Finding More Needles in the Haystack: A Simple and Efficient Method for Improving High-Throughput Docking Results

Anthony E. Klon, Meir Glick, Mathis Thoma, Pierre Acklin, and John W. Davies*

Novartis Institute for Biomedical Research, 100 Technology Square, Cambridge, Massachusetts 02139

Received July 30, 2003

The technology underpinning high-throughput docking (HTD) has developed over the past few years to where it has become a vital tool in modern drug discovery. Although the performance of various docking algorithms is adequate, the ability to accurately and consistently rank compounds using a scoring function remains problematic. We show that by employing a simple machine learning method (naïve Bayes) it is possible to significantly overcome this deficiency. Compounds from the Available Chemical Directory (ACD), along with known active compounds, were docked into two protein targets using three software packages. In cases where HTD alone was able to show some enrichment, the application of naïve Bayes was able to improve upon the enrichment. The application of this methodology to enrich HTD results can be carried out without a priori knowledge of the activity of compounds and results in superior enrichment of known actives compared to the use of scoring methods alone.

Introduction

High throughput docking (HTD) is a commonly utilized technique in modern drug discovery for screening large compound databases with the aim of eliciting novel drug candidates for a given therapeutic target.[1;2] The successful application of HTD is dependent upon a three-dimensional structure of a given target and an algorithm that orients the candidate molecule in the active site in the proper binding pose. Most docking algorithms perform well in this task. However, the accurate determination of the binding energy is notoriously difficult and can introduce artifacts which bias the results. The evaluation of binding affinities is commonly determined through knowledge-based potentials, scoring functions derived from molecular mechanics force fields, or empirical energy functions.[3] These scoring functions are typically a summation of approximate terms related to binding, such as atomic pairwise potentials.

Most HTD applications assume a rigid binding site and lack explicit solvent, which are crucial for the accurate calculation of binding free energy. Scoring functions vary greatly in the accuracy of their predictions, and all of them introduce separate biases to the problem of producing an accurate ranked list of compounds. Even the best scoring functions still rank some potential drug candidates poorly as a result of their inherent biases.[4] Attempts have been made to reduce the biases and weakness associated with scoring functions by employing consensus scoring approaches.[5−7] Although improved hit rates and enrichments have been observed, the approach makes no attempt to improve scoring per se, but instead simply attempts to reduce the bias of any one function by polling votes for those candidates which consistently score well. It has been argued that the reason for the successful implementation of consensus scoring approaches is for the simple statistical reason that the mean of repeated measurements tends to be closer to the actual value.[8]

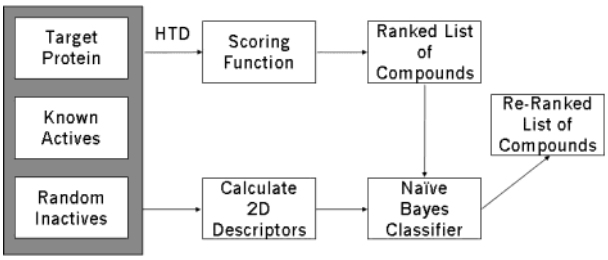

Figure 1. Diagram of the analysis workflow described in this paper.

High throughput docking is in many ways analogous to high throughput screening (HTS). In both cases the number of compounds evaluated is typically on the order of hundreds of thousands. In addition, artifacts in both technologies make it difficult to accurately and consistently identify true drug candidates without duplicate screening or slower but more accurate docking calculations. Machine learning techniques have been successfully employed in HTS data analysis to minimize these artifacts and enrich hit rates.[9] Here we describe the successful application of one such machine learning technique, naïve Bayes,[10] to enrich the data from HTD and overcome some of the weaknesses in the application of scoring functions.

Methods

Workflow Followed in This Paper. The procedure followed in this paper, for retrospective examples, is exemplified in the workflow in Figure 1. A method is described in which the ranked list of scored compounds from HTD can be used to train a modified naïve Bayesian (NB) classifier which predicts and reranks the compounds with superior results. This is achieved by training the classifier with those few ligand structures from the top scored poses from docking as "good" or well-

* Corresponding author. E-mail: john.davies@pharma.novartis.com, phone: 617-871-7127, fax: 617-871-7042.

10.1021/jm030363k CCC: $27.50
Published on Web 04/28/2004

Table 1. The Number of Known Active Compounds for Each HTD Target

protein target	number of actives	structural classes
PTP-1B[a]	1327	6
PKB	266	4

[a] References 24−28.

docked, and the rest of the ligand structures as "bad" or poorly docked. In this way the docking score is used to predetermine, in the absence of known activity, what are likely to be active or inactive molecules. The NB classifier takes as input the 2D fingerprints of the ligand structures, which can be precalculated prior to docking. The NB classifier predicts an improved ranking of the compounds based on structural elements in the top scored docking poses, by elimination of false positives and the identification of false negatives.

Preprocessing of the Compound Database for High-Throughput Docking. Along with a 3D structure of the target in question, a test set consisting of compounds with known activities and inactive compounds from the Available Chemicals Directory (ACD)[11] was generated for testing and validation purposes. The set of random inactive compounds from the ACD was filtered such that salts were stripped and duplicates were removed. Unity[12] was then used to further filter compounds which were either present in mixtures, contained metals, were isotopes, did not possess any carbon atoms, or had a ClogP greater than 5.0. This resulted in a database containing 179 805 compounds.

For FlexX and Dock, the database was ionized and Gasteiger−Marsili[13] partial charges were assigned. To prepare the database for docking with Glide, the ionization and assignment of partial charges was carried out using scripts supplied by Schrödinger.[14]

The series of known active compounds were also appropriately protonated and partial charges were assigned. The known active compounds were taken from literature references, patents, or in-house projects at Novartis for which either IC_{50} data was available for the inhibition of the target protein, or validated hits from HTS campaigns. Table 1 shows the number of active compounds and the number of representative structural classes for each of the active series.

Preparation of the Protein Targets for High-Throughput Docking. Cocrystal structures of each of the targets of interest, protein kinase B (PKB, PDB id #1O6K),[15] and protein-tyrosine phosphatase 1B (PTP-1B, PDB id # 1C88),[16] with bound inhibitors and cofactors were taken from the Protein Data Bank.[17] Water molecules were removed from all three crystal structures, and hydrogen atoms were added using either Sybyl[12] (for docking using FlexX[18] and Dock[19]) or Maestro[14] (for docking using Glide[14]). Gasteiger−Marsili[13] partial charges were added using Sybyl in order to prepare the structures for Dock.

Docking of the Test Set against the Protein Targets Using Dock. For each protein target, a set of spheres was generated from the heavy atom positions of the ligand in the cocrystal structure with the receptor obtained from the Protein Data Bank. A 0.3 Å resolution grid was calculated for energy scoring using an all-atom model with a 10 Å distance cutoff, a distance dependent dielectric constant ($\epsilon = 4$), and a bump overlap of 0.5 Å. Docking was then carried out by matching heavy atoms from the ligands in the sample database with the sphere centers. A flexible docking procedure was followed in which multiple anchors, each consisting of at least 10 heavy atoms, were generated and placed automatically for each ligand. The ligands were built iteratively after placement of the initial anchor, and the torsion drive option was used within Dock to sample the low-energy torsion angles. Twenty-five conformations were retained in each cycle of anchor search/torsion drive. Minimization of the entire molecule was carried out using 10 cycles consisting of 100 steps of simplex minimization with a convergence of 0.1 kcal/mol. Only the top ranking conformation from each ligand, corresponding to the best Dock energy score, was retained and written out to a multi-mol2 file.

Docking of the Test Set against the Protein Targets Using FlexX. The default parameters for FlexX were used as distributed by Tripos in Sybyl 6.9 for carrying out the flexible docking. The protein receptor model was generated by including all residues containing any atoms within 6.5 Å of the ligand in the cocrystal structure in the receptor description file for each protein. The best scoring conformation for each ligand docked, according to the FlexX score, was written out to a multi-mol2 file.

Docking of the Test Set against the Protein Targets Using Glide. The default input parameters were used for the generation of the command files for docking small molecule databases as implemented in the 5.1 release of Maestro.[14] However, for the generation of the scoring grids, a value of 1.0 × the van der Waals radii of the protein atoms was used instead of 0.9. From previous studies this value generated better results with respect to enrichment and binding poses. The maximum number of heavy atoms was set to 120, and the maximum number of rotatable bonds allowed was 30. The top scoring pose was retained for each ligand and written to a maestro-formatted output file.

Evaluating Enrichment of the Known Actives after High-Throughput Docking. For each of the HTD runs, all of the resulting poses were ranked according to their energies as calculated by the appropriate scoring function (Dock, FlexX, or Glide score) from the program which generated their final docking pose. These ranked lists were then subsequently used to generate the enrichment and Receiver Operating Characteristic (ROC) curves.[20] A 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 avoid false negatives. The area below a ROC curve can be used to quantify the observed enrichment. A ROC value greater than 0.9 is considered excellent, and a value below 0.6 represents no enrichment.

Extended-Connectivity Fingerprints (ECFPs). Extended-connectivity fingerprints (ECFPs) were used as structural descriptors for training the NB classifier, both of which are implemented in Pipeline Pilot.[21] The ECFPs are a new class of fingerprints for molecular characterization developed by Scitegic (Rogers and Hahn, unpublished results) that rely on the Morgan algorithm.[22] The ECFP features correspond to the presence of an exact structure (not a substructure) with

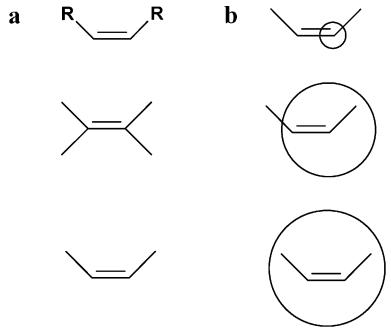

Figure 2. Extended-connectivity fingerprint structure searches. (a) A substructure search using the fragment shown at the top will return the other two compounds shown, while a substructure search using ECFPs will return only the bottom compound. (b) The circle around a single atom in the structure shown at the top illustrates an ECFP with a neighborhood size of 0, ECFP_0. Subsequent iterations update this atom's code to include molecules two (middle) and three (bottom) bonds distant, corresponding to ECFP_1 and ECFP_2, respectively.

limited specified attachment points (Figure 2). In generating the fingerprints, the program assigns an initial code to each atom. The initial atom code is derived from the number of connections to the atom, the element type, atomic charge, and atomic mass. This corresponds to an ECFP with a neighborhood size of zero. These atom codes are then updated in an iterative manner to reflect the codes of each atom's neighbors (Figure 2). In the next iteration, a hashing scheme is employed to incorporate information from each atom's immediate neighbors. Each atom's new code now describes a molecular structure with a neighborhood size of one. This process is carried out for all atoms in the molecule. When the desired neighborhood size is reached, the process is complete and the set of all features is returned as the fingerprint. For the ECFPs employed in this paper, a neighborhood size of six was used to generate the fingerprints. The resulting ECFPs can represent a much larger set of features than other fingerprints and contain a significant number of different structural units crucial for the molecular comparison, among the compounds. To accommodate the large amount of resulting information, Pipeline Pilot uses a 32 bit hashing scheme.

Naïve Bayes Classification. The naïve Bayes classifier is a statistical modeling method based upon Bayes's rule of conditional probability. The formula takes the following form:

$$P(\mathrm{A}|\mathrm{B}) = P(\mathrm{B}|\mathrm{A})\frac{P(\mathrm{A})}{P(\mathrm{B})}$$

where $P(\mathrm{A}|\mathrm{B})$ is the probability that event A will occur given the condition that event B occurred. In the cases studied in this paper, event A refers to the activity of a given compound, while event B refers to the presence of a certain ECFP bit. $P(\mathrm{A})$ represents the probability that event A will occur in a given dataset. $P(\mathrm{B})$ is the probability that a compound with a given feature occurs in the dataset. $P(\mathrm{A}|\mathrm{B})$ therefore is the probability that a given compound in the dataset will bind to the active site of the protein given that it has a particular feature. This probability is predicted from $P(\mathrm{B}|\mathrm{A})$, the probability

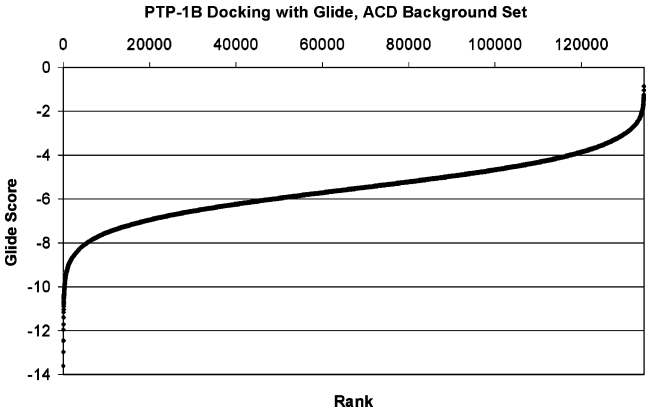

Figure 3. A plot of Glide docking energy versus compounds rank for the combined PTP-1B active set and ACD database.

that a compound with a given 2D-descriptor is active, along with $P(\mathrm{A})$, and $P(\mathrm{B})$:

$$P(\mathrm{active}|\mathrm{feature}) = P(\mathrm{feature}|\mathrm{active})\frac{P(\mathrm{active})}{P(\mathrm{feature})}$$

The Naïve Bayes classifier is referred to as naïve because it "naively" assumes independence among events. If this is true, then it is valid to multiply probabilities. Because each compound has n fingerprints $P(\mathrm{active}|\mathrm{feature})$ becomes:

$$P(\mathrm{active}|\mathrm{feature}) = P(\mathrm{feature}_1|\mathrm{active}) \times P(\mathrm{feature}_2|\mathrm{active}) \times P(\mathrm{feature}_3|\mathrm{active}) \times \ldots P(\mathrm{feature}_n|\mathrm{active})\frac{P(\mathrm{active})}{P(\mathrm{feature})}$$

When training the naïve Bayes classifier in Pipeline Pilot, each bin contains the number of occurrences of a given hashed fingerprint bit string. The normalized probability is then calculated to provide a final contribution of the feature to the total relative estimate.

Results

Defining the "Good" and "Bad" Compounds a priori To Train the Naïve Bayes Classifier. To correctly derive a NB classifier, and in the case of no a priori knowledge of the activity of the docked structures, one must define what are the representative active ("good") compounds and representative inactive ("bad") compounds. In this study the only way to determine an appropriate differentiation of "good" and "bad" compounds from HTD was to utilize the scoring function. It was observed that a plot of a given compound's docking score versus its rank is sigmoidal. In the authors' experience, this relationship holds true regardless of the docking software, the protein target, or the background databases used. Figure 3 depicts one such plot showing the results of docking the PTP-1B active set and the ACD inactive set with Glide. It can be observed from Figure 3 that docked structures with a favorable Glide score are in a region of the graph at the far left with a large positive slope. The slope is nearly flat throughout most of the dataset until the tail end of the plot is reached. For the docked structures that were scored very poorly by Glide, the slope once again increases sharply.

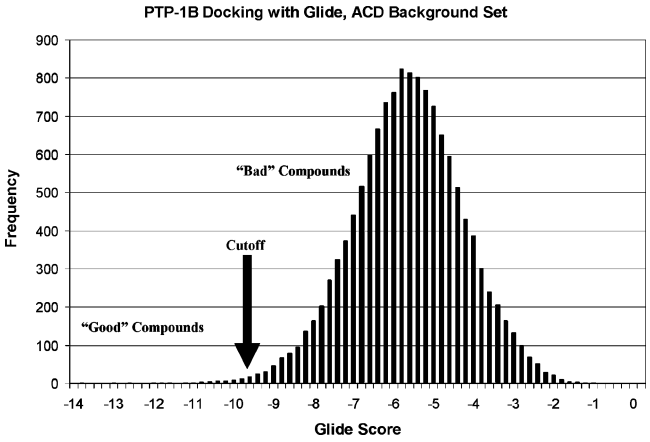

Figure 4. Histogram of Glide docking energy versus compounds rank for the combined PTP-1B active set and ACD database.

It seems logical, from the plot shown in Figure 3, that one could define the "good" and "bad" compounds by assigning cutoffs based solely on the energies generated by the scoring functions. However, it was unclear whether to use two separate cutoffs or a single cutoff to partition the dataset. If two cutoffs were applied, three categories of compounds were generated. All compounds with energies below the lower cutoff would be passed to the NB classifier as the "good" compounds. Similarly, all compounds with energies greater than the upper cutoff would be passed to the module as the "bad" compounds. The remaining set of compounds between the lower and upper cutoffs would not be used by the classifier. By applying a single cutoff, all compounds with energies less than the threshold were considered "good", while all compounds with energies greater than the threshold were considered "bad".

After studying the effect on enrichment it was determined that a single cutoff was found to be the optimal solution (data not shown). The value for that cutoff was determined to be at a point roughly in the region where the slope of the plot in Figure 3 drops from its initial large positive value. To determine this value arithmetically we calculated frequencies of the scores generated by Dock, FlexX, and Glide. Figure 4 shows an example of one such histogram for the docking of the PTP-1B active set and the ACD as inactive set using Glide. It was observed that in the case of Glide and FlexX, the cutoff point was approximately three standard deviations below the mean energy. For Dock, the cutoff point was approximately one standard deviation below the mean energy.

A clear difference was noted between Dock and the other two programs in how many standard deviations were required to generate appropriate cutoff values. Upon analysis of the kurtosis and skewness of the energies generated by the various scoring functions (Table 2), an interesting trend was noted that held for all of the cases examined in this paper. For a normal distribution, the value for kurtosis is 0 (normalized[23]) and skewness is 0. Unlike FlexX and Glide, Dock had values for the kurtosis and skewness which were significantly larger than would be expected for a normal distribution. Indeed, the distributions of Dock scores were leptokurtic and positively skewed. Figure 5 is a box plot showing the distribution of docking scores

Table 2. Descriptive Statistics for the Docking Energies in the ACD Test Case

target	program	mean	median	standard deviation	kurtosis	skewness
PTP-1B	Dock	−18.84	−20.92	23.78	525.72	18.86
	FlexX	−14.69	−14.15	7.65	2.10	−0.28
	Glide	−5.58	−5.60	1.33	0.51	−0.12
PKB	Dock	−17.75	−23.34	39.26	216.32	12.64
	FlexX	−17.56	−17.40	8.08	1.43	0.07
	Glide	−6.67	−6.62	2.05	0.39	−0.38

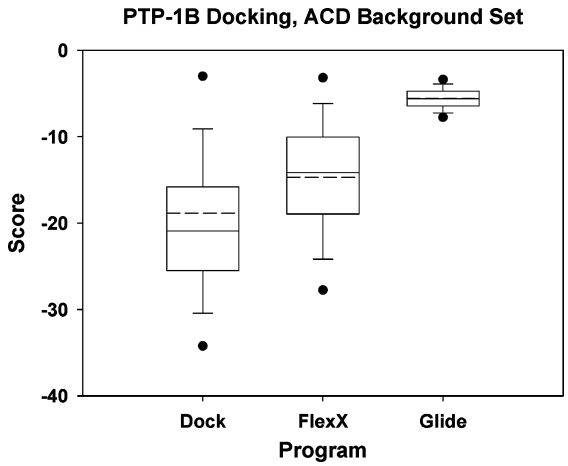

Figure 5. Box plot of docking scores for Dock, FlexX, and Glide. The box spans 50% of the scores observed. The area between the whiskers includes 95% of all scores. The first outliers above and below the 95% and 5% cutoff are shown. The median for each sample is shown as a solid line and the mean as a dashed line.

generated by Dock, FlexX, and Glide. For a normal distribution, the median and the mean should be equal. In cases where a normal distribution is not followed and the data is skewed, the mean is shifted in the direction of the outliers. Taken together, these results suggest that even for very large datasets, the distribution of scores generated by the Dock scoring function do not fit a normal distribution. A possible explanation for this is related to the types of scoring functions used by the docking software. In the case of FlexX and Glide, the scoring functions are empirical in nature, while the Dock scoring function is based upon a molecular mechanics force field not initially parametrized for scoring docking poses.[4]

The Naïve Bayes Classifier Improves the Enrichment of High-Throughput Docking Results Generated for PTP-1B by Dock, FlexX, and Glide. Figure 6 show three enrichment plots (a. Glide, b. FlexX, c. Dock) of the ACD set of compounds seeded with PTP-1B active compounds from both in-house data and published data sets.[24−28] The plots show enrichment curves before and after employing the NB classifier. All three docking systems were able to provide initial enrichment of the known actives. Subsequent application of the NB classifier to each ranked list resulted in improved enrichment of the known actives. Table 3 shows the dramatic increase in the total number of actives captured in the top 10% (i.e. 14 039 compounds) of the database for Dock and FlexX. In the case of Dock, application of the NB classifier resulted in an additional 23% of known actives being captured in the top 10% of the database. For FlexX, an additional 22% were

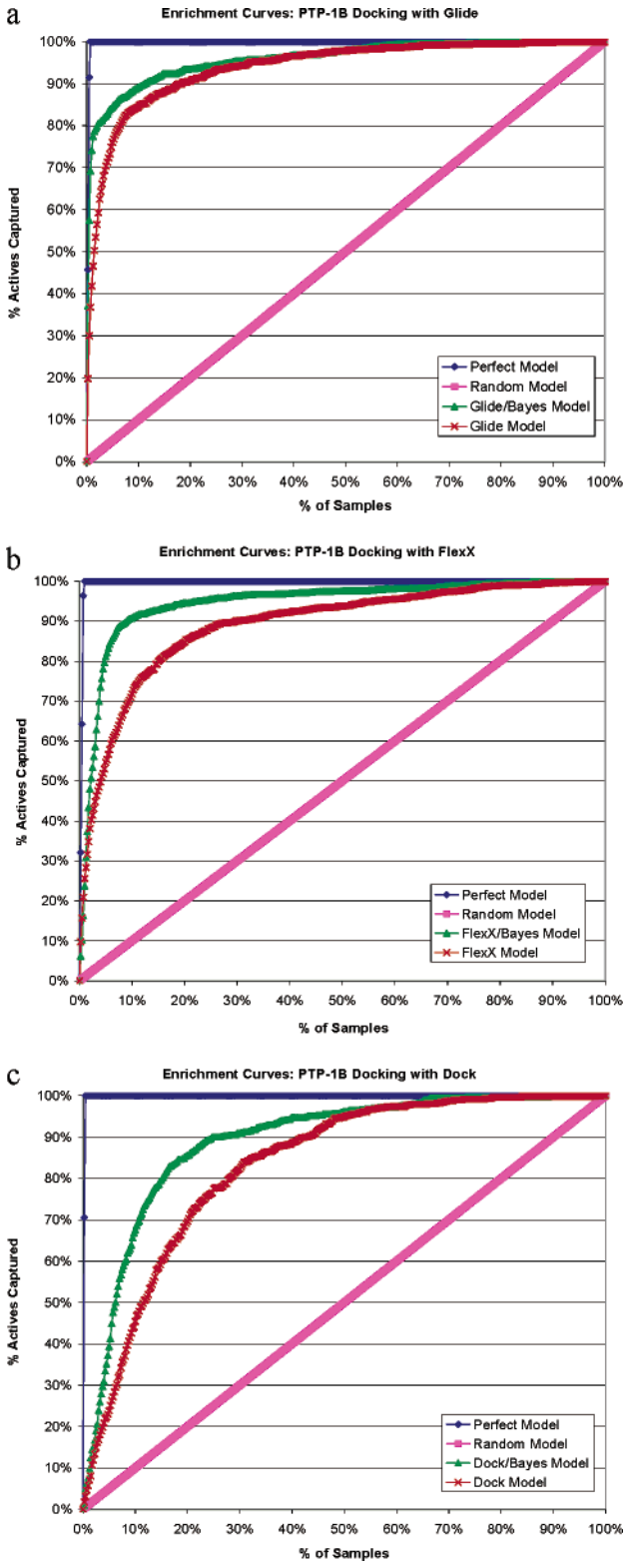

Figure 6. Enrichment plots from docking known inhibitors of PTP-1B using (a) Glide, (b) FlexX, and (c) Dock with the ACD database, before and after the application of NB.

captured. Glide, which already yielded an enrichment of 84% after docking, was improved to 91% after applying the NB classifier. The area under the ROC curve shows that the predictive model used (Dock/NB, FlexX/NB, or Glide/NB) is excellent in all three cases.

The Naïve Bayes Classifier Fails To Improve the Enrichment of High-Throughput Docking Results Generated for PKB. Figure 7 illustrates three enrichment curves generated in the same manner. The corresponding calculated values for the area under the ROC curve are shown in Table 3. From docking and scoring alone, and prior to the NB classifier, application of Dock does not result in any enrichment of the dataset. Similarly, FlexX shows only poor enrichment of the data. However, Glide shows the best result with good enrichment and a ROC curve value of 0.84. After application of the NB classifier to these docking results, the Dock/naïve Bayes model still fails to produce any enrichment, as expected. The FlexX/naïve Bayes model also fails to provide any enrichment, which might be expected based upon the accuracy of the original rankings. It was initially surprising, however, to observe that the Glide enrichment results degenerated from the previously good ROC area of 0.84 to 0.18, after the application of the NB classifier. This observation can be explained by closer examination of the original enrichment curve prior to the application of the NB classifier. Only those compounds ranked near the top of the list by the scoring function are used to train the NB classifier. Compounds are chosen that possess energies lower than three standard deviations below the mean energy. This results in the selection of compounds from the top 1% of the compounds docked. Examination of this portion of the enrichment curve in Figure 7, for both Glide and FlexX, clearly shows that this region has a negative deflection due to a number of false positives. The majority of these very highly ranked compounds possess negative formal charges (carboxylates etc), which interact with the cationic manganese in the active site of PKB. Consequently, these compounds receive significantly higher scores than would otherwise be expected. Although the enrichment curve for Glide continues to improve (and to a lesser extent for FlexX), this early region of the curve is precisely where the NB classifier takes the majority of its "good" compound examples from. As a result, the NB classifier eventually generates a predictive model for active compounds based upon data containing many false positives.

Table 3. Fraction of Actives Captured and the Area under the Corresponding ROC Curves, before and after the Application of NB

		fraction of total actives in top 10% of database screened		area under ROC curve[a]	
target	program	before NB	after NB	before NB	after NB
PTP-1B	Dock	0.45	0.68	0.83	0.89
	FlexX	0.72	0.91	0.89	0.95
	Glide	0.84	0.89	0.94	0.96
PKB	Dock	0.02	0.00	0.20	0.47
	FlexX	0.10	0.00	0.65	0.26
	Glide	0.48	0.02	0.85	0.22

[a] Qualitative interpretation of the area under ROC curves is as follows: 0.0−0.6 fail; 0.6−0.7 poor; 0.7−0.8 fair; 0.8−0.9 good; 0.9−1.0 excellent.

Negative Enrichment. Of the six experiments based on the two targets and three software packages studied in this paper, one (PKB:Dock) resulted in significant negative enrichment after initial docking. The application of naïve Bayes makes a marginal improvement to the enrichment results from negative to random. As previously discussed, there are specific artifacts in the docking model due to the presence of the two Mn^{2+} ions in the PKB active site. In addition, structural rear-

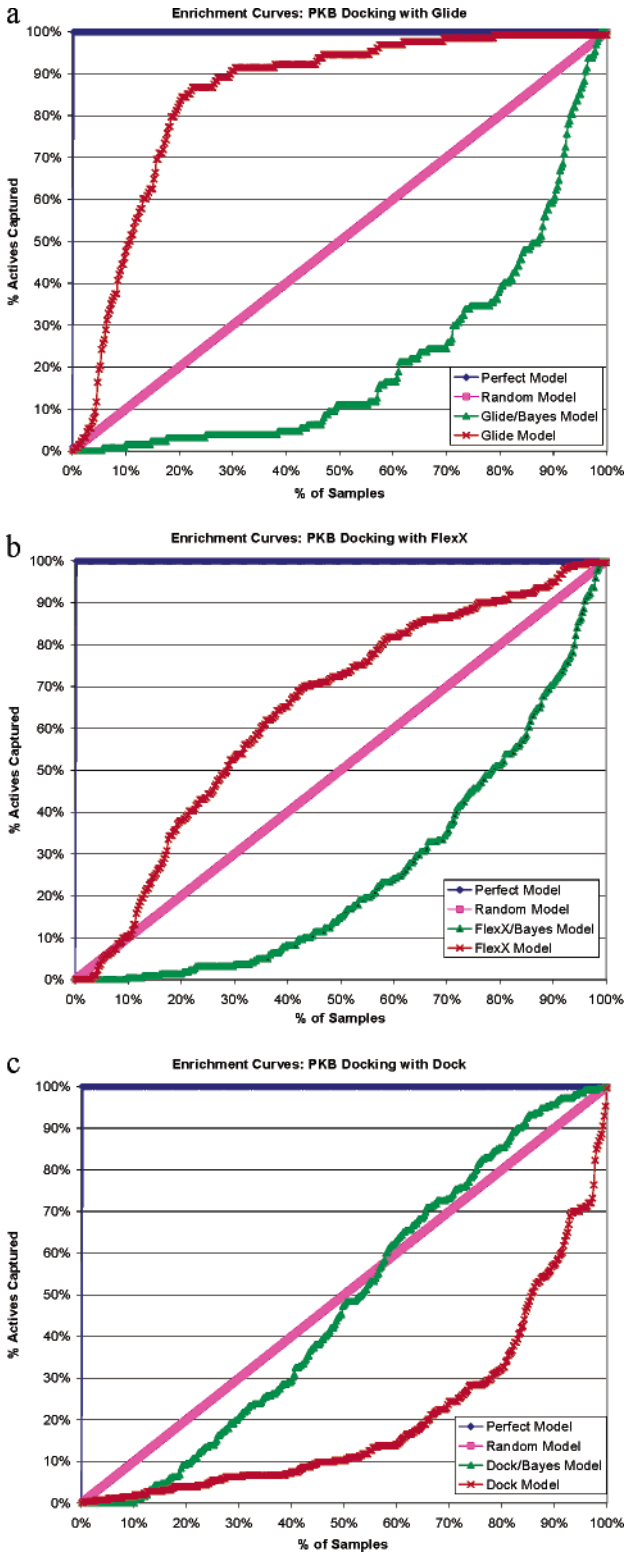

Figure 7. Enrichment plots from docking known inhibitors of PKB using (a) Glide, (b) FlexX, and (c) Dock with the ACD database, before and after the application of NB.

rangements due to global movements of PKB upon binding to the substrate biased the docking results toward low molecular weight compounds that were negatively charged. Visual inspection of the docking poses shows that Dock's objective function seems to favor ionic interactions over other types of hydrogen bonding and van der Waals interactions. The resulting failure of naïve Bayes to significantly enrich these results can therefore be explained. After initial ranking on the basis of the poses generated by Dock, the structural features of all the known inhibitors of PKB determined by the ECFPs were found to be randomly scattered throughout the dataset. Application of the naïve Bayes classifier therefore reranked the compounds in the dataset in a random manner.

It is important to note that although the application of the naïve Bayes classifier improved the enrichment from negative to random in the case of PKB:Dock, this is not always the case. In other instances studied where high-throughput docking generated negative enrichment, the application of the naïve Bayes classifier made the enrichment significantly worse (data not shown). Some caution is therefore required in applying the classifier in situations where docking alone fails to provide enrichment.

Discussion

We have demonstrated a novel procedure using the NB classifier to further enrich results from high-throughput docking. This approach is essentially an alternative consensus scoring method using two distinct techniques to rank compounds in virtual screening. Unlike traditional consensus scoring approaches that combine two or more 3D based scoring functions, here we combine a 3D scoring function with a machine learning method in a 2D space. In the first phase of virtual screening of a compound database, HTD software uses information on the three-dimensional structure of a protein, as well as other associated information such as ionization states, and atomic charges to place the compounds in the target active site. The compounds were then ranked based upon their scores calculated from their resulting poses generated by the docking program by whatever scoring function the user chooses. In the cases presented here, the scoring functions provided by each program were used to score the poses. The second phase of this virtual screening method utilizes two-dimensional fingerprints (ECFPs) to train a naïve Bayes classifier based upon the top scoring compounds and to then rerank all of the compounds in the database.

A motivation for this work is based on trying to use HTD to reduce the number of compounds needed to be screening with HTS. HTD can be used prior to HTS by generating a focused screening collection based on highly ranked compounds. It is crucial that any technology that attempts to improve screening efficiency does so by capturing known actives within the first few percent of the database in order to reduce the costs associated with cherry picking a large subset of the database.

On the basis of these results, we suggest using naïve Bayes, combined with ECFPs, routinely to improve the enrichment results from HTD. In cases where HTD is already successful, naïve Bayes will provide significantly improved results. However, there are some limitations. In the case of PKB, the application of naïve Bayes did not improve the ranking of known actives. This was due to the fact that the "good" compounds from the original scoring used to create the naïve Bayes model contained 100% of false positives. In cases where HTD alone does not produce significant enrichment, it is not reasonable

to expect the classifier to produce reliable results. In other words, the classifier is incapable of rescuing poor HTD results since the classifier must be trained with meaningful data. A careful examination of the top scoring poses after HTD for potential artifacts, which are likely to lead to false positives, would alleviate this problem.

The work described in this paper relies on the hypothesis that the machine learning method, naïve Bayes, is tolerant toward noise.[9] Results obtained from HTD are often noisy due to the misclassification of compounds as false negatives or positives; particularly the numbers of false positives. The improved enrichments indicate that naïve Bayes combined with ECFPs can be successfully used to filter false positives and pinpoint the false negatives.

This method was developed to achieve the best enrichment of true positive compounds to the very top of the docked list of structures. Although this method is not intended to improve all aspects of HTD such as generating more accurate binding poses, as more accurate methods do, it is a computationally inexpensive, effective, and efficient tool, requiring only a few minutes on a desktop PC, which can be easily and generically applied to docking results for the purposes of creating, for example, a more target-specific focused library.

Experimental Section

Docking Hardware. All high-throughput docking calculations were carried out on a 200 processor Linux cluster consisting of 100 dual-processor Intel Pentium III CPUs (850 MHz) using the Linux 2.4.19 operating system.

Desktop Hardware. The Pipeline Pilot calculations in this paper were carried out on a desktop PC with a 2.0 GHz Intel Pentium 4 CPU with 1.00 GB of RAM operating under Microsoft Windows XP, version 2002.

Software. Software versions described in this paper were: Dock 4.0, FlexX 1.1, Glide 2.5, Sybyl 6.9, Unity 4.4, Maestro 5.1 and Pipeline Pilot 3.0.

Acknowledgment. The authors would like to thank Goran Pocina from Informatics and Knowledge Management at Novartis Research for the maintenance of the Linux cluster used to carry out the docking studies described in this paper.

References

(1) Abagyan, R.; Totrov, M. High-throughput docking for lead generation. *Curr. Opin. Chem. Biol.* **2001**, *5*, 375−382.
(2) Schneider, G.; Bohm, H. J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, *7*, 64−70.
(3) Gohlke, H.; Klebe, G. Statistical potentials and scoring functions applied to protein−ligand binding. *Curr. Opin. Struct. Biol.* **2001**, *11*, 231−235.
(4) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(5) 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.
(6) 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.
(7) Clark, R. D.; Strizhev, A.; Leonard, J. M.; Blake, J. F.; Matthew, J. B. Consensus scoring for ligand/protein interactions. *J. Mol. Graph. Model.* **2002**, *20*, 281−295.
(8) 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.
(9) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enrichment of extremely noisy high throughput screening data using a Naive Bayes classifier. *J. Biomol. Screen.* **2003**, *9*, 32−36.
(10) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, 2001.
(11) MDL Information Systems, Inc. Available Chemicals Directory. 2003. 14600 Catalina St., San Leandro, CA 94577.
(12) Tripos, Inc. 2003. 1699 South Hanley Rd., St. Louis, MO 63144.
(13) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.
(14) Schrodinger, L. L. C. 2003. 32nd Floor, Tower 45, 120 West Forty-Fifth Street, New York, 10036.
(15) Yang, J.; Cron, P.; Good, V. M.; Thompson, V.; Hemmings, B. A.; Barford, D. Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP−PNP. *Nat. Struct. Biol.* **2002**, *9*, 940−944.
(16) Iversen, L. F.; Andersen, H. S.; Branner, S.; Mortensen, S. B.; Peters, G. H.; Norris, K.; Olsen, O. H.; Jeppesen, C. B.; Lundt, B. F.; Ripka, W.; Moller, K. B.; Moller, N. P. Structure-based design of a low molecular weight, nonphosphorus, nonpeptide, and highly selective inhibitor of protein-tyrosine phosphatase 1B. *J. Biol. Chem.* **2000**, *275*, 10300−10307.
(17) Berman, H. M.; Battistuz, T.; Bhat, T. N.; Bluhm, W. F.; Bourne, P. E.; Burkhardt, K.; Feng, Z.; Gilliland, G. L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J. D.; Zardecki, C. The Protein Data Bank. *Acta Crystallogr. D. Biol. Crystallogr.* **2002**, *58*, 899−907.
(18) 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.
(19) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des* **2001**, *15*, 411−428.
(20) Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*; Morgan Kaufmann Publishers: New York, 1999.
(21) Scitegic, Inc. 2003. 9665 Chesapeake Dr., Suite 401, San Diego, CA 92123.
(22) Morgan, H. L. The generation of a unique machine description for chemical structures − a technique developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107−113.
(23) Microsoft Corporation. Excel. 2002.
(24) Akamatsu, M.; Roller, P. P.; Chen, L.; Zhang, Z. Y.; Ye, B.; Burke, T. R., Jr. Potent inhibition of protein-tyrosine phosphatase by phosphotyrosine- mimic containing cyclic peptides. *Bioorg. Med. Chem.* **1997**, *5*, 157−163.
(25) Andersen, H. S., Hansen, T. K., Lau, J., Moller, N. P. H., Olsen, O. H., Axe, F. U., Bakir, F., Ge, Y., Holsworth, D. D., Judge, L. M., Newman, M. J., Uyeda, R. T., Shapira, B. Z. Modulators of protein tyrosine phosphatases (PTPases). PCT/DK01/00451[WO 02/04459]. 1−17−2002. 6−28−2001.
(26) Johnson, T. O.; Ermolieff, J.; Jirousek, M. R. Protein tyrosine phosphatase 1B inhibitors for diabetes. *Nat. Rev. Drug Discovery* **2002**, *1*, 696−709.
(27) Leblanc, Y.; Dufresne, C.; Wang, Z.; Li, C. S.; Gauthier, J. Y.; Therien, M.; Roy, P. Phosphonic acids derivatives as inhibitors of PTP-1B. PCT/CA99/00864[WO 00/17211]. 3−30−2000. 9−21−1999.
(28) Vlattas, I.; Wennogle, L. P.; Sytwu, I. I.; Liang, H.; Mandiyan, S.; Yuryev, A. Combinatorial approaches to elucidate the subsite specificity of tyrosine phosphatases. *Proceedings of the European Peptide Symposium*; Bajusz, S., Hudecz, F., Eds.; Akademiai Kiado: Budapest, Hungary, 1998; pp 766−767.

JM030363K