	80–84	[24]
(c) Binding energies				
Rank correlation of predicted and experimental binding energies	Sum-rank	0.13–0.92	0.54–0.85	[14]
Rank correlation of predicted and experimental binding energies	CScore	0.13–0.92	0.60–0.86	[14]
Correlation (r^2) between predicted and experimental binding energies	Average rank	0.16–0.32	0.34	[29]
Correlation (r^2) between predicted and experimental binding energies	PLS	0.10–0.56	0.68	[31]
RMS error (kJ/mol) between predicted and experimental binding energies	Average rank	3.00–4.93	2.49	[29]

[a] Numbers presented are not meant for comparing different consensus scoring methods, only to give a sense of their expected performance. The actual numbers depend at least as much on the performance of the docking engines and the datasets used as on their consensus arrangement. See also Box 2 for issues to consider when comparing methods.

and the applied scoring functions can be sensitive to these differences. For this reason, it is important to distinguish the consensus of results from different docking experiments and the consensus of several scoring functions applied for a single docking experiment (rescoring). Similarly, relaxing the ligand in the pocket before scoring might improve or harm the accuracy of the results, depending on the way the scoring function was originally developed and optimized.

As mentioned earlier, consensus scoring is usually applied for virtual library enrichment, predictions of binding poses and binding affinity. Clearly, the separation of these areas is artificial. Nonetheless, for clarity, the simplest consensus methods will be discussed using these application categories. To demonstrate the kinds of gains expected when using these methods, the performance of selected examples is shown in Table 1.

Virtual library enrichment and focusing

Virtual library enrichment and focusing is currently the primary application of consensus scoring. It was first described by Charifson *et al.* [12] who, using the intersection approach, observed better and more consistent hit rates than those obtained using individual scoring functions. The authors also identified the disadvantage of their intersection approach: the intersection of hit lists is, by definition, smaller than the original lists. Hence the number of consensus hits might be small when many scoring functions are used, especially in the case of highly nonoverlapping hit lists. The authors partly attributed their observed low false-positive rates to their careful property filtering procedure of the actives and the randomly selected inactives before virtual screening.

Stahl and Rarey [13] also tested the intersection approach and observed an improvement in hit rates and a significant decrease in the actual number of hits when combining several scoring functions. Therefore, instead of combining the scores, they combined selected individual terms from these functions. Specifically, they combined the localized and directed FlexX hydrogen bonding terms with the hydrophobic terms in the PLP function. Although, on average, they observed some improvement compared to the individual constituent scoring functions, there was a marked decrease in hit rates for polar targets. Also, when tested on a dataset of 200 complexes, this combined scoring function was inferior to a single scoring function (DrugScore).

To overcome some of the issues with the intersection approach, Clark *et al.* [14] investigated other consensus techniques. They tested three rank-based methods (sum-ranks, worst-best ranks and deprecated sum-ranks) as well as CScore, a method based on coarse quantiles (see Box 1 for a description of these methods). The performance of these approaches was established using Spearman rank correlation of the consensus scores with experimental affinities. During the generation of the individual scores, only those conformations that were within a certain root mean square (RMS) distance from the experimental crystal conformation were considered. The authors found that CScore and sum ranks were more robust and reliable than any of the individual scoring functions. CScore was found to be a reasonable compromise; although the intersection approach has a lower false positive rate than CScore, this is often at the cost of not being able to identify any binders for certain receptors. In comparison, using a three-vote CScore leads to a modest increase in false-positive rate but at the same time a reduced false-negative rate (i.e. fewer missed leads). The CScore approach was also found to improve the affinity ranking of protein–ligand complexes with ligand affinities in the nanomolar to micromolar range.

Bissantz *et al.* [17] evaluated several popular scoring functions and their combinations and concluded that a voting scheme leads to

substantial improvements in hit rates. They found that when using a consensus list common to two scoring functions, the hit rate among the top 5% scorers increased from 10% to 25–40% and by using three scoring functions, hit rates in the 65–70% range were obtained. They recommended a two-stage protocol for large datasets. In the first step, the optimal docking and consensus scoring scheme should be established on a small subset of the data and, in the second step, this combination would be applied to the full dataset. They concluded that even though the hit rates were high, the method was not yet suitable for lead optimization because the correct pose and binding energy predictions were still relatively unreliable.

Krovat and Langer [18] evaluated different scoring functions and their combinations seeking full recovery of their actives. Their dataset comprised ten low or subnanomolar inhibitors and 1000 assumed inactives, generated as a diverse virtual combinatorial library. From the seven scoring functions they considered, the best four retrieved all of the actives in the top 20% of the compound database. It was found that by using a voting scheme, nine out of ten hits were among the top 1.4% of the database. As expected, slightly worse enrichments were obtained when only three scoring functions were applied in the voting scheme; however, this did allow the retrieval of all the actives in the top 8.4% of the database.

For phosphodiesterase-4B ligands, Mpamhanga *et al.* [19] attempted to combine three scoring functions on a 'rational basis' using average ranks to bring together one knowledge-based, one empirical and one force-field-based function. They found that such a combination generally performed better than most random combinations. They also studied the performance of different scoring functions across different chemical classes and observed that there is a natural bias towards the chemotype of the ligand present in the crystal structure. It was concluded that consensus scoring is an appropriate tool in the decision about the suitability of compound classes for lead optimization.

Bissantz *et al.* [20] applied consensus scoring to screen homology models of G-protein-coupled receptors. The process involved generating pairwise combinations of hit lists (and combinations of three ranking lists) using several scoring functions in a voting arrangement where ligands in the top 15–25% of the database for each scoring function received a vote, depending on the quality of the receptor model. It was concluded that antagonist receptor models are suitable for simple virtual screening, even if the binding site is generated using single-ligand energy minimization, whereas the virtual screening of agonists requires a more complex procedure. In the antagonist case, substantial improvements in hit rates were observed when using consensus scoring. The application of consensus scoring for homology models has also been reported by other authors [21,22].

Lyne *et al.* [23] recently undertook a study in which they applied consensus scoring by considering several docked poses for Chk-1 kinase. Each pose was scored using their z-score (Equation 1, where x is the raw score from the given scoring function, x_{av} is the average raw score for all the poses, and σ is the standard deviation of the raw scores for all the poses).

$$z = \frac{(x - a_{av})}{\sigma} \quad (1)$$

The consensus Z-score of a pose was then defined as the sum of z-scores over all applied scoring functions and this was applied in enrichment studies. It was concluded that 50–100 docked poses needed to be retained and scored using this consensus scheme to recover all the actives. The scoring functions and parameters applied were determined for a related kinase (Cdk-2) for which abundant data are available and were subsequently applied for selecting novel hits for Chk-1 with good efficiency.

Prediction of binding conformations and poses

Different consensus scoring schemes, studied by Clark *et al.* [14] for database enrichment, were described above. The same methods have also been tested to predict binding conformations or poses. It was found that, in most cases, the sum-rank, deprecated sum-rank and CScore (coarse quantiles) approaches can all be applied successfully for selecting conformations or poses that are similar to the crystal structure. It was argued that false-positives (i.e. previously undetected conformations that were ranked highly using consensus scoring) might also represent viable binding hypotheses because independent methods agree on their suitability and their thermodynamic accessibility, and hence might represent novel starting points in the design of new structural templates.

Paul and Rognan [15] developed a procedure called ConsDock for finding consensus poses of conformations generated by different docking programs. This method does not introduce its own score; instead it selects the most appropriate scores from the applied docking programs. ConsDock is a postprocessing procedure performed in four steps. First, all the poses generated by each docking tool are hierarchically clustered. These clusters are then used to identify consensus pairs of poses (i.e. binding hypotheses that are common between at least two programs). These consensus pairs are clustered into classes and each class is represented by a mean structure. The clusters are ranked according to the number of means they include. Because mean structures do not necessarily correspond to energy minima, they are minimized in a force field or, alternatively, a real docked structure closest to the mean is chosen. It was shown using 100 ligand–protein complexes that this consensus docking approach outperformed any single docking tool in the quality of the top ranked pose.

Wang *et al.* [24] compared 11 scoring functions and their combinations on 100 protein–ligand complexes to evaluate their performance in predicting experimental conformations and poses. In this process, each scoring scheme was applied to rank all conformations of the ligand and the final rank was the average of the ranks received from each scoring function. Using a maximum atomic deviation of 2 Å RMS between the experimental and predicted structure as a selection criterion, the success rates were 66–76% for the individual scoring functions, 76–80% for the double scoring systems and 80–84% for the triple schemes. Thus, the authors found that there was a clear improvement when using multiple scoring functions. They also found that the performance range (difference between best and worst performance) decreases, indicating that the actual selection of scoring functions is less important in multiple combinations.

It is worth mentioning at this point that there are significant developments in the area of using consensus scoring to predict ligand poses for flexible targets [25–27]. Although the consensus methods applied in such cases are often similar to those discussed above (e.g. sum ranks, deprecated sum ranks, regression methods), these are applied to find consensus among different receptor

conformations, rather than different scoring methods and, thus, they are not discussed further here.

Prediction of binding affinity

Methods to predict binding affinities have been extensively reviewed by Gohlke and Klebe [28] and the underlying scoring functions were discussed in a more recent article [4]. It has been pointed out that scoring functions usually describe binding affinity as a sum of independent terms and thus they all suffer from the same disadvantages [4]: size dependence (the larger the molecule, the higher the probability of a favorable score); they usually ignore entropic effects (rigid receptor and single ligand binding modes with no ensemble averages); and many scoring functions ignore solvation and desolvation effects.

It was concluded that most scoring functions assess receptor–ligand complementarity, rather than the free energy of binding. This complementarity is a prerequisite for ligand-binding but not a sufficient condition on its own. Because consensus scores are derived from individual scores that suffer from these shortcomings, these issues are unlikely to be addressed by such combinations.

Marsden *et al.* [29] tested two consensus methods for combining individual binding-energy predictions: average rank and multilinear regression, with the weights in the multilinear regression optimized to achieve minimum RMS error in the predicted distribution coefficient (log K_d). It was found that the application of both consensus combinations led to modest improvements in predictivity – at major computational expense. It appears that multilinear regression generally performs somewhat better if there are no major outliers, whereas the combination of ranks has a stronger inherent averaging effect, reducing the effect of major outliers. However, it was concluded that even with consensus scoring, binding energy predictions were still quite unreliable: on a set of ~100 ligand–receptor complexes, average rank and multilinear regression consensus scores had Pearson correlation coefficients (r^2) = 0.33 and 0.36 and RMS errors in binding energy of 2.49 kJ/mol and 2.64 kJ/mol, respectively.

Instead of using established scoring functions, Wang *et al.* [30] developed three novel ones, each of which approximated hydrophobic interactions differently. These functions were, of course, highly correlated with each other. The authors argued that as separate scoring functions, their predictions for protein–ligand complexes would be different because they recognized different geometric features of the target. Each of the scoring functions was individually optimized and tested on a set of 200 protein–ligand complexes. This showed that in only ~40% of complexes was the difference between the highest and lowest predictions less than 0.5 pK_d units, and in almost 20% of the compounds it was greater than 1.0 pK_d unit. The authors, not too surprisingly, found that if the binding affinity predictions closest to the experimental value were correlated with the experimental values, the standard deviation of the prediction greatly improved in comparison with using predictions with any single scoring function. To generate a consensus score, the three scoring functions were averaged. Although the average score, by definition, was always a medium performer, it had two advantages: firstly, it helped to identify the expected level of accuracy of these functions and, secondly, it reduced the likelihood of large errors. The average accuracy of binding free energy predictions was ~2 kcal/mol. This scoring function has been made available to the public under the name X-Score.

Less common ways of generating consensus scores

Terp *et al.* [31] applied multivariate statistics [principal component analysis (PCA) and projection to latent structures (PLS)] to score and predict the best ligand poses in a pocket, as well as protein–ligand binding affinity. The selection of the binding conformation was based on the assumption that the best inhibitor conformation corresponds to the structure with the most favorable predicted binding energy. From results with eight scoring functions, a one-component PCA model was successfully applied to predict ligand binding modes in 18 different complexes. It was claimed that this process was superior to using a simple consensus score (CScore) because the latter does not distinguish between the top scoring solutions. For binding affinity predictions, the results from eight different scoring functions were combined and, after variable selection, a one-component PLS model was derived using the five most predictive scoring functions correlated with the experimental pK_i. The correlation coefficient for fitting all data points was $r^2 = 0.78$; significantly better than the performance of any single scoring function.

Jacobson *et al.* [32] have also applied multivariate statistics to generate classifiers that could discriminate between active and inactive compounds from virtual screens. Results from seven scoring functions were combined using PLS discriminant analysis, rule-based methods and Bayesian classification, all of which require an extra training set. In comparison with single scoring functions and traditional consensus scores, multivariate statistical methods were found to improve the discriminative power of consensus scoring, especially the rule-based method. Three reasons were given for this: the use of extra training set made the classifier more aware of the target; it was possible to include information on inactives; and the ability of multivariate statistics to take correlation between the scoring functions into account quantitatively. However, it must be noted that because the binding conformation of inactive molecules cannot be ascertained, it is questionable how much the consideration of inactive data might help in the process. Because these multivariate methods were unable to predict actual activities, it was suggested that they could be applied as a first step in a virtual screening cascade, reducing the number of molecules that need to be considered for more accurate and time consuming binding affinity predictions.

Seeing that the original scoring functions were unable to explain activity variations, Prathipati and Saxena [33] combined docking scores from the MOE and LUDI scoring functions using binary QSAR. Combination of the individual terms of the two scoring functions using PLS was attempted, but only marginally improved the results. A classification model was built using binary QSAR methodology trained on binders and nonbinders, using either the individual scoring function terms or the final scores from the two scoring functions – the individual scoring function terms provided superior performance. The discriminatory power of this combination appeared to be high. This combination was also compared to multivariate statistical methods [32], using a common dataset, and was found superior.

A naïve Bayes classifier has also been applied in combination with rank-by-median consensus scoring [34]. In this approach, the

median rank across five scoring functions was calculated. The top 1% of compounds were passed to a naïve Bayes classifier as 'good binders', the rest as 'bad binders'. It appears that, at least in the two studied examples, adding this classifier to the consensus scheme led to some minor improvement in the overall performance.

Mestres *et al.* [35] combined ligand- and target-based measures. In addition to docking, similarity to a reference ligand was considered using field-based similarity involving the electrostatic potential and molecular shape (the reference was either the cocrystallized ligand, or a pharmacophore-fitted ligand placed in the active site). Target and ligand-based scores were combined by ranking compounds on both measures and accepting the top 10% of the solutions. It was found that the similarity measure (MIMIC) and the scoring function (DOCK) agreed in almost half of the top scoring molecules. The proportion of false-positive and false-negative results using this consensus score was not investigated. It was concluded that the two methods might complement each other: 3D similarity to a reference structure might identify molecules more easily if a single receptor conformation or imperfect scoring functions present an obstacle, whereas docking is independent of the reference structure and, hence, will enforce complementarity with the receptor in regions not explored by the reference ligand.

A simple multiplication to combine ligand similarity and docking scores was suggested by Fradera *et al.* [36]. In this case, similarity essentially acts as a weighting factor for the docking results, generating a preference in the cavity for ligand orientations that are similar to those of the reference ligand. This should create a balance in which favorable interactions with the receptor are maintained without deviating too much from the observed binding mode of the reference ligand or the pharmacophore model. The combination of the docking and similarity scores can be performed either as part of the actual docking process (similarity-guided docking) or by rescoring all of the docked poses and conformations (similarity-penalized docking). It was found that the former approach yields better binding mode predictions for the example set of 32 thrombin inhibitors with an average improvement of ~1 Å RMS distance over the docking engine (DOCK) alone.

Binding energy predictions using QSAR and traditional scoring functions were compared and combined for glycogen phosphorylase inhibitors by So and Karplus [37]. The applied ligand-based methods included hologram QSAR, receptor-surface models, CoMFA and genetic neural networks. Target-based scoring functions were LUDI and a structure-based binding energy predictor using force fields, solvation and an estimate for entropy changes. It was found that by averaging the results obtained by different prediction methods, predictivity, as measured by the q^2 value, increases with the number of methods considered. In this context it is interesting to note that the five ligand-based methods performed considerably better than the two structure-based methods.

How does consensus scoring work?

Although consensus scoring has gained wide acceptance from the virtual screening community, relatively few studies deal with the question of how consensus scoring enriches datasets. As pointed out by Clark [14], if different scoring functions estimate a property independently and that property relates smoothly to an experimentally determined quantity (such as the free energy of binding or RMS distance from a crystal pose), then the mean from several scoring functions should be a better predictor than each individual score, if all functions have comparable precisions. Because different scoring functions can have different scales, their combination is nontrivial. Therefore, the use of ranks often improves performance compared with the combination of raw scores. Also, because each scoring function contributes something unique to the consensus (at least in an ideal case) but will miss certain effects, it might be beneficial if they have different weights in the final consensus score.

Wang and Wang [38] studied the behavior of consensus scores using idealized computer simulations. They attributed the performance increase in consensus scoring simply to the fact that the mean of repeated samplings tended to be closer to the true value than any single sampling. They also found that performance improvements were proportional to the square root of the number of applied scoring functions and that no improvement beyond the variance of the data can be expected if more than three or four scoring functions are simultaneously used. It was pointed out by Baber *et al.* [39] that some of the underlying assumptions might not be satisfied in realistic cases and, therefore, consensus scoring might lead to improvements beyond those seen from the simulations. They also pointed to an apparent contradiction: real scoring functions were likely to be somewhat correlated if they had better than random performance (they have an increased likelihood of selecting the same actives) but the applied methods must not be strongly correlated, particularly in their errors, otherwise they would no longer be independent. It is likely that properly considering the level of correlation when selecting methods that work well together is essential for improving operational performance in consensus scoring, especially by ensuring that systematic errors are not correlated between methods.

Although most authors observed significant improvements when using consensus scoring, cases with only marginal improvement have also been reported [40]. More surprisingly, some performance decrease has been observed recently when using consensus scoring [41]. This failure can be understood by recognizing that there was a systematic component to the error, common across all methods shown in Ref. [41] – a single docking engine was used to dock molecules, and hence errors in ligand placement affected all scoring functions. For example, any inactives for which 'good' poses were erroneously found were likely to be recognized by the different scoring functions as such, introducing some correlation in false positives. At the same time, there was little correlation in true positives between many of the applied scoring functions [41]. Similar problems might be expected in all cases when a single docked pose is rescored using several scoring functions.

Yang *et al.* [42] have recently investigated how consensus scoring operates, based on four receptors and a dataset comprising ten actives and 990 random compounds. In their study, they looked at all combinations of five different scores derived from two docking programs. They found that although the average accuracy improves with the number of applied methods, maximum accuracy occurred when pairs of methods were combined. This was not so surprising given that molecules were docked with two different methods and that all scoring functions originated from two families of related functions. They concluded that consensus scoring improved screening performance only when each of the individual scoring function had relatively good performance and the scoring characteristics of the individual functions were differ-

ent. They also laid out quantitative criteria for these conditions. Under these conditions, rank combinations were found to perform at least as well as other score combinations.

Baber *et al.* [39] recently summarized factors that might contribute to the performance of ligand-based consensus scoring; most of the conclusions are also likely to apply to the target-based process. These include:

(i) Statistical improvement, arising from the fact that the mean of repeated samplings is closer to the true value than the individual measurements.

(ii) Actives are clustered more tightly than inactives and thus multiple samplings will recover more actives than inactives.

(iii) Ranking concordance – different methods agree more on the ranking of actives than on the ranking of inactives. This arises as different scoring functions focus on different aspects of ligand-binding (or have individual deficiencies associated with them) and thus lead to different false positives.

(iv) Consensus results are not only better but also more consistent across receptor systems, and the user is thus less dependent on picking the correct method for each receptor.

Conclusions

Since its first application less than a decade ago, target-based consensus scoring has come a long way and has become an important tool in the computer-aided discovery of new pharmaceutics. Through the number of published and commercially available methods, it has become easily accessible to practicing computational chemists. The majority of published work indicates that consensus scoring leads to tremendous improvements through the improved quality of the results and through less reliance on the choice of the particular scoring function. The amount of improvement varies greatly between different studies and depends on the application, the consensus method and the conditions of the computational tests. It appears likely that the scoring functions themselves must be reasonable; a blind combination of some arbitrarily chosen scoring functions will not necessarily lead to better results. However, it is still somewhat unpredictable how much improvement can be expected from a combination of different scoring functions and how different factors affect this improvement. It is expected that, as the field of consensus scoring of virtual screening results matures, a greater understanding of these effects will emerge.

Acknowledgements

Christian Baber (Neurocrine Biosciences, San Diego, CA, USA) is acknowledged for the careful reading of the manuscript.

References

1 Jorgensen, W.L. (2004) The many roles of computation in drug discovery. *Science* 303, 1813–1818

2 Krovat, E.M. *et al.* (2005) Recent advances in docking and scoring. *Curr. Comp. Aided Drug Des.* 1, 93–102

3 Mohan, V. *et al.* (2005) Docking: successes and challenges. *Curr. Pharm. Des.* 11, 323–333

4 Schulz-Gasch, T. and Stahl, M. (2004) Scoring functions for protein-ligand interactions: a critical perspective. *Drug Discov. Today: Technol.* 1, 231–239

5 Kitchen, D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949

6 Lyne, P.D. (2002) Structure-based virtual screening: an overview. *Drug Discov. Today* 7, 1047–1055

7 Kitchen, D.B. *et al.* (2004) Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* 3, 935–948

8 Ajay, (1994) On better generalization by combining two or more models: a quantitative structure-activity relationship example using neural networks. *Chemometr. Intell. Lab.* 24, 19–30

9 So, S. and Karplus, M. (1996) Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks. *J. Med. Chem.* 39, 1521–1530

10 Rogers, D. and Hopfinger, A.J. (1994) Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J. Chem. Inf. Comput. Sci.* 34, 854–866

11 Ginn, C.M.R. *et al.* (1996) Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* 37, 23–37

12 Charifson, P.S. *et al.* (1999) Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* 42, 5100–5109

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

14 Clark, R.D. *et al.* (2002) Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* 20, 281–295

15 Paul, N. and Rognan, D. (2002) ConsDock: A New Program for the Consensus Analysis of Protein-Ligand Interactions. *Proteins* 47, 521–533

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

17 Bissantz, C. *et al.* (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* 43, 4759–4767

18 Krovat, E.M. and Langer, T. (2004) Impact Of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. *J. Chem. Inf. Comput. Sci.* 44, 1123–1129

19 Mpamhanga, C.P. *et al.* (2005) Retrospective docking study of PDE4B ligands and an analysis of the behavior of selected scoring functions. *J. Chem. Inf. Model.* 45, 1061–1074

20 Bissantz, C. *et al.* (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G-potein coupled receptors suitable targets? *Proteins* 50, 5–25

21 Liu, Z. *et al.* (2005) Virtual screening of novel noncovalent inhibitors for SARS-CoV 3C-like proteinase. *J. Chem. Inf. Model.* 45, 10–17

22 Ashton, M. *et al.* (2004) The selection and design of GPCR ligands: from concept to the clinic. *Comb. Chem. High Throughput Screen* 7, 441–452

23 Lyne, P.D. *et al.* (2004) Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *J. Med. Chem.* 47, 1962–1968

24 Wang, R. *et al.* (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* 46, 2287–2303

25 Cavasotto, C.N. and Abagyan, R.A. (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* 337, 209–225

26 Barril, X. and Morley, S.D. (2005) Unveiling the full potential of flexible receptor docking using multiple crystallographic structures. *J. Med. Chem.* 48, 4432–4443

27 Merlitz, H. *et al.* (2004) Impact of receptor conformation on *in silico* screening performance. *Chem. Phys. Lett.* 390, 500–505

28 Gohlke, H. and Klebe, G. (2002) Approaches to the description and prediction of binding affinity of small molecule ligands to macromolecular receptors. *Angew. Chem. Int. Ed. Engl.* 41, 2644–2676

29 Marsden, P.M. *et al.* (2004) Predicting protein-ligand binding affinities: a low scoring game? *Org. Biomol. Chem.* 2, 3267–3273

30 Wang, R. *et al.* (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comp. Aided Mol. Des.* 16, 11–26

31 Terp, G.E. *et al.* (2001) A new concept for multidimensional selection of ligand conformations (MultiSelect) and multidimensional scoring (MultiScore) of protein-ligand binding affinities. *J. Med. Chem.* 44, 2333–2343

32 Jacobsson, M. *et al.* (2003) Improving Structure-Based Virtual Screening by Multivariate Analysis of Scoring Data. *J. Med. Chem.* 46, 5781–5789

33 Prathipati, P. and Saxena, A.K. (2006) Evaluation of binary QSAR models derived from LUDI and MOE scoring functions for structure based virtual screening. *J. Chem. Inf. Mod.* 46, 39–51

34 Klon, A.E. *et al.* (2004) Combination of a naïve Bayes classifier with consensus scoring improves enrichment of high-throughput docking results. *J. Med. Chem.* 47, 4356–4359
35 Mestres, J. and Knegtel, R.M.A. (2000) Similarity versus docking in 3D virtual screening. *Persp. Drug Disc. Des.* 20, 191–207
36 Fradera, X. *et al.* (2000) Similarity-driven flexible ligand docking. *Proteins* 40, 623–636
37 So, S. and Karplus, M. (1999) A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Comp. Aided Mol. Des.* 13, 243–258
38 Wang, R. and Wang, S. (2001) How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment *J. Chem. Inf. Comput. Sci.* 41, 1422–1426
39 Baber, C. *et al.* (2006) The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model* 46, 277–288
40 Cummings, M.D. *et al.* (2005) Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* 48, 962–976
41 Xing, L. *et al.* (2004) Evaluation and application of multiple scoring functions for a virtual screening experiment. *J. Comp. Aided Mol. Des.* 18, 333–344
42 Yang, J.-M. *et al.* (2005) Consensus scoring criteria for improving enrichment in virtual screening. *J. Chem. Inf. Model* 45, 1134–1146
43 Livingstone, D. (1995) *Data analysis for chemists. Applications to QSAR and chemical product design.* Oxford University Press