Library Fingerprints: A Novel Approach to the Screening of Virtual Libraries

Anthony E. Klon* and David J. Diller

Department of Molecular Modeling, Pharmacopeia, P.O. Box 5350, Princeton, New Jersey 08543-5350

Received January 22, 2007

We propose a novel method to prioritize libraries for combinatorial synthesis and high-throughput screening that assesses the viability of a particular library on the basis of the aggregate physical−chemical properties of the compounds using a naïve Bayesian classifier. This approach prioritizes collections of related compounds according to the aggregate values of their physical−chemical parameters in contrast to single-compound screening. The method is also shown to be useful in screening existing noncombinatorial libraries when the compounds in these libraries have been previously clustered according to their molecular graphs. We show that the method used here is comparable or superior to the single-compound virtual screening of combinatorial libraries and noncombinatorial libraries and is superior to the pairwise Tanimoto similarity searching of a collection of combinatorial libraries.

INTRODUCTION

The increased use of combinatorial chemistry methods to synthesize chemical libraries has led to a dramatic increase in the size of compound collections at pharmaceutical companies as well as chemical vendors that supply them. Although combinatorial methods are a powerful tool for the exploration of chemistry space, considerable amounts of time are required for library synthesis and quality control. Recent approaches for the design of combinatorial libraries have included the use of structural information to design target-specific libraries, as well as the calculation of physical−chemical parameters to create libraries with desirable druglike characteristics.[1−5] Structure-based library design uses high-throughput docking, high-resolution X-ray crystal structures, and NMR to create models of protein−ligand interactions that are then used to engineer combinatorial collections that have a high probability of containing compounds that are likely to bind to the target class of interest.[6] Although this approach is useful in cases where protein structural information is available, it is not feasible in instances where no high-resolution structure is available, as in the case of G-coupled protein receptors (GPCRs). Another technique applied to library design is based upon the statistical knowledge of drug- and leadlikeness of compounds such as Lipinski's "Rule of Five"[7] or Oprea's leadlike criteria.[8] These statistical models are based upon historical data gleaned from public sources such as scientific literature, patents, and approved drugs, as well as proprietary information such as internal high-throughput screening (HTS) assays and lead optimization projects. This approach that may be used concurrently with structure-based library design calculates the physical−chemical properties of the compounds prior to their synthesis and applies filters on the basis of properties such as the molecular weight, the number of hydrogen bond acceptors and donors, the number of rotatable bonds, the calculated logP, and the polar surface area.[9,10]

* Corresponding author phone: (609) 452-3676; e-mail: aklon@pcop.com.

The definition of an activity threshold once an HTS campaign has been completed is a binary response that is somewhat arbitrary and is often influenced by factors such as the composition and size of the chemical libraries being screened, the details of the biological assay, or mechanisms of nonspecific inhibition such as the aggregation of compounds in the well. The activity threshold may therefore change depending on the library being screened, the assay format, or even how many "hits" can be reasonably investigated on the basis of the resources available for lead optimization. Several reports in recent years have proposed methods to utilize the inherent structure−activity relationship data available from primary HTS screening campaigns and the structural redundancy available in the libraries being screened to identify false negatives or false positives.[11−13] For combinatorial libraries, the inherent structural similarity between compounds is already immediately apparent. For noncombinatorial libraries, the molecular frameworks, scaffolds, or chemical graphs[14] may be used in conjunction with clustering techniques to classify the compounds in a given library into a set of clusters. Statistical tools are then used to identify and evict false positives, primarily on the basis of the molecular similarity principle. If a given cluster has very few active compounds or a single active identified from the HTS assay, then the likelihood of any compound in the cluster being a true positive is poor, and it may be evicted. Similarly, compounds classified as having no activity that are found in clusters with a high number of true actives may have been incorrectly classified as negatives.

We demonstrate that the method presented here takes advantage of this structural degeneracy to prioritize in advance which combinatorial libraries are the most likely to contain compounds with the desired biological activity. In this retrospective study, a naïve Bayesian classifier is used to classify the likelihood of screening libraries to be active against one of six targets recently screened at Pharmacopeia using literature compounds and publicly available databases. The compounds in non-combinatorial libraries are clustered

10.1021/ci7000204 CCC: $37.00
Published on Web 06/27/2007

and statistical tools are used to predict which clusters or combinatorial libraries are most likely to be active.

The utility of this approach has benefits for both screening and lead optimization. For existing combinatorial and noncombinatorial libraries that are used in high-throughput screening assays, the number of unique compounds may easily run into the millions. In environments where efficient cherry-picking methods are available, smaller libraries can be created from much larger noncombinatorial libraries in preparation for HTS assays. Similarly, only combinatorial libraries or sublibraries of interest need be screened against a biological target. Because a smaller set of compounds representing higher structural redundancy would be screened, a higher confidence could be applied to the screening results as opposed to compounds selected by single-molecule virtual screening methods. Although it may be argued that it is more straightforward for screening organizations to test an entire compound collection running into the millions in lieu of cherry picking, additional resources must be expended after screening to pursue secondary assays, resynthesis of the confirmed hits, and lead optimization. Undesirable compounds or false positives that progress further down the drug discovery pipeline before failing are a waste of valuable chemical and biological resources for an organization. Even though the approach we present here can be used to prioritize sublibraries for screening, it retains its utility if an entire compound collection is screened. The models built prior to the HTS screen can be used when analyzing the active compounds to determine which ones to nominate for lead optimization programs on the basis of the molecular similarity principle.

The Pharmacopeia (PCOP) internal screening collection is composed of combinatorial libraries that have been generated using ECLiPS technology.[15] Each library is created using solid-phase synthesis on a polymer support by three to five combinatorial reactions, resulting in approximately 20 000 to 100 000 compounds per library,[16] with the entire Pharmacopeia collection comprising ~7.5 million discreet chemical entities in all. Individual compounds are linked to a single bead, which are attached to a series of aromatic tags. Individual beads are plated into single wells, and the compounds are cleaved from the beads and transferred to a screening plate. After an HTS assay identifies active compounds, the structures are decoded by detaching and analyzing by gas chromatography the aromatic tags from the original beads in the corresponding wells. Although this approach enables HTS assays to rapidly and efficiently identify active compounds, there is no way to cherry pick individual compounds prior to screening because the identity of each chemical entity is not known until its aromatic tags have been determined. Library prioritization using conventional techniques that rely on prioritizing individual compounds and are supported by efficient cherry-picking capabilities are therefore not applicable. Virtual screening techniques in this environment must therefore focus on the prioritization of combinatorial libraries. If we assume that the average combinatorial library contains ~60 000 compounds, prioritization of individual libraries as single records (125) would therefore be expected to be ~60 000 times faster than prioritizing the same collection by individual compounds.

Table 1. Total Number of Known Actives Represented in the Training and Test Sets for Each Target

	number of literature compounds added to NIH training set (active/inactive)	number of active PCOP compounds added to MDDR test set
A2A	374[30−42] (186/188)	339
cysteine protease	306[43−52] (161/145)	337
GPCR	67 (21/46)	125
CCR1	121[53,54] (40/81)	272
CXCR3	32[55] (32/0)	97[16]
kinase	208 (111/97)	271

The naïve Bayesian models used to prioritize libraries on the basis of their fingerprints were tested against two different types of libraries. The Bayesian models were used to predict the active libraries from a collection of combinatorial libraries used for HTS assays against six protein targets of interest. In order to evaluate the effectiveness of the library fingerprinting approach beyond combinatorial libraries, the MDL Drug Data Report (MDDR) was used as a set of random inactive compounds and seeded with known actives for the six targets. Two different clustering methods were employed in order to emulate sublibraries. In both cases, fingerprints were then calculated for each of the clusters, and the Bayesian models were used to predict which clusters were likely to contain active compounds against the targets of interest.

METHODS

Data Set Preparation. To construct the training set, the medicinal chemistry literature was mined for compounds with measured K_i or IC_{50} values against one of six targets recently screened at Pharmacopeia: adenosine receptor 2A (A2A); chemokine receptors CCR1 and CXCR3, as well as an additional GPCR; kinase; and cysteine protease (Table 1). Active compounds were defined as those with a reported K_i value $\leq$ 100 nM. The known active compounds were then seeded into a set of compounds from the NIH screening collection[17] that were used as inactives. The NIH screening collection was filtered to remove compounds that contained atoms other than H, C, O, N, S, F, or Cl. Compounds containing less than eight or more than 45 non-hydrogen atoms were also removed. In cases where an entry contained multiple fragments, the largest one was retained. These filters resulted in 71 387 compounds that were presumed inactive.

All compounds in Pharmacopeia's internal collection from all libraries were used as a combinatorial test set. The library fingerprints were calculated for a particular combinatorial library using all compounds in the library. The known active libraries were taken from the results of internal HTS assays against six protein targets. An active library is considered to be any library that contained any active compounds in an HTS assay whose structure had been confirmed by a chemical resynthesis and whose activity had been confirmed in an assay with a measured K_i or IC_{50} value.

For a noncombinatorial library test set, the MDDR[18] was taken to be the set of presumed random inactive compounds. The MDDR was filtered to remove duplicate entries, salts, and metal-containing compounds, resulting in 97 475 compounds. This data set was then seeded with known actives for each of the targets taken from Pharmacopeia's historical data (Table 1), resulting in a total of 98 914 compounds.

Library Fingerprints. Library fingerprints for each library or cluster were calculated within MOE[19] using scripts written in Scientific Vector Language (SVL). For each compound in a library or cluster, the MACCS keys used as feature counts, the log of the octanol/water partition coefficient (SlogP),[20] molecular weight (MW), topological polar surface area contributed by nitrogen and oxygen atoms (TPSA),[21] number of rotatable bonds, and number of hydrogen bond donors and acceptors were calculated. The values for each descriptor were then averaged over each combinatorial library (PCOP collection) or cluster (MDDR test set), creating a single nonlibrary fingerprint representation for that set of compounds. The resulting library fingerprints are composed of 172 numerical descriptors.

Database Clustering. Two different approaches were used to cluster the compounds in the noncombinatorial library test set into clusters, for which library fingerprints were subsequently calculated.

Principle components analysis (PCA) was used to cluster compounds on the basis of the MACCS feature counts, MW, SlogP, TPSA, number of rotatable bonds, and number of hydrogen bond donors and acceptors within MOE. A total of 55 principle components were used to account for 90% of the variance in the observed data. The maximum number of conditions for the principle components transform was set to 1000; the descriptors were decorrelated, and each axis was divided into two equiprobable subdivisions. This resulted in a total of 128 clusters, with the smallest cluster containing 102 compounds and the largest containing 1808 compounds. Figure 1A shows the resulting distribution of cluster frequencies as a function of the population size after PCA clustering.

The noncombinatorial test set was separately clustered using the molecular graphs of the compounds in the data set. The graph frameworks were calculated as described previously by Bemis and Murcko[14] using an SVL script within MOE resulting in a total of 17 219 unique chemical graphs. An example of the chemical graphs generated using this procedure for four CXCR3 agonists[16] present in the MDDR test set is shown in Figure 2. These graphs were then clustered in MOE according to their bit-packed MACCS structural keys using the Tanimoto coefficient as the similarity metric. The compounds in the test set were ultimately segregated into 812 clusters ranging in size from 170 singletons to the largest cluster, which contained 9442 compounds. The known active compounds were included in 17 clusters, with only three singletons. Figure 1B shows the distribution of cluster frequencies as a function of the population size after chemical graph clustering. It is important to note that, while the clustering itself was carried out using the chemical graphs of individual compounds, the calculated library fingerprints for each cluster are the averaged values of the MACCS feature counts, MW, SlogP, TPSA, number of rotatable bonds, and number of hydrogen bond donors and acceptors.

Bayesian Model Construction. Naïve Bayesian models were created for each of the six protein targets from the training set using previously described methods.[23] These models were then used to create two different kinds of predictions for Pharmacopeia's internal collection. Library fingerprints generated from the 172 descriptors as described above were used to prioritize combinatorial libraries, and individual compounds were prioritized using the same 172 descriptors. The same six Bayesian models were also used to prioritize the MDDR test set in three different ways. Two different methods were used to cluster the MDDR test set on the basis of the results of PCA clustering or chemical graph clustering using Tanimoto similarity. Library fingerprints were calculated for each cluster, and the clusters were prioritized using the previously constructed Bayesian models. Individual compounds in the MDDR test set were also prioritized using the same 172 descriptors. It is important to note that the same six models were used in a total of five tests to prioritize the active compounds, clusters, or combinatorial libraries.

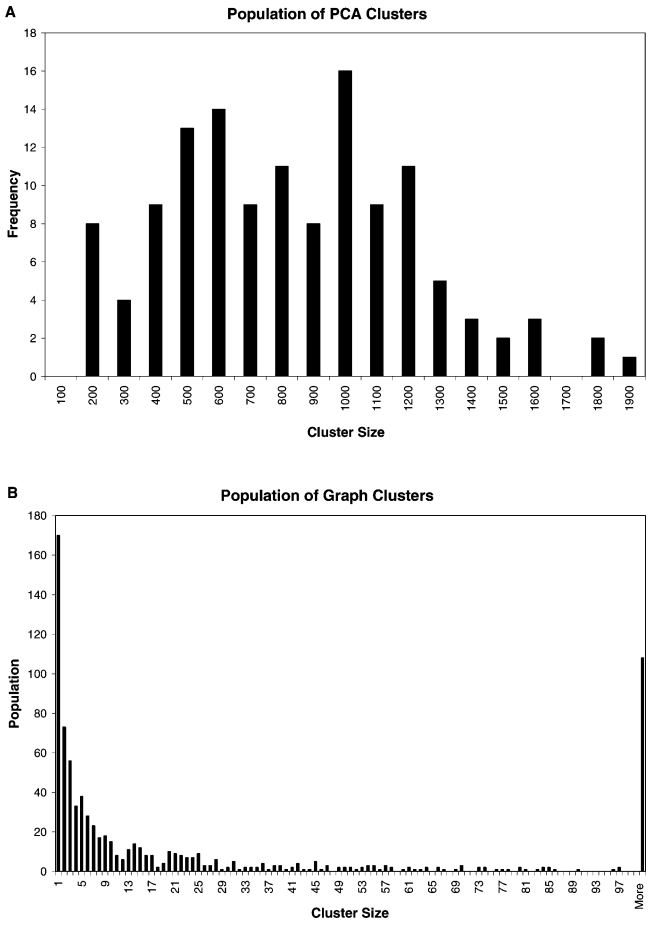

Figure 1. Histogram of population sizes for clusters created by clustering the principle components (A) and chemical graphs (B) of the MDDR test set.

Ranking and Scoring. Combinatorial libraries, PCA clusters, and graph clusters were ranked according to their calculated Bayesian probability of containing one or more active compounds. For predictions on individual compounds from the MDDR test set, the compounds were ranked according to their individual probabilities of being active against the given target protein.

For the test case where Pharmacopeia's internal collection was classified by the Bayesian models, combinatorial libraries were also ranked by the number of compounds in each library predicted to be active against a given protein target, divided by the total number of compounds in that library. An individual compound was considered active if the corresponding Bayesian model assigned a probability of activity of 50% or greater.

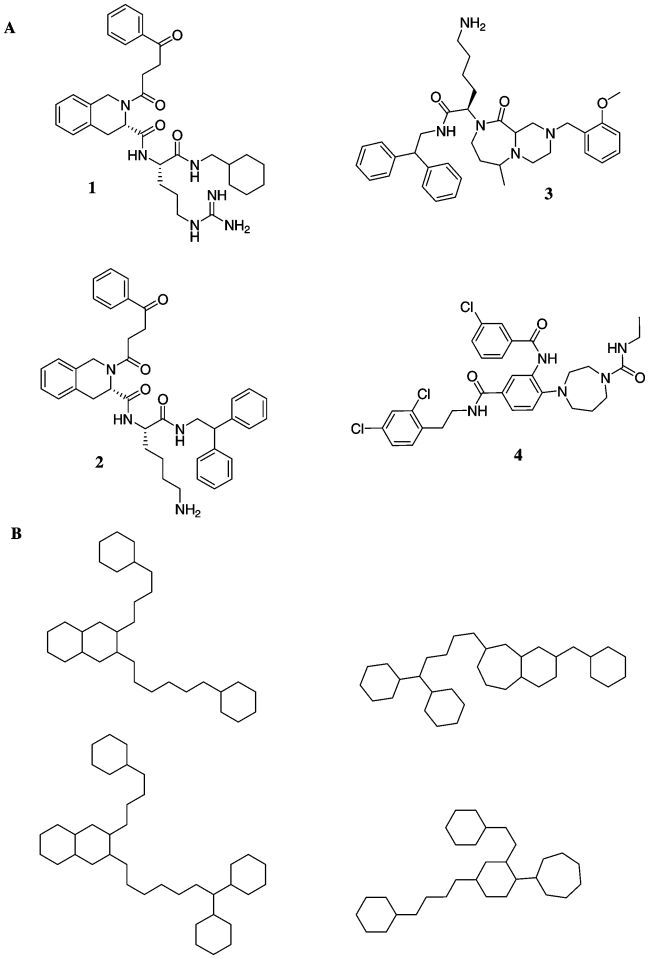

Figure 2. Four CXCR3 agonists (A), along with their calculated chemical graphs (B).

The performance of the Bayesian models was compared to similarity searching by calculating the Tanimoto distance between the known active compounds for a given target and all 7.5 million compounds in the collection using the MACCS fingerprints as the bitstring. The MAX approach[24] was used to score individual compounds in the combinatorial libraries by assigning the highest Tanimoto similarity to any of the known actives for a given compound as its score. In keeping with the logic used to designate an active library in an HTS assay, the score assigned to each combinatorial library was taken to be the highest MAX score found in that library (the MAX of MAX scores). We also calculated the average Tanimoto score over the entire library (the average of all MAX scores over all compounds), which demonstrated inferior performance with respect to the identification of known active libraries when compared to the MAX of all MAX scores (data not shown). The resulting similarity score for a given library is therefore taken to be the highest value found in the Tanimoto distance matrix calculated for the set of all known actives against a particular protein target versus all compounds in a single combinatorial library. The same procedure was used with the same compounds in the training set to prioritize the compounds in the MDDR test set.

The performance of each Bayesian model as well as the similarity searching was assessed by calculating the area under the receiver operating characteristic (ROC)[25] curves using the ranked compound, cluster, or library lists.

Descriptor Selection. Our initial selection of 172 descriptors to include in the library fingerprints was based upon two considerations. The six physicochemical descriptors used were based upon the fact that Pharmacopeia's internal libraries were designed specifically with favorable values for these parameters in mind. Combinatorial libraries are designed by computational and combinatorial chemists with in-house informatics software that includes plots of various physicochemical parameters including logP, PSA, molecular weight, the number of rotatable bonds, and the number of hydrogen bond acceptors and donors for a prospective library. Histograms for these properties were generated on the fly for each library during the design stage, revealing a normal distribution. These distributions are modeled by the library fingerprinting method we describe in this paper. The use of these descriptors aids combinatorial chemists in library design and medicinal chemists in lead optimization programs, so we include these in the model because of their relevance to our chemistry colleagues as well as because of their interpretability.

The MACCS keys were used in addition to the above set of six descriptors for the generation of library fingerprints because their use represents a well-validated fingerprinting methodology. Because the Bayesian classifier models numerical descriptors as a normal distribution, we use the keys as feature counts to calculate the number of times a given key is observed in a particular compound. We have previously shown that using the MACCS keys as feature counts results in better classification than Bayesian models that model the same keys as binary data.[23]

A total of 418 descriptors were calculated with MOE, and 1630 were calculated with DRAGON.[26] It should be noted that, although there is some overlap between the two sets of descriptors, identical values were not obtained for many of these cases, likely due to differences in implementation. Correlations were calculated between the experimentally observed activity of the training compounds from the literature and the complete set of 2048 descriptors. For A2A, CCR1, CXCR3, and the cysteine protease targets, none of the descriptors were found to correlate with measured activity. In the case of the undisclosed GPCR target, two atom-centered fragments[27] calculated with DRAGON were found to correlate to activity. These two descriptors were each found to be highly correlated with several MACCS keys and, so, were not added to the library fingerprints.

For the lone kinase we examined, a total of 13 descriptors calculated with DRAGON were observed to correlate with the measured activity. One of these was a functional group count describing the number of carboxylic acids, and three were atom-centered fragments.[27] Of the remaining nine descriptors, the highest correlation was observed with the Balaban distance connectivity index.[28] The remaining eight descriptors are based upon the Balaban distance metric and were found to be highly correlated ($r > 0.70$) with this index, as was the number of carboxylic acids. The Balaban index was found to be highly correlated ($r > 0.70$) with 20 of the MACCS structural keys, the TPSA, the number of rotatable bonds, and the number of hydrogen bond donors and acceptors and inversely correlated ($r < -0.70$) with 15 MACCS structural keys and SlogP. Since the Balaban index is highly correlated with 40 of the descriptors used to derive the library fingerprints, it was not included.

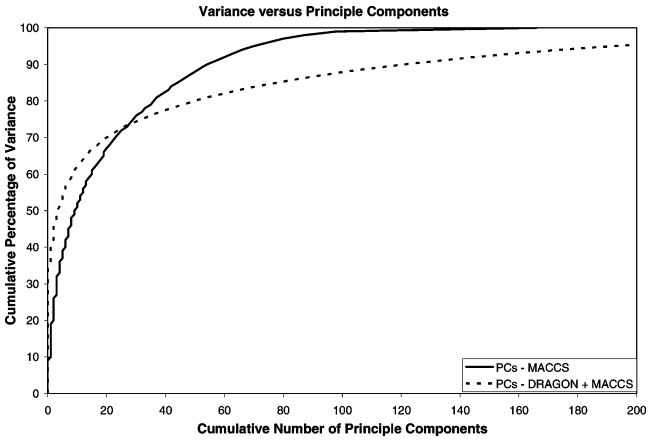

Figure 3. Plot of the cumulative percentage of variance versus the principle components calculated from the PCA of the 166 MACCS feature keys (solid line) and the MACCS keys with 1630 descriptors available in DRAGON (dashed line).

Table 2. Calculated Area under Receiver Operating Characteristic (AUROC) Curves for the Six Bayesian Models Used to Screen Libraries Using Library Fingerprints and Individual Prioritization, as Well as Tanimoto Similarity[a]

Target	AUROC		
	Library FP (Bayesian)	Individual Compound FP (Bayesian)	Individual Compound (Tanimoto)
A2A	0.82	0.67	0.75
Cysteine Protease	0.53	0.49	0.48
GPCR	0.63	0.67	0.42
CCR1	0.67	0.58	0.46
CXCR3	0.75	0.77	0.65
Kinase	0.72	0.76	0.68

[a] Results are color coded as follows: Best performing model (green), poorest performing model (red), and intermediate models (yellow).

Although calculating the correlations of descriptors to the activity is a straightforward way to evaluate the potential utility of >2000 descriptors, failure to find a satisfactory correlation does not mean that any particular descriptor does not add useful information. In order to evaluate whether any additional information could be gained by including any of the DRAGON descriptors, a principle component analysis was carried out on a random subset of the NIH screening collection consisting of 10 000 compounds. The 166 MACCS keys were used as feature counts, and PCA showed that a total of 166 principle components (PCs) were necessary to account for all the variance in the data. All 1630 descriptors available in DRAGON were calculated for the same subset of the NIH screening collection and were added to the MACCS keys. The PCA was repeated with this larger set of descriptors, resulting in a total of 1371 PCs necessary to account for all the variance in the calculated values. As can be observed in Figure 3, however, only ~30 principle components are required from both sets of descriptors to account for 75% of the total variance in the data. Above this threshold, a progressively increasing number of PCs is needed to account for variance in the values calculated by DRAGON, indicating that the additional descriptors obtained from DRAGON have not added significantly to the information encoded in the original MACCS keys.

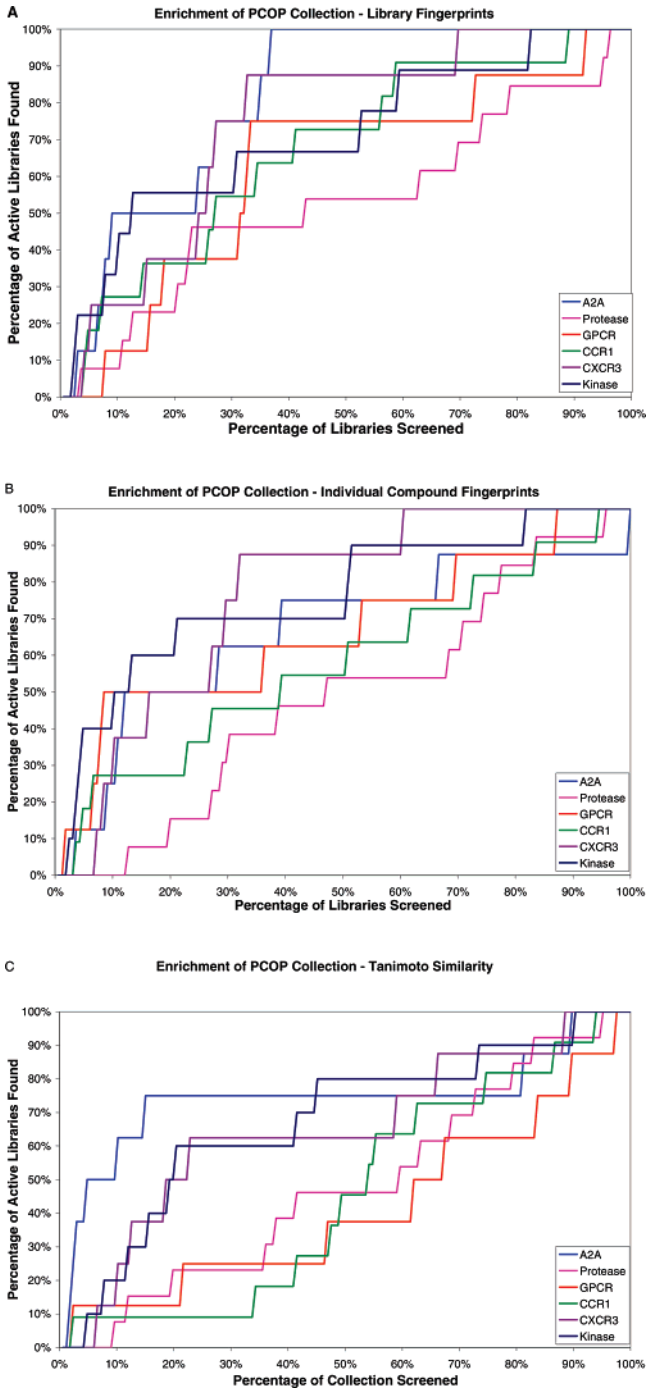

Figure 4. Enrichment curves for the active combinatorial libraries retrieved using the Bayesian models to predict activity on the basis of library fingerprints (A) and ~7.5 million individual compounds (B). Also shown are the enrichment curves for the retrieval of active compounds on the basis of Tanimoto similarity (C).

RESULTS AND DISCUSSION

The Performance of Bayesian Models in Prioritizing Combinatorial Libraries by Fingerprints Is Comparable to the Prioritization of Individual Compounds. The library fingerprints were designed to reduce the time it takes to prioritize a collection of combinatorial libraries for virtual screening and reduce statistical biases that may occur through prioritizing individual compounds. Below, we examined the ability of the library fingerprints to recall the active libraries for six targets retrospectively.

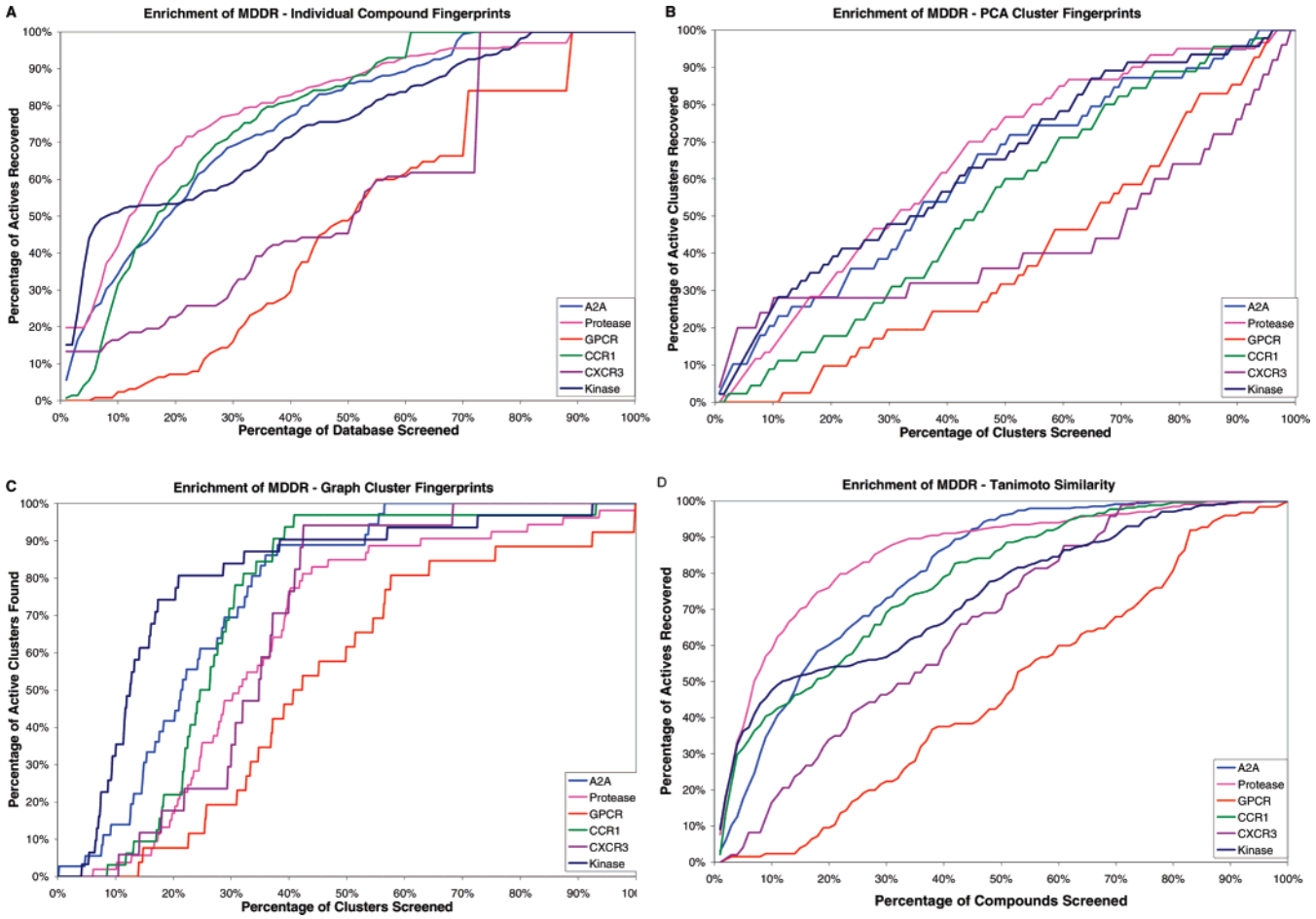

Figure 5. Enrichment curves for the predicted probability of active compounds in the MDDR test set with individual fingerprints (A), PCA cluster fingerprints (B), and chemical graph cluster fingerprints (C) calculated using the six Bayesian models. Also shown is the enrichment obtained by calculating the MAX Tanimoto distances from the known active compounds (D).

Table 3. Number of Clusters Containing Known Active Compounds after Clustering the MDDR Test Set on the Basis of Principle Components or Chemical Graphs

target	PCA	graph
A2A	39	36
cysteine protease	60	53
GPCR	41	26
CCR1	45	32
CXCR3	25	17
kinase	46	31

Table 4. Cluster Sizes and Distribution of Sample CXCR3 Agonists

	cluster size	
compound	PCA	graph
1	1490	287
2	494	287
3	494	928
4	566	21

Table 2 shows the calculated area under the ROC curves for the six Bayesian models that were used to predict active libraries from Pharmacopeia's compound collection using either library fingerprints or single compounds representing a random subset of the libraries. Figure 4 shows the enrichment curves resulting from the ranking of the active libraries according to their calculated Bayesian probability of being active (library FPs), or their fraction of predicted actives (random sampling). The ROC curves show that, in the cases of A2A and CCR1, there is a clear improvement in performance when using the library FPs over individual compound FPs, corresponding to an increased area under the ROC curves of 15% and 9%, respectively. In the case of the cysteine protease, there is a 4% improvement in the models' predictive ability, but the model is still considered to have no predictive ability. The GPCR model's performance is slightly worse by 4% when compared to the prioritization of individual compounds, but both models may be considered to have poor enrichment overall. In the cases of CXCR3 and the kinase, the library FPs are slightly outperformed by single-compound Bayesian prioritization by 2% and 4%, respectively, but do not demonstrate substantially inferior performance. In each case, the models may be considered to have fair performance regardless of the whether individual compounds or libraries are prioritized.

Bayesian Classifiers Outperform Tanimoto Similarity Searching for Prioritizing Active Libraries. As mentioned in the Methods section, using the MAX algorithm in conjunction with the Tanimoto similarity proved to be more successful than using the average of the MAX Tanimoto scores. Despite this, using the Tanimoto distances to select active libraries was substantially less successful than using the Bayesian classifier to prioritize libraries using either library fingerprints or the frequency of predicted actives in a given library obtained by prioritizing compounds individually (Table 2). Only in the case of A2A was similarity searching able to obtain better results than Bayesian priori-

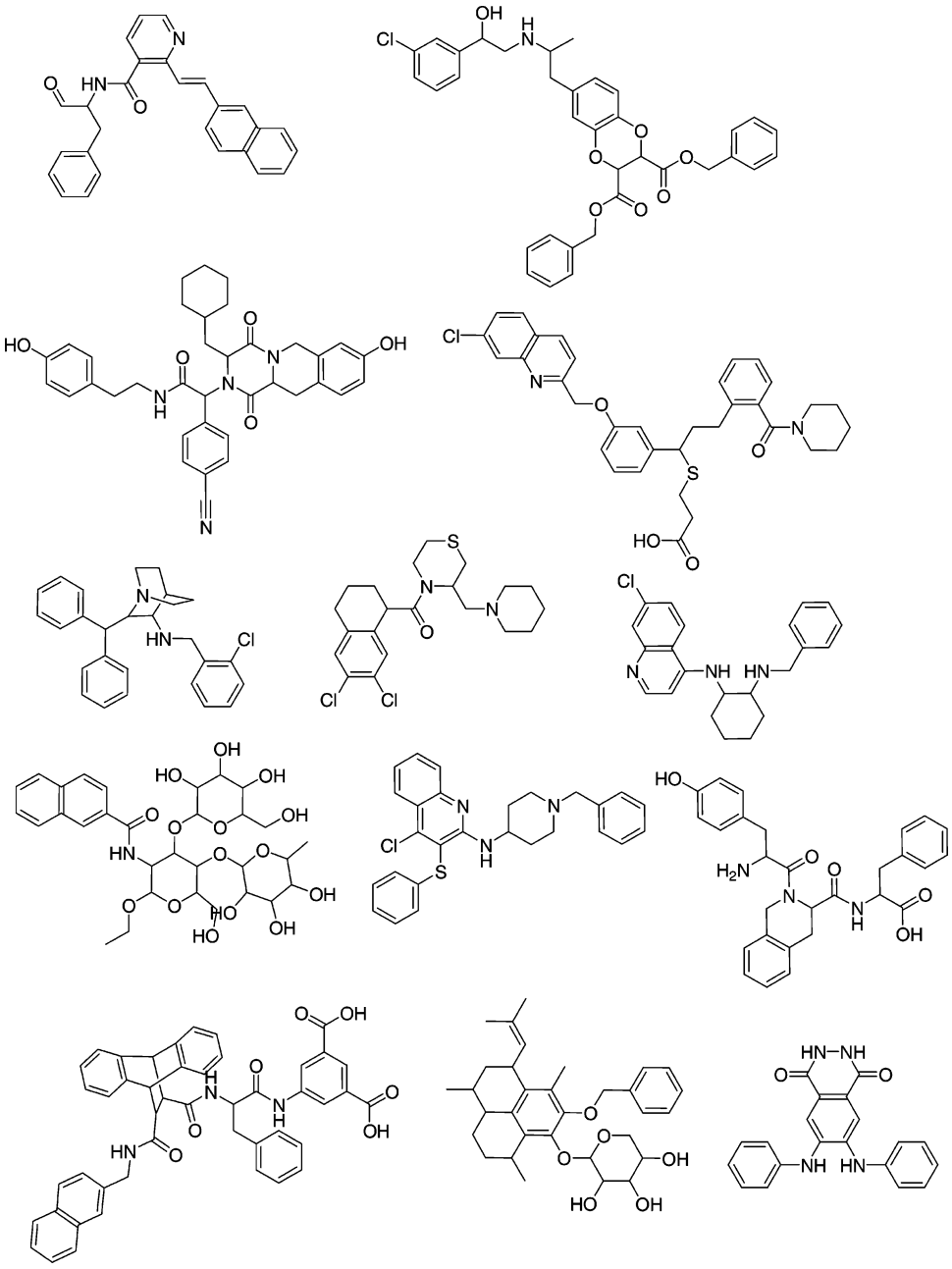

Figure 6. Representative compounds taken from the MDDR that occupy the same cluster as compounds **1** and **2**.

tization of individual compounds, but in the same case, it was outperformed by the library fingerprints. The drawback of using the Tanimoto distance metric hinges on the fact that, although the best performance is obtained through the MAX algorithm, this increases the likelihood that an inactive library may be highly ranked due to the presence of a single compound that has a high degree of similarity to a single known active compound in the training set. If this compound is on the fringe of the chemistry space occupied by the library rather than the center, the resulting Tanimoto score for that library would be artificially high. If such a library compound is not truly active, then the combinatorial library it is a member of will receive an erroneously high score. On the other hand, Bayesian models using library fingerprints will predict a combinatorial library to be active only if the overall characteristics of the library are consistent with the known actives.

Bayesian Prioritization of Chemical Graph Clusters Outperforms Prioritization of PCA Clusters. We next considered whether library fingerprints could be used to prioritize a noncombinatorial library for HTS. After seeding the MDDR with confirmed actives discovered from previous HTS assays in-house, we explored whether clustering the resulting test set could be used as an analogue for the prioritization of a collection of combinatorial libraries.

Figure 5 shows the enrichment curves resulting from ranking the PCA clusters, graph clusters, or individual compounds according to their calculated Bayesian probabilities of being active. All six Bayesian models demonstrate superior performance in prioritizing the active chemical graph clusters compared to the active PCA clusters. The fact that graph clustering yields more clusters, including many with few compounds, suggests that the success of the Bayesian models in identifying clusters as shown in Table 3 may be

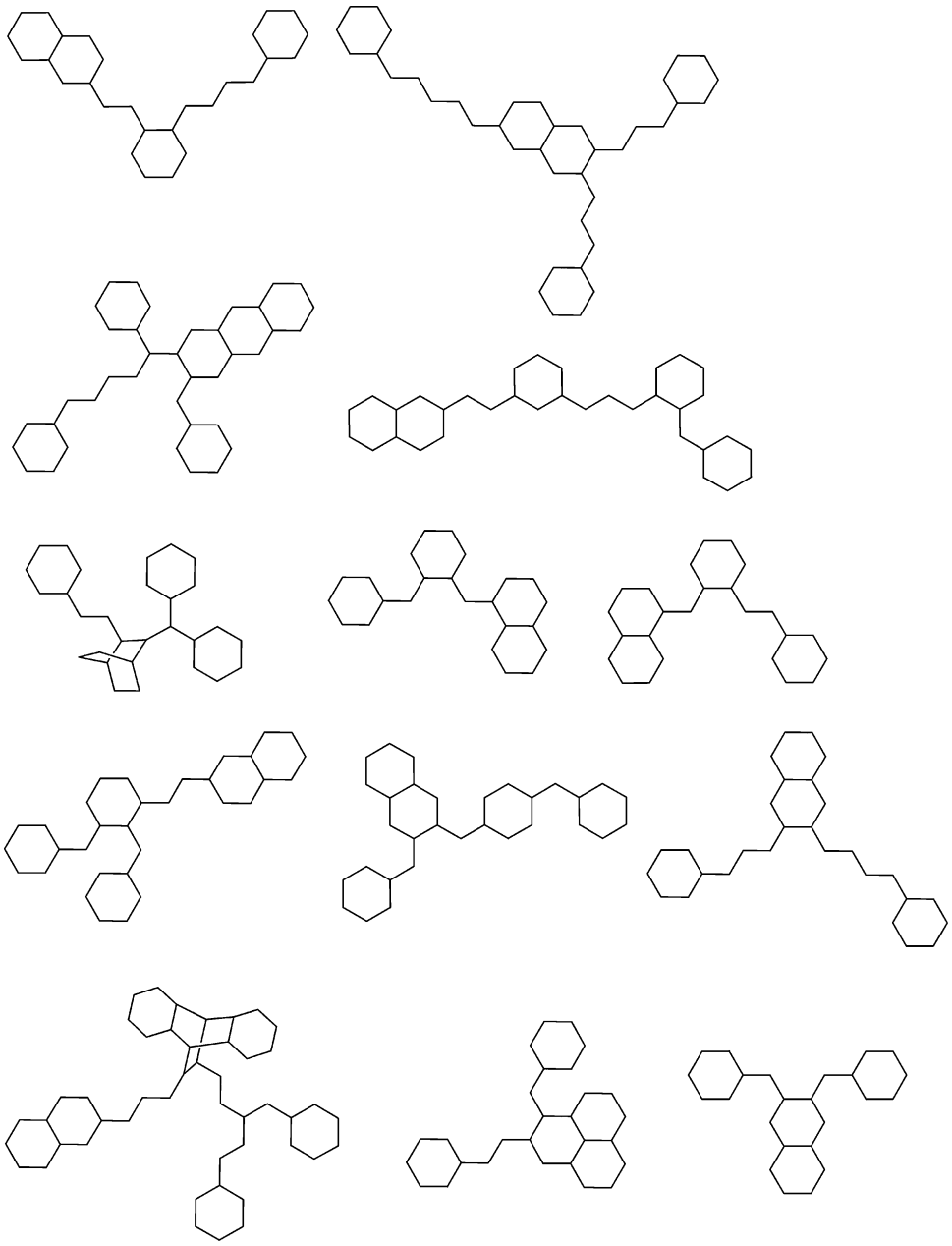

Figure 7. Chemical graphs corresponding to the structures shown in Figure 6.

due to more gradations in the test set, particularly if the known active molecules are found in sparsely populated clusters. Conversely, it could be argued that, because the known actives were taken from combinatorial libraries, they may tend to be grouped into a few clusters dominated by these compounds. The bar for success would be substantially lower in this case because the Bayesian model need only identify a handful of active clusters out of the hundreds that are present after clustering. Table 3 shows the number of active clusters for each target resulting from the clustering of the MDDR test set on the basis of either the principle components of the descriptors or the chemical graphs of the compounds. The proportion of active PCA clusters is substantial, comprising between 19.5% and 47% of the total number of clusters. The distribution of known actives in the graph clusters is better, with the active clusters comprising between 2% and 6.5% of the total number of clusters for each target. Because the MDDR was seeded with known active molecules identified through the screening of our internal combinatorial libraries, the compounds used in the MDDR test set are instead more likely to be grouped into larger graph clusters, yet there remains sufficient structural diversity that the known active compounds do not occupy an unreasonably small proportion of the total number of graph clusters.

Table 4 demonstrates the resulting cluster sizes observed for the clusters occupied by the four CXCR3 agonists shown in Figure 2 after clustering the MDDR test set using the principle components and chemical graphs. Compounds **1** and **2** occupy the same cluster when clustered on the basis of their chemical graphs, while compounds **2** and **3** are clustered together when clustering is carried out using the principle components. Figure 6 shows some representative compounds taken from the graph cluster occupied by compounds **1** and **2**. The corresponding chemical graphs for these representatives are shown in Figure 7. Figures 6 and 7 demonstrate the structural diversity that may be observed

Table 5. MACCS Structural Keys Present in Chemical Graphs Shown in Figures 2 and 6

of atoms in a four member rings
of atoms in a seven member rings
of atoms in a three member rings
of non-ring bonds connecting rings
of atoms in five member rings
of atoms in eight member or higher rings
of methylene carbons separated by four bonds
of methylene carbons separated by three bonds
of methylene carbons attached to methylene carbons
of non-ring methylene carbons
of atoms in six member rings
of ring atoms

among compounds and even their graphs within a single cluster.

Tanimoto Clustering of Chemical Graphs Does not Discard Information Vital to Bayesian Classification. A majority of the MACCS fingerprint bits are empty when clustering the chemical graphs of the compounds in the MDDR test set because all atom- and bond-type information is discarded. Table 5 lists the only bits that are set when the 17 219 chemical graphs are clustered into the 812 clusters using this protocol. While it may be implied that there is some correlation between the remaining 154 MACCS structural keys and the 12 shown in Table 5, it appears that most information describing the compounds in each cluster is discarded. To test whether this information loss while using this clustering approach was significant, new fingerprints were generated that contained only MW, SlogP, TPSA, the number of rotatable bonds, the number of hydrogen bond donors and acceptors, and the feature counts for the 12 descriptors shown in Table 5. This resulted in a reduced set of 18 descriptors instead of 172 when including all MACCS feature keys and physical descriptors. If clustering the chemical graphs results in significant information loss, then Bayesian prioritization of the graph clusters using only the reduced set of descriptors should result in models with poor predictive ability. Indeed, the calculated area under the ROC curves for the resulting predictions were all below 0.6, with five of them below 0.4 (data not shown), constituting a failure of all models to prioritize the clusters containing active compounds. These results suggest the data encoded in the remaining 154 MACCS keys are not lost upon clustering. These remaining keys encode more detailed information about the atom, and bond types of the compounds in the graph clusters are needed by the Bayesian classifiers to discriminate between active and inactive clusters.

Bayesian Prioritization of Chemical Graph Clusters Is Comparable to the Prioritization of Individual Compounds in the MDDR Test Set. Table 6 shows the calculated area under the ROC curves for the six Bayesian models that were used to predict either active clusters or active compounds from the MDDR test set seeded with known actives from internal HTS screening at Pharmacopeia. In the case of GPCR and CXCR3, the prediction of active graph clusters was more effective than the prediction of either active PCA clusters or active individual compounds. The graph clusters used to predict the GPCR model resulted in only slightly improved performance when compared with individual compound predictions but may still be considered a failure. In the case of CXCR3, predictions based upon the graph clusters resulted in an improved performance versus single compound predictions by 11%, whereas the predictions based upon the PCA clusters resulted in a 12% decreased predictive ability when compared to single compound predictions. In the cases of A2A, CCR1, and the protein kinase, the Bayesian predictions of the active graph clusters were comparable to the predictions for the individual compounds, showing differences of + 1%, −5%, and + 5%, respectively, relative to the single compound predictions. In each of these three cases, the predictions of the graph clusters were superior to the predictions of the active PCA clusters by 14%, 19%, and 16%, respectively. Only in the case of the cysteine protease did the predictions of the active graph clusters compare to the predictions of the active PCA clusters, and they were substantially poorer than the predicted active individual compounds by 17%. These results suggest that, in most cases, clustering a virtual library on the basis of the chemical graphs yields comparable or superior predictive ability when compared with predicting active compounds individually.

Table 6. Calculated Area under ROC Curves for the Six Bayesian Models Used to Identify Known Active Compounds or Clusters from the MDDR Test Set[a]

Target	AUROC			
	Library FP, PCA Clustering, Bayesian Prioritization	Library FP, Graph Clustering, Bayesian Prioritization	Individual Compound, Bayesian Prioritization	Tanimoto Similarity
A2A	0.62	0.76	0.77	0.81
Cysteine Protease	0.65	0.64	0.81	0.86
GPCR	0.40	0.53	0.48	0.48
CCR1	0.54	0.73	0.78	0.78
CXCR3	0.44	0.67	0.56	0.66
Kinase	0.64	0.80	0.75	0.74

[a] Results are color-coded as in Table 2.

Tanimoto Similarity Searching of the MDDR Test Set Outperforms Bayesian Classification. Prioritization of individual compounds in the MDDR test set using the Tanimoto distance metric was substantially more successful than that for the Bayesian models overall. Only in the cases of GPCR and protein kinase were the Bayesian models able to outperform Tanimoto similarity by 5% and 6%, respectively, when using library fingerprints to prioritize clusters on the basis of chemical graphs. In these two cases, however, the overall improvement relative to the Tanimoto similarity was modest. In the case of CXCR3, the Bayesian model was able to provide comparable enrichment to similarity searching when predicting active graph clusters using library fingerprints. In the cases of CCR1 and A2A, similarity searching was only modestly superior to Bayesian prioritization of graph clusters, yielding a better enrichment by only 5%. Only in the case of the cysteine protease was similarity searching greatly superior (22%) to Bayesian models built using library fingerprints, but it was only modestly superior (5%) to the Bayesian prioritization of individual molecules.

CONCLUSIONS

We have demonstrated a new approach to the virtual screening of chemical libraries through the use of library fingerprints. These fingerprints are straightforward to calculate and occupy less disk space than the individual compound fingerprints for an entire collection. Although we calculated library fingerprints using 172 descriptors based upon the MACCS keys and six physicochemical parameters, library fingerprints could be calculated using any descriptor. We have shown that, when used to prioritize combinatorial libraries for HTS, the performance is more consistent and often superior than pairwise similarity searching. The ability to prioritize libraries containing tens or hundreds of thousands of compounds by using library fingerprints also offers a dramatic reduction in the computation time relative to prioritizing all compounds in a collection individually.

The fingerprints may be used during the rational design of combinatorial libraries to create libraries for a particular class of protein targets, provided that some information about known active compounds for the target class of interest is available. The results reported here indicate that prioritizing combinatorial libraries on the basis of their library fingerprints using Bayesian models is equally as or more effective than prioritizing libraries by attempting to predict whether individual compounds within the library are active. The substantial increase in speed obtained when using the method means that many more combinatorial libraries may be tested in a virtual environment prior to synthesis. The results reported here suggest that a further increase in throughput could be obtained through the method by testing subsets of very large combinatorial libraries. This approach could be used as a first pass to prioritize ideas for combinatorial libraries that could be narrowed down to a smaller set of library proposals that could receive greater scrutiny at the design stage.

The fingerprints may also be used to prioritize collections for HTS assays by selecting combinatorial libraries that are more likely to contain active compounds. A similar approach may be employed with noncombinatorial libraries where the compounds have been segregated into clusters on the basis of their chemical connectivity. In this case, library fingerprints may be useful in cherry picking a subset of compounds to screen from a large compound collection where screening the entire collection may not be feasible.

Library fingerprints have the added advantage in that they are not biased by the size of a given combinatorial library or cluster, in contrast to traditional virtual screening methods that test compounds individually. Traditional methods such as Tanimoto similarity or previously presented applications of Bayesian classifiers that prioritize individual compounds encounter issues when prioritizing combinatorial libraries for screening because arbitrary decisions are made about what cutoff is to be used to decide whether a compound is active. The libraries that are then prioritized on the number of active compounds found may change depending upon the cutoff being used, thus altering the prioritization of the combinatorial libraries being evaluated. The situation becomes further complicated when comparing combinatorial libraries of differing sizes. Does one select the library with the largest number of predicted actives or the largest number of predicted actives scaled by library size? While the latter might be considered more "efficient", is the screening organization more concerned about the hit rate or the total number of hits obtained? Typically, when prioritizing a collection for screening using methods that prioritize individual compounds, test sets are constructed from a set of presumed actives and seeded with known active compounds in an attempt to identify what this cutoff should be a priori and evaluate the model's performance. While suitable for diverse collections of compounds, there is no guarantee that this approach is appropriate for combinatorial libraries where there is expected to be less chemical diversity. The method presented here has the advantage that decisions about what cutoff to use for active compounds do not need to be made. Instead, entire libraries are prioritized using a single score obtained from the Bayesian algorithm, removing the possibility of altering the relative prioritization of libraries under consideration by differences in the cutoff.

It was not surprising that the Bayesian models were able to use the library fingerprints to generate comparable or even superior enrichment to the predictions on the basis of individual compounds to prioritize a collection of combinatorial libraries. The initial assumption was that combinatorial libraries have much less structural diversity than a collection of noncombinatorial compounds, exemplified here by the MDDR. We have shown, however, that the Tanimoto similarity scores calculated using the MACCS keys indicate a large inter- and intralibrary variation in structural diversity. By contrast, it would be expected that the Bayesian prioritization of chemical graphs should be outperformed by the prioritization of individual compounds in the MDDR test case because this data set was seeded with known true actives with measured potencies taken from Pharmacopeia's internal programs. These results suggest that, in the test cases evaluated here, Bayesian methods in general are not able to identify the most potent individually active compounds but instead are able to identify an active region of chemistry space.

SOFTWARE

Descriptor calculations were carried out using MOE version 2006.08 and DRAGON 5.0. SVL scripts were written for the calculation of the library fingerprints in MOE. The chemical structures in Figures 2, 6, and 7 were generated using the 2D structure depiction algorithm in MOE.[29] The naïve Bayesian classifier used was written in C++ and implemented the algorithm described previously.[23]

REFERENCES AND NOTES

(1) Lowrie, J. F.; Delisle, R. K.; Hobbs, D. W.; Diller, D. J. The Different Strategies for Designing GPCR and Kinase Targeted Libraries. *Comb. Chem. High Throughput Screening* **2004**, *7*, 495−510.

(2) Matter, H.; Baringhaus, K.-H.; Naumann, T.; Klabunde, T.; Pirard, B. Computational Approaches Towards the Rational Design of Drug-Like Compounds. *Comb. Chem. High Throughput Screening* **2001**, *4*, 453−475.

(3) Miller, J. L. Recent Developments in Focused Library Design: Targeting Gene-Families. *Curr. Top. Med. Chem.* **2006**, *6*, 19−29.

(4) Klabunde, T.; Hessler, G. Drug Design Strategies for Targeting G-Protein-Coupled Receptors. *ChemBioChem* **2002**, *3*, 928−944.

(5) Crossley, R. The Design of Screening Libraries Targeted at G-Protein Coupled Receptors. *Curr. Top. Med. Chem.* **2004**, *4*, 581−588.

(6) Orry, A. J.; Abagyan, R. A.; Cavasotto, C. N. Structure-Based Development of Target-Specific Compound Libraries. *Drug Discovery Today* **2006**, *11*, 261−266.

(7) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Developmental Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.
(8) Oprea, T. I. Property Distribution of Drug-Related Chemical Databases. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 251−264.
(9) Delisle, R. K.; Lowrie, J. F.; Hobbs, D. W.; Diller, D. J. Computational ADME/Tox Modeling: Aiding Understanding and Enhancing Decision Making in Drug Design. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 325−345.
(10) Cheng, A.; Diller, D. J.; Dixon, S. L.; Egan, W. J.; Lauri, G.; Merz, K. M., Jr. Computation of the Physio-Chemical Properties and Data Mining of Large Molecular Collections. *J. Comput. Chem.* **2002**, *23*, 172−183.
(11) Glick, M.; Klon, A. E.; Acklin, P.; Davies, J. W. Enriching Extremely Noisy High-Throughput Screening Data Using a Naive Bayes Classifier. *J. Biomol. Screening* **2004**, *9*, 32−36.
(12) Diller, D. J.; Hobbs, D. W. Deriving Knowledge through Data Mining High-Throughput Screening Data. *J. Med. Chem.* **2004**, *47*, 6373−6383.
(13) Yan, S. F.; Asatryan, H.; Li, J.; Zhou, Y. Novel Statistical Approach for Primary High-Throughput Screening Hit Selection. *J. Chem. Inf. Model.* **2005**, *45*, 1784−1790.
(14) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(15) Ohlmeyer, M. H.; Swanson, R. N.; Dillard, L. W.; Reader, J. C.; Asouline, G.; Kobayashi, R.; Wigler, M.; Still, W. C. Complex Synthetic Chemical Libraries Indexed with Molecular Tags. *Proc. Natl. Acad. Sci. U. S. A.* **1993**, *90*, 10922−10926.
(16) Stroke, I. L.; Cole, A. G.; Simhadri, S.; Brescia, M. R.; Desai, M.; Zhang, J. J.; Merritt, J. R.; Appell, K. C.; Henderson, I.; Webb, M. L. Identification of CXCR3 Receptor Agonists in Combinatorial Small-Molecule Libraries. *Biochem. Biophys. Res. Commun.* **2006**, *349*, 221−228.
(17) NIH Roadmap for Medical Research. Molecular Libraries and Imaging Overview. http://nihroadmap.nih.gov/molecularlibraries/index.asp. (accessed August 9, 2006).
(18) *MDL Drug Data Report*; Elsevier MDL: San Leandro, CA, May 25, 2006.
(19) *Molecular Operating Environment*, version 2006.08; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2006.
(20) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.
(21) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.
(22) Reference deleted in press.
(23) Klon, A. E.; Lowrie, J. F.; Diller, D. J. Improved Naive Bayesian Modeling of Numerical Data for Absorption, Distribution, Metabolism and Excretion (ADME) Property Prediction. *J. Chem. Inf. Model.* **2006**, *46*, 1945−1956.
(24) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.
(25) Witten, I. H.; Frank, E. ROC Curves. In *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 1st ed.; Morgan Kaufmann Publishers: New York, 2000; pp 141−147.
(26) Todeschini, R.; Consonni, V.; Mauri, A.; Pavan, M. *DRAGON*, version 5.; Talete srl: Milano, Italy, 2004.
(27) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occurring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163−172.
(28) Balaban, A. T. Highly Discriminating Distance-Based Topological Index. *Chem. Phys. Lett.* **1982**, *89*, 399−404.
(29) Clark, A. M.; Labute, P.; Santavy, M. 2D Structure Depiction. *J. Chem. Inf. Model.* **2006**, *46*, 1107−1123.
(30) Novellino, E.; Cosimelli, B.; Ehlardo, M.; Greco, G.; Iadanza, M.; Lavecchia, A.; Rimoli, M. G.; Sala, A.; Da Settimo, A.; Primofiore, G.; Da Settimo, F.; Taliani, S.; La Motta, C.; Klotz, K.-N.; Tuscano, D.; Trincavelli, M. L., Martini, C. 2-(Benzimidazol-2-yl)quinoxalines: A Novel Class of Selective Antagonists at Human A_1 and A_3 Adenosine Receptors Designed by 3D Database Searching. *J. Med. Chem.* **2005**, *48*, 8253−8260.
(31) Matasi, J. J.; Caldwell, J. P.; Hao, J.; Neustadt, B.; Arik, L.; Foster, C. J., Lachowicz, J.; Tulshian, D. B. The Discovery and Synthesis of Novel Adenosine Receptor (A_{2A}) Antagonists. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 1333−1336.
(32) Matasi, J. J.; Caldwell, J. P.; Zhang, H.; Fawzi, A.; Cohen-Williams, M. E.; Varty, G. B.; Tulshian, D. B. 2-(2-Furanyl)-7-phenyl[1,2,4]-triazolo[1,5-*c*]pyrimidin-5-amine Analogs: Highly Potent, Orally Active, Adenosine A_{2A} Antagonists. Part 1. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3670−3674.
(33) Manetti, F.; Schenone, S.; Bondavalli, F.; Brullo, C.; Bruno, O.; Ranise, A.; Mosti, L.; Menozzi, G.; Fossa, P.; Trincavelli, M. L.; Martini, C.; Martinelli, A.; Tintori, C.; Botta, M. Synthesis and 3D QSAR of New Pyrazolo[3,4-*b*]pyridines: Potent and Selective Inhibitors of A_1 Adenosine Receptors. *J. Med. Chem.* **2005**, *48*, 7172−7185.
(34) Minetti, P.; Tinti, M. O.; Carminati, P.; Castorina, M.; Di Cesare, M. A.; Di Serio, S.; Gallo, G.; Ghirardi, O.; Giorgi, F.; Giorgi, L.; Piersanti, G.; Bartoccini, F.; Tarzia, G. 2-*n*-Butyl-9-methyl-8-[1,2,3]-triazol-2-yl-9*H*-purin-6-ylamine and Analogues as A_{2A} Adenosine Receptor Antagonists. Design, Synthesis, and Pharmacological Characterization. *J. Med. Chem.* **2005**, *48*, 6887−6896.
(35) Catarzi, D.; Colotta, V.; Varano, F.; Lenzi, O.; Filacchioni, G.; Trincavelli, L.; Martini, C.; Montopoli, C.; Moro, S. 1,2,4-Triazolo-[1,5-a]quinoxaline as a Versatile Tool for the Design of Selective Human A_3 Adenosine Receptor Antagonists: Synthesis, Biological Evaluation, and Molecular Modeling Studies of 2-(Hetero)aryl- and 2-Carboxy-substitued derivatives. *J. Med. Chem.* **2005**, *48*, 7932−7945.
(36) Peng, H.; Kumaravel, G.; Yao, G.; Sha, L.; Wang, J.; Van Vlijmen, H.; Bohnert, T.; Huang, C.; Vu, C. B.; Ensinger, C. L.; Chang, H.; Engber, T. M.; Whalley, E. T.; Petter, R. C. Novel Bicyclic Piperazine Derivatives of Triazolotriazine and Triazolopyrimidines as Highly Potent and Selective Adenosine A_{2A} Receptor Antagonists. *J. Med. Chem.* **2004**, *47*, 6218−6229.
(37) Alanine, A.; Anselm, L.; Steward, L.; Thomi, S.; Vifian, W.; Groaning, M. D. Synthesis and SAR Evaluation of 1,2,4-Triazoles as A_{2A} Receptor Antagonists. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 817−821.
(38) Fredholm, B. B.; Cunha, R. A.; Svenningsson, P. Pharmacology of Adenosine A_{2A} Receptors and Therapeutic Applications. *Curr. Top. Med. Chem.* **2002**, *3*, 413−426.
(39) Vu, C. B.; Peng, B.; Kumaravel, G.; Smits, G.; Jin, X.; Phadke, D.; Engber, T.; Huang, C.; Reilly, J.; Tam, S.; Grant, D.; Hetu, G.; Chen, L.; Zhang, J.; Petter, R. C. Piperazine Derivatives of [1,2,4]Triazolo-[1,5-*a*][1,3,5]triazine as Potent and Selective Adenosine A_{2A} Receptor Antagonists. *J. Med. Chem.* **2004**, *47*, 4291−4299.
(40) Jeong, L. S.; Lee, H. W.; Jacobson, K. A.; Kim, H. O.; Shin, D. H.; Lee, J. A.; Gao, Z. G.; Lu, C.; Duong, H. T.; Guanga, P.; Lee, S. K.; Jin, D. Z.; Chun, M. W.; Moon, H. R. Structure−Activity Relationships of 2-Chloro-N6-substituted-4′-thioadenosine-5′-uronamides as Highly Potent and Selective Agonists at the Human A3 Adenosine Receptor. *J. Med. Chem.* **2006**, *49*, 273−281.
(41) Jacobson, K. A.; Gao, Z.-G. Adenosine Receptors as Therapeutic Targets. *Nat. Rev. Drug Discovery* **2006**, *5*, 247−264.
(42) Moro, S.; Gao, Z.-G.; Jacobson, K. A.; Spalluto, G. Progress in the Pursuit of Therapeutic Adenosine Receptor Antagonists. *Med. Res. Rev.* **2005**, *26*, 131−159.
(43) Palmer, J. T.; Bryant, C.; Wang, D.-X.; Davis, D. E.; Setti, E. L.; Rydzewski, R. M.; Venkatraman, S.; Tian, Z. Q.; Burrill, L. C.; Mendonca, R. V.; Springman, E. McCarter, J.; Chung, T.; Cheung, H.; Janc, J. W.; McGrath, M.; Somoza, J. R.; Enriquez, P.; Yu, Z. W; Strickley, R. M.; Liu, L.; Venuti, M. C.; Percival, M. D.; Falgueyret, J. P.; Prasit, P.; Oballa, R.; Riendeau, D.; Young, R. N.; Wesolowski, G.; Rodan, S. B.; Johnson, C.; Kimmel, D. B.; Rodan, G. Design and Synthesis of Tri-Ring P_3 Benzaminde-Containing Aminonitriles as Potent, Selective, Orally Effective Inhibitors of Cathepsin K. *J. Med. Chem.* **2005**, *48*, 7520−7534.
(44) Marquis, R. W.; Ru, Y.; LoCastro, S. M.; Zeng, J.; Yamashita, D. S.; Oh, H. J.; Erhard, K. F.; Davis, L. D.; Tomaszek, T. A.; Tew, D.; Salyers, K.; Proksch, J.; Ward, K., Smith, B.; Levy, M.; Cummings, M. D., Haltiwanger, R. C.; Trescher, G.; Wang, B.; Hemling, M. E.; Quinn, C. J.; Cheng, H. Y.; Lin, F.; Smith, W. W.; Janson, C. A.; Zhao, B.; McQueney, M. S.; D'Alessio, K.; Lee, C. P.; Marzulli, A.; Dodds, R. A.; Blake, S.; Hwang, S. M.; James, I. E.; Gress, C. J.; Bradley, B. R.; Lark, M. W.; Gowen, M.; Veber, D. F. Azepanone-Based Inhibitors of Human and Rat Cathepsin K. *J. Med. Chem.* **2001**, *44*, 1380−1395.
(45) Marquis, R. W.; Ru, Y.; Zeng, J.; Trout, R. E. L.; LoCastro, S. M.; Gribble, A. D.; Witherington, J.; Fenwick, A. E.; Granier, B.; Tomaszek, T.; Tew, D.; Hemling, M. E.; Quinn, C. J.; Smith, W. W.; Zhao, B.; McQueney, M. S.; Janson, C. A.; D'Alessio, K.; Veber, D. F. Cyclic Ketone Inhibitors of the Cysteine Protease Cathepsin K. *J. Med. Chem.* **2001**, *44*, 725−736.
(46) Altmann, E.; Renaud, J.; Green, J.; Farley, D.; Cutting, B.; Jahnke, W. Arylaminoethyl Amines as Novel Non-Covalent Cathepsin K Inhibitors. *J. Med. Chem.* **2002**, *45*, 2352−2354.

(47) Altmann, E.; Cowan-Jacob, S. W.; Missbach, M. Novel Purine Nitrile Derived Inhibitors of the Cysteine Protease Cathepsin K. *J. Med. Chem.* **2004**, *47*, 5833−5836.
(48) Loser, R.; Schilling, K.; Dimmig, E.; Gutschow, M. Interaction of Papain-Like Cysteine Proteases with Dipeptide-Derived Nitriles. *J. Med. Chem.* **2005**, *48*, 7688−7707.
(49) Tavares, F. X.; Deaton, D. N.; Miller, A. B.; Miller, L. R.; Wright, L. L.; Zhou, H. Q. Potent and Selective Ketoamide-Based Inhibitors of Cystein Protease, Cathepsin K. *J. Med. Chem.* **2004**, *47*, 5049−5056.
(50) Tavares, F. X.; Boncek, V.; Deaton, D. N.; Hassell, A. M.; Long, S. T.; Miller, A. B.; Payne, A. A.; Miller, L. R.; Shewchuk, L. M.; Wells-Knecht, K.; Willard, D. H., Jr.; Wright, L. L.; Zhou, H. Q. Design of Potent, Selective, and Orally Bioavailable Inhibitors of Cystein Protease Cathepsin K. *J. Med. Chem.* **2004**, *47*, 588−599.
(51) Falgueyret, J.-P.; Desmarais, S.; Oballa, R.; Black, W. C.; Cromlish, W.; Khougaz, K.; Lamontagne, S.; Masse, F.; Riendeau, D.; Toulmond, S.; Percival, M. D. Lysosomotropism of Basic Cathepsin K Inhibitors Contributes to Increased Cellular Potencies against Off-Target Cathepsins and Reduced Functional Selectivity. *J. Med. Chem.* **2005**, *48*, 7535−7543.
(52) Marquis, R. W.; James, I.; Zeng, J.; Trout, R. E. L.; Thompson, S.; Rahman, A.; Yamashita, D. S.; Xie, R.; Ru, Y.; Gress, C. J.; Blake, S.; Lark, M. A.; Hwang, S.-M.; Tomaszek, T.; Offen, P.; Head, M. S.; Cummings, M. D.; Veber, D. F. Azepanone-Based Inhibitors of Human Cathepsin L. *J. Med. Chem.* **2005**, *48*, 6870−6878.
(53) Ng, H. P.; May, K.; Bauman, J. G.; Ghannam, A.; Islam, I.; Liang, M.; Horuk, R.; Hesselgesser, J.; Snider, R. M.; Perez, H. D.; Morrissey, M. M. Discovery of Novel Non-Peptide CCR1 Receptor Antagonists. *J. Med. Chem.* **1999**, *42*, 4680−4694.
(54) Naya, A.; Sagara, Y.; Ohwaki, K.; Saeki, T.; Ichikawa, D.; Iwasawa, Y.; Noguchi, K.; Ohtake, N. Design, Synthesis, and Discovery of a Novel CCR1 Antagonist. *J. Med. Chem.* **2001**, *44*, 1429−1435.
(55) Medina, J. C.; Johnson, M. G.; Collins, T. L. CXCR3 Antagonists. *Annu. Rep. Med. Chem.* **2005**, *40*, 215−225.

CI7000204