Using Ensembles to Classify Compounds for Drug Discovery

J. Kevin Lanctot,* Santosh Putta,† Christian Lemmen,‡ and Jonathan Greene§

Deltagen Research Laboratories, Inc., 740 Bay Road, Redwood City, California 94063

Received June 24, 2003

This paper introduces Signal, a novel method for classifying activity against a small molecule drug target. Signal creates an ensemble, or collection, of meaningful descriptors chosen from a much larger property space. The method works with a variety of descriptor types, including fingerprints that represent four-point pharmacophores or shape descriptors. It also exploits information from both active and inactive compounds and generates predictive models suitable for high throughput screening data analysis. Given the fingerprints and activity data for a set of compounds, Signal is a two step process. The first step is to *Evaluate the Descriptors*: for each descriptor in the fingerprint, quantify and rank the correlation between the activity of the compounds and the presence of that descriptor. The second step is to *Create an Ensemble Model*: use the high ranking descriptors to create a model of activity against the biological target. For the first step, two possible ranking strategies were investigated: mutual information and chi-square. For the second step, two types of ensemble models were investigated: high ranking and a novel method called high ranking set cover. Of the four possible pairings, the combination of chi-square and high ranking set cover performed the best on a Thrombin data set.

INTRODUCTION

An integral task in modern drug discovery is the analysis of high throughput screening data with the goal of finding information to suggest new compounds for synthesis. One possible approach is to synthesize analogues around the most promising hits, which has a high likelihood of discovering additional actives but a low likelihood of finding novel leads. Another approach is to identify molecular properties that appear among many actives and use combinations of these properties to suggest new compounds for synthesis. While this search may be performed by humans, computers handle the tedious task with the capacity to process the large amount of data that may be generated by combinatorial chemistry. The goal of these algorithms is to narrow down, if not identify, the molecular properties that are relevant for activity.

Machine learning is one class of algorithms that have been explored to solve this problem. Jurs[1−5] has applied a variety of standard machine learning techniques, such as perceptrons, neural nets, and genetic algorithms to determine structure-activity relationships in molecule data. Muggleton and co-workers proposed Inductive Logic Programming to discover pharmacophore models,[6−9] while Young and co-workers applied recursive partitioning with the same objective.[10] Although most of these approaches were implemented using a pharmacophore model represented as pharmacophoric descriptors in 3-D space with the distances specified between them, in principle these methods are extendable to other types of descriptors. Using real-valued descriptors to discover correlations with activity has been implemented using neural nets,[3,11] partial least squares,[9,12] and nearest neighbor approaches.[13−17] Another strategy besides regression and classification is the rank-ordering of virtual compounds.[18,19] A few other approaches which do not use machine learning techniques have also been suggested for the elucidation of pharmacophores;[20−22] however, they are constrained to a specific type of pharmacophore and do not generalize to other types of descriptors.

Quantitative structure-activity relationships, as an entire discipline, have developed around regression-type methods. However, generally regression is more difficult to solve than classification, as has been pointed out by Vapnik.[23] Additionally, many approaches that use real-valued data show numeric instability. That is, slight changes in the input cause large changes in the output or the algorithms may not converge to a solution. Furthermore, in many applications a regression model is unnecessary. If the goal is to select compounds from a virtual library, classification or rank ordering should suffice. The methods presented here agree with this philosophy and classify compounds into two classes, active or inactive.

This paper introduces a general framework, called Signal, which analyzes any property space that can be binned into a binary vector or fingerprint. Signal selects relevant descriptors from the fingerprint, a task known as the *feature-selection problem* in the machine learning literature.[24] The relevant descriptors are collected into an *ensemble*, which is a subset of descriptors from the fingerprint that represents a model for activity. The more descriptors from an ensemble that a test molecule has, the more likely it is active against the target of the training set. Using this approach, Signal was developed with three major goals. First, it must be capable of evaluating a variety of descriptor types. Other than the two types of fingerprint presented below, pharma-

* Corresponding author phone: (519)888-4659; fax: (519)885-1208; e-mail: klanctot@bioinformatics.uwaterloo.ca. Current address: Department of Computer Science, University of Waterloo, Waterloo, ON, N2J 3G1, Canada.

† Current address: Rational Discovery, 34336 Portia Terrace, Fremont, CA, 94444.

‡ Current address: BioSolveIt GmbH, An der Ziegelei 75, 53757 Sankt Augustin, Germany.

§ Current address: Cambios Computing, LLC, 1481 Pitman Avenue, Palo Alto, CA 94301.

10.1021/ci034129e CCC: $25.00
Published on Web 09/18/2003

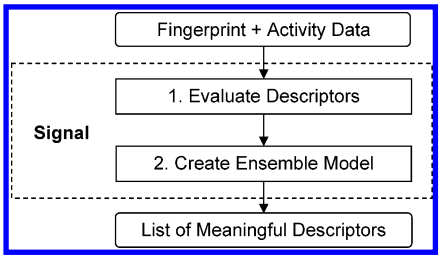


Figure 1. It is a two step process to convert the input, a set of fingerprints, and their corresponding activities, into a model suitable for screening a virtual library. First, Evaluate Descriptors: Signal evaluates each descriptor's correlation with activity. Second, Create Ensemble Model: Signal creates an ensemble, or a set of the most promising descriptors as a model of activity. A virtual library could then be screened to identify compounds that have an unusually large number of descriptors in common with the ensemble.

cophore and shape-feature, alternatives are in development and more are commercially available. Second, the models that Signal generates must be applicable in an automated fashion such as searching a virtual library. Finally, Signal should exploit information from both biologically active and inactive compounds. For example, if two pharmacophores are present in an equal fraction of active molecules, but one is present in only a few inactives and the other in many, then the prevalence among the inactives of the latter suggests that it is likely to be less relevant. Overall, these goals were met by developing a method that makes few assumptions except that the property space is encoded as a collection of binary fingerprints. Therefore continuous values, such as volumes or interatomic distances, must be binned.

METHODS

The overall process is illustrated in Figure 1. For a given set of compounds, Signal takes as input their fingerprints and activity data. To illustrate its flexibility, Signal is demonstrated using two different binary descriptor spaces: one is a pharmacophore fingerprint which represents all possible combinations of up to four pharmacophoric descriptors and the distances between them[25] and the other is a shape-feature fingerprint.[26,27] Both descriptors were generated for a Thrombin data set consisting of 41 active compounds and 6468 inactive compounds in a lead evolution stage of the drug design process. This data set consists of 38 known Thrombin actives, a library of 5837 informative[28] compounds from the Universal Informer Library,[29,30] and a set of 634 compounds selected from a virtual library of 285 872 compounds.

Generating Descriptors. In addition to using a common set of compounds for the pair of descriptors spaces, conformations were generated and pharmacophoric descriptors were identified using the same method, components of the Discovery Engine.[30] Given two-dimensional structure information, conformational analysis was performed to generate a collection of low-energy three-dimensional conformers for each compound in the data set using CONAN.[31] It typically generates 30−200 low-energy conformations per compound. The conformations are chosen to cover the conformational space of the compound. For the rest of the paper, the term conformers will refer to the conformers generated by CONAN. Next using a rule-based method, the compound's substructure was analyzed, and atoms in each conformation were classified into pharmacophoric descriptors. Six were used: hydrogen bond donor, hydrogen bond acceptor, positive charge, negative charge, hydrophobic group, and aromatic ring centroids. From this common starting point two different types of descriptors were generated for each compound.

Pharmacophore Descriptors. These descriptors are the classical pharmacophores where the distances between the points are discrete. Given a set of pharmacophoric descriptors, typically six, and a set of distance bins, typically between 8 and 14, all possible four-point pharmacophores are generated. This combination constitutes the descriptor space of potential pharmacophores. Some combinations of distance bins are physically impossible, such as those that violate the triangle inequality, and are discarded. For the remainder, on the order of tens to hundreds of millions of them, a unique offset is assigned corresponding to a position in a binary fingerprint. For the Thrombin data set, the range 1.5−24 Å was divided up into 11 distance bins resulting in a fingerprint with roughly 33 million positions. These positions in the binary fingerprint represent indicator variables for pharmacophores: if any conformer of a compound possesses pharmacophore *i*, then the *i*th position is set to one, otherwise it remains zero.[32]

Shape-Feature Descriptors. Given a set of active compounds, this fingerprint is created by considering all possible shapes that arise from the conformations of these compounds. This collection is called a shape catalog. If two shapes in the catalog are similar, within a user-defined threshold, then one of them is removed in order to reduce the size and redundancy of the catalog. These shapes are represented by placing the Thrombin conformation of interest into a three-dimensional grid so that the positive charge is at the origin, the centroid falls on the *x*-axis, and the heavy atom farthest from the *x*-axis is on the *xy* plane. Various higher order moments are calculated to determine if the centroid is placed on the positive or negative *x*-axis, and if the heavy atom should occur on the positive or negative *y*-axis. Putta et al.[26] give complete details of both the alignment procedure used to compare shapes and shape-feature descriptors. In short, the descriptor space consists of all possible combinations of shapes in the catalog, grid locations, and pharmacophoric descriptors at that grid location.[26,27] For example, for the Thrombin data, 38 different Thrombin ligands generated 3096 different conformations (between 3 and 334 per compound) which when duplicates were removed, resulted in a catalog of 196 shapes. Each shape was placed into a box with 8550 grid locations. The same six pharmacophoric descriptors as above were used, resulting in a binary fingerprint that was 10 054 800 bits long.

Evaluating Descriptors. To create a model for activity, Signal quantifies how well individual descriptors (*D*) correlate with activity (*A*) by tracking the number of active (*a*) and inactive (*i*) compounds in which *D* is either present (*p*) or not (*n*). A particular descriptor is said to be present in the compound's fingerprint, if the corresponding position in the fingerprint is set to one (in short, *D covers* this compound). Signal quantifies the relative merit of each descriptor's correlation with activity by evaluating these counts with a *ranking equation*. In general, their value increases as a descriptor covers more active or fewer inactive compounds. The relative tradeoff assigned to the number of active versus inactive compounds covered by the descriptor varies for

Table 1. Ranking Equation Parameters for a Descriptor *D*

symbol	meaning
N	total number of compounds
N_a	number of active compounds
N_i	number of inactive compounds
N_p	number of compounds with *D* present
N_n	number of compounds with *D* not present
N_{ap}	number of active compounds with *D* present
N_{an}	number of active compounds with *D* not present
N_{ip}	number of inactive compounds with *D* present
N_{in}	number of inactive compounds with *D* not present

different ranking equations. To capture this variation, two equations which have been well characterized in the literature are compared: *mutual information* from information theory[33] and *chi-square* from, among other places, categorical data analysis.[34] While the significance of the differences between them in a drug design application will be presented in the Results section, the formulas are described below.

Mutual Information. The first ranking equation (eq 1) is called the mutual information, *I(A, D)*, of the activity of an individual descriptor.[32] Its first term (eq 2) is the Shannon entropy of the activity data and is a function solely of the proportion of active and inactive compounds in the data set. It reaches its maximum when the number of active compounds in a data set equals the number of inactives. The second term (eq 3) is a function of the probability of a compound being active given the presence or absence of the descriptor *D*. The parameters used in eqs 1−7 are listed in Table 1.

$$I(A, D) = H(A) - H(A|D) \tag{1}$$

where

$$H(A) = -\left(\frac{N_a}{N}\log\left(\frac{N_a}{N}\right) + \frac{N_i}{N}\log\left(\frac{N_i}{N}\right)\right) \tag{2}$$

$$H(A|D) = -\left(\frac{N_{ap}}{N}\log\left(\frac{N_{ap}}{N_p}\right) + \frac{N_{an}}{N}\log\left(\frac{N_{an}}{N_n}\right) + \frac{N_{ip}}{N}\log\left(\frac{N_{ip}}{N_p}\right) + \frac{N_{in}}{N}\log\left(\frac{N_{in}}{N_n}\right)\right) \tag{3}$$

If the descriptor and the activity are completely independent, then the compounds are just as likely to be active whether *D* is present or not. Hence the three fractions in eq 4 are all equal.

$$\frac{N_a}{N} = \frac{N_a}{N_p} = \frac{N_{an}}{N_n} \tag{4}$$

Therefore the first term of $H(A)$ cancels out the first two of $H(A|D)$, as shown in eq 5.

$$\frac{N_a}{N}\log\left(\frac{N_a}{N}\right) - \frac{N_{ap}}{N}\log\left(\frac{N_{ap}}{N_p}\right) - \frac{N_{an}}{N}\log\left(\frac{N_{an}}{N}\right) = 0 \tag{5}$$

With a similar argument, the second term of $H(A)$ cancels out the last two terms of $H(A|D)$. Therefore $I(A,D)$ reaches its minimum, zero, in the case where *A* is independent of *D*. So mutual information varies between zero and the entropy in the data set, depending on how well the presence of the descriptor being ranked correlates with the activity of the compounds. Another ranking equation with similar properties, the Kullback-Leibler distance, has also been used in drug design.[35]

Chi-Square. The second ranking equation considered is the chi-square statistic:

$$\chi^2 = \frac{(N_{ap}N - N_aN_p)^2}{N_aN_p} + \frac{(N_{ip}N - N_iN_p)^2}{N_iN_p} + \frac{(N_{an}N - N_aN_n)^2}{N_aN_n} + \frac{(N_{in}N - N_iN_n)^2}{N_iN_n} \tag{6}$$

Each of the four terms in the sum quantifies the correlation of a different combination of whether a descriptor is present (*p*) or not (*n*) in the compounds and whether the compounds are active (*a*) or inactive (*i*). The first term in the sum quantifies the first combination: to what extent is *p* independent of *a*? If completely independent, then the probability of finding actives in the whole data set equals the probability of finding them among those covered by the descriptor; that is, eq 7 would be satisfied. The further from independence, the more positive or negative the difference is.

$$\frac{N_{ap}}{N} - \frac{N_aN_p}{N\,N} = 0 \tag{7}$$

The numerator of the first term in eq 6 is obtained by multiplying both the numerator and denominator of eq 7 by N^2. This value is squared ensuring that the difference is always positive, and it is normalized by dividing by N_aN_p making the value independent of the sample size. In summary, in the case of independence between *p* and *a*, the numerator of the first term of eq 6 equals zero. With similar arguments, the other numerators of eq 6 equal zero as well. On the other hand, the stronger the correlation between the presence of the descriptor and activity, the larger the value of the chi-square statistic.

Creating an Ensemble Model. With both chi-square and mutual information, a descriptor that is present in all actives and absent in all inactive compounds receives the highest possible rank. However, in practice, a single descriptor rarely possesses such strong correlation. Using an ensemble, or collection, of descriptors to model activity can mitigate this limitation. Researchers have observed that combining classifiers in a meaningful way increases their accuracy.[36,37] Hansen and Salamon[38] have shown that as long as each descriptor in an ensemble has better than a 50% accuracy and the errors of each descriptor are uncorrelated, then an ensemble of descriptors can achieve arbitrarily high accuracy. Dietterich[39] has noted three reasons for this improvement. First, when there is insufficient data, which is typical in the early stages of drug design, there may be many equally good models of activity and using an ensemble of them minimizes the risk that the wrong one is selected. Flexible ligands further aggravate the problem when it is unclear which conformation accounts for activity. Second, many search strategies halt at a local minima, so tracking many potential solutions reduces the risk that the global minima is missed. Having many potential solutions at hand is also useful when one promising avenue turns out to have pervasive pharmacokinetic problems. Third, simple models, such as a single pharmacophore, may not account for the complexity of

Table 2. Descriptor Example[a]

descriptor	A_1	A_2	A_3	I_1	I_2	I_3	I_4	I_5	I_6	$I(A, D)$	χ^2
D_1	1	1	0	1	0	0	0	0	0	2.250	0.179
D_2	1	1	0	1	1	0	0	0	0	0.900	0.073
D_3	1	0	1	0	0	1	1	1	0	0.225	0.018
D_4	0	0	1	0	0	0	0	0	1	0.321	0.025

[a] A cartoon example where the columns represent the fingerprints of three active compounds, A_1-A_3, and six inactive compounds, I_1-I_6. The rows represent the presence or absence of four descriptors D_1-D_4. The last two columns, respectively, represent the mutual information and chi-square value for these four descriptors.

binding, such as different ligands having different binding modes or overlapping but nonidentical sets of interactions within the same pocket. An ensemble model can address this complexity because different binding modes are modeled by the ability of the each mode to match a different subset of the ensemble.

In this paper, two ensemble methods were investigated. The first, *high ranking ensemble*, is simple and generally quite effective. The second, *high ranking set cover*, is a novel method specifically developed to address the limitations of the first when it is applied to early stage computational drug discovery. In principle, either ensemble method could be used in conjunction with either chi-square or mutual information.

High Ranking Ensemble (HRE). A straightforward approach for creating an ensemble is to select descriptors with the largest chi-square or mutual information values. Typically the user would decide on an ensemble size, *n*, and select the top *n* ranking descriptors. A simple example involving three active compounds, A_1-A_3, and six inactive compounds, I_1-I_6, characterized by four descriptors, D_1-D_4, is given in Table 2. Both mutual information and chi-square ranking equations would rank the descriptors in the order D_1, D_2, D_4, and D_3. For an ensemble of size two, the two highest ranking descriptors, D_1 and D_2, would be chosen.

High Ranking Set Cover (HRSC). Unfortunately, HRE possesses limitations. For many data sets HRE only covers a fraction of the active compounds. An example is presented in the Results and Discussion section. The basis for an alternate approach for selecting the ensemble, called *set covering*, originates from Dash.[40] A descriptor is said to cover a set of compounds if it differentiates one class (such as the actives) from another (inactive compounds). For example, in Table 2 descriptor D_1 differentiates A_1 and A_2 from I_2-I_6 because D_1 has different values for these active versus inactive compounds, so it is said to cover $\{A_1, A_2, I_2, I_3, I_4, I_5, I_6\}$. Similarly, D_2 covers $\{A_1, A_2, I_3, I_4, I_5, I_6\}$ and D_3 covers $\{A_1, A_3, I_1, I_2, I_6\}$. The subset $\{D_1, D_2, D_3\}$ of descriptors covers all the compounds as does the subset $\{D_1, D_3\}$. Dash's algorithm searches for the minimum number of descriptors that covers the data set so it would select $\{D_1, D_3\}$.

A novel variation of Dash's algorithm called *high ranking set cover* overcomes the limitations of HRE. As in HRE, the descriptors are first ranked using one of the two equations presented earlier. The descriptors are then examined in descending order, and one is added to the ensemble only if it covers active compounds that are not already covered by at least *d* descriptors, where *d* is a user-specified parameter establishing the depth of coverage. The algorithm continues to evaluate the remaining descriptors until all the actives are covered by at least *d* descriptors. For example, using the data in Table 2 with a coverage of $d = 1$, the algorithm runs as follows. The descriptors rank D_1, D_2, D_4, and D_3 by either ranking equation. D_1 is added to the ensemble because it covers actives A_1 and A_2. D_2 is skipped because it covers only actives that are already covered. D_4 is added because it covers A_3. With that choice, all actives are covered at a depth of at least one, so the algorithm halts selecting the ensemble $\{D_1, D_3\}$.

The high ranking set cover approach differs from Dash's set cover approach in three ways. First Dash's algorithm covers each compound once, whereas the high ranking set cover specifies a depth of coverage, *d*. Dash's algorithm was not developed specifically for early stage drug discovery, so while it mitigates the problem of certain subsets dominating the ensemble, a single descriptor per active is usually insufficient at this stage. Second, Dash's algorithm targets covering the entire data set, both actives and inactives; in contrast, HRSC covers only the active compounds. It takes into account information from the inactives implicitly via the ranking equation rather than by explicitly tracking the descriptors that correlate with inactivity. These are often difficult to discover from the data or even impossible at the early stages of drug design when the inactives are typically quite numerous and diverse. Third, Dash's algorithm tries to cover the compounds with as few descriptors as possible, whereas the goal of HRSC is coverage with high ranking descriptors. The belief being that the high ranking descriptors afford a computationally less expensive yet more effective tool for distinguishing the actives from the inactive compounds.

Applying an Ensemble Model. Given an ensemble, one possible application is to search a virtual library. A compound is *scored* against the ensemble by counting the number of descriptors a compound shares in common with the ensemble. The score may be interpreted as follows: the higher the score, the more likely that the compound is active. That is, the score correlates with the likelihood of activity, not necessarily the magnitude of activity.

Assessing the Performance. One would typically search a virtual library and synthesize either the top scoring compounds or compounds that score above a certain threshold. Even with an excellent model, a tradeoff exists: lowering the threshold discovers more actives (true positives) at the expense of including more inactive compounds (false positives). This tradeoff is assessed using a score plot, such as in Figure 2. Here the *x*-axis shows all possible thresholds. Signal predicts that a compound is active when its score is greater than or equal to the threshold. The *y*-axis shows the fraction of molecules nominated active. For example in Figure 2 using a threshold of 15 out of a possible 100, 96% of the actives and only 7% of the inactives in the test set have a score greater than 15. As the threshold increases less compounds, both active and inactive, are nominated active. Ultimately, the user decides on a suitable tradeoff between false negatives and true positives when setting the threshold. To assess the model's quality, the score plot includes curves for the performance on both training and test data. A significant difference between the two suggests that Signal is overfitting the data.

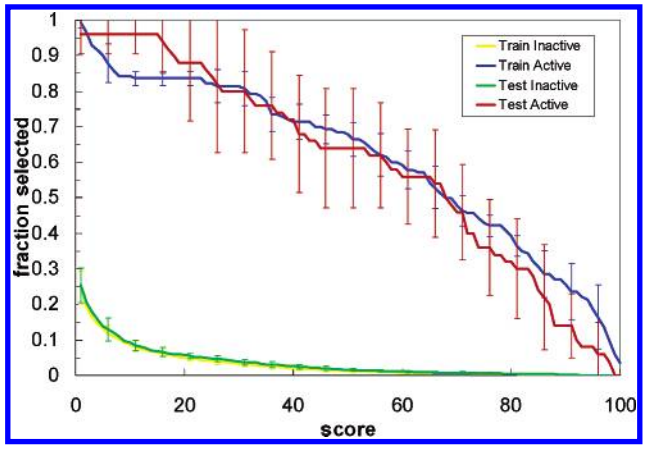


Figure 2. Sample score plot with a Thrombin data set: The *x*-axis shows all possible thresholds for an ensemble of 100 descriptor components. A molecule with a score greater than or equal to a particular threshold is nominated active. The *y*-axis is the fraction of molecules nominated active for each threshold value. The plot shows the performance curves for both training and test sets which are further subdivided into active and inactive compounds. The curve corresponding to train inactive and test inactive lie on top of each other.

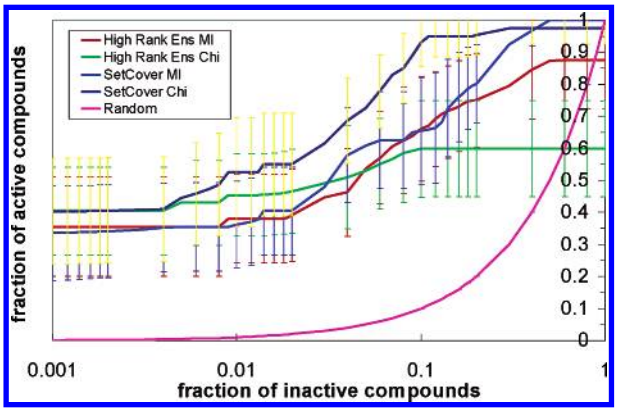


Figure 3. Pharmacophore descriptors with a Thrombin data set: This plot displays the log of the false positive probability on the *x*-axis against the true positive probability on the *y*-axis, as the threshold for classifying a compound in the test set as active changes. There is a performance curve for each of the four combinations of ranking equations (MI: mutual information vs Chi: chi-square) and ensemble methods (high ranking ensemble vs high ranking set cover). The combination of chi-square with set cover is the highest curve and hence the best combination.

While the score plot assesses the performance of a particular method, it is not well suited for comparing different methods. To facilitate the latter, the performance is quantified on log-linear ROC plots, such as in Figure 3 or Figure 4. A log-linear ROC plot displays the log of the false positive probability (fraction of inactives selected) on the *x*-axis against the true positive probability (fraction of actives selected) on the *y*-axis, as the threshold for classifying a test compound as active is decreased.[41] However, unlike the score plot, here the threshold is not explicitly shown on the graph. Rather, on an ROC plot the performance of a single method on the test set is depicted as a single curve. Different methods or parameter settings may be compared by plotting their performance curves on the same graph. If one curve is always above another, it means for a given fraction of inactives that pass the threshold, the method corresponding to the higher curve discovers more active molecules; that is, it is superior on that particular data set.

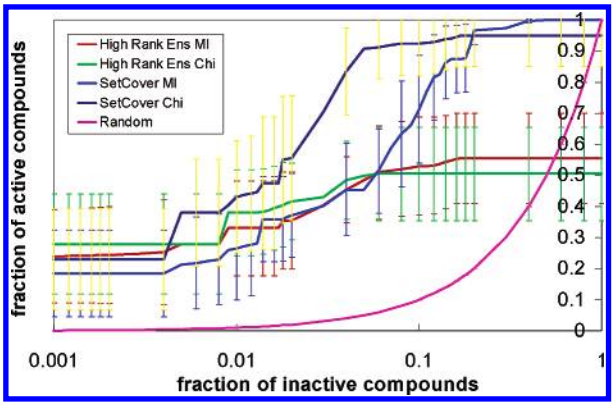


Figure 4. Shape-feature descriptor with a Thrombin data set: This chart displays the log of the false positive probability on the *x*-axis against the true positive probability on the *y*-axis, as the threshold for classifying a compound in the test set as active changes. There is a performance curve for each of the four combinations of ranking equations (MI: mutual information vs Chi: chi-square) and ensemble methods (high ranking ensemble vs high ranking set cover). The combination of chi-square with set cover is the highest curve and hence the best combination.

The error bars on each curve in Figures 2−4 represent the 95% confidence interval associated with each point (rather than the whole curve), and they are calculated from the distribution of 10 cross-validation trials. One method outperforms another if its ROC curve is predominately over the others. With enough samples, if one curve is completely above another, but always within its confidence interval, it could still be significantly better than the other as proved by a simple argument. If two curves are identical, and noise from the same distribution is added to each, then on average one curve would be above the other curve 50% of the time. The odds of having one curve above the other in seven out of seven data points by chance would be one-half to the exponent seven, which is less than one percent.

RESULTS AND DISCUSSION

The use of Signal is illustrated on a Thrombin data set consisting of 6509 compounds in a lead evolution stage of the drug design process. It took roughly 20 min to process the data set on a 400 MHz Pentium III computer. For each of the two types of descriptors, shape-feature and pharmacophore, a graph of all four combinations of the two types of ensemble models, high ranking set cover (HRSC) versus high ranking ensemble (HRE), and the two types of ranking equations, mutual information (MI) vs chi-square (Chi), are plotted along with a curve that represents the average performance of a random selection. All the curves are the average performance from 10-fold stratified cross-validation. Here, the cross-validation is stratified by activity level so that each training sample is guaranteed to have nine-tenths of both the active and inactive compounds. This strategy is necessary when inactive compounds vastly outnumber the active ones because picking samples completely randomly could result in a training sample with few if any active compounds. Figure 3 illustrates the performance for pharmacophore descriptors and Figure 4 for shape-feature descriptors. In both of these plots, the combination of using high ranking set cover with the chi-square ranking equation works best on the range 0.003−0.3. Experience with other data sets suggests that this combination is generally superior,

Table 3. Mutual Information's Highest Ranked Descriptors[a,b]

descriptor rank	N_{ap}	N_{ip}	mutual information
1	35	487	0.016141
2	35	500	0.015946
3	39	827	0.015897
4	39	828	0.015887
5	39	829	0.015877

[a] N_{ap} − number of active positive compounds. [b] N_{ip} − number of inactive positive compounds.

Table 4. Chi-Square's Highest Ranked Descriptors[a,b]

descriptor rank	N_{ap}	N_{ip}	chi-square
1	15	6	1687.1
2	11	1	1591.8
3	11	1	1591.8
4	14	6	1542.4
5	15	8	1538.3

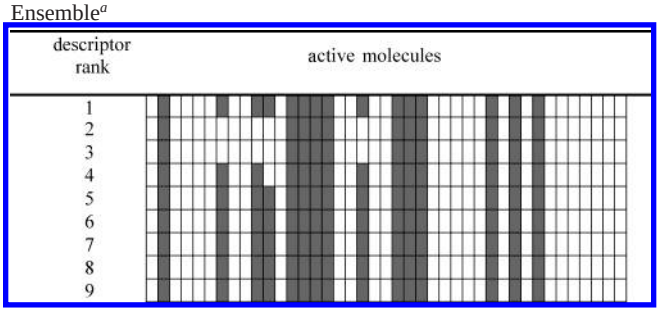
[a] N_{ap} − number of active positive compounds. [b] N_{ip} − number of inactive positive compounds.

sometimes equivalent, but never significantly worse than any of the other three combinations. To understand why, one must understand the limitations of the descriptors used, and how the ranking equations and model building techniques that Signal uses mitigate these limitations.

Comparing the Ranking Equations. When mutual information and chi-square were used to rank order descriptors on the Thrombin data set mentioned above, the results were quite different, the top five of which are listed in Tables 3 and 4. Clearly the different rankings favored different proportions of true positive, N_{ap}, at the expense of false positives, N_{ip}. In the case of mutual information, the high ranking descriptors are the ones that select almost all true positives at the expense of a large number of false positives, whereas descriptors ranked highly by chi-square have a greater ratio of true positives to false positives. This contrast is to be expected because mutual information is a measure of association[42] which evaluates how well the descriptor correlates with the any compound being active or inactive. We call this approach a whole model bias. Chi-square is a measure of statistical significance and is related to the likelihood that a random descriptor is positive for many active compared to inactive compounds. We call this approach a bias toward significant components.

Comparing Ensembles Methods. Both ranking equations suffer from the same problem during drug discovery on a real target. Typically the top ranking ensemble covers only a fraction of the active compounds; however, in an ideal situation, all the active compounds would be equally well represented. Table 5 illustrates this limitation. It shows which active molecules in the Thrombin data set are covered by the nine descriptors with the highest chi-square values. Each column under Active Molecules represents a single active compound. Each row represents a single high ranking descriptor, with the highest ranking descriptor in row one. When a cell is filled, then the corresponding descriptor covers the corresponding active compound. So the first active compound (or column) does not contain any of the nine highest ranked descriptors, and the second compound contains all nine of them. Table 5 demonstrates that the same 15 compounds are covered by the nine top ranking descriptors. So a high ranking ensemble of size nine accounts for only 15 out of the 41 active compounds. This is an example of a set of active compounds, typically from the same congeneric series, dominating the ensemble. When descriptors from a single, or few, congeneric series dominates an ensemble, it is less likely that this ensemble accounts for all the different reasons a compound may be active, reducing the likelihood that novel scaffolds will be discovered.

Table 5. Set of Compounds Dominating a High Ranking Ensemble[a]



[a] Each row represents a single high ranking descriptor. Each column represents a single active compound. If a cell is filled in then the corresponding descriptor is present in the corresponding active compound. This table illustrates that the top ranking descriptors cover only 15 of the 41 active compounds in a Thrombin data set.

The first version of Signal used the combination of the mutual information ranking equation paired with high ranking ensemble. After about a half dozen early stage drug discovery projects involving Signal, many users observed the phenomena described above of a small set of active compounds dominating the ensemble. Using the chi-square ranking equation, or others like it, only exacerbated the problem, because it would focus even more on the significant components. However, the combination of chi-square with set cover alleviated the problem, as is illustrated in Figures 3 and 4, where the curve corresponding to high ranking set cover plus chi-square, for the most part, dominates all other curves. The reason for this result is that the two methods are complimentary. While chi-square ranking ensures that few inactives are picked (reducing false positives), the set cover algorithm, by design, covers as many actives as possible (increasing true positives).

CONCLUSION

This paper presents a novel machine learning approach to distinguish active molecules from inactives. Within a general framework, two different types of 3-D descriptors were used as a basis for discrimination. The first is a traditional pharmacophore type, and the second encodes the shapes and features of a compound. Both are encoded as large binary vectors. The model itself is an ensemble of the binary descriptors whose presence correlates strongly with the activity. Correlation is evaluated using either mutual information or chi-square ranking equations, and the high ranking descriptors are then collected into an ensemble based solely on their ranking or by covering the set of active compounds. Cross-validation experiments with this method on a Thrombin data set show reasonably accurate classification of the active and inactive compounds in the test set. Further, the combination of the high ranking set cover with the chi-square ranking equation seems to yield the best results.

Several positive features of this approach may be noted. Signal is robust in the sense that it works with a variety of descriptors, and it is economical because it exploits information from both active and inactive compounds. It is efficient because it can process hundreds of thousands of compounds a day, making it appropriate for automated searching through virtual libraries of compounds. Finally, the combination of the chi-square ranking equation plus the novel method, high ranking set cover, helps alleviate the problem of a small number of active compounds having a large influence on the final predictive model.

ACKNOWLEDGMENT

The authors thank the other members of the Computation Sciences Team who have provided valuable insight and suggestions for this project. In particular we acknowledge the contributions Jeff Blaney, Erin Bradley, Karen Bradshaw, Scott Cheng, John Eksterowicz, Erik Evensen, Joel Galloway, Peter Grootenhuis, Michelle Lamb, Connie Oshiro, David Spellmeyer, Jayashree Srinivasan, and Rob Stanton.

REFERENCES AND NOTES

(1) Bakken, G. A.; Jurs, P. C. Classification of Multidrug-Resistance Reversal Agents Using Structure-Based Descriptors and Linear Discriminant Analysis. *J. Med. Chem.* **2000**, *43*, 4534−4541.

(2) Goll, E. S.; Jurs, P. C. Prediction of Normal Boiling Points of Organic Compounds from Molecular Structure with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.

(3) Sutter, J. M.; Jurs, P. C. Neural Network Classification and Quantificaton of Organic Vapors Based on Fluorescence Data from a Fiber-Optic Sensor Array. *Anal. Chem.* **1997**, *69*, 856−862.

(4) Sutter, J. M.; Jurs, P. C. Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.

(5) Xu, L.; Ball, J. W.; Dixon, S. L.; Jurs, P. C. Quantitative Structure−Activity Relationships for Toxicity of Phenols Using Regression Analysis and Computational Neural Networks. *Environ. Toxicol. Chem.* **1994**, *13*, 841−851.

(6) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure−activity relationships by neural networks and inductive logic programming. I. The inhibition of dihydrofolate reductase by pyrimidines. *J. Comput.-Aided. Mol. Des.* **1994**, *8*, 405−420.

(7) Hirst, J. D.; King, R. D.; Sternberg, M. J. E. Quantitative structure−activity relationships by neural networks and inductive logic programming. II. The inhibition of dihydrofolate reductase by triazines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 421−432.

(8) King, R. D.; Srinivasan, A. The discovery of indicator variables for QSAR using inductive logic programming. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 571−580.

(9) King, R. D.; Muggleton, S. H.; Srinivasan, A.; Sternberg, M., J. Structure activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *PNAS* **1996**, *93*, 438−442.

(10) Chen, X.; Rusinko, A.; Tropsha, A.; Young, S. S. Automated pharmacophore identification for large chemical data sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 887−896.

(11) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; John Wiley & Sons: 1999.

(12) Carey, R. N.; Wold, S.; Westgard, J. O. Principal component analysis: an alternative to "referee" methods in method comparison studies. *J. Anal. Chem.* **1975**, *47*, 1824−1829.

(13) Ghuloum, A. M.; Sage, C. R.; Jain, A. N. Molecular hashkeys: a novel method for molecular characterization and its application to predicting important pharmaceutical properties of molecules. *J. Med. Chem.* **1999**, *42*, 1739−1748.

(14) Hoffman, B.; Cho, S. J.; Zheng, W.; Wyrick, S.; Nichols, D. E.; Mailman, R. B.; Tropsha, A. A QSAR modeling of dopamine D1 agonists using comparative molecular field analysis, genetic algorithms-partial least-squares, and K nearest neighbor methods. *J. Med. Chem.* **1999**, *42*, 3217−3226.

(15) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E.; Bauer, B. E.; Webster, T. A.; Lozano, P., T. Compass: A shape-based machine learning tool for drug design. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 635−652.

(16) Poetter, T.; Matter, H. Random or rational design? Evaluation of diverse compound subsets from chemical structure databases. *J. Med. Chem.* **1998**, *41*, 478−488.

(17) Stanton, D. T.; Morris, T. W.; Roychoudhury, S.; Parker, C. N. Application of nearest-neighbor and cluster analyses in pharmaceutical lead discovery. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 21−27.

(18) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 4−15.

(19) Lemmen, C.; Zien, A.; Zimmer, R.; Lengauer, T. Application of Parameter Optimization to Molecular Comparison Problems. Pacific Symposium on Biocomputing (PSB'99); 1999; pp 482−493.

(20) Martin, Y. C.; Bures, M. G.; Danaher, E. A.; DeLazzer, J.; Lico, I.; Pavlik, P. A. A fast new approach to pharmacophore mapping and its application to dopaminergic and benzodiazepine agonists. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 83−102.

(21) Marshall, G. R.; Barry, C. D.; Bosshard, H. D.; Dammkoehler, R. A.; Dunn, D. A.; Olson, E. C.; Christoffersen, R. E. The Conformational Parameter in Drug Design: The active analogue approach. *Comput.-Assisted Drug Des.* **1979**, 205−222.

(22) Jones, G.; Willett, P.; Glen, R. C. A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 532−549.

(23) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

(24) Liu, H.; Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*; Kluwer: NY, 1998.

(25) Bradley, E. K.; Beroza, P.; Penzotti, J. E.; Grootenhuis, P. D. J.; Spellmeyer, D. C.; Miller, J. L. A rapid computational method for lead evolution: description and application to α1-adrenergic antagonists. *J. Med. Chem.* **2000**, *43*, 2770−2774.

(26) Putta, S.; Lemmen, C.; Beroza, P.; Greene, J. A Novel Shape-Feature Based Approach to Virtual Library Screening. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1230−1240.

(27) Srinivasan, J.; Castellino, A.; Bradley, E. K.; Eksterowicz, J. E.; Grootenhuis, P. D. J.; Putta, S.; Stanton, R. V. Evaluation of a Novel Shape-Based Computational Filter for Lead Evolution: Application to Thrombin Inhibitors. *J. Med. Chem.* **2002**, *45*, 2494−2500.

(28) Teig, S. L. Informative libraries are more useful than diverse ones. *J. Biomol. Screening* **1998**, *3*, 85−88.

(29) Saunders: J.; Myers, P. L.; Barnum, D.; Greene, J. W.; Teig, S. L. Drug Discovery Development of a Universal Informer Library: Data Derived from the Training Set. *Genetic Eng. News* **1997**, *17*, 35−36.

(30) Myers, P. L.; Greene, J. W.; Saunders: J.; Teig, S. L. Rapid reliable drug discovery. *Today's Chemist at Work* **1997**, *6*, 46−48.

(31) Smellie, A.; Stanton, R. V.; Henne, R. M.; Teig, S. L. Confomational Analysis by Intersection. *J. Comput. Chem.* **2003**, *24*, 10−20.

(32) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(33) Cover, T. M.; Thomas, J. A. *Elements of Information Theory*; Wiley-Interscience: New York, 1991.

(34) Chap, T. L. *Applied Categorical Data Analysis*; John Wiley & Sons: 1998.

(35) Barnum, D.; Greene, J. W.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.

(36) Breiman, L. Bagging Predictors. *Machine Learning* **1994**, *24*, 123−140.

(37) Freund, Y.; Schapire, R. Experiments with a new boosting algorithm. Proceeding of the Thirteenth International Conference of Machine Learning; Morgan Kaufmann: Bari, Italy, 1996; pp 148−156.

(38) Hansen, L.; Salamon, P. Neural network ensembles. *IEEE Trans. Pattern Analysis Machine Intelligence* **1990**, *12*, 993−1001.

(39) Dietterich, T. G. Ensemble Methods in Machine Learning. *First International Workshop on Multiple Classifier Systems, Lecture Notes in Computer Science*; Springer-Verlag: 2000; pp 1−15.

(40) Dash, M. Feature Selection via Set Cover. *IEEE Knowledge and Data Engineering Exchange Workshop*; Newport Beach, CA, 1997; pp 165−171.

(41) Swets, J.; Dawes, R. M.; Monahan, J. Better Decisions through Science. *Sci. Am.* **2000**, *283*, 82−87.

(42) Press: W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: New York, 1993.

CI034129E