How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment

Renxiao Wang and Shaomeng Wang*

Institute of Cognitive and Computational Science and Lombardi Cancer Center, Departments of Oncology and Neuroscience, Georgetown University Medical Center, 4000 Reservoir Road, Washington, D.C. 20007

Received March 15, 2001

It has been reported recently that consensus scoring, which combines multiple scoring functions in binding affinity estimation, leads to higher hit-rates in virtual library screening studies. This method seems quite independent to the target receptor, the docking program, or even the scoring functions under investigation. Here we present an idealized computer experiment to explore how consensus scoring works. A hypothetical set of 5000 compounds is used to represent a chemical library under screening. The binding affinities of all its member compounds are assigned by mimicking a real situation. Based on the assumption that the error of a scoring function is a random number in a normal distribution, the predicted binding affinities were generated by adding such a random number to the "observed" binding affinities. The relationship between the hit-rates and the number of scoring functions employed in scoring was then investigated. The performance of several typical ranking strategies for a consensus scoring procedure was also explored. Our results demonstrate that consensus scoring outperforms any single scoring for a simple statistical reason: the mean value of repeated samplings tends to be closer to the true value. Our results also suggest that a moderate number of scoring functions, three or four, are sufficient for the purpose of consensus scoring. As for the ranking strategy, both the rank-by-number and the rank-by-rank strategy work more effectively than the rank-by-vote strategy.

INTRODUCTION

Structure-based virtual screening of chemical libraries has become an extremely valuable tool for identifying lead compounds in the case that the three-dimensional structure of the target has been determined.[1] This approach helps to narrow the size of the chemical library under investigation and thus results in a generally improved efficiency in drug discovery. With the completeness of the human genome sequencing, the number of potential pharmaceutical targets will increase dramatically.[2] Undoubtedly this technique will become even more useful in the drug discovery processes in the future.

Flexible docking programs are usually adopted in virtual library screening approaches to predict the receptor−ligand binding modes. Examples of the most popular docking programs are DOCK,[3] AutoDock,[4] GOLD,[5] and FlexX.[6] Although the conformational searching is usually addressed adequately in a docking program, the scoring method remains as a weakness. A scoring method implemented in a docking program is expected to meet two different purposes: during the docking process, it is used to detect the correct bound conformation among the false ones; after the completion of docking, it is used to estimate the binding affinities of the candidate molecules. Furthermore, since library screening usually needs to process hundreds of thousands of compounds in a relatively short time period, the scoring method has to be fast enough. The scoring methods used in docking programs nowadays are based on either force field calculations,[3−5] empirical scoring functions,[6−9] or knowledge-based potential of mean force.[10−12] It is true that these scoring methods have been applied successfully to many drug discovery projects. However, due to the inadequate understanding of the elegant physics embedded in the receptor−ligand binding process, these scoring methods are still far from accurate. Finding better scoring methods are still a haunting goal for the researchers who are working on structure-based drug design.

A very interesting method called consensus scoring appeared recently.[13] In such an approach, a docking program is still employed to fit compounds to the target receptor. However, the best docked conformer of each compound is reevaluated with multiple scoring functions. Only the top scored compounds common to each scoring function will be identified as candidates for bioassay. Compared to single scoring procedure, it was shown that false positives in virtual library screening were largely reduced and hence the hit-rates were improved. More applications of consensus scoring have begun to appear in the literature.[14,15] It seems that this method is quite independent to the target receptor, the docking method, or even the scoring functions under investigation. In fact, some commercial molecular modeling software, such as SYBYL,[16] has included consensus scoring as part of the drug design package.

Consensus scoring thus appears to have a clear advantage in structure-based virtual library screening approaches. All the consensus scoring approaches reported so far focus on testing a number of scoring functions on some specific biological targets. This kind of approach is valuable if one wants to explore the most efficient combination of scoring

* Corresponding author phone: (202)687-2028; fax: (202)687-4032; e-mail: wangs@georgetown.edu.

10.1021/ci010025x CCC: $20.00
Published on Web 07/17/2001

functions in a particular drug design project. However, since there are many scoring functions in existence and their performance actually varies from case to case, it is not clear whether the conclusions drawn from one project are transferable to another project. In addition, although consensus scoring is repeatedly shown to work more robust than any single scoring, its mechanism has not been adequately addressed.

In this paper we present an idealized computer experiment designed to investigate the theoretical basis of consensus scoring and explore the best strategy of combining multiple scoring functions in virtual library screening. A hypothetical set of 5000 compounds is used to represent a chemical library under screening. The binding affinities of all its member compounds to a certain biological target are assigned by mimicking a real situation. Based on the assumption that the error given by a scoring function in binding affinity estimation is a random number in a normal distribution, the calculated binding affinities of these compounds are also generated by computer. Then, with all these simulated binding affinities, the performance of the consensus scoring procedure is explored.

METHODS

Basic Assumptions. The central idea of our approach is to design an idealized procedure to simulate a virtual library screening, in which no specific biological target, chemical library, scoring function, or docking program is involved. Some basic assumptions therefore must be applied here. (i) All the compounds under screening are perfectly docked to the target. The error occurred in binding affinity estimation is solely determined by the accuracy of the scoring function. (ii) The error given by any scoring function in binding affinity estimation is a random number in a normal distribution. The normal distribution is centered at zero, which means there is no systematic error. For the convenience of computation, the accuracy of all the scoring functions is assumed to be at the same level. (iii) All the scoring functions are independent to each other. The error from one scoring function has no correlation with the error from another.

Data Preparation. A hypothetical chemical library that contained 5000 compounds was constructed first. Each individual candidate was assigned a floating random number between 0.00 and 10.00, representing its experimentally determined binding affinity to a certain biological target (in $-\log K_i$ units, i.e., negative logarithm of the dissociation constant). This set of random numbers, $X(i)$ ($i = 1, 2, ..., 5000$), will be referred as the "*experimental data set*" in this paper. To reflect the common sense that most compounds in a chemical library are inactive to the target, these random numbers were generated from a normal distribution $N(0.00, 4.00^2)$. Therefore, only a small portion of them shows high binding affinity (see Figure 1).

Another set of data, $Y(i)$, was generated based on the experimental data set. Each number in this data set was generated by adding a random number, $E(i)$, to the corresponding one in the experimental data set, i.e., $Y(i) = X(i) + E(i)$. The number $E(i)$ was generated from a normal distribution $N(0.00, 2.00^2)$. This set of data, $Y(i)$, represented the predicted binding affinities for the 5000 candidates given by a certain scoring function. Note that the mean value of $E(i)$ is 0.00 because the scoring function is supposed to have no systematic error. The variance of 2.00 log units reflects the average accuracy level that existing scoring functions can reach nowadays. Such a data set will be referred as "*predicted data set*" in this paper. In our experiment, 10 predicted data sets were independently generated in this way, i.e., $Y_1(i)$, $Y_2(i)$, ..., $Y_{10}(i)$. They were considered as the predicted results given by 10 individual scoring functions.

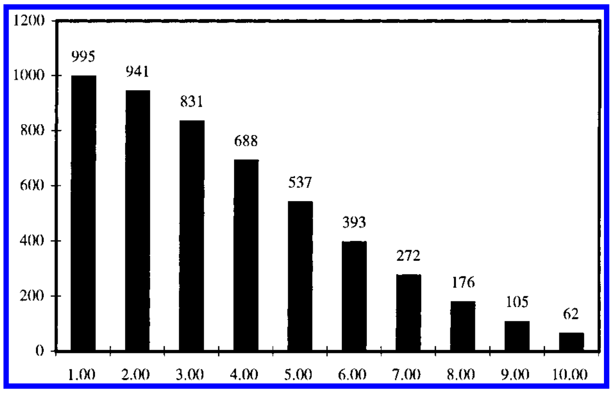

Figure 1. Distribution histogram of the experimental data set (X axis: binding affinity; Y axis: population).

Data Analysis. A predicted data set will, of course, deviate from the experimental data set. Therefore, objective criteria for evaluating the overall quality of a given predicted data set are needed. Our first method is to count the total number of "misranks". To count this number, first we ranked all the candidates according to the predicted values, and then we checked every each pair of candidates to see whether it was ranked correctly. For example, suppose candidate i ranks higher than candidate j in the experimental data set. While if candidate i ranks lower than candidate j in the predicted data set, it will be considered as a "misrank". For all 5000 candidates, there are $5000*4999/2 = 12\,497\,500$ pairs in total. We checked them all and counted the number of misranks. Obviously, the fewer misranks are observed, the better is the prediction. In the case that multiple scoring functions were combined to make prediction (consensus scoring), we simply ranked the candidates according to the average predicted results given by all the scoring functions and counted misranks in the same way.

Another method to evaluate the overall quality of a predicted data set is to count the hit-rates, which relates more directly to "real" library screening approaches. We defined arbitrarily that the top 2% candidates in the experimental data set were truly active compounds (hits). Since the size of the data set is 5000, there are exactly 100 of them. By checking the candidates appearing on the top of a predicted data set, e.g. top 2%, the number of true hits can be counted. In this way, the higher is the hit-rate, the better is the prediction. For consensus scoring, there is a special problem in how to rank candidates according to the predicted results since more than one scoring function is involved in this process. In our experiments, we have tested three ranking strategies. (1) All the candidates are ranked according to the average predicted values given by all the scoring functions. This is a straightforward way to combine multiple scoring functions in prediction. We will refer this as the "*rank-by-number*" strategy in this paper. (2) All the candidates are ranked by the average ranks predicted by all the involved scoring functions. For example, if a candidate ranks no. 10

according to scoring function A and ranks no. 20 according to scoring function B, then its average rank will be (10+20)/2 = 15. This strategy uses relative ranks rather than absolute binding affinities for ranking. We will refer this as the "*rank-by-rank*" strategy in this paper. (3) If a candidate is predicted to be on the top, e.g. 2%, by a certain scoring function, then it gets a "vote" from that scoring function. The final score of a candidate compound is the number of votes gathered from all the scoring functions, which may range from 0 to the total number of scoring functions. All the candidates will be ranked according to their final scores. We will refer to this as the "*rank-by-vote*" strategy in this paper.

All the possible scoring procedures, i.e., from single scoring to decadal scoring, have been tested. All the methods described above were used for data analysis. Another point should be addressed here: since we have used random numbers to generate the experimental data set and the predicted data sets, each experiment (experimental data set generation + 10 predicted data sets generation + data analysis) was repeated for 100 times to flatten the fluctuations in the statistical results. If not specified, all the data reported in this paper are the average results out of 100 individual runs.

RESULTS AND DISCUSSION

Rationale in an Idealized Computer Experiment. Before we proceed to present and analyze the results, we must explain the rationale in our idealized computer experiment: how closely does it resemble the reality? Let us go through the basic assumptions underlying our simulation one by one.

First, we assume that the docking procedure is perfect and the errors in binding affinity estimation are raised all by scoring functions. As mentioned in the Introduction section, in practice consensus scoring is performed after the docking procedure is completed. The orientation/conformation of the docked compound will certainly affect the result in binding affinity calculation. However, exploring how the docking procedure affects the performance of a virtual library screening approach is another issue. To concentrate on studying the consensus scoring itself, this assumption is essential.

Second, we assume that the error of each scoring function in binding affinity estimation is a random number in a normal distribution that is centered at zero. This is a reasonable assumption since an ideal scoring function should be so. A real scoring function is close to this ideality, too. Existing scoring functions are usually validated by a variety of receptor−ligand complexes. In a statistical sense, they are able to reproduce the known binding affinities with an acceptable accuracy. Furthermore, in practice one can test all his scoring functions in advance to see which ones work well for the given biological target and only apply to the good ones. By doing so, scoring functions that exhibit significant systematic errors will not enter the consensus scoring procedure.

Third, we assume that all the scoring functions involved in consensus scoring are independent to each other. This will not be a problem if the scoring functions have different origins: for example, one is an empirical scoring function, while the other is a potential of mean force approach. For the scoring functions falling in the same category, they are not necessarily correlated either. For example, the empirical scoring functions may use different master equations to describe the binding free energy. Even for the same term in the master equation, such as hydrogen bonding and hydrophobic effect, the algorithms vary. For the potential of mean force approaches, they may differentiate each other in many aspects, such as choosing atom types, setting preference state, deriving potentials, and so on. In both cases, it is not unusual that the difference is more than trivial. Therefore it is reasonable to believe that the internal independence of all the involved scoring functions is not a demanding request. For our idealized computer experiment, it is not a problem at all. However, in a consensus scoring practice, special attention should be paid to the selection of scoring functions: if one scoring function is simply an enhanced or simplified version of another, one would better to avoid including them both.

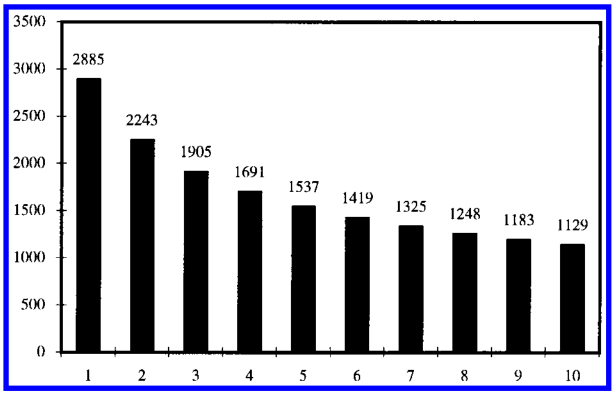

Figure 2. Relationship between the number of scoring functions used in consensus scoring (X axis) and the total number of misranks (in thousand) observed in the predicted data set (Y axis).

Given these assumptions, the consensus scoring procedure can then be simulated with an idealized computer experiment. The beauty of such an experiment is that it is based purely on numbers. Not a single specific object, such as the biological target molecule, the chemical library, the docking program, or the scoring function, needs to be involved. As a result, the conclusions derived from this experiment will not be biased to any of them.

How Does Consensus Scoring Work? We have counted the numbers of misranks observed in the predicted data sets from single scoring to decadal scoring. The results are shown in Figure 2. There is a clear trend that the overall errors in prediction decrease with the increase in the number of scoring functions. Compared to the results given by single scoring, double scoring drops the number of errors by 23%, triple scoring by 34%, and quadruple scoring by 41%. The number of errors continues to drop when even more scoring functions are employed in prediction. However, the dropping speed will slow significantly after three or four scoring functions are applied. We have performed the same kind of experiment with data sets sizing from 1000 to 10 000. It was found that the above observations are independent to the size of the data set.

If one accepts the basic assumptions underlying our simulation, then a simple explanation for the performance of consensus scoring exists: the mean value of repeated samplings tends to be closer to the true value than any single sampling. Let random variables E_1, E_2, ..., E_n denote, respectively, the errors of n scoring functions in binding

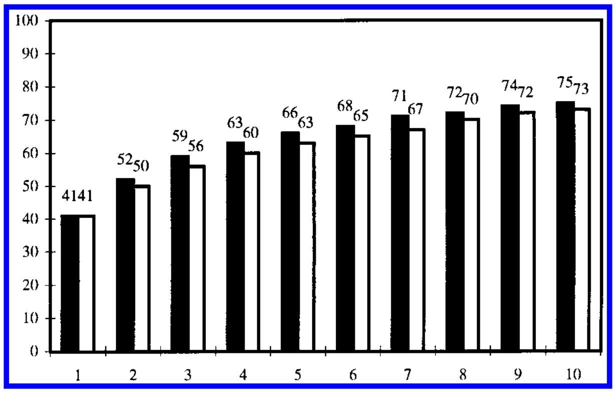

Figure 3. Relationship between the number of scoring functions used in consensus scoring (*X* axis) and the hit-rates observed among top 100 candidates (*Y* axis). (The solid bars denote for the "rank-by-number" strategy, while the opened bars denote for the "rank-by-rank" strategy.)

affinity prediction. According to the mathematical statistics theory,[17] the larger *n* is, the closer the maximum likelihood estimator ($\bar{E}$) will be to the expectation value (μ). Since $\bar{E}$ is also under a normal distribution, e.g. $N(\mu,\sigma^2/n)$, if we define ($\bar{E} - \mu \leq \sigma$) is a successful prediction and ask the question how many samplings are required to achieve a successful prediction with the confidence interval of 95%, we need to solve the following equation:

$$\Pr\left(-\delta < \frac{\bar{E} - \mu}{\sigma/\sqrt{n}} < \delta\right) = 0.95 \qquad (1)$$

By looking up the normal distribution integration table,[17] the value of δ that fits the above equation is 1.96. Therefore, we can get $n \geq 4$ from this equation. Note that this conclusion is independent of the variance (σ). If the confidence interval changes to 90%, then the minimal *n* drops to 3. For consensus scoring the above conclusion can be interpreted as that, if the prediction is required to be accurate enough (close to the true value within the range of variance), then three or four scoring functions are required. As illustrated in Figures 2 and 3, introducing more scoring functions into a consensus scoring procedure becomes less effective. As indicated by eq 1, the efficiency is scaled with $\sqrt{n}$.

The above discussion will not be restricted to virtual library screening approaches. We would like to point out that, for any computational approach in which a method is employed to evaluate a certain feature of the system under investigation, if there is more than one method available and they are simply complementary to each other, then combining them in application will generally lead to improved results. But in other areas, it is not always referred as consensus scoring.

Which Ranking Strategy Is the Best? As mentioned in the Methods section, we have tested three different ranking strategies of consensus scoring, which are possible in virtual library screening approaches. The hit-rates observed among the top 2% of the predicted data set by adopting the "rank-by-number" strategy and the "rank-by-rank" strategy are shown together in Figure 3. As implied in this figure, (1) hit-rates can be increased steadily by using more scoring functions; (2) once again three or four scoring functions seem to be sufficient for improving the accuracy in prediction; and (3) the "rank-by-number" strategy slightly outperforms the "rank-by-rank" strategy. As having been described above, the "rank-by-number" strategy is simply ranking all the candidates according to the mean values of all the scoring functions. Unfortunately this strategy cannot be applied if not all the scoring functions give results in a compatible unit. For example, empirical scoring functions give, as claimed, absolute binding free energies; force field based scoring functions give force field energies; while some other scoring functions give potentials of mean force. Thus, if the scoring functions employed in consensus scoring come from different categories, it is not proper to mix up their results. Nevertheless, the "rank-by-rank" strategy will work perfectly in such cases since it uses the relative ranks for instead. This strategy also has another potential advantage in practice: for a given compound, if the prediction of one scoring function deviates from all the others too much, using the rank as score will help to narrow the difference.

Table 1. Hit-Rates Observed in the "Rank-by-Vote" Experiments (Voting for Top 2%)[a]

	number of scoring functions				
votes	1	2	3	4	5
0	59(4900)[b]	36(4825)	22(4767)	13(4720)	8(4681)
1	41(100)	46(150)	41(174)	33(190)	25(197)
2		18(25)	30(50)	34(66)	34(79)
3			7(9)	17(21)	23(32)
4				3(3)	9(10)
5					1(1)

[a] If a compound is predicted to be on the top 2% among all the 5000 candidates by a certain scoring function, it will get the vote from that scoring function. [b] The first number denotes for the true hits, while the number in brackets denotes for the population. For example, this cell should be read as "by single scoring, among the 4900 candidates with zero vote, there are 59 true hits".

Another choice is the "rank-by-vote" strategy, which has been adopted, for example, by the consensus scoring module in the SYBYL software.[16] Since the score of a given candidate could range only from 0 to the number of scoring functions, the idea of this strategy is to rank all the candidates in a semiquantitative manner. It is more complicated than the other two strategies since there are two adjustable variables embedded in its procedure: the voting criterion and the minimal votes needed to qualify a good candidate. We have tested different voting criteria with this ranking strategy in our experiment: (1) a candidate will get a vote if it is enlisted on the top 2%; (2) a candidate will get a vote if it is enlisted on the top 5%; and (3) a candidate will get a vote if it is enlisted on the top 10%. The results given by these three voting criteria are listed in Tables 1−3, respectively. From these tables, it is clear that the hit-rates are very high among the compounds with full votes. But the drawback is that such compounds become so rare when the number of scoring functions exceeds three or four. If one lowers the threshold to consider all the compounds with at least one vote, the false negatives can be eliminated efficiently. However, by doing so more compounds need to be tested in bioassay. As indicated in Tables 1−3, variation of voting criteria also shows the same tendency: by using "tighter" voting criterion (e.g. voting for top 2%), fewer false positives are found among the tested compounds, while by using "looser" voting criterion (e.g. voting for top 10%), fewer false negatives are found among the untested compounds. There is always a tradeoff between the number of false positives and the number of false negatives. Therefore, if

Table 2. Hit-Rates Observed in the “Rank-by-Vote” Experiments (Voting for Top 5%)[a]

	number of scoring functions				
votes	1	2	3	4	5
0	33(4750)[b]	12(4594)	4(4482)	2(4397)	0(4326)
1	67(250)	43(313)	23(335)	11(343)	6(350)
2		45(93)	42(134)	28(152)	17(162)
3			31(49)	38(79)	31(93)
4				21(29)	31(51)
5					15(18)

[a] If a compound is predicted to be on the top 5% among all the 5000 candidates by a certain scoring function, it will get the vote from that scoring function. [b] The first number denotes for the true hit, while the number in brackets denotes for the population. For example, this cell should be read as “by single scoring, among the 4750 candidates with zero vote, there are 33 true hits”.

Table 3. Hit-Rates Observed in the “Rank-by-Vote” Experiments (Voting for Top 10%)[a]

	number of scoring functions				
votes	1	2	3	4	5
0	14(4500)[b]	3(4234)	0(4054)	0(3920)	0(3813)
1	86(500)	24(532)	6(537)	1(537)	0(536)
2		73(234)	30(264)	10(269)	3(272)
3			64(145)	35(172)	13(177)
4				54(102)	37(126)
5					47(76)

[a] If a compound is predicted to be on the top 10% among all the 5000 candidates by a certain scoring function, it will get the vote of that scoring function. [b] The first number denotes for the true hits, while the number in brackets denotes for the population. For example, this cell should be read as “by single scoring, among the 4500 candidates with zero vote, there are 14 true hits”.

the “rank-by-vote” strategy is applied to a consensus scoring approach anyway, how to set these voting criteria properly should depend on the primary concern of the researcher.

By comparing the data in Figure 3 and Tables 1−3, we can see that the “rank-by-vote” strategy is basically inferior to the “rank-by-number” or the “rank-by-rank” strategy as far as the hit-rates are concerned. For example, by using four scoring functions and voting for top 10%, the hit-rate given by the “rank-by-vote” strategy is 54 among the top 102 candidates. While by the “rank-by-number” and the “rank-by-rank” strategy, the hit-rates are 63 and 60 among the top 100 candidates, respectively. The relatively poor performance of the “rank-by-vote” strategy is due to its semiquantitative nature: suppose *n* scoring functions are employed in the consensus scoring procedure, all the candidate compounds will be simply divided into $n+1$ classes. Such a classification is probably too rough when only three or four scoring functions are actually applied. Some quantitative information will be lost during this conversion.

CONCLUSIONS

We have demonstrated through an idealized computer experiment that consensus scoring improves the hit-rates in virtual library screening because of a simple reason: the mean value of repeated samplings tends to be closer to the true value. Therefore, in a statistical sense, consensus scoring is more robust and accurate than any single scoring procedure. Our results also suggest that a moderate number of scoring functions, three or four, are sufficient for improving the results significantly. The idea of consensus scoring will be applicable not only to virtual library screening approaches but also to other similar approaches in which certain methods are needed to make predictions. As for the strategies of combining multiple scoring functions, the “rank-by-number” strategy works reliably. The “rank-by-rank” strategy can be used as an alternative when it is not practical to apply the “rank-by-number” strategy. The “rank-by-vote” strategy is generally not recommended.

ACKNOWLEDGMENT

This research is financially supported by the Department of Defense (Grant No. DOD DAMP17-93-V-3018). The authors are also grateful to the reviewers’ comments, which make this paper more informative than its original form.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening: An Overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(2) Drews, J. Drug Discovery: A Historical Perspective. *Science* **2000**, *287*, 1960−1964.

(3) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269−288.

(4) Morris, G. M.; Goodsell, D. S.; Halliday, R.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639−1662.

(5) Jones, G.; Wilett, 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.

(6) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(7) 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.

(8) Bohm, H.-J. Prediction of binding constants of protein ligands: A fast method for the priorization of hits obtained from de novo design or 3D database search programs. *J. Comput-Aided. Mol. Des.* **1998**, *12*, 309−323.

(9) Wang, R.; Liu, L.; Lai, L.; Tang, Y. SCORE: A new empirical method for estimating the binding affinity of a protein−ligand complex. *J. Mol. Model.* **1998**, *4*, 379−394.

(10) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: A simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.

(11) Mitchell, J. B. O.; Laskowski, R. A.; Alex, A.; Thornton, J. M. BLEEP−Potential of mean force describing protein−ligand interactions: I. Generating potential. *J. Comput. Chem.* **1999**, *20*, 1165−1176.

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

(13) 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.

(14) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.

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

(16) The SYBYL software; Tripos Associates: St. Louis, MO. http://www.tripos.com/.

(17) Hogg, R. V.; Craig, A. T. In *Introduction to Mathematical Statistics*, 5th ed.; Pirtle, R. W., Ed.; Prentice Hall: NJ, 1995.

CI010025X