Combination of a Naive Bayes Classifier with Consensus Scoring Improves Enrichment of High-Throughput Docking Results

Anthony E. Klon,* Meir Glick, and John W. Davies

Novartis Institute for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139

Received January 9, 2004

Abstract: We have previously shown that a machine learning technique can improve the enrichment of high-throughput docking (HTD) results. In the previous cases studied, however, the application of a naive Bayes classifier failed to improve enrichment for instances where HTD alone was unable to generate an acceptable enrichment. We present here a protocol to rescue poor docking results a priori using a combination of rank-by-median consensus scoring and naive Bayesian categorization.

We have previously described a conceptually simple, computationally inexpensive approach to improve the enrichment obtained from the high-throughput docking (HTD) of a database against a protein target.[1] This approach used a modified naive Bayes classifier[1,2] (NB) to rerank the compounds generated by HTD. The only required inputs to the naive Bayes classifier are a ranked list of compounds from HTD (scored according to their binding pose) and their chemical structure. Our previous work showed that in cases where HTD alone was able to provide enrichment of a database, the application of the naive Bayes classifier significantly improved the results. However, a weakness of the method lies in its inability to rescue HTD when the initial enrichment was no better than random. The procedure depends on using the top ∼1% of scored and ranked poses from HTD to generate a model for compounds that are defined as "good" binders to the active site of the protein. It is in cases where these "good" binders contain a large number of false positives that the application of naive Bayes will actually make the enrichment poorer.

It is obviously of great interest to know a priori whether HTD or the application of the naive Bayes classifier is unlikely to generate a good enrichment. In principle, one can seed the virtual screening collection with known binders to the target protein in order to generate an enrichment curve after docking or naive Bayes categorization. From this curve, an assessment could be made regarding the ability of these computational approaches to identify active compounds in the database being tested. Unfortunately, this is not possible in cases where there are no known actives. It is therefore imperative to develop a protocol that will be robust enough to generate an enrichment of screening databases for a variety of protein targets and docking algorithms.

* To whom correspondence should be addressed. Phone: 617-871-7132. Fax: 617-871-4088. E-mail: anthony.klon@pharma.novartis.com.

A protocol, described here, has been developed in response to the situations where application of the naive Bayesian classifier resulted in decreased enrichment in cases where there was an *initial* negative enrichment from HTD. In this protocol, consensus scoring[3] (CS) is applied after high-throughput docking but prior to the application of the naive Bayes classifier. A test case is reported where the use of consensus scoring rescued poor docking results. The improved scored and reranked list of compounds from the application of CS then acted as a suitable input to the naive Bayes classifier, which further improved the enrichment. A test case is also presented showing that in an ideal case where HTD and naive Bayes were previously shown to succeed the consensus scoring step does not have an adverse effect on the subsequent application of naive Bayes.

The details of database preparation, high-throughput docking, and naive Bayes classification have been described previously.[1] Two test cases were selected from the previous study in order to demonstrate the utility of using consensus scoring: the docking of protein tyrosine phosphatase 1B (PTP-1B) and protein kinase B/Akt (PKB) with FlexX.[4] The poses generated by FlexX for these test cases were rescored using the Gold,[5] PMF,[6] Dock,[7] and ChemScore[8] scoring functions as implemented in CScore.[4,9]

Consensus scoring approaches have been classified into three different categories: rank-by-vote, rank-by-number, and rank-by-rank.[10] Under this scheme, the program CScore is classified as a rank-by-vote method. Rank-by-vote gives each compound a vote from a particular scoring function if it appears in the top *n*% of the database according to that scoring function. The compounds are then ranked according to how many votes they receive. This method is not recommended because it lacks sufficient granularity to differentiate between binders and nonbinders for very large data sets. For example, CScore uses four different scoring functions, and so there are only five possible ranks for each compound: 0−4. For a large database of hundreds of thousands of compounds, there will be many compounds with the same rank. This shortcoming can be overcome by decreasing the value of *n*, thus increasing the precision of this approach while decreasing the recall rate. The latter situation occurs if one or more of the scoring functions fail to perform well against the target in question, thus guaranteeing that some true positives will never get the maximum possible rank in this approach and may therefore be overlooked. Rank-by-number approaches average the score of several different scoring functions. This was noted by Wang and Wang to be valid only in cases where the different scoring functions calculate measurements on a comparable scale but was regarded to be the superior approach.[10] The rank-by-number approach could not be applied in this case because the five scoring functions constitute three different classes of scoring functions: empirical, knowledge-based, and those based on molecular mechanics.[11] The FlexX and Gold scoring functions are empirical scoring functions, while the ChemScore and PMF scoring functions are knowledge-based scoring

10.1021/jm049970d CCC: $27.50
Published on Web 07/27/2004

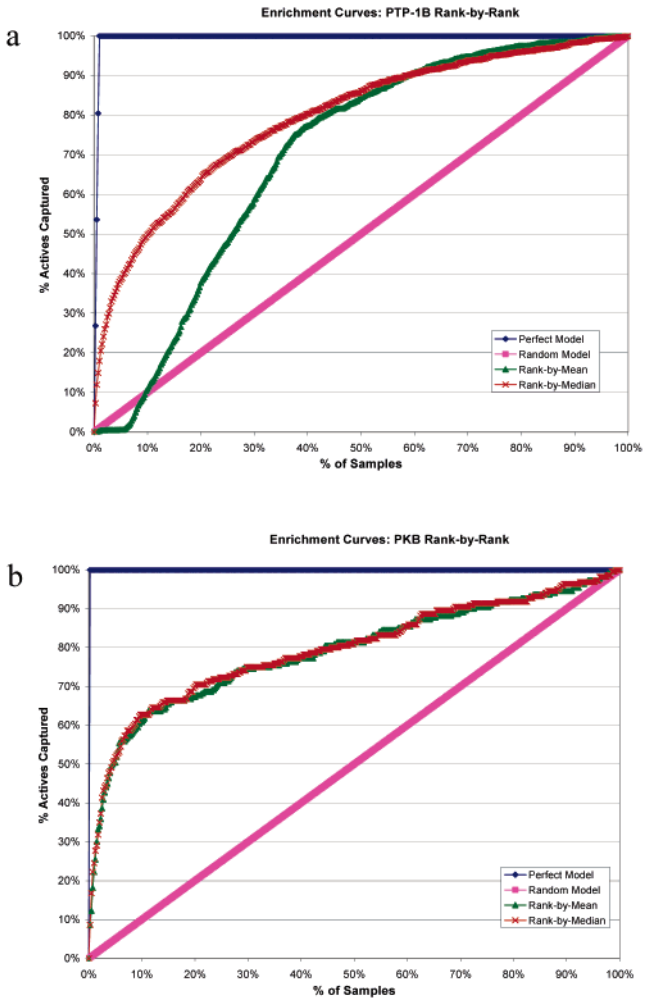

Figure 1. Comparison of enrichment curves calculated for FlexX docking results generated by the rank-by-mean (green) and rank-by-median (red) consensus scoring approaches for (a) PTP-1B and (b) PKB.

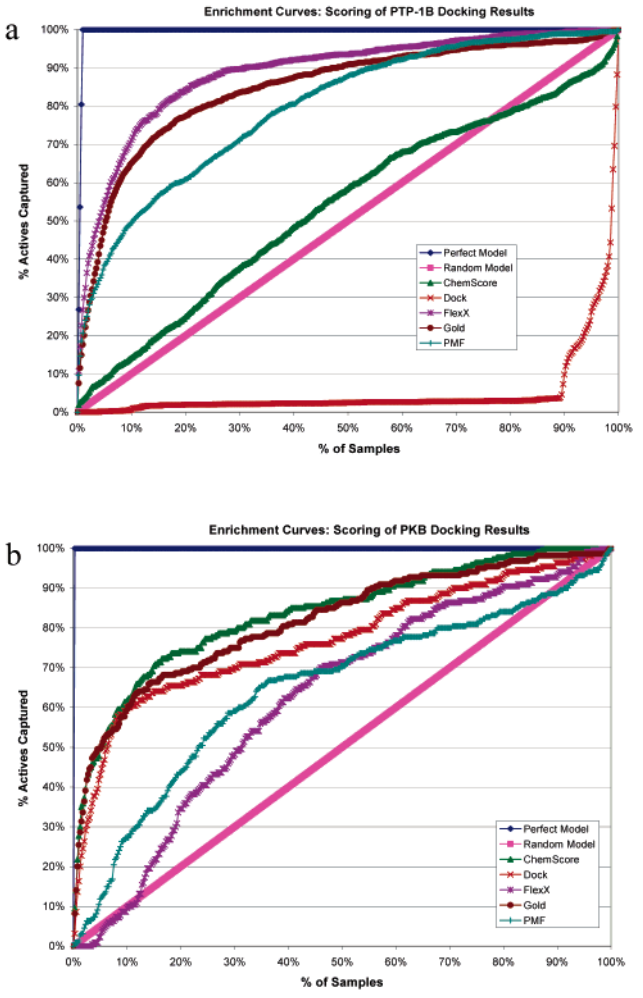

Figure 2. Enrichment curves calculated for scoring the poses generated by FlexX using the ChemScore (green), Dock (red), FlexX (purple), Gold (brown), and PMF (light blue) scoring functions for (a) PTP-1B and (b) PKB. Also shown for comparison are the predicted curves for the random (pink) and perfect (dark-blue) cases.

functions. The lone molecular mechanics based scoring function included in this study was Dock. The last consensus scoring approach, rank-by-rank, averages the ranks given to a single compound for each scoring function. To quantitatively estimate the amount of enrichment observed, we employed the receiver operating characteristic (ROC) curve.[12] An ROC curve describes the tradeoff between sensitivity and specificity. Sensitivity is defined as the ability of the model to detect true positives, while specificity is its ability to identify true negatives. The area below an ROC curve can be used to quantify the observed enrichment. An ROC score greater than 0.9 is considered excellent, and a value below 0.6 represents no enrichment.

In this study, the rank-by-rank method was chosen because of the fact that the five scoring functions are not on a comparable scale. One modification to the rank-by-rank method described by Wang and Wang was introduced. The ranking of the compounds calculated according to the four scoring functions implemented in CScore and with the FlexX[13] scoring function were used as input to the rank-by-rank method. It was found that enrichment curves generated based on the rank-by-rank method of consensus scoring gave initially poor results in the case of the PTP-1B target (Figure 1a). Further analysis of enrichment plots generated from ranked lists

Table 1. Area under ROC Curves Corresponding to the Enrichment Calculated by Five Different Scoring Functions for the Docking of a Subset of the ACD Using FlexX

scoring function	target	
	PTP-1B	PKB
ChemScore	0.54	0.84
Dock	0.05	0.77
FlexX	0.89	0.62
Gold	0.85	0.82
PMF	0.80	0.65

of the individual scoring functions showed that the Dock scoring function was generating a negative enrichment for this particular target, while the ChemScore scoring function was generating a random enrichment (Table 1, Figure 2a). Curiously, consensus scoring seemed to perform well for the PKB test case (Figure 1b), and the individual scoring functions seemed to perform fairly well overall (Table 1, Figure 2b). To address these seeming inconsistencies, the median, and not average, rank for each compound over all five scoring functions was taken because calculation of the median value is less sensitive to outliers than calculation of the mean. This method substantially improved the enrichment curve in the case of PTP-1B. The two rank-by-rank

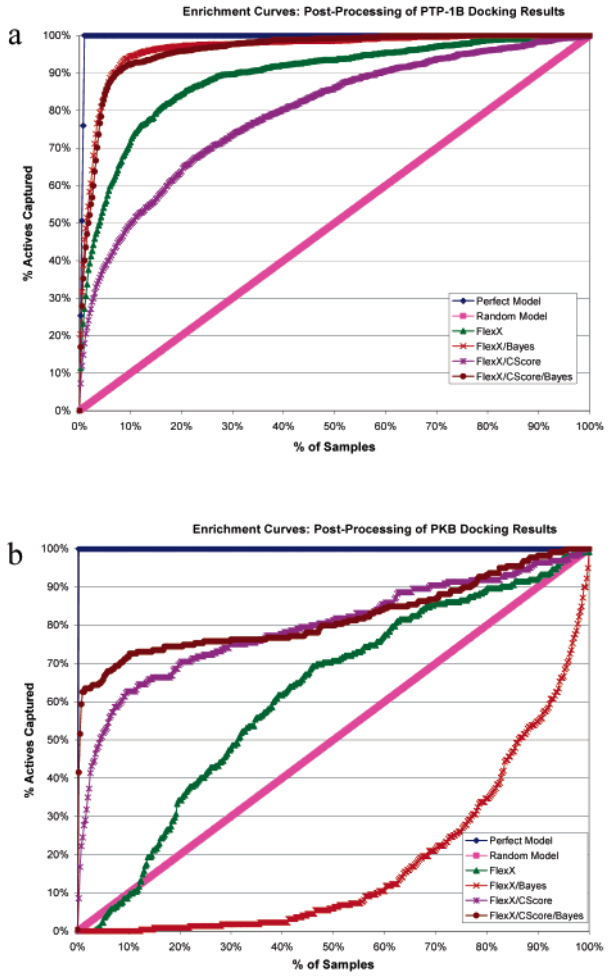

Figure 3. Enrichment curves calculated from docking poses generated by FlexX alone (green), FlexX/naive Bayes (red), FlexX/rank-by-median (purple), and FlexX/rank-by-median/naive Bayes (brown) for (a) PTP-1B and (b) PKB.

methods described here will hereafter be defined as rank-by-mean and rank-by-median.

To apply naive Bayes categorization on the data set, a cutoff must be established to define which compounds are "good" and "bad" binders. In our previous work a cutoff of three standard deviations (~1% of the database) below the mean energy[1] was used. In the case of consensus scoring, the rank-by-median approach does not calculate an energy for the docking score but instead produces a median rank for each compound across the five scoring functions. The approach used to define a cutoff based on the statistical distribution of the calculated energies of docked poses cannot be used in this case. Consequently, a more straightforward criterion was applied to the rank-by-median results in which the top 1% of compounds was passed to the naive Bayes classifier as the "good" binders, while the remaining 99% were defined as the "bad" binders.

The resulting enrichment curves for protein targets PTP-1B and PKB after HTD with FlexX alone, HTD followed by either naive Bayes or consensus scoring, and HTD followed by combined consensus scoring and naive Bayes are shown in Figure 3. The area under the corresponding ROC curves is shown in Table 2. As reported previously for the case of PTP-1B, HTD using FlexX alone generates a good initial enrichment (green line), and the application of the naive Bayes classifier improves this result significantly[1] (red line). However, by use of the rank-by-median consensus scoring approach, the overall enrichment is reduced (purple line). This initially surprising result is caused by the failure of the Dock and ChemScore scoring functions to accurately identify those active compounds that were well-placed in the active site by FlexX. Because the overall enrichment is still good, particularly in the top 1% of the database, the subsequent application of the naive Bayes classifier is able restore the enrichment to the level observed previously for HTD and naive Bayes alone (brown line).

Table 2. Area under ROC Curves Corresponding to the Enrichment after Each Step in the Protocol Discussed in This Paper

method	target	
	PTP-1B	PKB
FlexX[a]	0.89	0.62
FlexX/naive Bayes[a]	0.97	0.18
FlexX/rank-by-median	0.80	0.80
FlexX/rank-by-median/naïve Bayes	0.96	0.82

[a] Reference 1.

The enrichment curves calculated for PKB, shown in Figure 3b, are more revealing of the utility of the protocol outlined in this paper. As previously reported, docking the compounds in the database using FlexX generates an overall enrichment but a negative initial enrichment[1] (green line). Compounds in this region of the database are precisely those used to train the naive Bayes classifier; the application of naive Bayes immediately following HTD reduces the enrichment (red line). Application of the rank-by-median consensus scoring method after docking improved the enrichment results significantly (purple line). Importantly, there was a very large improvement in the number of active compounds reranked near the top of the database after rank-by-median. When the compounds in the top 1% of this new ranked list were used as "good" binders to train the naive Bayes classifier, a further enrichment was observed (brown line). Although there was only a marginal improvement in the value for the area under the ROC curve (Table 2), the improvement in the shape of the enrichment curve is very significant. The number of active compounds found in the top 1% of the database after rank-by-median increased from 24% to 64% after naive Bayes. This dramatic improvement in the initial enrichment of the database is extremely important when relying on in silico methods either to provide a target-focused library for high-throughput screening or to select a very small number of compounds out of a very large data set containing hundreds of thousands for testing.

A more robust approach to scoring the resulting poses generated through high-throughput docking has been presented. This approach takes advantage of both consensus scoring as well as naive Bayes categorization to further enrich docking results. Significantly, the order in which these methods are applied has a crucial impact on the quality of the results obtained. The successful application of naive Bayes requires that there is an initial positive enrichment in the database. Consensus scoring is shown to have the ability to rescue poor HTD results in cases where the docking algorithm generates

reasonable poses for the compounds that bind to the active site. In the cases presented here, rank-by-median was found to be more effective than the rank-by-mean and rank-by-vote approaches to consensus scoring because it is less sensitive to scoring functions that perform poorly against a particular protein target. Because information about a given scoring function's performance with respect to a specific target may not be known a priori, it is better to use a variety of scoring functions and calculate the median rank of each compound. The application of the naive Bayes classifier after consensus scoring is then capable of improving upon this enrichment in cases where consensus scoring is able to place a higher percentage of good binders in the top 1% of the ranked list. The overall protocol of high-throughput docking → rank-by-median → naïve Bayes classification is robust enough to ensure maximum enrichment in the cases presented here.

The development of our approach was motivated by the need to reliably increase the chance of identifying specific binders to the active site of a protein target by the high-throughput docking of a screening database. Unlike other methods that have used statistical approaches to increase the effectiveness of scoring functions,[14] our method is effective prospectively where no a priori information about known actives or experimentally determined complex structures existed. The method can ideally be used to direct a high-throughput screening campaign through the generation of focused libraries or iterative screening.

References

(1) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: a simple and efficient method of improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743−2749.

(2) *Pipeline Pilot 3.0*; *Scitegic, Inc.* (9665 Chesapeake Drive, Suite 401, San Diego, CA 92123), 2003.

(3) 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.

(4) Tripos, Inc. (1699 South Hanley Road, St. Louis, MO 63144), 2003.

(5) 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.

(6) 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.

(7) 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.

(8) 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.

(9) 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.

(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) Bohm, H. J.; Stahl, M. The use of scoring functions in drug discovery applications. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons: Hoboken, NJ, 2002; pp 41−87.

(12) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann Publishers: New York, 1999.

(13) 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.

(14) Jacobsson, M.; Liden, P.; Stjernschantz, E.; Bostrom, H.; Norinder, U. Improving structure-based virtual screening by multivariate analysis of scoring data. *J. Med. Chem.* **2003,** *46*, 5781−5789.

JM049970D