Application of Machine Learning To Improve the Results of High-Throughput Docking Against the HIV-1 Protease

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

Lead Discovery Center, Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue,
Cambridge, Massachusetts 02139

Received July 6, 2004

We have previously reported that the application of a Laplacian-modified naïve Bayesian (NB) classifier may be used to improve the ranking of known inhibitors from a random database of compounds after High-Throughput Docking (HTD). The method relies upon the frequency of substructural features among the active and inactive compounds from 2D fingerprint information of the compounds. Here we present an investigation of the role of extended connectivity fingerprints in training the NB classifier against HTD studies on the HIV-1 protease using three docking programs: Glide, FlexX, and GOLD. The results show that the performance of the NB classifier is due to the presence of a large number of features common to the set of known active compounds rather than a single structural or substructural scaffold. We demonstrate that the Laplacian-modified naïve Bayesian classifier trained with data from high-throughput docking is superior at identifying active compounds from a target database in comparison to conventional two-dimensional substructure search methods alone.

INTRODUCTION

High-throughput docking (HTD) is a commonly used technique in modern drug discovery in which a database of hundreds of thousands or millions of compounds is screened against a protein target of interest in an attempt to identify novel compounds which elicit a desired biological response under physiologically relevant conditions. One of the difficulties in applying HTD in a consistent manner to identify such compounds is due to the inability of extant scoring functions to adequately and consistently discriminate between compounds which are binders and those which are nonbinders.

A variety of methods has been attempted in recent years to improve the performance of scoring functions after virtual screening has been carried out. Among these are the consensus scoring of HTD results,[1−3] sum-ranked fusion methods to combine the results of separate virtual screening approaches,[4] and the application of a naïve Bayes (NB) classifier to enrich the results of a single HTD experiment.[5] The success of each of these approaches has been described in the literature.

Here we show that the use of a Laplacian-modified naïve Bayes classifier[6] trained on the data from HTD is capable of not only identifying structurally similar compounds but also discriminating between active and inactive compounds within a given chemical class. This is due to the fact that the structural basis for the activity range observed for a given compound class cannot be represented by a single structural scaffold. The range of activities associated with compounds belonging to a single structural class must therefore be represented by a larger set of structural elements in order to account for this granularity of activity. This paper aims to explain why the unique combination of a Bayesian model and extended connectivity fingerprints (ECFPs)[7,8] are so successful at enriching HTD scores. The results show that far from being simply a similarity method essentially akin to carrying out a substructure search of a chemical database, the naïve Bayes classifier is indeed capable of improving the enrichment of HTD results due to the subtleties of the chemical features used in the training set.

The test case presented here involves the application of a Laplacian-modified naïve Bayesian classifier[6] trained with ECFPs to enrich the HTD results obtained using three different docking programs in a virtual screening campaign against the HIV-1 protease. The three docking programs selected for this study were Glide,[9,10] FlexX,[11] and GOLD.[12] The database screened was composed of 179 805 compounds from the Available Chemicals Directory (ACD),[13] which was then seeded with a total of 424 Novartis compounds which were known inhibitors of the HIV-1 protease. After HTD was carried out, the top-ranked compounds were then designated as “good”, or active compounds, while all others were designated as “bad”, or inactive compounds. ECFPs were then calculated for all compounds in the entire data set, and the naïve Bayesian classifier was then trained using the ECFPs from the “good” and “bad” compounds. All of the docked compounds were then reranked according to the NB model. The workflow followed in this paper is outlined in Figure 1.

METHODS

Preprocessing of the Compound Database for High-Throughput Docking. A set of 179 805 compounds from the Available Chemicals Directory (ACD)[13] was used for the background data set, along with two sets of known HIV-1 protease inhibitors (Table 1), prepared according to the protocol described previously.[5] The two sets of known HIV-1 protease inhibitors (HIVSet I and HIVSet II) consisted of compounds from in-house projects for which IC_{50} values were available. The first set consisted of 424 proprietary

* Corresponding author phone: (617)871-7132; fax: (617)871-4088; e-mail: anthony.klon@pharma.novartis.com.

10.1021/ci0497861 CCC: $27.50
Published on Web 10/07/2004

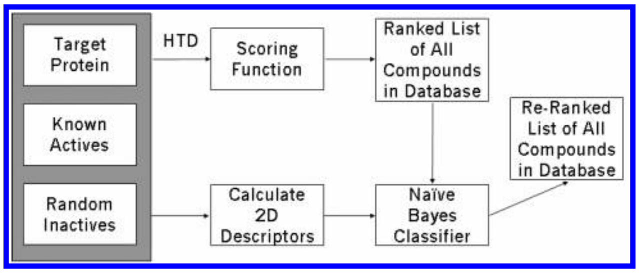


Figure 1. Schematic of the workflow followed in this paper.

compounds from three structural classes (HIVSet I). The second set consisted of 175 peptidic HIV-1 protease inhibitors from a single combinatorial series[14] (HIVSet II), which was a subset of HIVSet I. To prepare the database for docking with FlexX[11] and GOLD,[12] ionization was carried out using scripts provided by Tripos.[11] In addition, Gasteiger–Marsili[15] partial charges were added for all compounds in the database in preparation for docking with FlexX. To prepare the database for docking using Glide, the ionization of the compounds at pH 7.0 was carried out using scripts provided by Schrödinger.[10] For GOLD and Glide, it was unnecessary to assign formal and partial charges, respectively, because these programs assign the necessary atomic charges to compounds at run time.

Preparation of the HIV-1 Protease for High-Throughput Docking. A cocrystal structure of HIV-1 protease with a bound inhibitor, JE-2147 (PDB id # 1KZK),[16] was selected from the protein data bank.[17] JE-2147 is a tetrapeptide mimetic which is a picomolar inhibitor of the HIV-1 protease. This particular cocrystal structure was selected due to the fact that the resolution was 1.09 Å, the refined R_{free} was 15.2%, and the high data-to-parameter ratio made the refinement of the anisotropic B-factors possible. These characteristics made this the most accurate HIV-1 protease X-ray structure available at the time the HTD studies discussed in this paper were carried out.

All water molecules were stripped from the PDB file, and the inhibitor was removed. Hydrogen atoms were added using Sybyl[11] to prepare the protein for docking with FlexX and GOLD or Maestro[10] in order to prepare the target for docking with Glide. In both cases, residue Asp-25 on chain B was protonated in light of the crystallographic evidence of the role of Oδ_2 in hydrogen bonding with the bound inhibitor.[16]

Docking of the Test Set against the HIV-1 Protease Using FlexX. The default software parameters were used as supplied by Tripos with Sybyl 6.9. The residue description file contained all protein atoms within 6.5 Å of the ligand molecule from the cocrystal structure. The single best pose for each compound generated by the docking algorithm, as calculated using the FlexX scoring function, was retained and written out to a single multi-mol2 file.

Docking of the Test Set against the HIV-1 Protease Using Glide. The default software parameters supplied with Maestro 5.1 for the creation of command files for high-throughput docking with Glide were used. For generation of the scoring grids, the size of the box defining placement of the ligand center was left at the default value of 12 Å for the x, y, z axes, while the size of the enclosing box was 40 Å on each side in order to accommodate the larger compounds from the ACD as well as the HIV-1 protease inhibitors. The value of 0.9 × the van der Waals radius of the atoms was used for the protein, and 0.8 × the van der Waals radius of the atoms was used for the compounds. The scaling of the van der Waals radii is a crude approximation of ligand flexibility in the active site in order to ensure that the Glide scoring function is not too strict in rejecting ligands which may have plausible binding poses to the inhibitor-bound conformation of the HIV-1 protease from the X-ray crystal structure. The maximum number of heavy atoms permitted per compound was 120, and the maximum number of rotatable bonds allowed per compound was 30. The top scoring pose for each compound, as assessed by its Glide score, was exported to a Maestro-formatted output file.

Table 1. Number of Known Active Compounds for Each Data Set

active set	no. of active compds	no. of structural classes
diverse HIV-1 protease inhibitor set (HIVSet I)	424	3
peptidic HIV-1 protease inhibitor set[a] (HIVSet II)	175	1

[a] See ref 14.

Docking of the Test Set against the HIV-1 Protease Using GOLD. The active site of the HIV-1 protease was constrained to those atoms within 25 Å of the hydrogen atom attached to Oδ_2 of Asp-25 of chain B. This atom was selected due to its proximity to the center of the active site, and the 25 Å radius was used because of the large volume of the HIV-1 protease active site. The default software screening settings as described by Jones et al. were used for the parameters controlling GOLD's genetic algorithm.[12] A total of 100 000 genetic operations were carried out on five islands, each containing 100 individuals. The niche size was set to 2, and the value for the selection pressure was set to 1.1. Genetic operator weights for crossover, mutation, and migration were set to 95, 95, and 10, respectively. Docking of the database was carried out on 1800 CPUs within the Novartis Research environment using the United Devices Grid.[18] The top-scoring pose generated for each compound according to the highest GOLD score was written out to a single multi-mol2 file.

Training the Naïve Bayes Classifier with Extended-Connectivity Fingerprints. ECFPs are a new class of fingerprints developed by SciTegic[8] based upon the Morgan algorithm.[19] One ECFP corresponds to a single feature as defined in Figure 2. Several examples of individual features derived in this study are shown in Figure 6. Figure 1 describes the workflow followed in this paper.[5] After HTD, a ranked list was generated based upon the single best scoring docked pose for each compound, according to the Glide, FlexX, or GOLD scoring functions, respectively. The "good" compounds were defined to be those with a score three standard deviations below the mean in the cases of Glide and FlexX or three standard deviations above the mean in the case of GOLD. The remaining compounds are defined as "bad". The sets of "good" and "bad" compounds were passed to the naïve Bayesian classifier in Pipeline Pilot, which then calculated the ECFPs with a neighborhood size of six for each compound. The frequency of "good" and "bad" features present among the training sets were then used to train the NB classifier. The NB model was then used to rerank all compounds in the database according to their chemical structure.

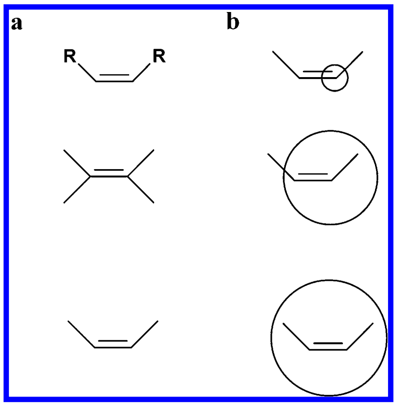


Figure 2. Using extended-connectivity fingerprints to carry out structural searches. (*a*) A substructural search using the top fragment would return compounds containing the middle and bottom features. A search using this fragment with specified attachment sites, **R**, will return only compounds containing the bottom feature. (*b*) The circle identifying the single atom *top* corresponds to an ECFP with a neighborhood size of 0 (ECFP_0). Subsequent iterations update this atom's code to include information about atoms one (*middle*) and two (*bottom*) bonds distant, corresponding to ECFPs with a neighborhood size of one (ECFP_1) and two (ECFP_2), respectively.

Evaluating the Enrichment of the Known Actives after High-Throughput Docking and Naïve Bayesian Classification. For each of the six HTD runs, all of the compounds successfully docked were ranked according to the scores assigned by each docking program (FlexX, Glide, or GOLD), and these rankings were used to calculate enrichment curves and the area under the receiver operating characteristic (ROC) curves[20] using Pipeline Pilot.[8] All six data sets were reranked according to the models created by the Laplacian-modified naïve Bayes classifier within Pipeline Pilot, and the enrichment curves and values for the area under the ROC curves were recalculated. We have previously described the protocol used for training the naïve Bayes classifier with the results from HTD and reranking the compounds using the trained model.[5]

Similarity Searching of the Database. Tanimoto similarity, maximal common substructure (MCSS) analyses, and substructure searches were all carried out using protocols implemented in Pipeline Pilot. For the substructure searches, the options "StereoAtomsCanMapNonStereoAtoms" and "HFillIfRAtomsFound" were turned on. The former feature was necessary due to the fact that in training the NB classifier, the three-dimensional poses generated by the docking programs were converted into SMILES strings,[21] resulting in a loss of chirality information. The second parameter ensured that in the cases where R groups were present when searching the data set using a feature from the NB model, the attachment sites were restricted. A constrained substructure search restricting attachment points at R groups ensures consistency with the features generated by the ECFPs in Pipeline Pilot.

The Laplacian-Modified Naïve Bayesian Classifier. The naïve Bayesian classifier is a statistical model based upon Bayes' rule of conditional probability

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

P(A|B) is the probability that event *A* will occur given that event *B* occurred. In the context of this paper, event *A* refers to a compound's activity, while event *B* refers to the presence of a given feature, defined by an ECFP. *P(A)* is the probability that a given compound in the data set will be active, while *P(B)* is the probability that a given feature will occur in the data set. *P(A|B)* is therefore the probability that a compound from the data set will be active given that it

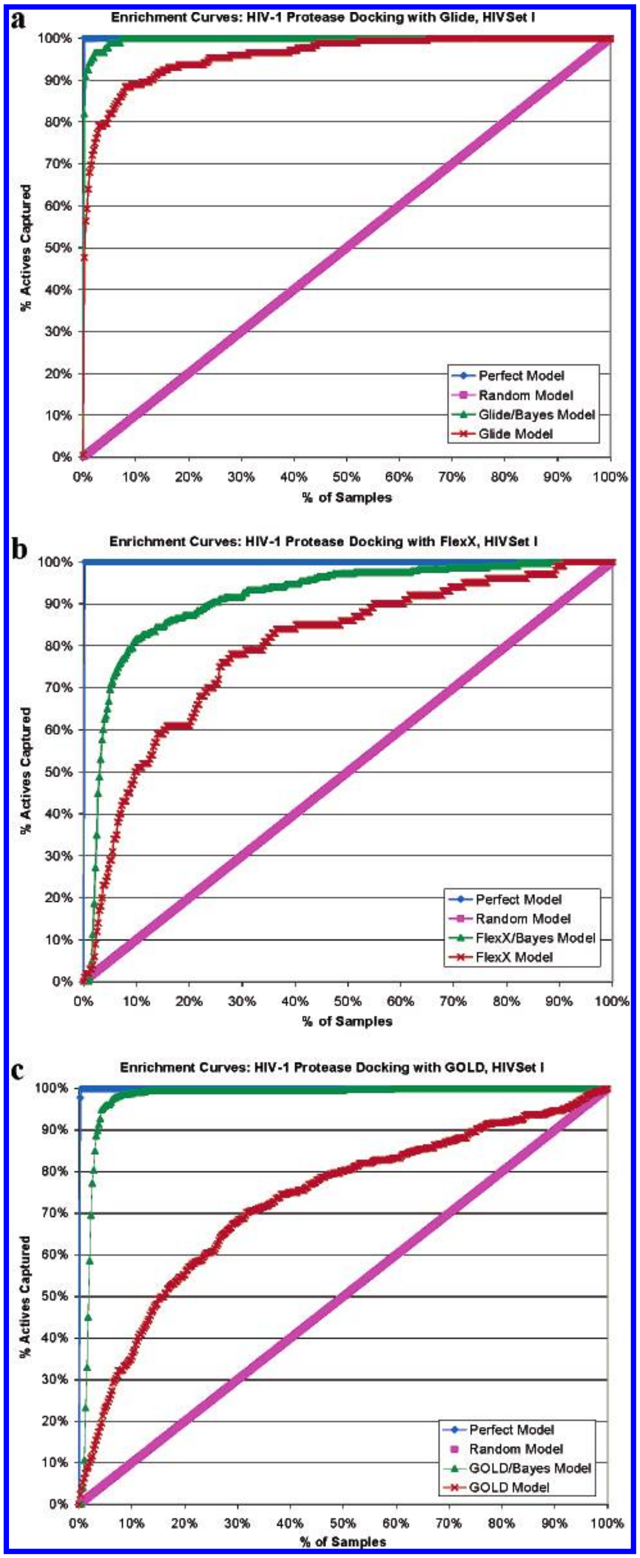


Figure 3. Enrichment curves calculated after HTD using (*a*) Glide, (*b*) FlexX, and (*c*) GOLD before (*red*) and after (*green*) application of the naïve Bayes classifier. The background set was seeded with 424 known HIV-1 protease inhibitors from HIVSet I.

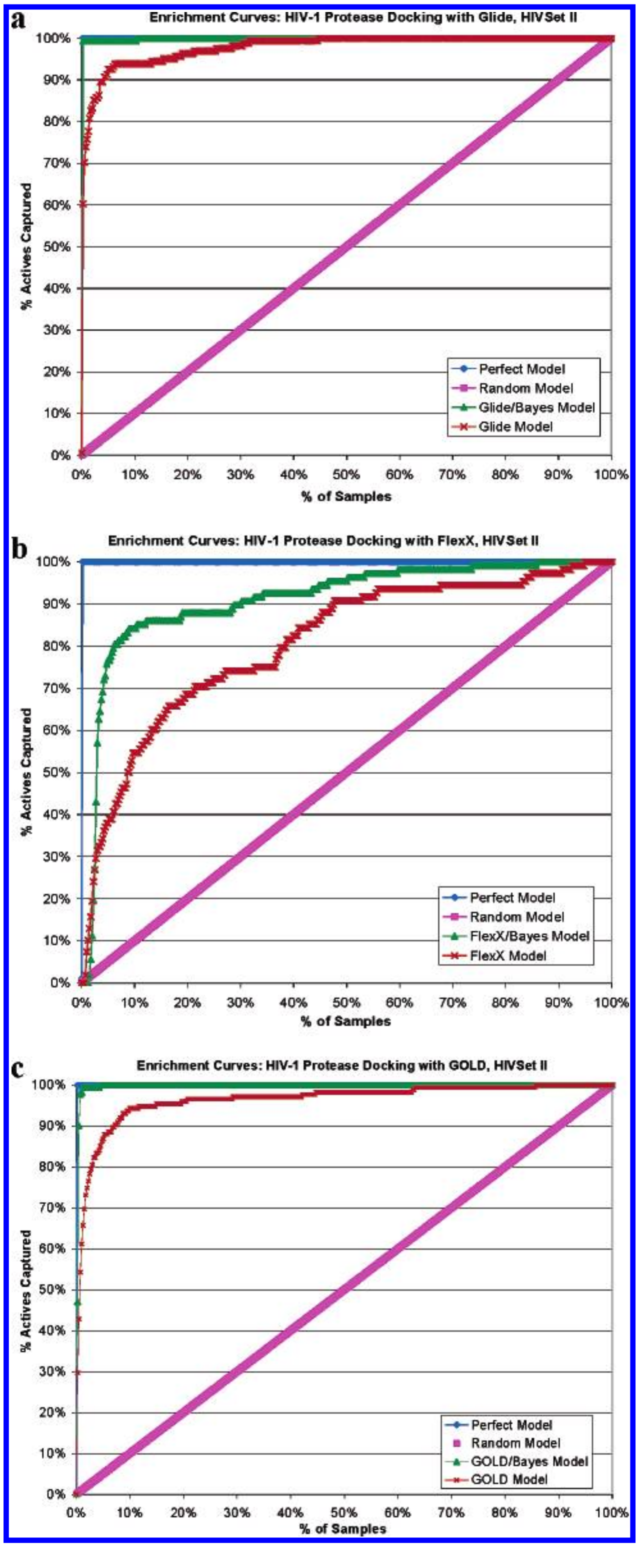


Figure 4. Enrichment curves calculated after HTD using (*a*) Glide, (*b*) FlexX, and (*c*) GOLD before (*red*) and after (*green*) application of the naïve Bayes classifier. The background set was seeded with set of 175 peptidic HIV-1 protease inhibitors from HIVSet II.

has a particular feature. Similarly, *P(B|A)* is the probability that a compound will have a certain feature given that it is active. This probability is predicted from *P(B|A)*, *P(A)*, and *P(B)*:

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

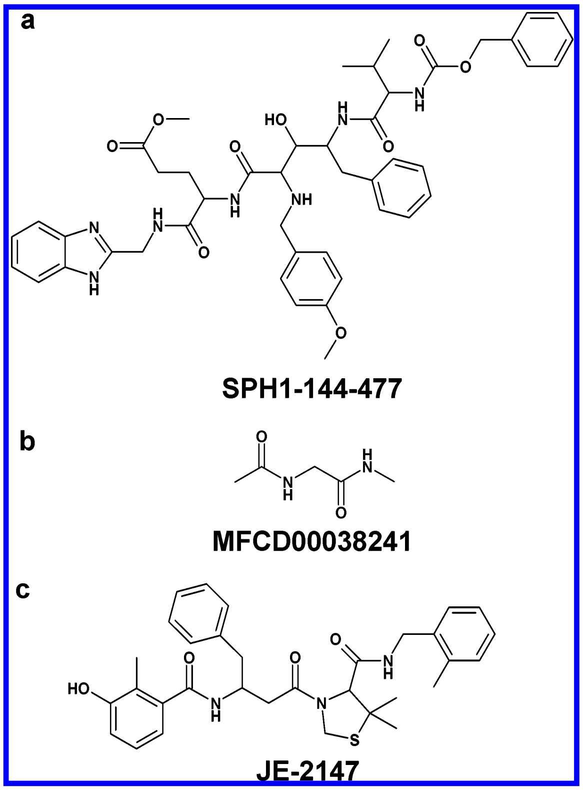


Figure 5. (*a*) The single known HIV-1 protease inhibitor found in the training set after HTD carried out using FlexX against the ACD seeded with 175 previously published inhibitors. (*b*) The simple dipeptide used for substructure searching as described in the text. (*c*) The HIV-1 protease inhibitor from the cocrystal structure reported by Reiling et al.

The Bayesian classifier is called naïve because it naively assumes the features are independent. From this assumption, it is valid to multiply probabilities of the individual events. Because each compound possesses *n* features, it follows that

$$P(active|feature) = P(feature_1|active) \times P(feature_2|active) \times P(feature_3|active) \times \ldots P(feature_n|active)\frac{P(active)}{P(feature)}$$

If we have a large sample size for a given feature,

$$P(active|feature) = \frac{feature_{act}}{feature_{tot}}$$

where $feature_{tot}$ is the total number of compounds with a given feature, and $feature_{act}$ is the number of compounds with a given feature that are active. However, as the number of samples becomes small, the estimate for *P(active|feature)* becomes overconfident. For example, if $feature_{tot} = feature_{act} = 1$, then

$$P(active|feature) = 1$$

The Laplacian modification corrects for this by adding a “virtual” sample:

$$P(active|feature) = \frac{(feature_{act} + P(active))}{(feature_{tot} + 1)}$$

As a result, as the number of samples containing a feature

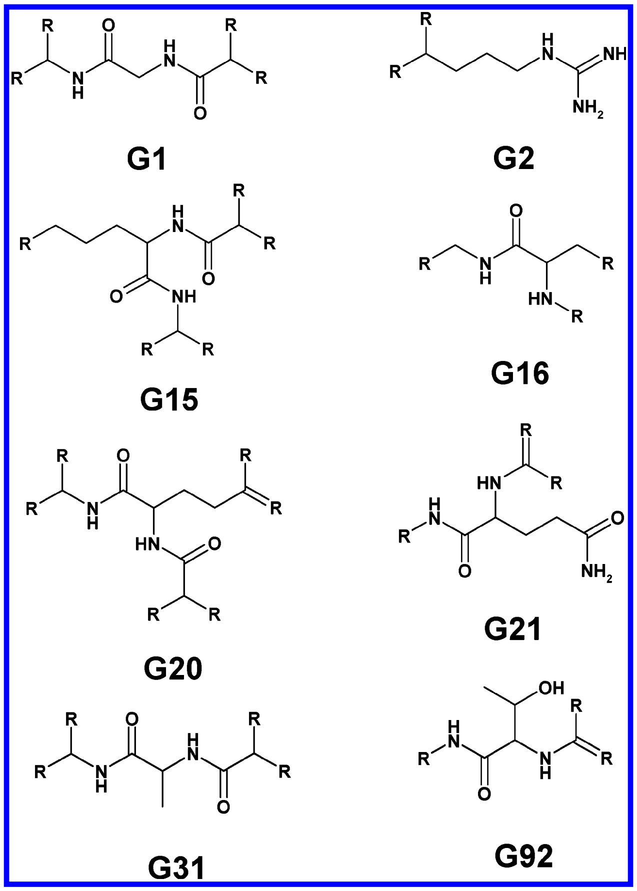


Figure 6. Eight of the top "good" features generated by the naïve Bayes classifier for the NB model based upon the top scoring 1027 compounds after HTD of the ACD database seeded with 175 peptidic HIV-1 protease inhibitors (HIVSet II) with FlexX.

(*feature$_{tot}$*) decreases, the feature's contribution to the probability converges to *P(active)*.[6]

RESULTS

Naïve Bayesian Enrichment of High-Throughput Docking Results. Table 1 shows the total number of compounds and structural classes used in the two HIV-1 test cases presented here. Figures 1 and 2 show the enrichment curves for docking of the test database against the HIV-1 protease using (*a*) Glide, (*b*) FlexX, and (*c*) GOLD, seeded with 424 (Figure 3) and 175 (Figure 4) active compounds. Also shown in each figure is the resulting enrichment curve after the application of the naïve Bayes classifier. In each case, NB is shown to improve upon the desired ranking of known inhibitors from HTD. Table 2 shows the calculated values for the areas under the ROC curves for all six test cases, before and after the application of NB.

As shown in Table 2, the initial high-throughput docking results (expressed as the area under the ROC curve and as the % of actives captured) prior to application of the naïve Bayes classifier are quite impressive for Glide. These results are consistent with those obtained by Halgren et al. for the performance of Glide in database screening against this target.[9] In both test cases, Glide generates excellent results, ranking 82% and 93% of the known inhibitors from the unpublished and published inhibitor sets in the top 2% of the database after docking. After application of the naïve Bayes classifier to the docking results obtained with Glide, 99% of the known inhibitors are ranked in the top 2% of the database in both cases.

Table 2. Area under ROC Curves and Percentage of Actives Recovered in Top 2% of Database Screened for All Six Test Cases, before and after Application of the Naïve Bayes Classifier

		area under ROC curve		percentage of actives captured in top 2% of database screened	
program	inhibitor set	before NB	after NB	before NB	after NB
Glide	HIVSet I	0.96	1.00	82	99
	HIVSet II	0.98	1.00	93	99
FlexX	HIVSet I	0.80	0.91	38	77
	HIVSet II	0.81	0.91	29	70
GOLD	HIVSet I	0.73	0.98	23	96
	HIVSet II	0.96	1.00	87	100

Table 3. Total Number of "Good" Compounds and Number of Known Inhibitors Used in the Naïve Bayes Training Sets for the Six Test Cases

program	inhibitor set	no. of "good" compds in training set	no./%age of known inhibitors present among the "good" compds (%)
Glide	HIVSet I	305	73/23.9
	HIVSet II	300	85/28.3
FlexX	HIVSet I	1031	7/0.7
	HIVSet II	1027	1/0.1
GOLD	HIVSet I	302	8/2.6
	HIVSet II	335	44/13.1

The more interesting cases are those results from the docking carried out using FlexX and GOLD. In both cases, FlexX and GOLD produce moderate enrichment against the target protein when the 424 compound data set is used. For the 175 peptidic inhibitors GOLD yields excellent enrichment prior to application of the naïve Bayes classifier, while FlexX continues to generate moderate results. In both of the FlexX cases, application of NB is able to yield excellent enrichment, as represented by Figures 3 and 4 as well as the values for the areas under the corresponding ROC curves shown in Table 2. The detailed discussions throughout the rest of this paper will concentrate on the specific case of the docking of the ACD seeded with 175 peptidic inhibitors (HIVSet II) against the HIV-1 protease using FlexX.

Analysis of the Naïve Bayes Training Set. As reported previously[5] and demonstrated in the previous section, naïve Bayes is capable of enriching HTD results in cases where docking already provides an initial positive enrichment.[5] This would seem to be intuitively obvious because the classifier is trained on only the top scoring docking results, and these compounds are structurally similar to other active compounds in the database. The naïve Bayesian classifier creates a model using a set of two-dimensional descriptors from the top scoring compounds, so the success of the ECFP/Bayes method might be similar to carrying out a similarity search of the target database. If this were the case, one could in principle simply search for a set of maximal common substructures from among the top scoring compounds and use these results to carry out substructure searches of the rest of the entire database. Tables 4 and 5 illustrate the weakness of this intuitive argument.

For the case of FlexX docking against the ACD seeded with the 175 peptidic inhibitors, a total of 1027 compounds was used in the training set for NB. However, only one compound, SPH1-144−477, was a true active (Figure 5a).

Table 4. Total Number of Compounds and Number of Known Inhibitors Retrieved Using MFCD0038241 as a Function of Tanimoto Similarity

Tanimoto similarity	no. of ACD compds retrieved with MFCD00038241	no. of HIVSet I inhibitors retrieved with MFCD00038241	no. of ACD compds retrieved with JE−2147	no. of HIVSet II inhibitors retrieved with JE−2147
0.6	1	0	0	0
0.5	2	0	0	0
0.4	13	0	0	0
0.3	57	0	0	0
0.2	692	0	427	21
0.15	4390	0	7672	105
0.1	31712	14	61790	106

Table 5. Number of Compounds Retrieved with Substructure Searches Carried out Using Several “Good” Features (Shown in Figure 6) Present in the Model Generated by the Naïve Bayes Classifier for the FlexX HTD Results

	entire data set		top 1027 compds	
feature	HIVSet II	ACD	HIVSet II	ACD
G1	1	639	0	274
G2	0	419	0	206
G15	0	326	0	178
G16	4	575	1	227
G20	2	186	1	125
G21	1	95	0	65
G31	0	411	0	186
G92	0	109	0	48

The question then arises: How is the application of the naïve Bayesian classifier able to enrich the results of HTD in a case where there appears to be a paucity of information about the chemical structure of the known actives? Two explanations are possible. In the first case, the top-ranking compounds have no structural relationship to the set of true actives. This presents difficulties, as NB would not appear to have sufficient structural information to generate a model about what a “good” compound looks like. In the second instance, the set of random inactive compounds which rank highly after HTD are structurally similar to the set of known inhibitors. But if the inactive set used to train the classifier is structurally similar to the active set, how is NB able to generate an improved enrichment?

Figure 6 shows some of the top-scoring features present in the NB model generated from the FlexX docking data for case when the ACD was seeded with 175 peptidic inhibitors. For this case, the NB classifier generated a total of 84,701 good and bad features for the model. The top-ranked feature present in the NB model, G1, is essentially a simple dipeptide bond. There is a single compound in the ACD which matches this exact structure, MFCD00038241 (Figure 5b). This compound moves from 69561 to 5962 after application of the NB classifier. This ACD structure was used for further studies with database searching using a maximal common substructure as well as Tanimoto similarity.

Retrieval of Known HIV-1 Protease Inhibitors Using a Substructure Search. A search was carried out using MFCD00038241 (equivalent to ECFP_6 fragment G1 shown in Figure 6) on the database containing the combined ACD and the 175 known peptidic HIV-1 protease inhibitors. For the entire database, 3075 compounds from the ACD as well as 122 known HIV-1 inhibitors were retrieved. A search with MFCD00038241 as a query was also carried out against the top 1027 compounds used in the training set for the NB classifier. This resulted in 750 compounds from the ACD being retrieved, along with a single compound from the set of known HIV-1 protease inhibitors, SPH1-144−477. Clearly, searching the entire database on the basis of a single substructure would be successful in retrieving those compounds with a peptide-like structure from the ACD, while it would not be successful in retrieving all of the known HIV-1 protease inhibitors. Because the set of known inhibitors are not strictly all dipeptides, this would require a method which is capable of searching a database for more diverse chemical structures.

Retrieval of Known HIV-1 Protease Inhibitors Using Tanimoto Similarity. MFCD00038241 was used to search the combined ACD/HIV-1 protease peptidic inhibitor database using the Tanimoto similarity in an attempt to retrieve known inhibitors of the HIV-1 protease. Table 4 shows the results for searching the combined database using this scaffold. Results are reported as the total number of compounds retrieved as well as the number of known HIV-1 inhibitors retrieved for a given Tanimoto similarity. The results show that even at a very low Tanimoto score of 0.1, only 14 of the known HIV-1 protease inhibitors are retrieved. By contrast, 31 712 compounds from the background ACD are retrieved. This extremely low hit rate demonstrates that using a single scaffold common to a large percentage of the known inhibitors in order to search the target database by Tanimoto similarity is not sufficient to recall the active compounds.

We repeated the search based upon Tanimoto similarity using JE-2147. As seen in the results shown in Table 4, using the cognate ligand yields superior results. However, the results are still quite poor, with only 105 inhibitors from HIVSet II being identified with a Tanimoto similarity of 0.15 or better. By contrast, the number of false positives retrieved from the ACD, 7672, is also considerably higher.

Retrieval of Known HIV-1 Protease Inhibitors Using the Top “Good” Features from the Naïve Bayes Model. The naïve Bayes classifier generated a model containing 84 701 “good” and “bad” structural features. To determine how many of the most highly ranked features would be required to described the entire set of known HIV-1 protease inhibitors used in this study, the model created by NB was used to screen the data set in an interactive fashion using the “Learned Feature Filter” in Pipeline Pilot.[8] It was determined that the top 277 “good” features in the NB model were necessary to capture all 175 inhibitors. To determine what fraction of the background test set this corresponded to, the top 277 features from the NB model were then used to screen the ACD for compounds which matched any of these features. A total of 12 647 compounds from the ACD were identified as containing at least one of the 277 “good” features. These results suggest that rather than a single structural feature defining the majority of the compounds in the set of known inhibitors, it is instead the presence of a large number of features which is necessary to describe the set of known inhibitors.

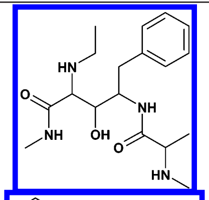
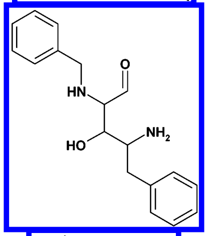
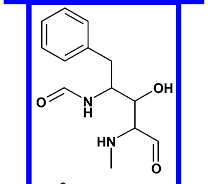
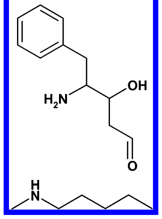
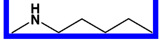
This situation is the opposite of the substructure searching methods described previously. Furthermore, although 277 features were required overall to describe the set of known inhibitors, this is a comparatively small number with respect to the total number of features generated by the NB classifier, corresponding to 0.327% of all the features generated by the classifier for the model in order to describe the entire data set.

Using the Top "Good" Features from the Naïve Bayes Model in a Substructure Search. A subset of the most important "good" features generated by the NB classifier for the naïve Bayesian model were selected for a substructure search of the database. These features are shown in Figure 6. These features were then used in separate substructure searches of the combined ACD/HIV-1 protease inhibitor database. By carrying out the substructure searches using these features, attachment points were permitted only at the locations of the R-groups. This differs from the substructure search described earlier which utilized a dipeptide bond as a substructure and did not specify attachment points. Table 5 reports the resulting number of compounds obtained from the ACD and HIV-1 inhibitor data sets after searching the entire database as well as the top 1027 compounds used to train the NB classifier. These results show that carrying out a substructure search using individual "good" features from the naïve Bayes model in this manner is not adequate to identify the known HIV-1 inhibitors in the database. Instead, a large number of structural features and their frequencies among the "good" vs "bad" compounds are required to adequately describe the set of known inhibitors. This confirms that NB is looking for combinations of favorable features rather than the presence of these features in isolation. This provides further support to the validity of the strategy of using a naive Bayes model trained on the top ranking compounds after HTD instead of simply carrying out a maximal common substructure search, followed by a search of the entire database with the MCSS.

Retrieval of Known HIV-1 Protease Inhibitors through the Use of Maximal Common Substructures. Several substructure searches were carried out on HIVSet II (175 compounds), which used the results of a MCSS on HIVSet II as a query. A range of values was used for the required proportion of the data set to be the accounted for by the largest maximal subgraph. The maximal subgraphs obtained were then used to carry out a substructure search against the entire ACD test set in order to determine how sensitive each substructure would be at retrieving the known inhibitors against the background database. The five maximal subgraphs calculated are shown in Table 6. The proportion of the known inhibitor set containing each subgraph as well as its frequency is shown in Table 6 along with the total number of ACD compounds retrieved by each after a substructure search.

The results clearly show that in order to retrieve a large number of compounds from HIVSet II, increasingly trivial scaffolds must be used as a search query. In particular, these scaffolds are not necessarily the intuitively obvious ones, such as a dipeptide fragment (MFCD00038241). Furthermore, the problem introduced by using increasingly trivial chemical scaffolds as a search query is that the number of false hits obtained from the background database increases quite substantially. This problem becomes exacerbated in the more realistic cases such as that shown with HIVSet I, where the true active compounds represent multiple chemical classes. In this case, decisions must be made in which separate chemical scaffolds must be defined as a search query in order to recover compounds from each class.

Table 6. Proportion of HIV-1 Protease Inhibitors Matching Results from Maximal Common Substructure Searches[a]

MCSS	frequency/ proportion of HIV-1 protease inhibitors	no. of ACD compds matching substructure query
	134/0.757	0
	158/0.90	0
	168/0.96	0
	173/0.989	3
	175/1.0	12754

[a] Also displayed are the total number of ACD compounds retrieved for each MCSS when used as a substructure query.

Naïve Bayes Avoids Artificial Enrichment due to the Presence of One-Dimensional Descriptor Similarity. We investigated whether the improved enrichments observed for the application of the NB classifier to the docking results were simply due to the identification of a set of one-dimensional descriptors from a subset of compounds in the ACD with properties similar to those observed in the known HIV-1 protease inhibitors. Verdonk et al. have recently cautioned that HTD using a set of random inactives as the background set could result in an artificial enrichment.[22] Consequently, we generated a filter which allowed 403 of the known inhibitors from HIVSet I (95%) to pass with the following characteristics: molecular weight between 350 and 950, number of hydrogen bond acceptors between four and ten, and number of hydrogen bond donors between two and nine. Applying this same set of filters to the ACD background set passed a total of 12 598 compounds.

The calculated areas under the ROC curves for the HTD results prior to the application of the NB classifier for this smaller set of compounds were 0.86, 0.42, and 0.48 for the Glide, FlexX, and GOLD results, respectively. We found it necessary when applying the NB classifier to define a new cutoff with this set of docked compounds. The three standard deviation cutoff described in the Methods Section resulted in 44, 68, and 0 compounds being passed to the NB classifier using the docking results from Glide, FlexX, and GOLD, respectively. Thirty-one of the compounds from the Glide results were known inhibitors, while none of the 68 compounds from the FlexX results were known inhibitors. No model could be generated from the GOLD results. The NB classifier trained on the Glide and FlexX results yielded values for area under the ROC curves of 1.00 and 0.34, respectively. Altering the cutoff for the docking results from three standard deviations to two standard deviations enabled the construction of a NB model for all three test cases, and the subsequent application of the NB classifier resulted in calculated areas under the ROC curves of 1.00, 0.50, and 0.87 for the Glide, FlexX, and GOLD results, respectively. The NB model was therefore able to distinguish between true inhibitors and background compounds with similar one-dimensional molecular properties and generate a significant enrichment in two of the three cases.

DISCUSSION

The use of a naïve Bayes classifier trained with extended connectivity fingerprints has several advantages over other similarity methods. There is no need to define a structural "probe" prior to carrying out the search operation. This results in substantially less bias being introduced on the part of the user. Naïve Bayes has previously been shown to perform well in cases where a high level of noise is present.[23] This is also true for the FlexX data presented here where a single true positive is present in the NB training set. The use of NB model trained with ECFPs results in a more detailed two-dimensional representation, resulting in a higher recall and higher precision rates compared with substructure search methods alone. This is exemplified by the FlexX case where the NB model was composed of a large number of features. This approach is analogous to NMR fragment-based screening methods.[24] The affinity of a compound for a protein is a sum of its substructural features. While it may not be possible to directly measure the IC_{50} data of a fragment, several fragments with weak binding affinities individually can be used to build higher MW compounds with a more favorable IC_{50} value.

We have investigated the details of the application of a Laplacian-modified naïve Bayes classifier as implemented in Pipeline Pilot[8] and used in a protocol to improve the results from high-throughput docking.[5] Although success might be expected in situations where a large percentage of the compounds are designated "good" compounds in the training set, the situation becomes less clear when the majority of the molecules in the "good" training set are false positives from HTD. Data presented here as well as in previous publications suggest that even in these cases, NB is capable of improving the results as long as there is at least modest initial enrichment after HTD. A comparison of the total percentage of "good" compounds present in the training sets shown in Table 3 with the corresponding areas under the ROC curves shown in Table 2 shows that there is no strong correlation. This indicates that NB can tolerate the very noisy data sets by improving the signal-to-noise ratio through the selection of features present among the "good" compounds.

We have carried out similar analysis of the best scoring features from the other five Bayesian models created for the Glide/HIVSet I, Glide/HIVSet II, FlexX/HIVSet I, GOLD/HIVSet I, and GOLD/HIVSet II cases (data not shown). The top scoring "good" features are the same or are structurally very similar among all cases studied, including the FlexX/HIVSet II case discussed in greater detail here. This comes as no surprise due to the fact that the same compounds and therefore the same structural features are present in all six test cases. In addition, all three software packages are successful in generating enrichment prior to application of NB, so the top-ranked "good" features are similar across all six test cases. The primary difference between the NB models generated for the Glide/HIVSet I, FlexX/HIVSet I, and GOLD/HIVSet I lies in the probability of whether a given feature is favorable or unfavorable. This same trend is observed for the Glide/HIVSet II, FlexX/HIVSet II, and GOLD/HIVSet II test cases.

We have shown that although the naïve Bayes classifier is trained using the extended connectivity fingerprints implemented in Pipeline Pilot, the method is more than a simple substructure search algorithm. Our data show that using a simple scaffold common to all, or a large percentage of known inhibitors of the HIV-1 protease, is not sufficient to adequately distinguish between these compounds and decoy molecules, particularly when the decoys possess similar substructures to those present in the set of known inhibitors. This assertion is further supported by the observation that using the top features present in the model generated by the NB classifier to carry out a substructure search of the database individually returns few of the known inhibitors. The large number of features which are required to extract the known inhibitors illustrate that it is not the presence of a single scaffold, or a small number of scaffolds, which the NB classifier is using to identify these inhibitors. Instead, the information contained in a large number of structural features is required to distinguish these molecules from the set of random compounds, many of which are structural decoys.

EXPERIMENTAL SECTION

Linux Cluster Hardware. High-throughput docking calculations with Glide and FlexX were carried out on a 92 processor Linux cluster consisting of 46 dual-processor Intel Xeon CPUs (2.80 GHz) operating under the Linux 2.4.22 operating system.

Desktop Hardware. All Pipeline Pilot protocols described in this paper were executed on a desktop PC with 2.0 GHz Intel Pentium 4 CPU with 1.00 GB of RAM operating under Microsoft Windows XP, version 2002. Docking calculations with GOLD were carried out over 1800 desktop PCs in the Novartis Research environment which were operating under the United Devices Grid.[18]

Software. Software versions used in this paper were as follows: Maestro 5.1, Glide 2.5, Sybyl 6.9, GOLD 2.0, FlexX 1.1, Unity 4.4, and Pipeline Pilot 3.0.

ACKNOWLEDGMENT

The authors would like to thank Dmitri Mikhailov from Informatics and Knowledge Management at Novartis for the maintenance of the Linux cluster used to carry out docking studies with Glide and FlexX, Justus Hainmueller for the maintenance of the United Devices Grid on which the GOLD studies were carried out, and Armin Widmer at Novartis

Research for the creation of the Perl scripts used to control job submission on the UD Grid.

REFERENCES AND NOTES

(1) 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.

(2) 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.

(3) 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.

(4) Raymond, J. W.; Jalaie, M.; Bradley, M. P. Conditional probability: A new fusion method for merging disparate virtual screening results. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 601−609.

(5) 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.

(6) Rogers, D. A Laplacian best-feature model for thrombin inhibitors. Ford, M., Livingstone, D., Dearden, J., Van de Waterbeemd, H., Eds.; Blackwell Publishing: Malden, MA, 2002; pp 281−283. Euro QSAR 2002: Designing drugs and crop protectants.

(7) Rogers, D.; Hahn, M. 2004, Unpublished manuscript.

(8) Scitegic, Inc. 2003. 9665 Chesapeake Drive, Suite 401, San Diego, CA 92123.

(9) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(10) Schrodinger, L. L. C. 2003. 32nd Floor, Tower 45, 120 West Forty-Fifth Street, New York, 10036.

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

(12) 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.

(13) MDL Information Systems, Inc. 2003. 14600 Catalina Street, San Leandro, CA 94577.

(14) Kroemer, R. T.; Ettmayer, P.; Hecht, P. 3D-quantitative structure−activity relationships of human immunodeficiency virus type-1 proteinase inhibitors: comparative molecular field analysis of 2-hetero-substituted statine derivatives- implications for the design of novel inhibitors. *J. Med. Chem.* **1995**, *38*, 4917−4928.

(15) Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity − a rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219−3288.

(16) Reiling, K. K.; Endres, N. F.; Dauber, D. S.; Craik, C. S.; Stroud, R. M. Anisotropic dynamics of the JE−2147-HIV protease complex: drug resistance and thermodynamic binding mode examined in a 1.09 A structure. *Biochemistry* **2002**, *41*, 4582−4594.

(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) United Devices. 2002. 12675 Research Building A, Austin, TX 78759.

(19) 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.

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

(21) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(22) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein−ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(23) 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. Screening* **2003**, *9*, 32−36.

(24) Jahnke, W.; Florsheimer, A.; Blommers, M. J. J.; Paris, G.; Heim, J.; Nalin, C. M.; Perez, L. B. Second-Site NMR Screening and Linker Design. *Curr. Top. Med. Chem.* **2003**, *3*, 69−80.

CI0497861