Journal of Computer-Aided Molecular Design, **13:** 533–545, 1999.
KLUWER/ESCOM
© 1999 *Kluwer Academic Publishers. Printed in the Netherlands.*

Investigation of classification methods for the prediction of activity in diverse chemical libraries

Steven L. Dixon* & Hugo O. Villar
Telik Inc., 750 Gateway Boulevard, South San Francisco, CA 94080, U.S.A.

Received 20 October 1998; Accepted 22 December 1998

Key words: chemoinformatics, classification, cluster analysis, discriminant analysis, recursive partitioning, topological descriptors, 2D structural keys

Summary

Classification methods based on linear discriminant analysis, recursive partitioning, and hierarchical agglomerative clustering are examined for their ability to separate active and inactive compounds in a diverse chemical database. Topology-based descriptions of chemical structure from the Molconn-X and ISIS programs are used in conjunction with these classification techniques to identify ACE inhibitors, β-adrenergic antagonists, and H_2 receptor antagonists. Overall, discriminant analysis misclassifies the smallest number of active compounds, while recursive partitioning yields the lowest rate of misclassification among inactives. Binary structural keys from the ISIS package are found to generally outperform the whole-molecule Molconn-X descriptors, especially for identification of inactive compounds. For all targets and classification methods, sensitivity toward active compounds is increased by making repetitive classifications using training sets that contain equal numbers of actives and inactives. These balanced training sets provide an average numerical class membership score which may be used to select subsets of compounds that are enriched in actives.

Introduction

Combinatorial chemistry and high-throughput screening (HTS) now serve as a standard paradigm in modern approaches to lead generation. Automated methods of synthesis yield chemical libraries that may contain tens of thousands of compounds, and robotic screening techniques allow these compounds to be assayed for in vitro potency in a matter of only hours or days. However, the availability of compounds and the means to rapidly screen them does not necessarily imply that huge chemical libraries should actually be created for each new target of interest. Indeed, the process of lead generation is becoming increasingly reliant upon the use of virtual libraries to guide efforts in combinatorial synthesis and screening [1]. Statistical data analysis techniques have proven to be powerful tools in this regard, and they are in widespread use for the selection of compounds which satisfactorily span any desired portion of chemical space.

If little or no structure–activity information is available, then design efforts may simply be aimed at maximizing the overall diversity of a screening library. The premise behind this approach is that the chances of finding an active compound are increased by considering the broadest possible range of chemical functionality. Diversity-based approaches to lead discovery have enjoyed a great deal of success, so it is not surprising that there has been a steady stream of literature in the field of chemical/molecular diversity [2–7].

Despite its increasing popularity as a tool in library design, chemical diversity still remains an elusive concept. Generally speaking, diversity is defined only in vague terms as the natural counterpart to similarity [8–11]. In this sense, chemical diversity approaches are ultimately grounded in the theory and framework of chemical similarity. Consequently, the evaluation of tools for grouping together chemicals with simi-

*To whom correspondence should be addressed. E-mail: sdixon@telik.com

lar pharmacological endpoints is relevant to the issue of chemical diversity and to the design of screening libraries. By using similarity to build certain biases into a virtual library, it is possible to increase the chances of finding compounds with any set of desired characteristics.

This paper is concerned with the use of similarity measures and chemoinformatic classification techniques for the separation of active and inactive compounds in diverse chemical libraries. A primary goal is to identify factors that are inherently important in maximizing sensitivity toward active compounds and in minimizing the number of incorrect predictions for inactives, all irrespective of the target being analyzed. These factors would clearly represent powerful tools in library design and lead optimization.

We note that the nature of modern chemical libraries poses a significant challenge to classification methodologies. Any useful approach must be able to distinguish actives from inactives, wherein both groups of compounds may contain a broad range of structural diversity. This differs from most classical structure–activity studies, where the data set is typically a series of analogs with a limited variety of substituents at a few specific positions. Thus, in moving from classical structure–activity data sets to large chemical libraries, the focus changes from determining how variations on a parent compound affect activity, to finding significant statistical differences between structurally diverse populations of actives and inactives.

Here we compare a variety of classification methods which involve linear discriminant analysis [12–14], recursive partitioning [15–17], and an adapted version of hierarchical agglomerative clustering [18]. These techniques are applied to the classification of angiotensin-converting enzyme inhibitors, β-adrenergic antagonists, and H_2 receptor antagonists, all of which appear in the MDL CMC database [19]. Classification studies are carried out using standard topological descriptors from the Molconn-X program [20], and binary structural keys from the ISIS package [21].

Classification methods

Standard approaches to predicting class membership typically involve the development of a single model that is able to satisfactorily classify a set of training compounds, followed by application of the model to a desired set of external prediction compounds. Under these circumstances, it is clear that the selection of an appropriate training set is critical to the accuracy of the predictions. The task can be particularly challenging when the model is intended for use on the wide range of structures typically found in large chemical libraries.

When classifying compounds on the basis of activity toward a particular target receptor or enzyme, it is important to recognize that the space of inactive compounds is generally much larger than the space of active compounds. It follows that any set of structures designed to represent the two populations as a whole should contain far more inactives than actives. Unfortunately, the use of such *unbalanced* training sets can sometimes lead to unbalanced predictions, especially when the classification algorithm is geared toward placing as many compounds as possible in the correct class. For a training set with a large ratio of inactives to actives, a high degree of accuracy is automatically achieved if the algorithm returns inactive status for all the compounds. As a result, models derived from such training sets may be prone to predictions that exhibit a higher percentage of misclassifications among actives than inactives and hence a low relative degree of sensitivity toward active compounds.

In order to improve sensitivity, the single unbalanced training set may be replaced with a large number of balanced training sets containing equal numbers of actives and inactives. The inactive compounds are selected randomly from a large population, and a faithful representation of the inactive space is obtained by generating a sufficiently large number of training sets. Prediction set compounds are then classified repeatedly, receiving in each trial a numerical classification score on the interval $[-1,1]$, where the endpoints correspond to maximum confidence in a prediction. An ‘average’ predicted class membership is obtained for each compound simply by taking the mean of the classification scores generated by the various balanced training sets. Here we investigate the use of both balanced and unbalanced training sets for class predictions on external compounds.

In the descriptions that follow, the symbol ‘$+$’ will denote active compounds and ‘$-$’ will denote inactive compounds. The variables $x_1, \ldots, x_m$ represent the set of numerical structural indices that are used in the various classification models.

Linear discriminant analysis

Each compound is classified in terms of a scalar value generated from a linear function of the structural descriptors [12–14]:

$$\mathrm{Discr}(i) = a_0 + a_1 x_1(i) + \cdots + a_m x_m(i) \qquad (1)$$

The discriminant parameters $a_0, a_1, \ldots, a_m$ may be defined so that compound *i* is considered to be active if Discr(*i*) $>$ 0 and inactive if Discr(*i*) $<$ 0. There are any number of ways to determine a reasonable set of parameters, and the actual values will depend on what sort of optimality criterion is used for classification. In the present work, we use the ***discr*** function from S-PLUS [22], which maximizes the ratio of the inter-class sum of squared distances to the intra-class sum of squared distances, with the discriminant function providing the distances.

The S-PLUS routine does not return a value for the intercept parameter a_0, so the discriminant function one must work with is actually given by

$$\boldsymbol{discr}(i) = a_1 x_1(i) + \cdots + a_m x_m(i) \qquad (2)$$

In order to achieve the algebraic $+/-$ designation, we define a_0 according to

$$a_0 = -1/2(\mu_+ + \mu_-) \qquad (3)$$

where μ_+ and μ_- are the active and inactive means for the training set compounds as calculated by Equation 2.

Because $a_0, a_1, \ldots, a_m$ are sensitive to the training set used, the range of values produced by Equation 1 can vary quite dramatically. Results, of course, depend on the number of descriptors used and the magnitudes of their values, but, for example, with models containing five variables, we have observed multiple balanced training sets to produce ranges in the discriminant function as low as 5 and as high as 65. Thus, when classifications are done in this manner, it is important to assign equivalent weights to each set of predictions. Accordingly, we define a sigmoidal transformation that maps the unbounded output of Equation 1 to the bounded interval ($-$1,1):

$$\mathrm{score}(i) = 2/\{1 + \exp[2\,\mathrm{Log}(3)\mathrm{Discr}(i)/(\mu_- - \mu_+)]\} - 1 \qquad (4)$$

This scoring function is designed to yield values of ±1/2 at the active and inactive means $\pm 1/2(\mu_+ - \mu_-)$ associated with Equation 1.

Recursive partitioning

Unlike discriminant analysis, where a single hyperplane divides descriptor space into exactly two regions, recursive partitioning [15–17] uses a decision tree to create a series of descriptor planes. These planes or partitions filter the compounds into cohesive blocks which contain as little inter-class mixing as possible.

At the base of the tree, all the compounds are mixed together in a single group. Then a node or branching point is defined by selecting a descriptor and a corresponding threshold at which to divide the data set. Compounds for which the descriptor value is above the threshold are assigned to one branch, and compounds below the threshold are assigned to the other branch. The choice of descriptor and threshold is designed to best ‘purify’ the mixture. In statistical terms, the split is made so as to minimize the sum of the squared deviances in the left (*L*) and right (*R*) branches [13]:

$$\mathrm{Minimize}\{\Sigma_{i\in L}(y_i - \mu_L)^2 + \Sigma_{i\in R}(y_i - \mu_R)^2\} \qquad (5)$$

For the active versus inactive case, y_i is assigned a value of 1 or 0, respectively, and the means μ_L and μ_R are simply the fraction of active compounds in each of the branches.

The partitioning procedure is repeated to form additional branches, with a new splitting criterion at each node. A single descriptor may be used more than once in the construction of the tree, but a unique threshold is selected each time. Partitioning continues until no further reduction in the deviance is possible. If perfect separation is achieved, the resulting tree will terminate in leaves consisting of groups of purely active or inactive compounds.

Once a tree has been grown using the training compounds, the same set of decisions may be used to assign external compounds into active or inactive leaves. Since it is rarely the case that all the leaves in a tree are pure, classification must be done in terms of probability. Accordingly, an active leaf is one which is more enriched in active compounds than the training set as a whole, and an inactive leaf is one which is more enriched in inactive compounds. When a balanced training set is used, the composition need only be greater than 50% in either direction in order to assign the leaf a class membership.

As indicated previously, when the class membership is computed from an average over multiple balanced training sets, it is desirable to define a scoring function which assigns the class membership as a number between -1 and $+1$. If prediction set compound i is assigned to leaf k, then we define its score as

$$\text{score}(i) = (n_{k+} - n_{k-})/(n_{k+} + n_{k-}) \tag{6}$$

where n_{k+} and n_{k-} are the numbers of active and inactive training set compounds contained in leaf k. With unbalanced training sets, no averaging is done for prediction set classifications, so a numerical score is not needed. In this case, compounds are classified simply on the basis of whether actives or inactives are over-represented in the leaf.

Hierarchical agglomerative clustering

Cluster analysis [18] is primarily a tool for unsupervised learning, and hence it is not specifically designed to yield predictive models. However, it is possible to capitalize on statistical similarities among active compounds to arrive at a cluster-based classification scheme. The method we describe here is appropriate only for balanced training sets, so a 50–50 mixture and repetitive classifications are assumed.

First of all, we employ hierarchical, agglomerative clustering (HAC) due to its demonstrated ability to uncover actual patterns in QSAR data [19]. The basic HAC algorithm consists of an iterative procedure wherein the two nearest clusters are joined at each step to form a single, larger cluster. Initially, each of the N compounds in a data set is treated as an individual cluster, and by the end of k iterations, the number of distinct clusters is $N - k$. The metric used here is just the Euclidean distance in the space of the descriptors. Inter-cluster distances are computed as the largest distance between any two compounds in the separate clusters. This approach tends to give rise to more compact, spherically shaped clusters [23].

The rationale behind our cluster-based classification scheme is that a prediction set compound, when considered together with equal numbers of actives and inactives, will tend to cluster with its own kind. Thus, if the compound falls into a cluster that is composed predominantly of actives, it will be classified as active. One admitted pitfall of this approach is that the inactive compounds, as a whole, should not really have any common structural characteristics, so they may have very little tendency to cluster. However, an underlying philosophy in QSAR is that inactive compounds are in fact similar because they are all missing one or more critical structural motifs that are responsible for activity. The use of descriptors that encode the presence or absence of these structural features should then lead to an increased clustering tendency among the inactive compounds. Binary structural keys are expected to perform well in this respect, provided that the relevant set of substructures is sufficiently well represented in the key string.

A primary issue of contention in HAC is deciding upon an appropriate number of clusters. There are statistical tests [24] which measure the probability for the existence of any particular number of clusters, but frequently no clear-cut optimum can be detected [23]. For classification purposes, it seems reasonable to select the number of clusters based on the degree of separation of actives from inactives within the training set. Much as with recursive partitioning, the goal is to form a set of clusters with some optimal degree of purity. With this in mind, we define the cohesivity C_k of cluster k using the numbers of active and inactive training set compounds contained in that cluster:

$$\begin{aligned} C_k = {} & (n_{k+}(n_{k+} - 1))/2 + (n_{k-}(n_{k-} - 1))/2 \\ & - n_{k+}n_{k-} \end{aligned} \tag{7}$$

A positive number is obtained for highly pure clusters, and a negative number results if significant inter-class mixing occurs. For a given degree of purity, the cohesivity is designed to increase with cluster size because large, highly pure clusters are far less likely to occur by chance. Many other functional forms were considered, but Equation 7 provided the most satisfactory classification results across a variety of data sets.

The number of clusters used in the classification is chosen to coincide with the greatest total cohesivity, which is obtained by summing Equation 7 over all clusters. The basic procedure is to cluster a single prediction set compound with a balanced training set of compounds, determine the optimal number of clusters based on the total cohesivity, and then, at the optimal clustering level, examine the composition of the cluster containing the compound to be predicted. If compound i falls into cluster k, then the numerical class score is computed according to

$$\text{score}(i) = 2(n_{k+} - n_{k-})/N \tag{8}$$

where N is the total number of compounds in the balanced training set. Thus, the largest possible score (±1) would occur if the training set split into exactly two clusters with perfect separation of actives and inactives. Under these circumstances, the highest possible level of confidence would exist for predictions. Other forms are certainly possible, such as replacing the normalization factor N by the total number of compounds in the cluster, which is analogous to the recursive partitioning score defined in Equation 6. However, the clustering process frequently gives rise to any number of very small clusters that contain peculiar compounds which are not really similar to the majority of compounds in either class. The purity of these clusters may be high, but they usually do not represent a statistically significant fraction of the training set, so predictions derived therefrom are less reliable. Any attempt to adjust this type of score for statistical significance ultimately requires something along the lines of multiplication by the size of the cluster, which then cancels out the original cluster size normalization factor.

The data set

With the aid of the ISIS package [21], three groups of active compounds were selected from the MDL CMC (Comprehensive Medicinal Chemistry) database [19] according to target or target family. These active sets consisted of 24 angiotensin-converting enzyme (ACE) inhibitors, 58 β-adrenergic antagonists, and 96 H_2 receptor antagonists. We note that the β-antagonists used here are for the most part non-selective and thus interact with both the β_1 and β_2 subtypes, but there should be enough structural similarity among these compounds to consider them as a single class.

An inactive set was generated from the same database by selecting at random 1000 compounds which were not labeled as being in any of the above three pharmacological classes. This is of course no guarantee that all the 1000 compounds are truly inactive toward the three targets, but care was taken to exclude, for example, full and partial agonists of the β-type receptors, as well as H_1 antagonists which could potentially exhibit activity toward the H_2 receptor.

Both the actives and inactives were randomly divided in half, and the duplexed $+/-$ compounds were combined to form equal-sized training and prediction sets with high fractions of inactive compounds. Thus, the ACE inhibitors gave rise to two sets of 512 compounds (12 actives and 500 inactives), the β-antagonists gave two sets of 529 compounds, and the antihistamines yielded two sets of 548 compounds. Balanced training sets used in repetitive classification procedures contained all the active training set members (12, 29, or 48), and an equal number of compounds selected at random from among the 500 inactive training set members. Once created, the data sets used for training and predicting were held fixed, so that classifications were performed on the same 512, 529, or 548 compounds.

Structural descriptors

While interactions between a drug and a target are certainly three-dimensional (3D) in nature, the ambiguity of 3D models and the computational expense of generating them for large chemical libraries make geometric characterizations of structure less desirable than topological characterizations. Moreover, in head-to-head comparisons of 2D and 3D descriptors, Brown and Martin [25] found fairly compelling evidence to support the use of purely topological indices in applications involving structure–activity data. This certainly does not imply that 2D descriptions of structure should always be preferred, but their simplicity and lack of ambiguity justify their use as an initial means of attack in most chemoinformatic applications. For these reasons, structures were characterized here using two different sets of purely topological indices.

The first set was comprised of 150 whole-molecule descriptors calculated using the Molconn-X program [20]. These descriptors primarily encode aspects of molecular size, shape, branching, and, to some extent, polarity. The set included number of atoms, molecular weight, subgraph counts, simple and valence connectivity indices, Wiener number, kappa indices, counts of hydrogen bonding groups, etc.

A second set of descriptors based on the searchable MDL keys [26] was employed to encode the presence or absence of 166 types of substructural patterns. These binary descriptors were extracted using the MOLSKEYS feature of the ISIS program [21]. It should be noted that no special consideration was made when performing mathematical analyses with the MOLSKEYS. Each substructure bit was treated as a separate descriptor variable which could take on the value of 1 or 0.

As is common when working with a large number of independent variables, it is necessary to extract a much smaller subset of them in order to avoid overfitting the training set data. To create a fair comparison among the various classification schemes, the same subsets of Molconn-X descriptors and MOLSKEYS bits were used to classify compounds against a given target, regardless of which classification technique was being applied. These two subsets of indices were arrived at by applying the standard stepwise regression procedure within S-PLUS [22] to the full set of training compounds, with the dependent activity variable defined as ±1. The F-to-enter and F-to-delete values were set at 2.0, and the five most statistically significant descriptors or substructure bits uncovered by the stepwise procedure were used. We note that, in every case, more than five descriptors passed the F-test and were thus retained by the stepwise procedure. For example, in the ACE inhibitor data set, a total of 45 MOLSKEYS bits were found to be statistically significant at the $F = 2.0$ level. However, using all 45 descriptors for classifications would be questionable since there are only 12 actives in the training set. We concede that the number of descriptors to use is just one of the many factors in this investigation which could be varied, but our goal was to simplify and hold constant the information supplied to each classification method in order to draw straightforward, fair comparisons among them. Using more advanced methods [27,28] of selecting descriptors might improve classifications, but these sorts of approaches are somewhat beyond the scope of the current investigation.

Descriptor scaling

Recursive partitioning relies only on the rank-ordering of the elements of a given variable, so it is insensitive to all monotonic transformations of the descriptor values. Linear discriminant analysis is simply a scaling procedure applied to the variables, so an appropriate global optimization algorithm guarantees insensitivity to additive and multiplicative transformations of the descriptors. By contrast, clustering methods usually rely on a distance (or similarity) measure which is highly sensitive to the length of the intervals on which the variables are defined. As a result, descriptors with a high variance will tend to exert a great deal of influence on the clustering process, regardless of their inherent ability to separate the data set into clusters.

The issue of scaling in cluster analysis was addressed by Bravi et al. [29] in a conformational study of linear and cyclic peptides. They employed a set of dihedral angles as variables and developed a similarity measure to compare conformations. The overall geometry was found to be much more sensitive to changes in certain dihedrals, so each angle was assigned its own difference threshold for use in conformational comparisons. Dissimilar pairs of structures were defined as those for which all the dihedral angles differed by more than their associated thresholds. Note that an equivalent approach would be to use a single threshold for all dihedrals, and simply scale the angles themselves.

In the present case, there is no truly objective way to determine an appropriate set of scales because the descriptors measure entirely different features of molecular structure. While it is common practice in cluster analysis to simply standardize the variables to zero mean and unit variance, this sort of transformation is completely arbitrary and it does not necessarily improve the separation of actives from inactives. Perhaps the most relevant information available is contained within the training set regression models of activity. Each regression coefficient measures the sensitivity of the activity indicator variable to a unit change in the associated descriptor value. Descriptors that have a large impact on activity for the training set should therefore be weighted more heavily in the distance measure used to cluster the compounds. Thus, before carrying out cluster analysis, descriptor values were multiplied by their associated regression coefficients from the five-variable training set models derived using the stepwise selection technique. We note that this procedure yielded more accurate classifications than either no scaling or the use of standardized descriptor values.

Computational details

Discriminant analysis was performed with the aid of the S-PLUS [22] statistical package and the ***discr*** function contained therein. S-PLUS was also employed for recursive partitioning through the use of the ***tree*** function. An in-house FORTRAN program was written to carry out classifications using hierarchical agglomerative clustering.

For balanced classifications, prediction set scores were computed from an average of 100 different training sets, each of which contained all the known actives and an equal number of randomly selected inactives. After each trial, the inactive compounds were returned to the training set pool. For the sake of consistency, all prediction set compounds associated with a given target were scored using the same 100 training sets. This removes any stochastic variability in the training sets among the different compounds being predicted.

Results and discussion

Table 1 contains a list of the structural descriptors that were used to classify compounds against the three targets. The MOLSKEYS are defined in terms of MDL query search language, so the structural features encoded are not necessarily obvious. However, sufficient information is provided within the table to allow interpretation of all symbols that appear.

There is of course some overlap of information provided by the two sets of descriptors. Cycle-based connectivity indices and cycle counts are probably encoding some of the same features as those MOLSKEYS bits which contain specific references to bonds contained in rings. It is also very likely that the difference connectivity indices DXP4 and DXP5 are accomplishing roughly the same thing as MOLSKEYS 53, 54, 97, and 116. All these indices are associated with paths of length 4 and 5 which contain heteroatoms.

We note that when regressions were built using the full training sets, the MOLSKEYS always produced better fits to the ±1 activity values than the Molconn-X descriptors. The MOLSKEYS R^2 values for the targets in Table 1 were 0.2067, 0.3925 and 0.2675, and the corresponding F values for the overall fits were 26.36, 67.58 and 39.60. This compares to the Molconn-X R^2 values of 0.1575, 0.1522 and 0.1331, and F values of 18.91, 18.78 and 16.65. Obviously, none of the associated regressions would be acceptable QSAR models in and of themselves, but they do indicate a mild statistical relationship between the descriptors and activity toward the targets.

Classification accuracy is expected to depend upon the degree of similarity among active compounds, and their overall dissimilarity to inactives. Table 2 summarizes this information for the various targets and the corresponding training and prediction set compounds. Average Tanimoto coefficients [30] computed with the full set of MOLSKEYS are reported for pairs of compounds in the training and prediction sets. Decreasing similarity among active compounds is observed as one progresses down the table from the ACE inhibitors to the antihistamines. Overall, the antihistamines are only slightly more similar to each other than they are to the inactive compounds. Accordingly, one would expect this data set to present the biggest challenge in terms of classifications.

For clarification, we pause here to note that in the discussions which follow, the terms ‘active misclassifications’, ‘misclassification of actives’, and so forth refer to instances wherein active compounds are incorrectly classified as being inactive. The converse applies to the terms ‘inactive misclassifications’, ‘misclassification of inactives’, etc.

Table 3 summarizes, for each target, the performance of the various classification methods using the top five descriptors selected from the Molconn-X and MOLSKEYS collections. The totals for the ACE inhibitors indicate a fairly drastic difference between these two sets of indices in the sensitivity toward active compounds. For this target, the MOLSKEYS correctly identify all the actives, while the Molconn-X descriptors misclassify them about half the time. Molconn-X sensitivity improves for the β-antagonists and the antihistamines; in the latter case, they outperform the MOLSKEYS. In terms of misclassifying inactive compounds, the MOLSKEYS are superior to the Molconn-X descriptors for all three targets. Our findings regarding the efficacy of binary structural keys thus concur with those of Brown and Martin [25].

Misclassification totals across all three targets appear in Table 4. Information is broken down in terms of descriptors and methods to allow overall head-to-head comparisons. As hypothesized, the use of balanced training sets gives rise to more sensitivity toward active compounds, independent of the classification method used or the descriptor set. Unbalanced training sets, however, perform substantially better in classifying inactive compounds. This is partially due to the fact that they are predisposed to yielding inactive predictions, a statistically favorable situation for a prediction set space that is predominantly inactive. Indeed, with unbalanced training sets, the overall fractional rate of inactive misclassifications is about 2.4 times lower than that observed for actives. These two rates are much more comparable in the case of balanced training sets.

In terms of the classification methods themselves, discriminant-based approaches appear to be the most sensitive for correctly identifying active compounds using either type of training set. Tree-based models, on the other hand, yield the lowest rate of misclassifications among inactives. Clustering exhibits sensitivity toward actives that is intermediate between the other

Table 1. Structural descriptors used for compound classifications

		Variable	Description
ACE inhibitors			
	Molconn-X[a]	XCH5	Simple connectivity, 5-atom cycle
		XVCH5	Valence connectivity, 5-atom cycle
		XVCH10	Valence connectivity, 10-atom cycle
		NXCH4	Number of 4-atom cycles
		SI	Shannon information index
	MOLSKEYS[b]	KEY075	A$A!S
		KEY090	QHAACH2A
		KEY092	OC(N)C
		KEY123	OCO
		KEY132	OACH2A
β-*Antagonists*			
	Molconn-X	DXP4	Difference simple connectivity, path-4
		DXP5	Difference simple connectivity, path-5
		IDCBAR	Bonchev–Trinajstic information index
		TM	Number of terminal methyls
		PHIA	Kappa flexibility index
	MOLSKEYS	KEY053	QHAAAHQ
		KEY054	QHAAQH
		KEY097	NAAAO
		KEY113	Onot%A%A
		KEY116	CH3AACH2A
H_2 *antihistamines*			
	Molconn-X	X0	Simple connectivity, path-0
		NXCH10	Number of 10-atom cycles
		SUMDELI	$\sum_i \Delta I_i$, where $\Delta I_i =$ perturbed atomic intrinsic state value
		DIAM	Graph diameter
		NUMHBA	Number of hydrogen bond acceptors
	MOLSKEYS	KEY065	C%N
		KEY074	CH3ACH3
		KEY086	CH2QCH2
		KEY147	ACH2CH2A
		KEY164	Oxygen

[a] See Reference 32 for more detailed definitions of the Molconn-X descriptors.

[b] Query language for MDL keys is outlined in Reference 22. Briefly, the syntax is *atom bond atom bond* etc. ‘A’ is an atom other than C; ‘Q’ is an atom other than C or H; ‘*n*H’ indicates that at least *n* hydrogens are attached to the previous atom in the list; ‘(*atom*)’ signifies an attachment to the left but not to the right; ‘$’ is a bond in a ring; ‘!’ is a bond in a chain; ‘%’ is an aromatic bond.

two balanced methods, but it is the least effective for identifying inactive compounds. The comparatively high inactive misclassification total for the clustering approach is primarily attributable to predictions from the Molconn-X descriptors. A review of the results in Table 3 reveals that twice as many inactive misclassifications arise in the cluster-based predictions with the Molconn-X descriptors than with the MOLSKEYS. This is a larger disparity than the overall Molconn-X/MOLSKEYS numbers in Table 4, indicating that inactive compounds probably do not cluster very well in the space of the Molconn-X descriptors. As mentioned previously, the lack of commonality in overall

Table 2. Average Tanimoto similarities[a] among active and inactive compounds

	Active–active[b]		Active–inactive[c]		Inactive–inactive[d]	
	All pairs[e]	Nearest neighbor[f]	All pairs[e]	Nearest neighbor[f]	All pairs[e]	Nearest neighbor[f]
ACE inhibitors						
Training set	0.6589	0.8160	0.3710	0.4430	0.3059	0.7038
Prediction set	0.7033	0.8885	0.3989	0.4746	0.3233	0.7138
β-*Antagonists*						
Training set	0.5793	0.7869	0.3342	0.4779	0.3059	0.7038
Prediction set	0.6646	0.8356	0.3568	0.4955	0.3233	0.7138
H_2 antihistamines						
Training set	0.4614	0.7668	0.3166	0.5268	0.3059	0.7038
Prediction set	0.4480	0.7560	0.3174	0.5191	0.3233	0.7138

[a] Tanimoto coefficients (Reference 30) were computed using the full set of MOLSKEYS.
[b] Comparisons between pairs of active compounds.
[c] Comparisons between active–inactive pairs.
[d] Comparisons between pairs of inactive compounds.
[e] Average similarity over all unique pairs.
[f] Average similarity between each compound and the compound that is most similar to it. For the active–inactive case, the inactive compound that is most similar to each active is identified, and vice versa.

structure among the inactives is likely the cause of this result.

It is important to remember that for the balanced approaches, compounds were classified simply according to the algebraic sign of the average score (Equations 4, 6 and 8). This hit-or-miss type of approach does not distinguish between high and low confidence predictions, the latter being associated with scores that are close to zero. By considering the magnitude of the score as well as the algebraic sign, it is possible to decrease either the number of active misclassifications or the number of inactive misclassifications.

A greater sensitivity toward actives may be achieved simply by reducing the active threshold from zero to some negative number. This effectively lowers the requirement for active status and correspondingly reduces the probability that active compounds will be overlooked. An increase in the number of inactive misclassifications is an obvious side-effect.

Conversely, if the goal is to reduce the number of inactive compounds that are mistaken for actives, then the threshold may be raised to some positive number. For methods that are highly sensitive to actives, yet prone to inactive misclassifications (e.g., clustering with the Molconn-X descriptors), this approach may be an effective way to improve the overall performance.

A related but more general approach is to simply sort the predicted activity scores in order of decreasing value. The effective threshold is thus high at the top of the list, maximizing the probability that these compounds will be active. If coupled to an HTS operation, this would provide a means for generating a relatively small, active-enriched core library for screening. Figures 1a–f show the simulated results of such a procedure applied to the prediction set compounds in the present study using the three balanced classification techniques. These curves indicate the fraction of active compounds that could be identified by a reliable screening procedure if the compounds were assayed in order of decreasing class score. The Molconn-X results are included for completeness, but screening according to the MOLSKEYS is more advantageous in most cases, so we focus on Figures 1d–f in the discussion.

In Figure 1d, the curves for the discriminant and tree-based methods are superimposed because the rank-ordered versions of these approaches exhibited the same hit rates with respect to the ACE inhibitors. The MOLSKEYS are clearly very effective for this target, and all three classification methods identify 100% of the actives after screening only about 6% of the compounds. These results are even more impressive considering the fact that only 12 actives were used for training purposes.

Table 3. Classification summaries for prediction set compounds

	Number (Fraction) of misclassifications			
	Molconn-X		MOLSKEYS	
	Actives	Inactives	Actives	Inactives
ACE inhibitors: 12 actives				
500 inactives				
Discriminant (unbalanced)	11 (0.917)	31 (0.062)	0 (0.0)	34 (0.068)
Discriminant (balanced)	4 (0.333)	142 (0.284)	0 (0.0)	48 (0.096)
Tree (unbalanced)	7 (0.583)	21 (0.042)	0 (0.0)	27 (0.054)
Tree (balanced)	3 (0.250)	143 (0.286)	0 (0.0)	150 (0.300)
Cluster (balanced)	6 (0.500)	190 (0.380)	0 (0.0)	98 (0.196)
Total	31 (0.517)	527 (0.211)	0 (0.0)	357 (0.143)
β-*Antagonists:* 29 actives				
500 inactives				
Discriminant (unbalanced)	1 (0.034)	101 (0.202)	2 (0.069)	59 (0.118)
Discriminant (balanced)	2 (0.069)	97 (0.194)	1 (0.034)	72 (0.144)
Tree (unbalanced)	10 (0.345)	37 (0.074)	2 (0.069)	35 (0.070)
Tree (balanced)	5 (0.172)	85 (0.170)	1 (0.034)	74 (0.148)
Cluster (balanced)	5 (0.172)	205 (0.410)	1 (0.034)	90 (0.180)
Total	23 (0.159)	525 (0.210)	7 (0.048)	330 (0.132)
H_2 *antihistamines:* 48 actives				
500 inactives				
Discriminant (unbalanced)	8 (0.167)	99 (0.198)	16 (0.333)	67 (0.134)
Discriminant (balanced)	6 (0.125)	115 (0.230)	6 (0.125)	171 (0.342)
Tree (unbalanced)	18 (0.375)	67 (0.134)	18 (0.375)	72 (0.144)
Tree (balanced)	13 (0.271)	106 (0.212)	18 (0.375)	75 (0.150)
Cluster (balanced)	6 (0.125)	150 (0.300)	16 (0.333)	78 (0.156)
Total	51 (0.213)	537 (0.215)	74 (0.308)	463 (0.185)

In moving to the β-antagonists (Figure 1e), it is possible to identify more than 90% of the actives by examining fewer than 10% of the compounds. The success rate drops noticeably after this point, and it becomes necessary to screen an additional 10–20% of the compounds in order to extract the remaining actives. For this target, the cost of screening would obviously have to be weighed against the need to identify a small number of additional hits.

Finally, in Figure 1f, results for the antihistamines are presented. Here, the majority of the chemical library would have to be screened in order to find all the actives. It would probably be more practical to examine only about 20% of the compounds, which would reveal close to 90% of the actives. The success rate would still be more than 4 times that of random sampling or any other ‘uninformed’ screen.

Implementation of this approach into an actual HTS operation would ideally involve continuous monitoring of the hit rate as compounds are assayed from high to low. Note that most of the curves presented here initially show a rapid rise in the number of actives found, followed by a fairly abrupt leveling-off. This behavior would be apparent when constructing such a curve in real time, so the screening process could be halted at the first sustained indication of diminishing returns.

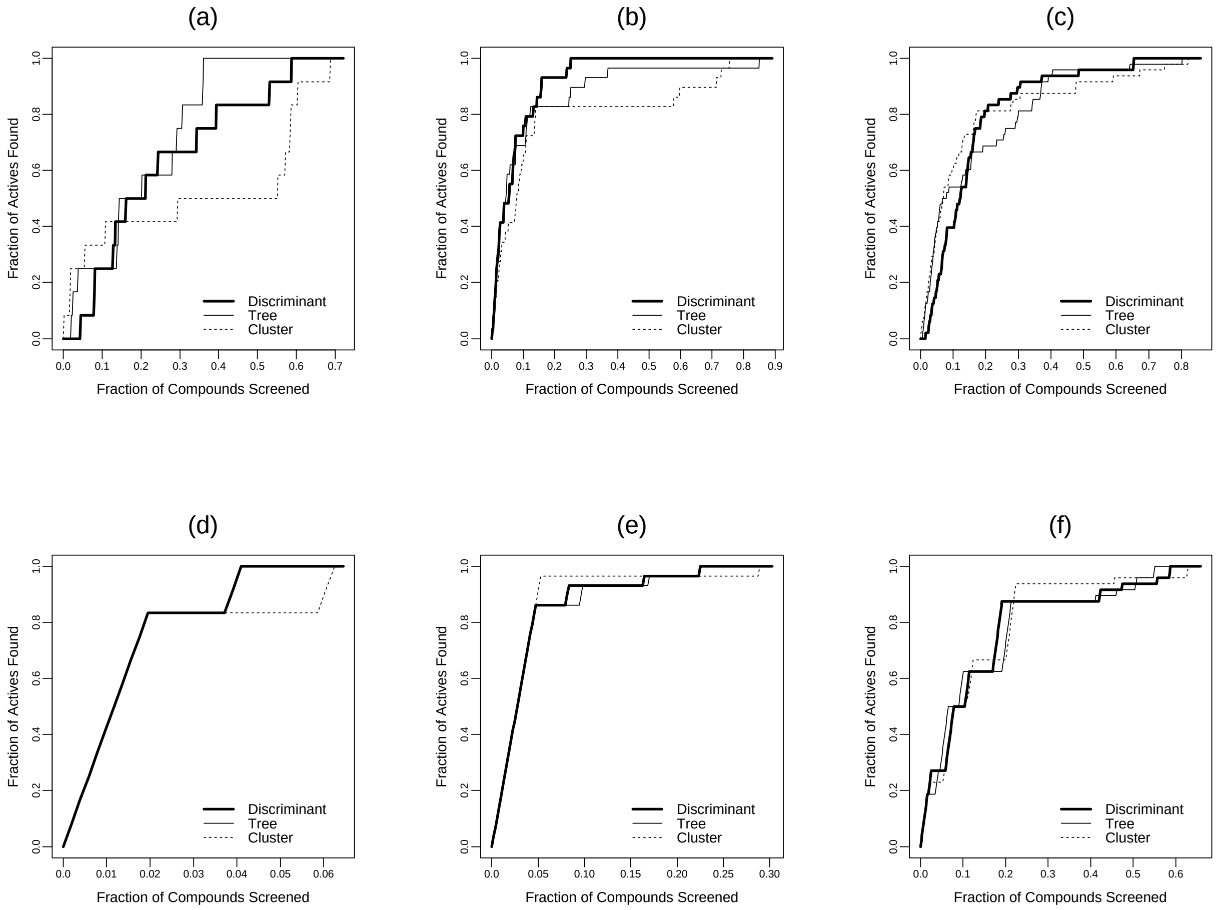

Figure 1. Results of a virtual screening procedure applied to the compounds in the prediction sets. Compounds are screened in order of decreasing score as computed by each balanced classification method. (a) ACE inhibitors, Molconn-X descriptors; (b) β-adrenergic antagonists, Molconn-X descriptors; (e) H_2 receptor antagonists, Molconn-X descriptors; (d) ACE inhibitors, ISIS MOLSKEYS; (e) β-adrenergic antagonists, ISIS MOLSKEYS; (f) H_2 receptor antagonists, ISIS MOLSKEYS.

Table 4. Classification totals across all targets

	Misclassification rates	
	Actives	Inactives
Molconn-X	105/445 = 0.236	1589/7500 = 0.212
MOLSKEYS	81/445 = 0.182	1150/7500 = 0.153
Discriminant (unbalanced)	36/178 = 0.213	391/3000 = 0.130
Tree (unbalanced)	55/178 = 0.309	259/3000 = 0.086
Total (unbalanced)	93/356 = 0.261	650/6000 = 0.108
Discriminant (balanced)	19/178 = 0.107	645/3000 = 0.215
Tree (balanced)	40/178 = 0.225	633/3000 = 0.211
Cluster (balanced)	34/178 = 0.191	811/3000 = 0.270
Total (balanced)	93/534 = 0.174	2089/9000 = 0.232

There is of course no reason to ignore information that becomes available as the screening proceeds, and it should be possible to refine a classification model accordingly. In general, robotic techniques do not allow ‘on-the-fly’ changes of the order in which compounds are assayed, so a library would have to be examined in blocks. If, for example, 1% of the compounds were set up for high to low screening, then the model could be updated at the end of this first block and used to recommend the best compounds or HTS plates to examine in the next block. A genuine improvement in the classification model would be manifested by an increase in the hit rate at the beginning of the second block as compared to the end of the first block. We have used this sort of approach in conjunction with ‘Affinity Fingerprints’ to identify actives in both functional assays, and in retrospective analyses of HTS data [31].

The ranking technique is directly applicable to combinatorial library design. Individual compounds could be recommended in cases where the reactions are fixed and the building blocks are easily varied. If the question becomes which reactions to employ, then libraries which yield the highest rankings overall could be made. For the iterative block-based approach, active discoveries would obviously provide the most valuable information, but it is also important to incorporate inactive points into the model to reduce the amount of effort devoted to synthesizing structurally similar compounds that ultimately exhibit no activity.

Conclusions

Combinatorial chemistry and high-throughput screening techniques have greatly facilitated the process of lead generation. There is, however, considerable room for improvement in these admittedly shotgun approaches to drug discovery. In particular, if even a small number of active compounds have been identified, then it should be possible to use the activity data in conjunction with classification techniques to selectively extract additional actives from a diverse chemical library. With the aid of standard 2D descriptors, we have employed classification schemes based on discriminant analysis, recursive partitioning, and hierarchical agglomerative clustering to identify ACE inhibitors, β-adrenergic antagonists, and H_2 receptor antagonists.

We have shown that replacing a single, predominantly inactive training set with a large number of smaller, balanced training sets yields greater sensitivity toward active compounds. This appears to be true irrespective of target or descriptor set, indicating a general applicability to any classification approach. Inactively biased or unbalanced training sets, on the other hand, do tend to return a smaller number of misclassifications among inactive compounds, but they also exhibit a large disparity between the rates of active and inactive misclassifications.

While analysis of the structural indices was not our primary goal, the ISIS MOLSKEYS appear to be more effective than the Molconn-X descriptors, especially in cluster-based classifications. This is probably attributable to the binary nature of the MOLSKEYS, which gives them an inherent ability to group together active compounds on the basis of structural features they share, and inactive compounds on the basis of structural features they lack.

By adjusting the activity threshold used in balanced classification approaches, it is possible to improve the sensitivity toward active compounds, or to reduce the number of inactive misclassifications. When screening order can be controlled, then it is probably best to examine the compounds sequentially according to decreasing activity score, until the hit rate drops abruptly. In this way, the majority of the actives may be extracted by screening only a fraction of a chemical library. This type of approach is well suited to the design of active-enriched virtual libraries for combinatorial synthesis.

Overall, we have seen that chemoinformatic approaches are quite capable of classifying highly diverse collections of compounds on the basis of activity. With binary structural keys and the proper training set design, correct classifications among actives and inactives are possible about 80% of the time, even when only a weak statistical relationship between structure and activity exists. The use of such techniques in library design could lead to significant improvements in modern approaches to lead generation.

References

1. Salemme, F.R., Spurlino, J. and Bone, R., Structure, 5 (1997) 319.
2. Martin, Y.C., Brown, R.D. and Bures, M.G., In Kerwin, J.F. and Gordon, E.M. (Eds.) Combinatorial Chemistry and Molecular Diversity in Drug Discovery, Wiley, New York, NY, 1998, pp. 369–385.
3. Ferguson, A.M., Patterson, D.E., Garr, C. and Underiner, T., J. Biomol. Screen., 1 (1996) 65.
4. Patterson, D.E., Cramer, R.D., Ferguson, A.M., Clark, R.D. and Weinberger, L.E., J. Med. Chem., 39 (1996) 3049.
5. Chapman, D., J. Comput.-Aided Mol. Design, 10 (1996) 501.
6. Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.
7. Shemtulskis, N.E., Dunbar, J.B., Dunbar, B.W., Moreland, D.W. and Humblet, C., J. Comput.-Aided Mol. Design, 9 (1995) 407.
8. Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Chapman and Hall, London, 1995.
9. Barnard, J.M. and Down, G.M., J. Chem. Inf. Comput. Sci., 32 (1992) 644.
10. Johnson, M.A. and Maggiora, G.M., Concepts and Applications of Molecular Similarity, Wiley, New York, NY, 1990.
11. Willett, P., Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, 1987.
12. Dillon, W.R. and Goldstein, M., Multivariate Analysis, Methods and Applications, Wiley, New York, NY, 1984.
13. Van de Waterbeemd, H., In van de Waterbeemd, H. (Ed.) Chemometric Methods in Molecular Design, VCH, New York, NY, 1995, pp. 283–293.

14. McFarland, J.W. and Gans, D.J., In Hansch, C., Sammes, P.G. and Taylor, J.B. (Eds.) Comprehensive Medicinal Chemistry, Vol. 4, Pergamon, New York, NY, 1990, pp. 667–689.
15. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., Classification and Regression Trees, Wadsworth International Group, Belmont, CA, 1984.
16. Hawkins, D.M. and Kass, G.V., In Hawkins, D.H. (Ed.) Topics in Applied Multivariate Analysis, Cambridge University Press, Cambridge, 1982, pp. 269–302.
17. Young, S.S. and Hawkins, D.M., J. Med. Chem., 38 (1995) 2784.
18. Murtagh, F., Multidimensional Clustering Algorithms, Vol. 4, Physica-Verlag, Heidelberg, 1985.
19. CMC database, MDL Information Systems, Inc., San Leandro, CA.
20. Molconn-X 2.0., Hall Associates Consulting, Quincy, MA.
21. ISIS^{TM}/Base 2.1.3, MDL Information Systems Inc., San Leandro, CA.
22. S-PLUS 3.4, StatSci Division, MathSoft Inc., Seattle, WA.
23. Brown, R.D. and Martin, Y.C., J. Chem. Inf. Comput. Sci., 36 (1996) 572.
24. S-Plus Guide to Statistical and Mathematical Analysis, Version 3.3, MathSoft Inc., Seattle, WA, 1995.
25. Banfield, J.D. and Raftery, A.E., Biometrics, 49 (1992) 803.
26. MACCS-II Menu Reference Version 2.2, MDL Information Systems, San Leandro, CA, 1994.
27. Sutter, J.M., Dixon, S.L. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 35 (1995) 77.
28. Furnival, G. and Wilson, R., Technometrics, 16 (1974) 499.
29. Bravi, G., Gancia, E., Zaliani, A. and Pegna, M., J. Comput. Chem., 18 (1997) 1295.
30. Hall, L.H. and Kier, L.B., Molconn-X User's Guide, Hall Associates Consulting, Quincy, MA, 1993.
31. Dixon, S.L. and Villar, H.O., J. Chem. Inf. Comput. Sci., 38 (1998) 1192.
32. Willett, P. and Winterman, V.A., Quant. Struct.–Act. Relatsh., 5 (1986) 18.