Virtual Screening of Molecular Databases Using a Support Vector Machine

Robert N. Jorissen and Michael K. Gilson*

Center for Advanced Research in Biotechnology, University of Maryland Biotechnology Institute,
9600 Gudelsky Drive, Rockville, Maryland 20850

Received November 30, 2004

The Support Vector Machine (SVM) is an algorithm that derives a model used for the classification of data into two categories and which has good generalization properties. This study applies the SVM algorithm to the problem of virtual screening for molecules with a desired activity. In contrast to typical applications of the SVM, we emphasize not classification but enrichment of actives by using a modified version of the standard SVM function to rank molecules. The method employs a simple and novel criterion for picking molecular descriptors and uses cross-validation to select SVM parameters. The resulting method is more effective at enriching for active compounds with novel chemistries than binary fingerprint-based methods such as binary kernel discrimination.

INTRODUCTION

Virtual screening refers to the use of a computer-based method to select compounds from a library or database of compounds in order to identify ones that are likely to possess a given activity, such as the ability to inhibit the action of a particular therapeutic target (see e.g. refs 1−3). Selection of molecules with a virtual screening algorithm should yield a higher proportion of active compounds, as assessed by experiment, relative to a random selection of the same number of molecules; i.e., the sample is enriched for active compounds. In this work, we are interested in the case where a heterogeneous set of active compounds is known (as could be obtained from a prior screening process or from the scientific literature), and we seek other molecules with the same activity, preferably from novel compound classes. These new compounds may include ones which have better pharmacokinetic properties than those previously known, are more “leadlike”,[4,5] and/or have not been previously patented.

The Support Vector Machine (SVM)[6,7] is an algorithm which has begun to receive attention in the cheminformatics field for its ability to classify objects into two classes as a function of their features. Several studies have shown the SVM to be among the best methods for correctly classifying molecules.[8−11] A standard application of the SVM algorithm involves defining two classes of objects, determining a set of numbers that characterize each object, and using the SVM algorithm to calculate a classification model for the objects. After this training step, the SVM model is used to classify other objects. In this work, the two classes of objects are active and inactive molecules, which are characterized by their molecular descriptors. Once trained, the model is then applied to a test set to predict which of its molecules are active.

A disadvantage of using a classifier such as the SVM is that it does not rank molecules according to their likelihood of being active. In practice, the number predicted to be active using a binary classification approach may significantly differ from the number of compounds that can be tested experimentally using available resources. Additionally, some of the compounds predicted to be active may not be available. In this work, we modify the SVM methodology to provide for the ranking of molecules. We also describe automated methods for choosing descriptors and appropriate parameters for using the SVM to enrich a selection of molecules for a desired activity.

METHODOLOGY

In this work, molecular descriptors of the active and inactive compounds in a training data set are used to train a Support Vector Machine.[6,7] The descriptors of these molecules can be represented as points in a multidimensional space where each dimension corresponds to one of the descriptors. The SVM seeks to find a boundary that best separates the two sets of points corresponding to the active and inactive compounds. The resultant SVM model then ranks a test set that consists of other active and inactives, and the recovery of the active compounds provides a measure of the performance of the SVM-based enrichment method. The molecular descriptors in the SVM model are chosen for their ability to cluster the training set actives in the descriptor space. As the SVM training process is affected by the choice of values for several SVM parameters, different SVM models can be obtained for a given training set. Here, a cross-validation-derived statistic is used to choose among the different possible models. The next few sections describe this process in more detail.

The Support Vector Machine. Each object i (molecule) to be classified by the SVM is described by a vector $\mathbf{x}_i$ of M real numbers (descriptors) and can be therefore represented as a point in an M-dimensional space. The objects in the first class (active molecules in the training set) are each assigned a value of $y_i = +1$, and the objects in the second class (inactive molecules) are assigned a value $y_i = -1$. In the linearly separable case, the SVM attempts to find an optimal hyperplane that perfectly separates the two classes of objects in the M-dimensional space. The optimal hyper-

* Corresponding author e-mail: gilson@umbi.edu.edu.

10.1021/ci049641u CCC: $30.25
Published on Web 04/16/2005

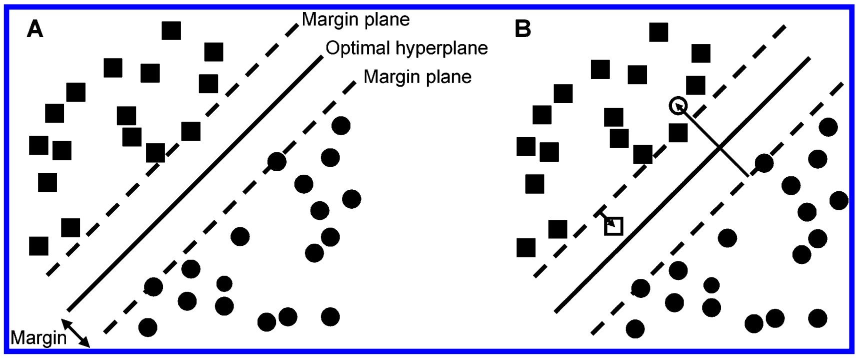

Figure 1. Separating hyperplane of the Support Vector Machine that maximizes the margin between two sets of perfectly separable objects, represented as circles and squares. (A) Optimal hyperplane that perfectly separates the two classes of objects. (B) Optimal soft margin hyperplane which tolerates some points (unfilled square and circle) on the "wrong" side of the appropriate margin plane.

plane is one which maximizes the margin, defined as the closest distance from any point to the separating hyperplane (Figure 1A). The points of each class then lie on or beyond one of two margin planes which are parallel to the separating plane (Figure 1A). The predicted class of another object (a molecule in the test set) thus depends on which side of the separating hyperplane the object's point is located.

A hyperplane is defined by a normal vector **w** and a scalar b such that any point on this plane obeys eq 1

$$\mathbf{w}\cdot\mathbf{x} + b = 0 \qquad (1)$$

and the equation of vector **w** that maximizes the margin is

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i x_i \qquad (2)$$

where N is the number of objects (molecules) and the α_i are coefficients obtained from the SVM training. Only the points on, or on the wrong sides of, the margin planes have nonzero values of α_i. These points are referred to as support vectors. The coefficients α_i are obtained by maximizing the following functional

$$W = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \qquad (3)$$

subject to the constraints

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

and

$$\alpha_i \geq 0 \text{ for } i = 1, ..., N$$

The value of b can be subsequently obtained by noting that for any object i

$$\alpha_i(y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1) = 0 \qquad (4)$$

Equation 3 has the form of a quadratic optimization problem and has a unique solution for a given system. To solve for b and the set of α_i, we used portions of code from the libsvm suite of programs[12] which employs a modified version of the Sequential Minimal Optimization (SMO) algorithm.[13,14] This method solves the SVM problem by iteratively solving for pairs of α_i while updating the values of upper and lower thresholds for b, until convergence of these values has been reached within a specified tolerance.

The SVM model consists of the values of b and the α_i, the support vectors and their assigned classes ($\mathbf{x}_i$ and y_i), and the SVM kernel and its associated parameters, which are described later. Once the SVM model has been obtained, the decision function f(x) can be used to predict whether an untested object, defined by its vector **x**, belongs to the +1 or −1 class:

$$\mathrm{f}(\mathbf{x}) = \mathrm{sgn}(\mathbf{w}\cdot\mathbf{x} + b) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i (\mathbf{x}_i \cdot \mathbf{x}) + b) \qquad (5)$$

In practice, the separation of the two sets of data points may not be perfect. Since the solution of the SVM training problem depends on the points closest to the decision boundary, one outlying point can greatly skew the position of the decision boundary with respect to the other points. Thus, allowing for a small number of training errors can lead to a decision boundary that provides for superior classification of objects in a test set. The L1 soft margin formulation of the SVM[15] allows for this possibility by allowing points to fall on the wrong side of the appropriate margin plane but penalizing each of these points by a constant multiplied by its distance from the margin plane (Figure 1B). This constant, C, controls the tradeoff between maximizing the margin and placing each point on the correct side of the relevant margin plane. The optimal separating plane is then found using the same procedure as before (eqs 3 and 4), but with the additional constraint that the values of α_i must be less than or equal to C. In this work, we used separate error weighting constants, $C+$ and $C-$, for each of the two classes of objects-active and inactive molecules, respectively. Since we are more concerned with the misclassification of the actives than the inactives, the value of $C-$ was restricted to be less than or equal to the value of $C+$.

The SVM methodology is not limited to the use of a planar separating boundary to classify the two classes of data points. A nonplanar decision boundary can be achieved by transforming the points into a higher dimensional space and performing a planar separation in this space. In practice, this is achieved using the elegant "kernel trick" in which the dot product in eqs 3 and 5 is replaced with a kernel function, K, which represents the dot product in the transformed space. This avoids explicit calculations involving the higher dimensions, which can even be infinite in number; the transformation is implicit in the choice of kernel. The generalized decision function then has the form

$$\mathrm{f}(\mathbf{x}) = \mathrm{sgn}(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b) \tag{6}$$

Within this framework, use of the so-called linear kernel recovers the linearly separable SVM:

$$K_{\mathrm{linear}}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1 \cdot \mathbf{x}_2 \tag{7}$$

In this work, we use the Gaussian or Radial Basis Function kernel:

$$K_{\mathrm{Gaussian}}(\mathbf{x}_1,\mathbf{x}_2) = \exp(-\gamma|\mathbf{x}_1 - \mathbf{x}_2|^2) \tag{8}$$

This kernel was chosen because it readily produces a closed decision boundary, which is consistent with the method used to select the molecular descriptors, as described later. It should be noted that when the value of γ is large, the separating boundary has a large number of support vectors and can become tortuous. This risks overfitting the training set data to yield an SVM model that is not robust. In contrast, a small value of γ can lead to separating boundaries described with a small number of support vectors but that may be too smooth to classify the training set examples with sufficient accuracy. Therefore, a suitable value of γ is needed for training the SVM.

In this work, we move away from the classification paradigm but retain other aspects of the SVM methodology. Removing the sgn function from eq 6 produces a function that generates a real number instead of -1 or $+1$:

$$\mathrm{f}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{9}$$

When applied to the descriptors of a set of untested molecules, the values produced by eq 9 can be used to rank the molecules. We refer to the number calculated by eq 9 as the SVM activity score, where higher scores are more favorable. The SVM is trained in the same way as if the resultant model were to be used for classification.

Evaluation of Support Vector Machine Parameters Using Cross-Validation. In this work, the derivation of an SVM model for a particular training set requires choosing appropriate values for the parameters γ, $C+$, and $C-$. To select values for these parameters, multiple SVM models are obtained from a systematic scan of combinations of parameter values, and the resulting models are each assessed by a cross-validation procedure to yield a statistic called AvRank (described below). The SVM parameter values that yield the lowest value of AvRank are then used to train a final SVM model but with a higher precision than was used in the scan (see below). This model is then validated by application to a test set.

The training set statistic, AvRank, is calculated as follows. The training set, which consists of k active molecules and $N-k$ inactive molecules, is subjected to a k-fold cross-validation procedure. Here, the training set is first divided into k groups of molecules, each containing one active and approximately $(N-k)/k$ inactive molecules, chosen randomly. The SVM is trained on the pooled set comprising the molecules in k-1 of these groups, and then the SVM activity scores (eq 9) of the molecules in the one left-out group are calculated using the SVM model. This process is performed k times so that each active and inactive is left out once. As a time-saving measure, the stopping tolerance of the SVM algorithm is increased from the usual value of 0.001 to a looser value of 0.01 during the cross-validation process. After one round of this cross-validation procedure, the molecules are ranked according to their cross-validated SVM activity scores, their fractional rankings (the ranking of the molecule divided by the total number of molecules) are determined, and the average of the fractional rankings of the actives is calculated. This procedure can yield slightly different results depending upon which inactive molecules are randomly assigned to each of the k groups, so it is performed five times to obtain five slightly varying values for the average fractional rank. We define our performance statistic AvRank as the average of these average fractional ranks.

In the systematic scan of SVM parameters, the values of $C+$ and $C-$ are set to 1, 10, 100, 1000, or 10000 with the restriction that $C-$ is less than or equal to $C+$. The values of the Gaussian kernel parameter γ are set to 0.01 or 0.1. (When the value of γ was 1.0 or more, the number of support vectors in the resultant model was found empirically to be greater than 85% of the total number of molecules in the training set, a situation which risks overfitting the training data.) Some combinations of parameters (same γ and different $C+$ and/or $C-$) produce identical models to those that have been produced earlier in the scan. Therefore, the SVM model resulting from a given combination of γ, $C+$, and $C-$ is compared to the previously generated models and then deleted if it is found to be a duplicate. The value of AvRank is calculated for each unique model. After the scan of parameters, the final SVM model is calculated for the training set data using the values of parameters that gave the lowest value of AvRank, but with the SVM stopping tolerance value set to 0.001. The set of SVM parameters obtained in this manner will differ according to the active and inactive compounds in the training set, and so the scan over candidate SVM parameters and concurrent calculation of AvRank values is considered part of the training process in this work.

Because we are using the SVM to rank compounds, rather than classify them, we could not apply standard SVM figures of merit to establish optimal SVM parameters. However, we did evaluate alternatives to the AvRank statistic. These include the fractional rank of the worst-ranked active compound, the average of the square root of the fractional ranks (a statistic which emphasizes the higher-ranked compounds), and these same quantities multiplied by functions of the number of support vectors. These alternatives did not improve the results. It is worth mentioning that the AvRank statistic devised for this work is similar to the sum of ranks of the training set actives (SumRank), obtained from a leave-one-out cross-validation procedure, which is used to select an appropriate parameter in binary kernel discrimination (BKD).[16,17]

Selection of Molecular Descriptors. Each molecule can be considered to correspond to a point in a descriptor space, where the molecule's coordinates are specified by the values of its descriptors. Descriptors are chosen with the aim of placing the active molecules in a small cluster whose volume excludes most of the inactive compounds. Thus, the preferred descriptors are those for which the values of the inactive molecules tend to lie outside of the range of the descriptor values of the actives. This is consistent with our use of the Gaussian kernel (eq 8) which readily leads to a closed boundary around the active compounds to the exclusion of many inactive compounds. A molecule whose descriptor values position it close to the center of the cluster will have a high SVM activity score (eq 9).

The first step in descriptor selection consists of removing descriptors which have the same value (usually zero) for at least half of the molecules in the training set. Discrimination scores for the remaining descriptors are then calculated, where the discrimination score for a given descriptor is defined as the fraction of inactive molecules in the training set whose descriptor values lie outside of the range of descriptor values of the actives.

For a few descriptors, the values for one or more active molecules significantly deviate from the values for the other actives. Using such a descriptor would be inconsistent with the idea of grouping the active molecules into a single cluster in the descriptor space. To remove these descriptors, we first calculate a score for each descriptor which is the median absolute deviation of the descriptor values of all molecules divided by the range of descriptor values for the active molecules, where the median absolute deviation is the median of the absolute differences of each value from the median. This score originally arose as a candidate for the descriptor discrimination score which ranks descriptors. Descriptors with a value of this score less than 0.1 are eliminated from further consideration. Of the remaining descriptors, those with the highest discrimination scores are retained if the absolute value of the correlation with another descriptor with a higher discrimination score is less than a user-defined value, fixed at 0.8 for this work, which we term the correlation cutoff. In this way, a set of favorable descriptors is selected prior to the SVM training step.

The descriptor selection used in this work is a so-called "filter" method in which the descriptors are chosen prior to applying the support vector machine (e.g. see ref 18). This class of descriptor selection method was considered to be preferable to "wrapper" methods such as recursive feature elimination[19] in which the SVM is trained multiple times and the descriptors subsequently evaluated in light of the training. However, "wrapper" methods can be time-consuming when applied to a large number of available descriptors (511 for this work). We also avoided the use of descriptor selection methods such as Golub's feature selection criterion[20] which presume a separation of active and inactive compounds into separate parts of the space spanned by the descriptors. This class of methods conflicts with our conception of the optimal descriptors as those which place the actives in a small, closed region in a "sea" of inactive compounds in the space spanned by the descriptors. The discrimination score for choosing the descriptors is a simple measure which is consistent with this concept.

For each of the selected descriptors, the values from all of the training set molecules are linearly transformed so that the values for the active compounds in the training set occupy the interval [0, 1]. This process scales the descriptor data so that descriptors with large numerical values do not dominate the SVM model. The same transformation is also used to scale the descriptors in the test set.

Except where otherwise noted, all SVM models are derived using the 50 descriptors with the highest discrimination scores, subject to the correlation cutoff criterion. However, in one set of calculations, the number of descriptors was not held fixed but was allowed to vary in the scan along with the SVM parameters γ, $C+$, and $C-$. The candidate values of the number of descriptors were 10, 20, 30, 40, 50, 60, and 70.

Construction of Training and Test Data Sets. The molecules used in the present study comprise five sets of 50 molecules which each target a different protein and also a set of background molecules that are assumed to be inactive. The active molecules are reversible inhibitors of cyclin-dependent kinase 2 (CDK2), cyclooxygenase-2 (COX2), factor Xa (FXa), and phosphodiesterase-5 (PDE5) and reversible antagonists of the α_{1A} adrenoceptor (α_{1A} AR). (For convenience, all of these molecules will be referred to interchangeably as actives or inhibitors.) Each set of 50 molecules was collected from the scientific literature and covers a variety of chemical classes (Supporting Information). The Lewis structures of these molecules were sketched using IsisDraw 2.4[21] and saved as MDL Mol files.[22,23] The background set molecules used for most of the calculations described in this paper were drawn from the National Cancer Institute (NCI) diversity set of chemical compounds.[24] For one set of calculations, however, molecules from the August 1999 release of the Maybridge database[25] were used instead. In addition, when one set of 50 actives was studied, all of the other 200 known inhibitors were assumed to be inactive against the target of interest as were the compounds from the background set (NCI or Maybridge).

To test the SVM-based enrichment method, the inhibitors and background molecules must be divided into two sets: one set that is used for training an SVM model, and a test set to which the SVM model is applied. Each of the five sets of 50 inhibitors was divided into two equal-sized data sets in two different ways. Prior to the division, the compounds with similar chemistries, as judged by one of us (R.N.J.), were grouped. The first split of each set of 50 compounds placed the first, third, fifth, etc. compounds of the ordered list into one set, termed ODD, and the second, fourth, sixth, etc. compounds into another set, termed EVEN. The ODD and EVEN sets each contain representatives from all of the different chemical classes of the inhibitors, except in the few instances where an inhibitor was not grouped with another one. Thus, most of the inhibitors in one of these sets will have at least one similar compound in the other set. The second separation of each set of 50 inhibitors placed the first 25 compounds into one set (called 1ST), and the second 25 compounds into another set (2ND). These two complementary sets of compounds contain inhibitors from nonoverlapping chemical classes and present a greater challenge to the algorithms described in this work.

Each of the four sets of compounds (ODD, EVEN, 1ST, and 2ND) was supplemented with a background set of molecules from the NCI diversity set. After filtering out unsuitable molecules, the odd entries from the remaining 1892 NCI molecules were added to both the ODD and 1ST data sets, and the even entries were used to supplement the EVEN and 2ND data sets. Each training and test set can thus be described in terms of the target of the inhibitors selected as the actives and the set of inhibitors used; the nomenclature is of the form FXa/1ST.

For one set of calculations, larger test sets were used in which the odd and even entries of the filtered Maybridge database (25 175 molecules each) replaced NCI diversity set molecules in the 1ST and 2ND sets. Smaller training sets were constructed from these test sets by retaining all of the known inhibitors but only a fraction of the Maybridge molecules. Starting from the first compound in the given test set, either every fifth or every 25th Maybridge molecule was retained for the training set. The resulting data sets contain 5035 and 1007 Maybridge molecules, respectively. This process was repeated to generate additional training sets but beginning from the second compound in the given test set. We refer to these training sets as 5−1, 25−1, 5−2, and 25−2, respectively, in addition to the nomenclature described previously.

Calculation of Molecular Descriptors. Prior to the calculations of their descriptors, the various sets of molecules (inhibitors, NCI and Maybridge) were edited and modified to put them in a suitable form for the calculations. After removal of counterions or other small molecules from each entry, molecules containing atoms other than H, C, N, O, P, and S and the halogens were removed. Deuterium atoms in several Maybridge database entries were manually changed to hydrogen atoms. Protonation of the molecules in the Maybridge database and the 250 inhibitors was performed by generating three-dimensional structures for these molecules with CORINA[26] and then applying a locally modified version of the molecular format interconversion program, Babel.[27] In this computer program, the rules for performing protonation were modified to better reflect the expected ionization state at physiological pH for functional groups including amidine, guanidinium, *N*-oxide, and tetrazole. The NCI diversity set is distributed with computed three-dimensional coordinates and protonation states that are representative of gas-phase conditions. Therefore, the hydrogen atoms of these compounds were removed and then added using our modified version of the Babel program. Molecules found at this stage to possess more than 150 atoms were not considered further as we could not calculate descriptors for molecules this large with the available software (see below). One more molecule, arbitrarily chosen, was omitted from each of these two sets to make an even number of molecules in both the filtered NCI diversity set (1892 molecules) and the filtered Maybridge set (50 350 molecules).

Molecular descriptors were calculated using version 2.1 of the DRAGON program.[28] In this work, 517 descriptors were calculated from descriptor categories 1−6 (constitutional descriptors, topological descriptors, molecular walk counts, BCUT descriptors, Galvez topological charge indices and 2D autocorrelations) and categories 17−18 (empirical descriptors and properties). Descriptors whose values depend on the three-dimensional coordinates of the molecules (descriptor categories 7−14) were not used. Additionally, descriptors that count functional groups and atom types (categories 15 and 16) were omitted since the descriptor values do not span a continuous and widely varying range of values and so are not well-suited to the methods described in this work. Also, the descriptors X0sol, X1sol, X2sol, X3sol, X4sol, and X5sol were omitted because their calculated values were found to vary depending on the order of the atoms in the input file for some test molecules.

Comparison of the SVM-Based Enrichment Method with Fingerprint-Based Ranking Methods. Using the 20 data sets, the SVM-based enrichment method was compared to four ranking methods which make use of pairwise similarities calculated from molecular fingerprints. In this work, the fingerprints used were chemical hashed fingerprints calculated from the GenerFP program in version 3.0.2 of JChem.[29] The fingerprints were generated using the following (default) parameters: fingerprint length 512 bits (64 bytes); two bits turned on for each pattern (where a pattern represents a unique path of atoms and bonds of a given length); and a maximum of five bonds for generating patterns. The pairwise similarities were calculated as the Tanimoto similarities between two fingerprints.[30]

Four of the fingerprint-based methods used in a comparative study of ranking methods[17] were selected for comparison against the SVM-based enrichment method. These methods are as follows: mean similarity of the test set compounds to the active compounds in the training set, S_A; the difference of mean similarities to the training set active and inactive compounds, S_{A-I}; the maximum similarity to a training set actives, S_{max}; and binary kernel discrimination (BKD).[16] These methods were implemented as described in the report of the comparative study.[17] An additional ranking function was generated by a “data fusion” method[33] where the test set SVM and BKD rankings of test set molecules were summed to generate a new score that ranks the test set molecules. These ranking methods were coded in the Java programming language. The BKD scoring of test set molecules was recoded in the C programming language in order to compare its speed with that of the SVM-based enrichment method.

Measures of Performance. To assess the performance of a particular virtual screening method, we determined the number of known active compounds that were retrieved in the top 2% and 10% of a ranked test set of compounds. Additional calculation of enrichment factors indicated the ratio of actives retrieved by the method relative to the expected number of actives in a randomly selected sample containing the same number of molecules (2% or 10% of the test set).[31] The enrichment factor, EF, is calculated using eq 10

$$\mathrm{EF} = \frac{\mathrm{Hits}_{\mathrm{sampled}}}{N_{\mathrm{sampled}}} \bigg/ \frac{\mathrm{Hits}_{\mathrm{total}}}{N_{\mathrm{total}}} \qquad (10)$$

where $\mathrm{Hits}_{\mathrm{sampled}}$ is the number of actives in the top-ranked sample of $\mathrm{N}_{\mathrm{sampled}}$ compounds, $\mathrm{Hits}_{\mathrm{total}}$ is the total number of actives, and N_{total} is the number of molecules in the test database.

Where the molecules were ranked using the SVM, we also calculated the modified enrichment factor, EF′, defined by Halgren et al.[32] as

$$\mathrm{EF} = \frac{50\%}{\mathrm{APR}_{\mathrm{sampled}}} \frac{\mathrm{Hits}_{\mathrm{sampled}}}{\mathrm{Hits}_{\mathrm{total}}} \qquad (11)$$

where $\mathrm{APR}_{\mathrm{sampled}}$ is the average percentage rank of the actives in the sample. We calculated the EF′ value for the retrieval of the top-ranked 18 of the 25 known actives (72% of actives) in a given test set, close to the 70% chosen by Halgren et

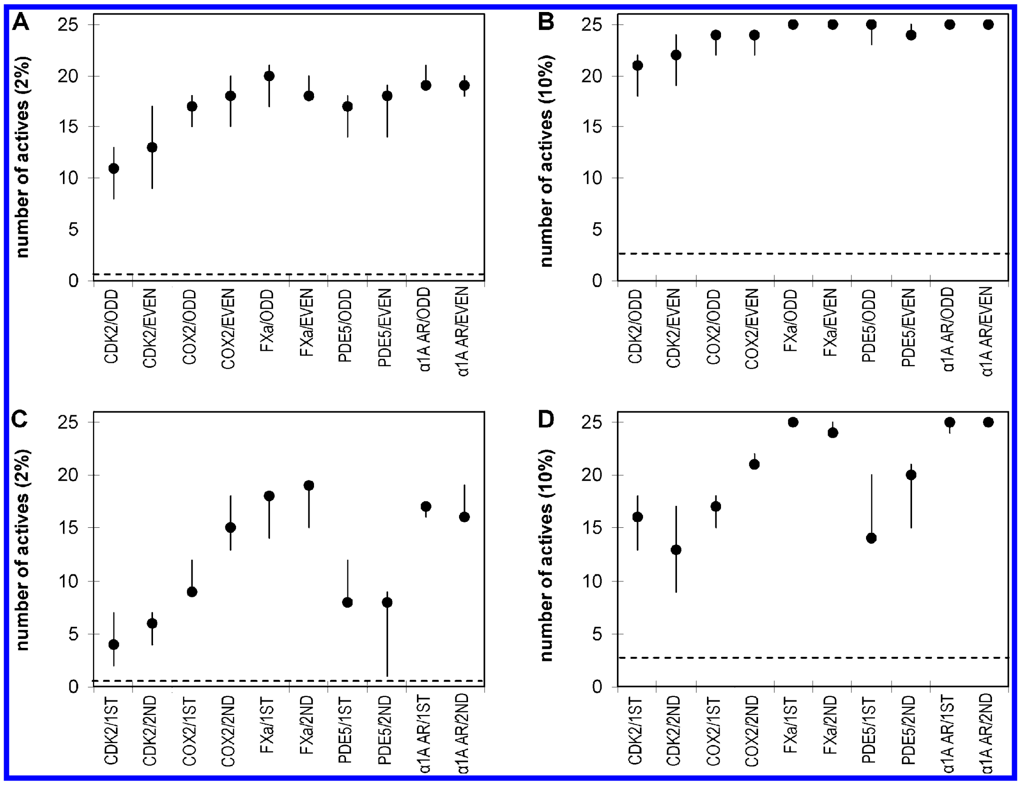

Figure 2. Retrieval of actives compounds from (A) the top-ranked 2% of the ODD and EVEN test sets of compounds, (B) the top-ranked 10% of the ODD and EVEN test sets, (C) the top-ranked 2% of the 1ST and 2ND test sets, and (D) the top-ranked 10% of the 1ST and 2ND test sets using SVM models that were trained using different parameters. The vertical lines span the minimum and maximum number of active compounds found in the top 2% or 10% of the test database for SVM models obtained using different values for the SVM parameters γ, $C+$, and $C-$. The filled circle indicates how many active compounds were retrieved using the SVM model with the lowest value of the AvRank statistic obtained from a scan of the SVM parameters. The best performance for AvRank occurs when the filled circle is placed at the top of the vertical line. The horizontal dashed line indicates the number of active compounds expected to be retrieved by random screening: 0.5 actives from 2% of the database and 2.5 actives from 10% of the database.

al.[32] The advantage of this EF′ statistic is that if two different sets of molecules' SVM activity scores give the same ranking for the 18th active, the modified enrichment factor will yield a higher value for the set of scores that rank the other 17 actives higher in one list than in the other; the conventional enrichment factor (eq 10) would not make this distinction.

We also use a statistic we call the relative performance (RP) to quantify how well each of the virtual screening methods performs over a number of test sets relative to the other methods:

$$\mathrm{RP}(j) = \frac{1}{N_i}\sum_i \frac{N_{\mathrm{actives}}(i,j)}{N_{\mathrm{actives}}(i)} \tag{12}$$

where

$$N_{\mathrm{actives}}(i) = \frac{1}{N_j}\sum_j N_{\mathrm{actives}}(i,j) \tag{13}$$

where the index i runs over the different test sets and the index j runs over the different ranking methods. Each term in the summation in eq 12 is the ratio of the number of actives retrieved from a test set (index i) using the specified virtual screening method (index j) divided by the number of actives retrieved from the same test averaged over all of the test set methods (eq 13). The best performing method has a relative performance greater than one, and the corresponding value for the worst-performing method will be less than one. Use of eq 12 instead of an unweighted average over the different test sets prevents the results from test sets with higher numbers of actives retrieved from dominating the number that is calculated.

RESULTS

We evaluated the SVM-based enrichment method by developing SVM models using training sets of active and inactive molecules and then using these models to rank the molecules in the corresponding test sets. Varying key parameters and data sets in the calculations provided insight about the factors that affect the performance of this virtual screening method.

Variation in the Performance of Candidate SVM Models. The first trials of our enrichment method explored its performance as a function of the SVM parameters. SVM models were trained using each of the combinations of the SVM parameters γ, $C+$, and $C-$, and these models were then applied to appropriate test sets. Figure 2 shows that there is generally some variation in the number of actives retrieved from the top-ranked 2% or 10% of the test sets. This variation tends to be lower when greater numbers of active compounds are retrieved from a given test set. All of the SVM models retrieve more active compounds than would be expected by a random sampling of compounds.

The AvRank statistic, described in the Methods section, represents an attempt to select optimal SVM parameters via cross-validation prior to applying the resultant model to test

Table 1. Results of the SVM-Based Enrichment Method[a]

	test	72% actives	2% database		10% database		test	72% actives	2% database		10% database	
actives	set	EF′	*n*	EF	*n*	EF	set	EF′	*n*	EF	*n*	EF
CDK2	ODD	17.0	11	22.4	21	8.4	1ST	8.4	4	8.2	16	6.4
CDK2	EVEN	26.5	13	26.5	22	8.8	2ND	5.4	6	12.2	13	5.2
COX2	ODD	37.1	17	34.7	24	9.6	1ST	9.3	9	18.4	17	6.8
COX2	EVEN	39.4	18	36.7	24	9.6	2ND	31.1	15	30.6	21	8.4
FXa	ODD	40.6	20	40.8	25	10.0	1ST	37.9	18	36.7	25	10.0
FXa	EVEN	27.8	18	36.7	25	10.0	2ND	39.7	19	38.8	24	9.6
PDE5	ODD	36.5	17	34.7	25	10.0	1ST	7.4	8	16.3	14	5.6
PDE5	EVEN	38.6	18	36.7	24	9.6	2ND	11.6	8	16.3	20	8.0
α_{1A}AR	ODD	40.1	19	38.8	25	10.0	1ST	38.3	17	34.7	25	10.0
α_{1A}AR	EVEN	38.6	19	38.8	25	10.0	2ND	33.4	16	32.6	25	10.0
max. values		40.6	21	42.8	25	10.0		40.6	21	42.8	25	10.0

[a] Fifty descriptors were used for these calculations. *n*: number of actives in sample. EF: enrichment factor (eq 10). EF′: modified enrichment factor (eq 11).

data. This statistic is calculated for each candidate SVM model obtained during a scan of the parameters γ, $C+$, and $C-$, and the parameters corresponding to the lowest value of AvRank are used to select a particular model. (The parameters selected using AvRank are listed in the Supporting Information.) With one exception, the SVM model with the lowest value of AvRank retrieved two or less fewer active compounds from the top-ranked 2% or 10% of the test compared to the best performing SVM models (Figure 2A,B). The lowest value of AvRank for the CDK2/ODD training set was for a model that retrieved 13 compounds from the top-ranked 2% of the CDK2/EVEN test set, midway between the minimum (9) and maximum (17) number of compounds retrieved by any of the SVM models. However, SVM models corresponding to AvRank values less than 2% higher than the lowest value retrieved 14−16 actives compounds from the same proportion of the test data set. Thus, for the ODD/EVEN split of molecules, the training-set derived AvRank statistic is a reasonably reliable predictor of performance for a test set.

The performance of the AvRank statistic is less impressive for the 1ST and 2ND data sets (Figure 2C,D). Use of this statistic avoided poorly performing models for the CDK2/2ND and PDE5/1ST training sets (CDK2/1ST and PDE5/ODD test sets, respectively) but did not help for some other training sets, notably COX2/2ND and PDE5/2ND (test sets COX2/1ST and PDE5/1ST, respectively).

Based on the results presented in this section, the remaining calculations in this report used the AvRank statistic to select the SVM parameters used in the training of SVM models.

Retrieval of Active Compounds against a Background of NCI Diversity Set Molecules. The SVM-based method provides substantial enrichment for the various test sets, as detailed in Table 1. At least 44% (11 molecules) of the 25 actives were found in the top 2% (21 molecules) of the test data sets when the ODD and EVEN data sets were used to train and test the SVM, and at least 84% (21 molecules) of the actives in a given test set were retrieved in the top 10% (107 molecules) of a given test set. The results are more varied for the 1ST and 2ND data sets, presumably because the active compounds in the training and test sets belong to nonoverlapping chemical classes, unlike the ODD and EVEN data sets. FXa inhibitors and α_{1A} AR antagonists were the easiest to retrieve from the 1ST and 2ND test sets, and the recoveries were almost as good as for the ODD and EVEN data sets. For the inhibitors of CDK2, COX2, and PDE5, the recoveries of active compounds from the 1ST and 2ND data sets test sets were lower than from the ODD and EVEN data sets (Table 1). Among these three sets of active compounds, the SVM models for the CDK2 inhibitors had the lowest enrichment values. In all cases, more than half of the active compounds were ranked in the top 10% of the test data sets. Thus, the SVM successfully retrieves test set active compounds which have chemistries different from the training set actives.

Table 2. Retrieval of Different Chemical Classes of Actives Using the SVM-Based Enrichment Method[a]

		number of chemical classes				number of chemical classes		
actives	test set	2% db[b]	10% db	100% db	test set	2% db	10% db	100% db
CDK2	ODD	6	9	9	EVEN	8	9	9
COX2	ODD	9	10	11	EVEN	10	11	11
FXa	ODD	9	11	11	EVEN	9	10	10
PDE5	ODD	6	8	8	EVEN	7	8	8
α_{1A} AR	ODD	10	11	11	EVEN	10	13	13
CDK2	1ST	3	4	5	2ND	3	3	4
COX2	1ST	4	4	4	2ND	5	6	7
FXa	1ST	6	6	6	2ND	5	5	5
PDE5	1ST	2	2	4	2ND	4	4	4
α_{1A} AR	1ST	6	6	6	2ND	6	7	7

[a] Fifty descriptors were used for these calculations. [b] X% db: top-ranked X% of the test database.

The high-ranking actives from a given test set themselves exhibit a range of different chemistries. Representatives of more than half of the different chemical classes of actives are found in the top-ranked 2% of the test set (Table 2) in 19 of the 20 cases. The exception was the 1ST/PDE5 test set from which seven of the eight PDE5 inhibitors in the top-ranked 2% of molecules belong to a single class of compounds.

It is also of interest to examine the background compounds in the test set that were highly ranked by the relevant SVM models. A few of the putative false positives retrieved from the CDK2 test sets are immediately recognizable as being similar to actives in the training set (Figure 3A); however, this very high level of similarity is not observed in the false positives when the active compounds are inhibitors of COX2, FXa, PDE5, or α_{1A} AR. Many of the high-ranking (assumed) inactive compounds exhibit a moderate degree of similarity to some of the training set compounds and have some

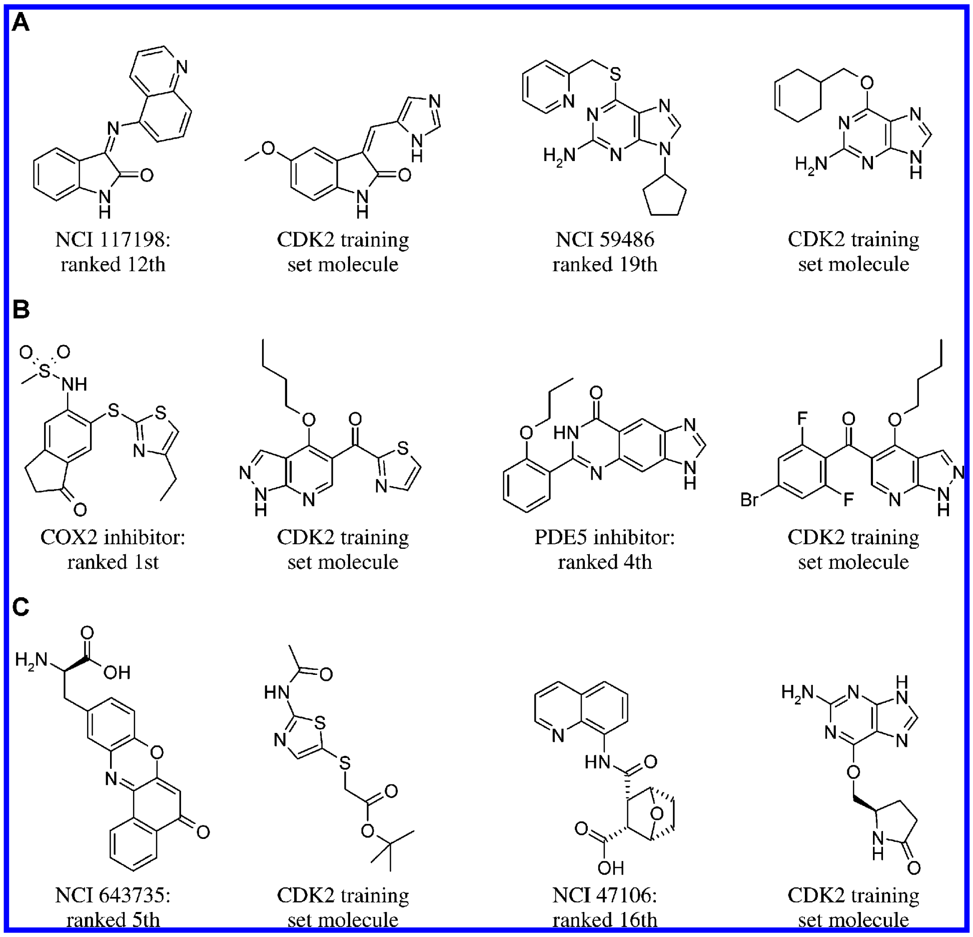
A
NCI 117198: ranked 12th
CDK2 training set molecule
NCI 59486 ranked 19th
CDK2 training set molecule
B
COX2 inhibitor: ranked 1st
CDK2 training set molecule
PDE5 inhibitor: ranked 4th
CDK2 training set molecule
C
NCI 643735: ranked 5th
CDK2 training set molecule
NCI 47106: ranked 16th
CDK2 training set molecule

Figure 3. Some molecules retrieved from the 1ST (test) set of molecules using an SVM model that was obtained after training with the 2ND data set of molecules with CDK2 inhibitors as the actives. The training set molecule considered to be the most similar to the test set molecule shown is also displayed. (A) Molecules from the 1ST data set that can be rationalized as having molecular features similar to those in one or more of the training set molecules. (B) Molecules from the 1ST data set with obviously similar molecules in the training set. (C) Molecules from the 1ST data set which do not have molecules that can be readily identified as having similar molecular features. Note that while the neutral forms of the molecules are shown, ionized forms that correspond to the expected protonation states at physiological pH were used in the calculations.

chemical features such as those that occur in some of the training set actives, as illustrated by the examples in Figure 3B. However, some of the highly ranked background compounds possess little obvious similarity to the training set actives (Figure 3C).

To examine the influence of changing the number of descriptors in the SVM models, we allowed the number of descriptors to vary in the scan of parameters during training runs where the 10 1ST and 2ND training sets were used. Thus, in the addition to varying the values of γ, $C+$, and $C-$, the number of descriptors was varied from 10 to 70 in increments of 10. As before, the combination of these parameters yielding the lowest value of AvRank was used to generate a final SVM model. The recoveries of active compounds from the top-ranked 2% and 10% of the various test sets using these models are similar to the corresponding recoveries using the models trained using 50 descriptors with one exception: the recovery of PDE5 inhibitors from the 1ST test set (Figure 4). The 50 descriptor SVM model selected by AvRank retrieved only 8 and 14 PDE5 inhibitors from the top-ranked 2% and 10%, respectively, of the 1ST test set and was the worst of the models with 50 descriptors (Figure 2). The 70 descriptor model obtained from the more extensive scan of parameters performed somewhat better, recovering 11 and 19 PDE5 actives from the top 2% and 10% of the 1ST test set. However, this was not as good as the best of the 50 descriptor models, which retrieved 12 and 20 actives, respectively, from the same percentages of the test set. Based on these observations, and the results from the other nine test sets, there appears to be little benefit from allowing the number of descriptors to vary during the search for optimal parameters.

Descriptor Usage. Although each set of 50 descriptors automatically selected for the various training sets was unique, some descriptors were more frequently selected than others, and some were never selected. An average of 11 descriptors was selected for all four training sets (ODD, EVEN, 1ST, and 2ND) for a given class of actives (inhibitors of CDK2, COX2, FXa, PDE5, or α_{1A} AR). Furthermore, certain descriptors were more frequently chosen than others across all training sets. Ten DRAGON descriptors were each selected for more than half of the 20 training sets and for at least one of all of the five classes of actives: ZM1V (first Zagreb index by valence vertex degrees); X5A (average connectivity index, chi-5); BEHm2 and BEHm4 (second- and fourth-highest eigenvalues from a mass-weighted Burden matrix/weighted by atomic masses); ATS5m, ATS6m, ATS7m, and ATS8m (Broto-Moreau autocorrelation − lags 5, 6, 7, and 8, respectively/weighted by atomic masses; and ATS5e and ATS8e (Broto-Moreau autocorrelation − lags 5

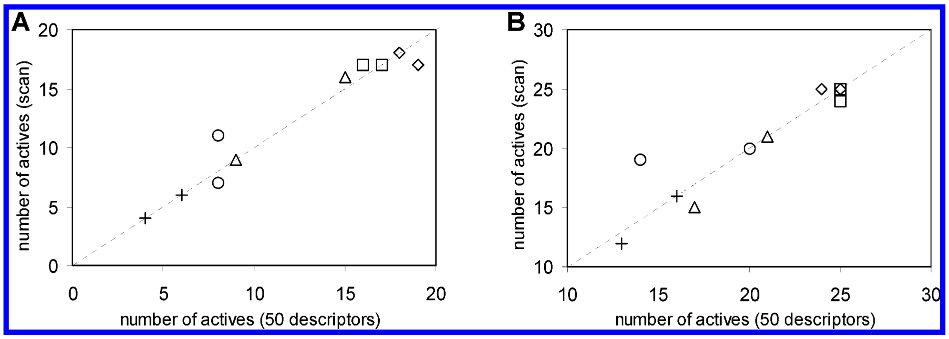

Figure 4. Number of actives from the top (A) 2% and (B) 10% of the 1ST and 2ND test sets retrieved using SVM models in which the number of descriptors was allowed to vary during the training run versus the number of actives retrieved using SVM models in which the number of descriptors was fixed at a value of 50. The active compounds for the training and testing procedures are indicated as follows: CDK2 inhibitors − plus signs; COX2 inhibitors − triangles; FXa inhibitors − diamonds; PDE5 inhibitors − circles; α_{1A}AR antagonists − squares. The outlying points in the graphs both correspond to the retrieval of actives from the PDE5/1ST test set.

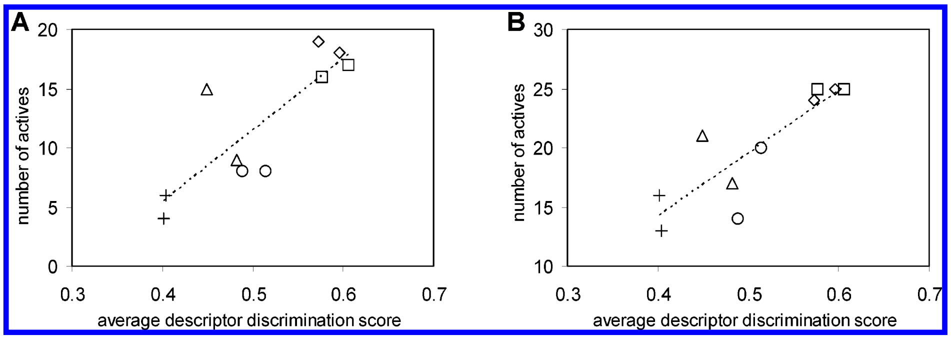

Figure 5. Number of actives retrieved from (A) the top-ranked 2% and (B) the top-ranked 10% of test sets versus the average descriptor score of the corresponding training sets. This evaluation was performed for various 1ST and 2ND data sets. The data points are indicated in the same way as for Figure 4. The fitted lines have R^2 values of (A) 0.69 and (B) 0.74. When one outlier point was removed from the graph in (A), the R^2 value of the best-fit line increased to 0.88.

and 8, respectively/weighted by Sanderson electronegativities). Interestingly, many of the descriptors are weighted by mass (BEHm2, BEHm4, ATS5m, ATS6m, ATS7m, and ATS8m) and/or are autocorrelation descriptors (ATS5m, ATS6m, ATS7m, ATS8m, ATS5e, and ATS8e).

Figure 5 shows graphs of the number of actives retrieved by a given SVM model (trained using 50 descriptors and one of the 1ST and 2ND training sets) plotted against the average of the descriptor discrimination scores from the training step. The graphs show a correlation between the performance of the SVM-based enrichment method and the average descriptor discrimination scores, especially when the top-ranked 2% of the test molecules are examined. This suggests that training sets whose descriptors have high discrimination scores are expected to lead to a better retrieval of actives from a test set than training sets whose descriptors have lower discrimination scores. In fact, the two sets of actives with the highest descriptor discrimination score, FXa inhibitors and α_{1A} AR antagonists, also have the highest enrichments (Figure 5).

Retrieval of Active Compounds from a Larger Background Set. We performed additional calculations using the SVM-based enrichment method to see if the good levels of enrichment obtained using the NCI diversity set are preserved when a larger background set is used. Calculations were performed using training and test sets based on the 1ST and 2ND splits of molecules where the NCI diversity set molecules were replaced by molecules from the Maybridge database. The test sets each contain 25 300 molecules, 25 of which are active in any given calculation. The training sets contained either 5160 molecules (125 inhibitors, 25 of which are active in a given run, plus 5035 Maybridge compounds for the 5−1 and 5−2 data sets) or 1132 molecules (125 inhibitors plus 1007 Maybridge compounds for the 25−1 and 25−2 data sets). In most cases, there was an increase in the number of actives retrieved from the top 2% of the test sets containing Maybridge molecules relative to the NCI diversity set-containing test set (Figure 6A). For the top-ranked 10% of compounds, the levels of enrichment from the data sets containing Maybridge compounds were similar to the corresponding data sets in which the background molecules were taken from the NCI diversity set (Figure 6B). Thus, the level of enrichment provided by the method does not appear to be greatly influenced by the number or composition of background compounds in the test set. Figure 6 also indicates that there was little benefit from training with one of the larger data sets (the 5−1 and 5−2 sets) relative to using the smaller training sets (the 25−1 and 2−5 sets) which are approximately one-fifth the size of the larger sets, and hence less computationally demanding.

CPU Requirements. Using a computer with a 2.5 GHz Pentium 4 CPU running the Windows XP operating system, the time required to train the SVM function for the data sets containing the background NCI diversity set molecules ranged from 1.5 to 10 min when the best 50 descriptors were used in the training. Of this time, 5 to 18 s was spent selecting the descriptors. When the number of descriptors was varied during the training process, the calculations generally took 7−8 times longer to execute. The training times for the various 5−1 and 5−2 training sets ranged between 15 and 49 min, whereas the training times were usually 8- to 10-fold quicker for the corresponding 25−1

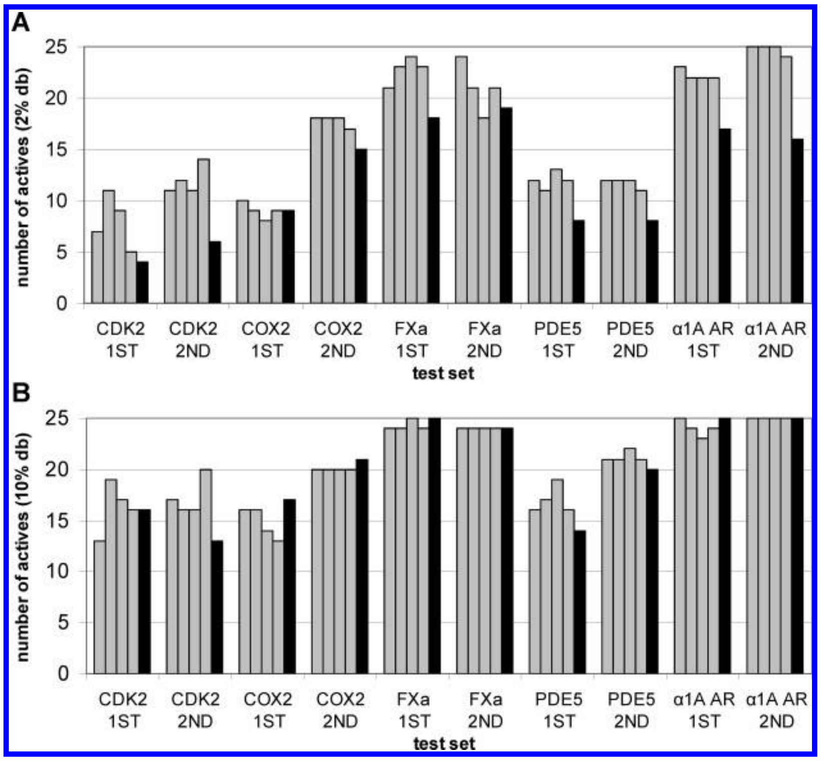

Figure 6. Number of active molecules retrieved in (A) the top-ranked 2% and (B) the top-ranked 10% of test database using the SVM-based enrichment method where the background molecules were from the Maybridge database (25 300 molecules in each test set, including 125 known inhibitors of which 25 were assigned as actives). The first four vertical bars correspond to models trained on the 5−1, 5−2, 25−1, and 25−2 types of training set, respectively. For reference, the results from using the NCI diversity background for both training and testing are colored black. As before, the expected numbers of active compounds obtained from random screening are 0.5 and 2.5 for the top-ranked 2% and 10% of the database, respectively.

Table 3. Comparison of SVM-Based Enrichment Method with Fingerprint-Based Ranking Methods for ODD/EVEN Data Sets

		actives in top ranked 2%						actives in top ranked 10%					
test set		SVM	S_A	S_{A-I}	S_{max}	BKD	SVM+ BKD[b]	SVM	S_A	S_{A-I}	S_{max}	BKD	SVM+ BKD[b]
CDK2	ODD	11	8	11	19	20	17	21	19	20	23	24	24
CDK2	EVEN	13	9	12	16	18	18	22	24	21	24	24	23
COX2	ODD	17	13	17	17	18	19	24	18	18	21	24	25
COX2	EVEN	18	10	13	18	20	20	24	20	19	21	23	25
FXa	ODD	20	12	14	16	19	21	25	22	21	24	25	25
FXa	EVEN	18	16	16	20	20	21	25	24	24	24	25	25
PDE5	ODD	17	15	16	17	20	20	25	23	23	25	25	25
PDE5	EVEN	18	16	16	16	21	21	24	20	20	25	25	25
α_{1A}AR	ODD	19	9	14	12	16	20	25	18	18	21	22	25
α_{1A}AR	EVEN	19	10	14	14	17	19	25	19	18	22	23	25
standard deviation		2.83	3.05	1.95	2.32	1.60	1.35	1.41	2.36	2.10	1.63	1.05	0.67
rel. performance[a]		1.08	0.74	0.91	1.06	1.21	[1.25][c]	1.08	0.92	0.90	1.03	1.07	[1.11][c]

[a] Calculated using eqs 12 and 13, except where noted. [b] Column of figures excluded from the calculation of the performance scores for the other ranking methods. [c] Modified performance score, whose calculation is described in the text.

and 25−2 training sets. For 5−1 and 5−2 training sets, the time to select the best 50 descriptors was approximately two minutes but was only 6−7 s for the 25−1 and 25−2 training sets. After the training procedure, the time needed to rank 25 300 molecules using a given SVM model was less than 15 s.

Comparison of SVM-Based Enrichment with Fingerprint-Based Ranking Methods. Using the ODD, EVEN, 1ST, and 2ND training and test sets, the SVM-based enrichment method was compared against four fingerprint-based ranking methods, S_A, S_{A-I}, S_{max}, and Binary Kernel Discrimination (BKD),[16] that were used in a previous comparison of ranking methods[17] (Tables 3 and 4). When applied to the ODD and EVEN training and test sets, BKD was the most effective in retrieving active compounds, followed by the SVM and S_{max} methods. However, the SVM method performed noticeably better than the other four methods when applied to the 1ST and 2ND data sets (Table 4), especially when the active compounds are inhibitors of FXa or α_{1A} AR. The performance of the S_{max} method is significantly worse for the 1ST and 2ND test sets relative to the ODD and EVEN sets. The relative performance measures for the S_A and S_{A-I} methods indicate that these approaches are not as effective as the other ranking methods.

We also combined the results of the SVM and BKD using a “data fusion” technique in which a new ranking function is generated by summing the SVM and BKD ranks of each molecule in the test database.[33] Application of this hybrid method to each test set recovered more active molecules than one or both of the SVM and BKD schemes (Tables 3 and

Table 4. Comparison of SVM-Based Enrichment Method with Fingerprint-Based Ranking Methods for 1ST/2ND Data Sets

		actives in top ranked 2%						actives in top ranked 10%					
test set		SVM	S_A	S_{A-I}	S_{max}	BKD	SVM+ BKD[b]	SVM	S_A	S_{A-I}	S_{max}	BKD	SVM+ BKD[b]
CDK2	1ST	4	3	3	9	5	5	16	13	11	15	13	19
CDK2	2ND	6	5	2	3	4	7	13	10	10	10	16	19
COX2	1ST	9	5	10	5	6	10	17	17	17	12	16	21
COX2	2ND	15	5	11	3	10	14	21	12	15	12	16	19
FXa	1ST	18	5	8	9	7	15	25	20	20	17	22	25
FXa	2ND	19	8	12	8	10	16	24	19	19	14	20	25
PDE5	1ST	8	9	11	3	6	10	14	19	19	18	18	18
PDE5	2ND	8	10	6	2	3	6	20	20	17	14	11	17
α_{1A}AR	1ST	17	6	9	3	8	17	25	16	18	12	21	25
α_{1A}AR	2ND	16	6	6	3	4	13	25	15	15	14	15	23
standard deviation		5.54	2.15	3.46	2.78	2.45	4.32	4.74	3.54	3.38	2.44	3.49	3.14
rel. performance[a]		1.56	0.89	1.01	0.69	0.85	[1.50][c]	1.20	0.97	0.97	0.84	1.02	[1.29][c]

[a] Calculated using eqs 12 and 13, except where noted. [b] Column of figures excluded from the calculation of the performance scores for the other ranking methods. [c] Modified performance score, whose calculation is described in the text.

4). To calculate an appropriate relative performance measure for this combined ranking method, the values for eq 13 (which specifies how to calculate weights in the weighted average of eq 12) were taken from the results of the five “unfused” ranking methods. The modified relative performance measures for the SVM+BKD ranking scheme usually exceeded the corresponding measures for the SVM and BKD, as shown in Tables 3 and 4.

The time taken to calculate scores for 25 300 test molecules using our implementation of the BKD method was approximately two minutes. This is somewhat slower than the than 15 s required to calculate scores for the same number of molecules using SVM-based enrichment.

DISCUSSION

The SVM-based method successfully enriches all of the test databases for active compounds, even when the actives in the test database are of different chemistries from those in the corresponding training set. The method recovers at least half of the active compounds in the top-ranked 10% of a test set of compounds, for more than 5-fold enrichment, and the recovery of active compounds from the top-ranked 2% of the test sets is 8−40 times better than that expected from random screening. The levels of enrichment are roughly independent of the size and composition of the inactive background molecules, as judged by the results of replacing the NCI diversity set background molecules with those from the much larger Maybridge database.

Several factors appear to affect the performance of the present method. One is the composition of the actives in the training and test sets. The retrieval of active compounds is generally higher for the various ODD and EVEN test sets than for the 1ST and 2ND data sets, presumably because of the higher degree of similarity between active compounds in the former two data sets compared to the latter two sets. Application of the method to the 1ST and 2ND splits represents an interesting challenge for the method since the test sets contain active molecules that are generally different in chemical class to the training set actives. Thus, the ability to enrich these types of data sets is one of the strengths of the method. Additionally, the actives that are retrieved in the top-ranked 2% or 10% of 9 of the 10 1ST and 2ND test sets span most of the different chemical classes present, instead of only one or two chemical classes (Table 2). This result suggests that the method can be used to discover a novel class of active compounds.

All of the SVM models trained using different parameters provide a measure of enrichment, although the enrichments vary from model to model (Figure 2). The ability of the cross-validation AvRank statistic to select among these models therefore affects the performance of the method. Encouragingly, the AvRank statistic usually selected one of the better performing models for the ODD and EVEN data sets (Figure 2A,B). This can be rationalized by noting that for these data sets, the active molecules in a given training set are representative of the types of active molecules present in the corresponding test set. The ability of the AvRank statistic to choose among the candidate SVM models was somewhat mixed for the 1ST and 2ND data sets (Figure 2C,D), presumably because the training set actives are less representative of the test set actives. Previous QSAR studies have shown a similar lack of correspondence between the predicted performance of a model, as indicated by the value of a cross-validation statistic obtained from the training set, and the performance of the model when applied to test set data.[34−36]

A third factor affecting the performance of the SVM-based enrichment method is the choice of descriptors. The automated selection of descriptors contrasts with a number of other chemistry papers in which the descriptors for the SVM appear to have been selected by hand.[8−11] In the present method, descriptors are selected for their ability to cluster the training set actives into a small region of descriptor space relative to the space spanned by the inactive compounds, as indicated by their discrimination scores. Thus, the preferred descriptors focus on properties that are common to the training set actives and distinguish them from the inactive molecules. Unsurprisingly, SVM models that were derived using descriptors with higher discrimination scores were generally able to retrieve more actives than models derived from descriptors with smaller discrimination scores. Thus, one avenue to improving the performance of the present method will be to find descriptors that are better able to capture the underlying common features of the training set active molecules.

The higher retrieval of active compounds from the 1ST and 2ND test sets using the SVM-based enrichment method, compared to several fingerprint-based ranking methods,

Table 5. Average Pairwise Similarities of Test Set Active Compounds to Training Set Active Compounds

split of compounds	targets of active compounds				
	CDK2	COX2	FXa	PDE5	α_{1A} AR
ODD/EVEN	0.45	0.40	0.45	0.50	0.44
1ST/2ND	0.42	0.37	0.44	0.47	0.43

further demonstrates the capacity of the SVM method for retrieving active compounds whose chemical classes are different to those contained in the training set. The differences in the representation (descriptors versus binary hashed fingerprints) of the molecules in the two types of methods may be an important factor in the performance of these methods in several respects. First, the set of molecular descriptors chosen for use by the SVM is system-specific. This provides a measure of flexibility to the representation of the molecules which is not afforded by binary fingerprints, where the encoding of molecular features is invariant with respect to the training set. Second, the molecular descriptors used here generally quantify aspects of the chemical structure of each molecule in its entirety, whereas the binary hashed fingerprints used in this work, whose maximum path length is five for the patterns coded, represent local features. Descriptors which encode nonlocal features may capture information about the types of groups which interact with the protein target (e.g. hydrogen bond acceptor, hydrogen bond donor, aliphatic, and aromatic interactions), even as the functional groups responsible for these interactions vary across the different chemical classes of a given set of inhibitors.

Interestingly, the variability in the number of active compounds retrieved from the 1ST and 2ND test sets using the SVM is greater that obtained using any of the fingerprint-based methods. The trends in the number of active compounds retrieved from these test sets using the SVM can be accounted for by noting their correlation with the corresponding training set descriptor discrimination scores (Figure 5). The lack of variation in the retrieval of actives by the fingerprint-based methods can be rationalized by noting the lack of variability of the average pairwise similarities of the training and test set actives across the different activity classes (inhibitors of CDK2, COX2, etc.) (Table 5).

Binary kernel discrimination (BKD) was the only fingerprint-based method tested that outperformed the SVM for the retrieval of active compounds from the ODD and EVEN test sets. BKD was also the best of the fingerprint-based methods in a previous comparative study of these and several other ranking methods.[17] The S_{max} method, which performed nearly as well as the SVM for the ODD and EVEN test sets, performed poorly for the 1ST and 2ND test sets. This can be understood by noting that the 1ST and 2ND test sets were constructed to exclude active compounds with chemistries similar to any of those of the active compounds in the corresponding training sets (2ND and 1ST, respectively), as illustrated in Table 6. The S_A and S_{A-I} methods were not among the best ranking methods in this work or in the previous study.[17]

The differences between the SVM and the various fingerprint-based methods suggest that these approaches to ranking molecules are complementary. This idea is supported by the performance of the combined SVM and BKD score which is generated by summing the SVM and BKD rankings of each molecule.[33] This combined ranking function compensates for instances where one of the methods did not perform particularly well, always performing better than at least one of SVM and BKD, and sometimes outperforming both methods.

Table 6. Average of the Maximum Pairwise Similarities between Each Test Set Active Compound with Any of the Training Set Actives

split of compounds (train/test)	targets of active compounds				
	CDK2	COX2	FXa	PDE5	α_{1A} AR
ODD/EVEN	0.78	0.73	0.75	0.82	0.70
EVEN/ODD	0.78	0.72	0.75	0.82	0.68
1ST/2ND	0.55	0.54	0.57	0.57	0.58
2ND/1ST	0.56	0.52	0.60	0.58	0.59

SUMMARY

This paper describes a novel method of applying the Support Vector Machine to the problem of enriching a database of molecules for active molecules. The SVM model generates substantial enrichment of active molecules with chemistries different from those in the training set. Comparison of the SVM-based enrichment method with ranking methods that use binary hashed fingerprints show that the SVM method is the best at finding active compounds which are chemically distinct from known actives. Interestingly, the best method tested was a hybrid of the SVM and the fingerprint-based method, Binary Kernel Discrimination. After training with known active and inactive molecules, the SVM and hybrid methods can be used to rapidly rank more than 10 000 compounds per minute.

ACKNOWLEDGMENT

We thank our colleagues Yuhmei Lin for assistance using the JChem package and Visvaldas Kairys for his comments on this manuscript. We also thank Chih-Chung Chang and Chih-Jen Lin for providing their libsvm package to the rest of the scientific community. This publication was made possible by Grants No. GM66524 and GM62050 from the National Institute of General Medical Sciences (NIGMS) of the NIH. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the NIGMS.

Supporting Information Available: Inhibitors of CDK2, COX2, FXa, PDE5, and α_{1A} AR as placed into the perceived chemical classes and the values of the γ, $C+$, and $C-$ parameters for the SVM models developed in this work. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Bajorath, J. Integration of virtual and high throughput screening. *Nat. Rev. Drug Discov.* **2002**, *1*, 882−894.

(2) Bajorath, J. Virtual screening: methods, expectations, and reality. *Curr. Drug Discovery* **2002**, *2*, 24−28.

(3) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discovery Today* **2002**, *7*, 903−911.

(4) Oprea, T. I.; Davis, A. M.; Teague, S. J.; Leeson, P. D. Is there a difference between leads and drugs? A historical perspective. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1308−1315.

(5) Oprea, T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 325−334.

(6) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998.

(7) Burges, C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowl. Discovery* **1998**, *2*, 121−167.

(8) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem*. **2001**, *26*, 5−14.

(9) Warmuth, M. K.; Liao, J.; Ratsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 667−673.

(10) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of linear and nonlinear classification algorithms for the prediction of drug and chemical metabolism by human UDP-glucuronosyltransferase isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019−2024.

(11) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048−2056.

(12) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines, 2001, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

(13) Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization; Microsoft Research Technical Report MSR-TR-98-14; 1998.

(14) Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.* **2001**, *13*, 637−649.

(15) Cortes, C.; Vapnik, V. Support vector networks. *Machine Learning* **1995**, *20*, 273−297.

(16) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1295−1300.

(17) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 469−474.

(18) Sindhwani, V.; Rakshit, S.; Deodhar, D.; Erdogmus, D.; Principe, J. C.; Niyogi, P. Feature Selection in MLPs and SVMs based on Maximum Output Information. *IEEE Trans. Neural Netw*. **2004**, *15*, 937−948.

(19) Guyon, I.; Weston, W.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389−422.

(20) Golub, T. R.; Slonim, D. K.; Tamayo, P.; Huard, C.; Gaasenbeek, M.; Mesirov, J. P.; Coller, H.; Loh, M. L.; Downing, J. R.; Caligiuri, M. A.; Bloomfield, C. D.; Lander, E. S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**, *286*, 531−537.

(21) IsisDraw 2.4, MDL Information Systems, Inc., 2001.

(22) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J.. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244−255.

(23) http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp.

(24) http://dtp.nci.nih.gov/docs/3d_database/Structural_information/structural_data.html.

(25) http://www.maybridge.com.

(26) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev*. **1993**, *93*, 2567−2581.

(27) Walters, P.; Stahl, M. http://smog.com/chem/babel.

(28) Todeschini, R.; Consonni, V.; Pavan, M. DRAGON 2.1. Milano Chemometrics and QSAR Research Group: Milan, Italy, 2002.

(29) Csizmadia F. JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 323−324.

(30) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 983−996.

(31) Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J. Med. Chem*. **2001**, *44*, 502−511.

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

(33) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 23−37.

(34) Kubinyi, H.; Hamprecht, F. A.; Mietzner, T. Three-dimensional quantitative similarity-activity relationships (3D QSiAR) from SEAL similarity matrices. *J. Med. Chem*. **1998**, *41*, 2553−2564.

(35) Golbraikh A.; Tropsha A. Beware of q2! *J. Mol. Graphics Mod*. **2002**, *20*, 269−276.

(36) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y. D.; Lee, K. H.; Tropsha, A. Rational selection of training and test sets for the development of validated QSAR models. *J. Comput.-Aided Mol. Des*. **2003**, *17*, 241−253.

CI049641U