Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the *k*-Nearest-Neighbor Principle

Weifan Zheng and Alexander Tropsha*

The Laboratory for Molecular Modeling, Division of Medicinal Chemistry and Natural Products,
School of Pharmacy, University of North Carolina, Chapel Hill, North Carolina 27599

Received March 25, 1998

A novel automated variable selection quantitative structure−activity relationship (QSAR) method, based on the *K*-nearest neighbor principle (kNN-QSAR) has been developed. The kNN-QSAR method explores formally the active analogue approach, which implies that similar compounds display similar profiles of pharmacological activities. The activity of each compound is predicted as the average activity of *K* most chemically similar compounds from the data set. The robustness of a QSAR model is characterized by the value of cross-validated R^2 (q^2) using the leave-one-out cross-validation method. The chemical structures are characterized by multiple topological descriptors such as molecular connectivity indices or atom pairs. The chemical similarity is evaluated by Euclidean distances between compounds in multidimensional descriptor space, and the optimal subset of descriptors is selected using simulated annealing as a stochastic optimization algorithm. The application of the kNN-QSAR method to 58 estrogen receptor ligands as well as to several other groups of pharmacologically active compounds yielded QSAR models with q^2 values of 0.6 or higher. Due to its relative simplicity, high degree of automation, nonlinear nature, and computational efficiency, this method could be applied routinely to a large variety of experimental data sets.

INTRODUCTION

The quantitative structure−activity relationship (QSAR) approach was first introduced by Hansch et al. in 1963,[1,2] on the basis of implications from linear free-energy relationships (LFER) in general and the Hammett equation in particular.[3] It is based upon the assumption that the difference in structural properties accounts for the differences in biological activities of compounds. According to this approach, the structural changes that affect the biological activities of a set of congeners are of three major types: electronic, steric, and hydrophobic.[4] These structural properties are often described by Hammett electronic constants,[5] Verloop STERIMOL parameters,[6] hydrophobic constants, etc.[5] The quantitative relationships between biological activity (or chemical property) and the structural parameters could be conventionally obtained using multiple linear regression (MLR) analysis. The fundamentals and applications of this method in chemistry and biology have been summarized by Hansch and Leo.[4]

Many different approaches to QSAR have been developed since Hansch's seminal work. These include both 2D (two-dimensional) and 3D (three-dimensional) QSAR methods. The major differences of these methods can be analyzed from two viewpoints: (1) the structural parameters that are used to characterize molecular identities and (2) the mathematical procedure that is employed to obtain the quantitative relationship between a biological activity and the structural parameters.

Most of the 2D QSAR methods are based upon graph theoretic indices, which have been extensively studied by Randic[7] and Kier and Hall.[8−10] Although these structural indices represent different aspects of molecular structures, their physicochemical meaning is unclear. The successful applications of these topological indices combined with multiple linear regression (MLR) analysis have been summarized recently.[10] In this same category, ADAPT system employs topological indices as well as other calculable structural parameters (e.g. steric and quantum mechanical parameters) and the MLR method for QSAR analysis. It has been extensively applied to QSAR/QSPR (QSPR = quantitative structure−property relationship) studies in analytical chemistry, toxicity analysis, and other biological activity prediction.[11−14] On the other hand, parameters derived from various experiments through chemometric methods have also been used in the study of peptide QSAR, where partial least squares (PLS)[15,16] analysis has been employed.[17]

With the development of accurate computational methods for generating three-dimensional (3D) conformations of chemical structures, QSAR approaches that employ 3D descriptors have been developed to address the problems of 2D QSAR techniques, e.g., their inability to distinguish stereoisomers. The examples of 3D QSAR include molecular shape analysis (MSA),[18] distance geometry,[19,20] and Voronoi techniques.[21] The MSA method utilizes shape descriptors and multiple linear regression analysis, while the other two approaches apply atomic refractivity as structural descriptors and the solution of mathematical inequalities to obtain the quantitative relationships. These methods have been applied to study structure−activity relationships of many data sets by Hopfinger and co-workers[22−27] and Crippen and co-workers,[29−32] respectively. Perhaps the most popular example of 3D QSAR is the comparative molecular field analysis (CoMFA) developed by Cramer et al.,[33] which has elegantly

* To whom correspondence should be addressed.

10.1021/ci980033m CCC: $19.00
Published on Web 11/19/1999

combined the power of molecular graphics and PLS technique and has found wide applications in medicinal chemistry and toxicity analysis.[34−45]

Recent trends in both 2D and 3D QSAR studies have focused on the development of optimal QSAR models through variable selection. This implies that only a subset of available descriptors of chemical structures, which are most meaningful and statistically significant in terms of correlation with biological activity, is selected. The optimum selection of variables is achieved by combining stochastic search methods with the correlation methods such as MLR, PLS analysis, or artificial neural networks (ANN).[46−51] More specifically, these methods employ either generalized simulated annealing,[46] genetic algorithms,[47] or evolutionary algorithms[48−51] as the stochastic optimization tool. Since the effectiveness and convergence of these algorithms are strongly affected by the choice of a fitting function, several such functions have been applied to improve the performance of the algorithms.[48,49] It has been demonstrated that these algorithms combined with various chemometric tools have effectively improved the QSAR models compared to those without variable selection.

The variable selection methods have also been adopted for region selection in the area of 3D QSAR. For example, GOLPE[52] was developed using chemometric principles, and q^2-GRS[53] was developed on the basis of independent CoMFA analyses of small areas (or regions) of near-molecular space to address the issue of optimal region selection in CoMFA analysis. More recently, a genetic algorithm based sampling of 3D regions of CoMFA fields was implemented.[54] Both of these methods have been shown to improve the QSAR models compared to the original CoMFA technique.

Most of the QSAR techniques (both 2D and 3D) assume the existence of a linear relationship between a biological activity and molecular descriptors, which may be an adequate assumption for relatively small data sets (dozens of compounds). However, the fast collection of structural and biological data, owing to recent development of combinatorial chemistry and high throughput screening technologies, has challenged traditional QSAR techniques. First, 3D methods may be computationally too expensive for the analysis of a large volume of data; and in some cases, an automated and unambiguous alignment of molecular structures is not achievable.[44] Second, although existing 2D techniques are computationally efficient, the assumption of linearity in the SAR may not hold true, especially when a large number of structurally diverse molecules are included in the analysis.

Several nonlinear QSAR methods have been proposed in recent years. Most of these methods are based on either artificial neural network (ANN)[55−62] or machine learning techniques.[63−66] Both back-propagation (BP-ANN) and counterpropagation (CP-ANN)[67] neural networks were used in these studies. Machine learning methods included inductive logic programming. Since optimization of many parameters is involved in these techniques, the speed of the analysis is relatively slow. More recently, Hirst reported a simple and fast nonlinear QSAR method,[68] where the activity surface was generated from the activities of training set compounds based on some predefined mathematical function.

These considerations provide an impetus for the development of *fast*, generally *nonlinear*, *variable selection* QSAR methods that can avoid the aforementioned problems of QSAR. In this paper, we report the development of a new method that adopts a *K*-nearest neighbor (kNN) principle to QSAR (kNN-QSAR). Formally, this method implements the active analogue principle that lies in the foundation of modern medicinal chemistry. The kNN-QSAR method employs multiple topological (2D) or topographical (3D) descriptors of chemical structures and predicts biological activity of any compound as the average activity of *K* most similar molecules. We first describe the kNN methodology and computational details, followed by the results of kNN-QSAR analysis of estrogen receptor ligands as well as several other data sets. We show that the application of this method to all test sets leads to robust QSAR models characterized by the high value (0.6 or greater) of cross-validated $R^2(q^2)$.

METHODOLOGY

The kNN technique is a conceptually simple approach to pattern recognition problems. In this method, an unknown pattern is classified according to the majority of the class memberships of its *K* nearest neighbors in the training set. The nearness is measured by an appropriate distance metric (e.g., a molecular similarity measure as applied to the classification of molecular structures). The standard kNN method[69] is implemented simply as follows: (1) calculate distances between an unknown object (*u*) and all the objects in the training set; (2) select *K* objects from the training set most similar to object *u*, according to the calculated distances (*K* is usually an odd number); (3) classify object *u* with the group to which a majority of the *K* objects belongs. An optimal *K* value is selected by the optimization through the classification of a test set of samples or by the leave-one-out cross-validation.

Many variations of the kNN method have been proposed in the past, and new and fast algorithms have continued to appear in recent years.[70,71] The applications of the kNN principle in chemistry have been summarized by Strouf.[72] In the area of biology, Raymer et al. has successfully applied a kNN pattern recognition technique with simultaneous feature selection and classification in the analysis of water distribution in protein structures.[73] In the area of QSPR, Basak et al. have applied this principle, combined with principal component analysis and graph theoretical indices, in the estimation of physicochemical properties of organic compounds.[74−77] However, no applications based on the kNN principle have been reported so far in the context of QSAR with simultaneous variable selections.

COMPUTATIONAL DETAILS

All chemical structures and their SMILES notations were generated using SYBYL software.[78] Molecular connectivity indices were generated using the MolConnX program.[79] Atom pair descriptors were generated using the *GenAP* program developed in this laboratory (see below). All calculations were performed on an SGI Indigo[2].

Generation of Molecular Descriptors. We have applied two types of molecular descriptors: molecular connectivity indices (MCI) and atom pair descriptors (AP). Various types of MCI descriptors have been developed over the years on the basis of chemical graph theory (see refs 7−10 for a complete description), and many of them can be calculated with the MolConnX program.[79] MolConnX produces over 400 different descriptors; most of them characterize chemical structure, but several depend upon the arbitrary numbering of atoms in a molecule and are introduced for bookkeeping purposes only. Thus, only 312 chemically relevant descriptors were used in this study. The AP descriptors were generated as follows, using an approach initiated by Carhart et al.[80]

The key components for defining a set of atom pair descriptors include the definition of atom types and the classification of distance bins. An atom pair is a simple type of substructure defined in terms of the atom types and the shortest path separation (or graph distance) between two atoms. The graph distance is defined as the smallest number of atoms along the path connecting two atoms in a molecular structure. The general form of an atom pair is as follows:

$$\text{atom type } i-(\text{distance})-\text{atom type } j$$

where (distance) is the graph distance between atom i and atom j in the case of a 2D atom pair description. (The distance can also be defined as the physical distance between atoms i and j in the case of a 3D atom pair description.)

In this study, SYBYL atom types (mol2 format)[78] were utilized as the starting point. In principle, all SYBYL atom types can be used in the generation of atom pair descriptors. However, in order to reduce the number of atom pair descriptors, we have used only 10 atom types: (1) C.ar, aromatic carbons; (2) C.na, nonaromatic carbons; (3) N.ar, aromatic nitrogen atoms; (4) N.na, nonaromatic nitrogen atoms; (5) O.3, oxygen atoms in the sp^3 hybridization state; (6) O.2, oxygen atoms in the sp^2 hybridization state; (7) S, all sulfur atoms; (8) P.3, phosphorus atoms; (9) X, halogen atoms; (10) other atoms. The total number of pairwise combinations of all 10 atom types is 55. Furthermore, 15 distance bins were defined in the interval between graph distance zero (i.e., zero atoms separating an atom pair) to 14. Thus, a total of 825 (55 × 15) atom pair descriptors were generated for each molecular structure.

General kNN-QSAR Algorithm. The kNN QSAR method employs the kNN classification principle combined with the variable selection procedure. For each predefined number of variables (*nvar*) it seeks to optimize the following: using stochastic sampling and simulated annealing as an optimization tool: (i) the number of nearest neighbors (k) used to estimate the activity of each compound and (ii) selection of variables from the original pool of all molecular descriptors that are used to calculate similarities between compounds (i.e., distances in *nvar*-dimensional descriptor space). Figure 1 shows the overall flow chart of the kNN-QSAR method, which involves the following steps.

(1) Select a subset of *nvar* descriptors randomly (*nvar* is a number between 1 and the total number of available descriptors) as a hypothetical topological pharmacophore (HTP). *nvar* is usually set to different values in several different runs.

(2) Validate this HTP by a standard leave-one-out cross-validation procedure as described below.

(3) Repeat 1−2, i.e., the procedure of generating trial HTP's and calculating corresponding q^2 values. The goal is to find the best topological pharmacophore that maximizes the q^2 value of the kNN-QSAR model. This optimization process is driven by a generalized simulated annealing (see below) using q^2 as the objective function.

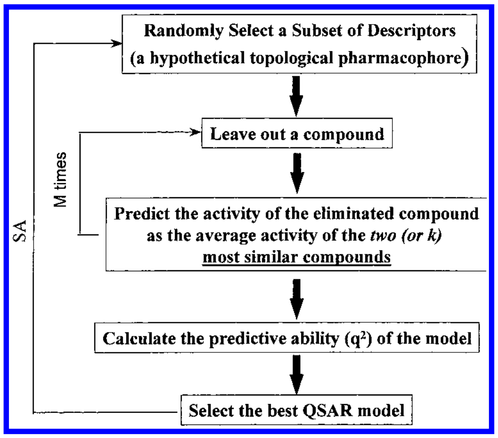
Randomly Select a Subset of Descriptors
(a hypothetical topological pharmacophore)
Leave out a compound
Predict the activity of the eliminated compound
as the average activity of the two (or k)
most similar compounds
Calculate the predictive ability (q²) of the model
Select the best QSAR model
M times
SA

Figure 1. Flow chart of the kNN-QSAR method.

Cross-Validation and the *k*-Nearest-Neighbor Principle. The standard leave-one-out procedure has been implemented as follows.

(1) Eliminate a compound in the training set and predict its biological activity on the basis of the kNN principle, i.e., as the average activity of k most similar molecules (k is set to 1 initially). The similarities are evaluated as Euclidean distances between compounds (eq 1) using only the subset

$$d_{i,j} = [\sum_{k=1}^{\text{nvar}} (X_{ik} - X_{jk})^2]^{1/2} \tag{1}$$

of descriptors that corresponds to the current trial HTP. The descriptors generated with MolconnX were autoscaled[69] prior to distance calculations, and no scaling was used for atom pairs. The reason for scaling the MolconnX descriptors was that their absolute ranges differ quite significantly, sometimes by orders of magnitude, unlike AP descriptors, which are integers ranging from zero to no more than a couple of dozens of AP counts. Thus, the scaling was used to avoid giving descriptors with significantly higher ranges a greater weight upon distance calculations in multidimensional MolconnX descriptor space.

(2) Repeat step 1 until every compound in the training set has been eliminated and its activity predicted once.

(3) Calculate the cross-validated R^2 (q^2) value using eq 2, where y_i and $\hat{y}_i$ are the actual and predicted activities of the ith compound, respectively, and $\bar{y}$ is the average activity of all compounds in the training set. Both summations are over

$$q^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{2}$$

all compounds in the training set. Since the calculation of the pairwise molecular similarities, and hence the predictions, are based upon the current HTP, the obtained q^2 value is indicative of the predictive power of the current kNN-QSAR model.

(4) Repeat steps 1−3 for k = 2, 3, 4, etc. Formally, the upper limit of k is the total number of compounds in the data set; however, the best value has been found empirically to lie between 1 and 5. The k value that leads to the highest q^2 value is chosen for the current kNN-QSAR model.

Simulated Annealing Based Optimization of the Variable Selection. The idea of simulated annealing is to simulate a physical process called annealing, in which a system is heated to a high temperature and then is gradually lowered to a preset temperature value (e.g., room temperature). During this process, the system samples possible configurations according to Boltzmann distribution. At equilibrium, low energy states will be mostly populated. The first implementation of the SA procedure was described by Metropolis et al.,[81] followed by the development of a more generalized mathematical optimization protocol.[82] The implementation of SA in this paper is as follows.

(1) Generate a trial solution to the underlying optimization problem; i.e., a kNN-QSAR model is built based on a random selection of descriptors (a trial HTP).

(2) Calculate the value of the fitness function, which characterizes the quality of the trial solution to the underlying problem, i.e., the q^2 value for a QSAR model built using only the HTP descriptors (q^2_{curr}).

(3) Perturb (i.e., slightly modify) the trial solution to obtain a new solution; i.e., change a fraction of the current HTP descriptors to other randomly selected descriptors and build a new kNN-QSAR model for the new trial HTP.

(4) Calculate the new value of the fitness function (q^2_{new}) for the new trial solution.

(5) Apply the optimization criteria: if $q^2_{\mathrm{curr}} \leq q^2_{\mathrm{new}}$ the new solution is accepted and used to replace the current trial solution; if $q^2_{\mathrm{curr}} > q^2_{\mathrm{new}}$, the new solution is accepted only if the following Metropolis criterion is satisfied; i.e.,

$$\mathrm{rnd} < e^{-({q_{\mathrm{curr}}}^2 - {q_{\mathrm{new}}}^2)/T} \tag{2}$$

where rnd is a random number uniformly distributed between 0 and 1 and T is a parameter analogous to the temperature in Boltzmann distribution law.

(6) Steps 3–5 are repeated until the termination condition is satisfied. The temperature lowering scheme and the termination condition used in this work have been adapted from Sun et al.[83] Thus, every time when a new solution is accepted or when a preset number of successive steps of generating trial solutions (100 steps) do not lead to a better result, the temperature is lowered by 10% (the default initial temperature is 1000). The calculations are terminated when either the current temperature of simulations is lowered to the value of $T = 10^{-6}$ or the ratio between the current temperature and the temperature corresponding to the best solution found is equal to 10^{-6}.

In summary, the kNN-QSAR algorithm generates both an optimum k value and an optimal subset of nvar descriptors, which together afford a QSAR model with the best predictive power in terms of q^2.

Data Sets. As a comprehensive test case for the kNN-QSAR technique, 58 estrogen receptor ligands were chosen. This data set was successfully analyzed earlier by Waller et al. using the CoMFA method.[35] In addition, we have considered several other QSAR data sets that were analyzed earlier, including 23 DHFR inhibitors,[84] 60 AChE inhibitors,[44] and 14 5HT-receptor agonists.[85] The chemical structures and biological activities of all compounds can be found in respective publications.

Statistical Significance of QSAR Models. To evaluate the statistical significance of a QSAR model for an actual data set, we have employed a standard hypothesis testing approach.[86] The robustness of the QSAR models for experimental training sets was examined by comparing these models to those derived for random data sets. Random sets were generated by assigning biological activities to the training set compounds randomly but restricting them to fall within the range of the actual activities of the original training set.

Table 1. Frequently Used α Values and the Corresponding Critical Values of Z_c for the One-Tail Test[86]

α	Z_c
0.10	1.28
0.05	1.64
0.01	2.33

According to the standard hypothesis testing approach two alternative hypotheses are formulated: for H_0 $h = \mu$ and for H_1 $h > \mu$, where μ is the average value of q^2's for random data sets and h is the q^2 value for the actual data set. Thus, the null hypothesis H_0 states that the QSAR model for the actual data set is not significantly better than random models, whereas the alternative hypothesis H_1 assumes the opposite; i.e., the actual model is significantly better than random models. The decision making is based on a standard one-tail test, which involves the following procedure:

(1) Determine the average value of q^2's (μ) and its standard deviation (σ) for random data sets.

(2) Calculate a Z score that corresponds to the q^2 value for the actual data set:

$$Z = (h - \mu)/\sigma$$

(3) Compare this Z score with the tabular critical values of Z_c at different levels of significance (α)[86] to determine the level at which H_0 should be rejected. If the Z score is higher than tabular values of Z_c (cf. Table 1), one concludes that at the level of significance that corresponds to that Z_c, H_0 should be rejected and, therefore, H_1 should be accepted. In this case, it is concluded that the result obtained for the actual data set is statistically much better than those obtained for random data sets at the given level of significance.

RESULTS AND DISCUSSIONS

In order to demonstrate the effectiveness of the method, a kNN-QSAR analysis of a set of estrogen receptor ligands was performed using both molecular connectivity indices (MCI) and atom pairs (AP) as molecular descriptors. To further substantiate its utility as a novel computational tool for QSAR analysis, we have also applied kNN-QSAR to the analysis of several other well-studied data sets. Results obtained for both description methods will be discussed below in terms of the q^2 values, variable selection, actual vs predicted activities, and statistical significance of the resulting QSAR models.

kNN-QSAR Analysis of Estrogen Receptor Ligands. This data set was characterized by both MCI and AP descriptors. Results will be discussed in the next two sections for MCI description and AP description, respectively.

MCI as Molecular Descriptors. In the kNN-QSAR method, nvar (the number of descriptors to be selected) can be set to any value that is less than the total number of descriptors generated by a particular molecular description

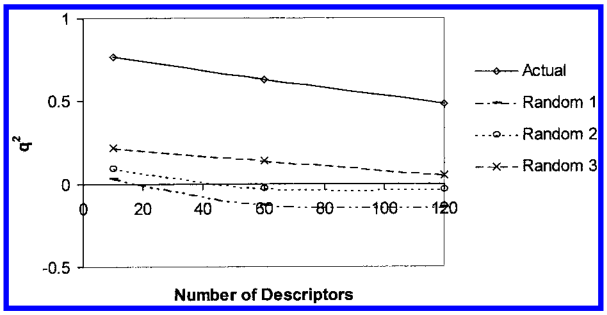

Figure 2. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR models for estrogen receptor ligands using MCI as molecular descriptors. The results for both the actual estrogen data set and three data sets with random activity values are shown.

method. Since the optimum number of nvar is not known a priori, several runs are needed to examine the relationship between the predictive power of a model (characterized by the q^2 value) and the number of descriptors selected (nvar). Figure 2 shows this relationship when MCI were used as descriptors.

Thus, when the real activity values for the estrogen receptor ligands were used in the kNN-QSAR analysis, the optimal q^2 values were 0.77, 0.63, and 0.48 for a 10-descriptor model, 60-descriptor model, and 120-descriptor model, respectively.

In order to prove the robustness of the kNN-QSAR model, one needs to demonstrate that no comparable q^2 values can be obtained for random data sets. Figure 2 also shows the q^2 vs nvar for three random data sets. Overall, these q^2 values are very low compared to those of the real data set. This suggests that the kNN-QSAR models obtained for the real data set are nonspurious.

One can observe that the q^2 values decrease somewhat when the number of descriptors increases. On the surface, this may be counterintuitive. The intuition may come from the fact that in multiple linear regression analysis the more descriptors are used, the higher regression coefficient is usually obtained. However, the kNN-QSAR is based not on a regression method but on the active analogue principle and variable selection. Theoretically, there should be no apparent trend in q^2 vs nvar relationships, although in many practical situations, q^2 tends to decrease somewhat when the number of descriptors increases. Conceivably, there should be one optimum number of descriptors, where either the q^2 is the highest or the separation between the q^2 for the real data set and those for random data sets is the largest.

The plots of predicted vs actual activity for the optimal cross-validated kNN-QSAR models are shown in Figure 3a−c for a 10-descriptor model, 60-descriptor model, and 120-descriptor model, respectively. It is important to keep in mind that each of the activity values was predicted when the corresponding compound was eliminated and treated as unknown, i.e., that the kNN-QSAR model is not based on any fitting as most linear QSAR techniques. In all three cases, the trend of the predicted values is similar to that of the real activity values although the 120-descriptor model (Figure 3c) should be regarded as rather poor. The values of the actual activity, the predicted activity from all three models, and the associated absolute errors are listed in Table 1. The results obtained in this work are better than those reported by Waller et al.[35] using the CoMFA method in terms of the q^2 values: we have obtained $q^2 = 0.77$ for nvar = 10 and $q^2 = 0.63$ for nvar = 60 as compared to $q^2 = 0.59$ obtained by Waller et al.

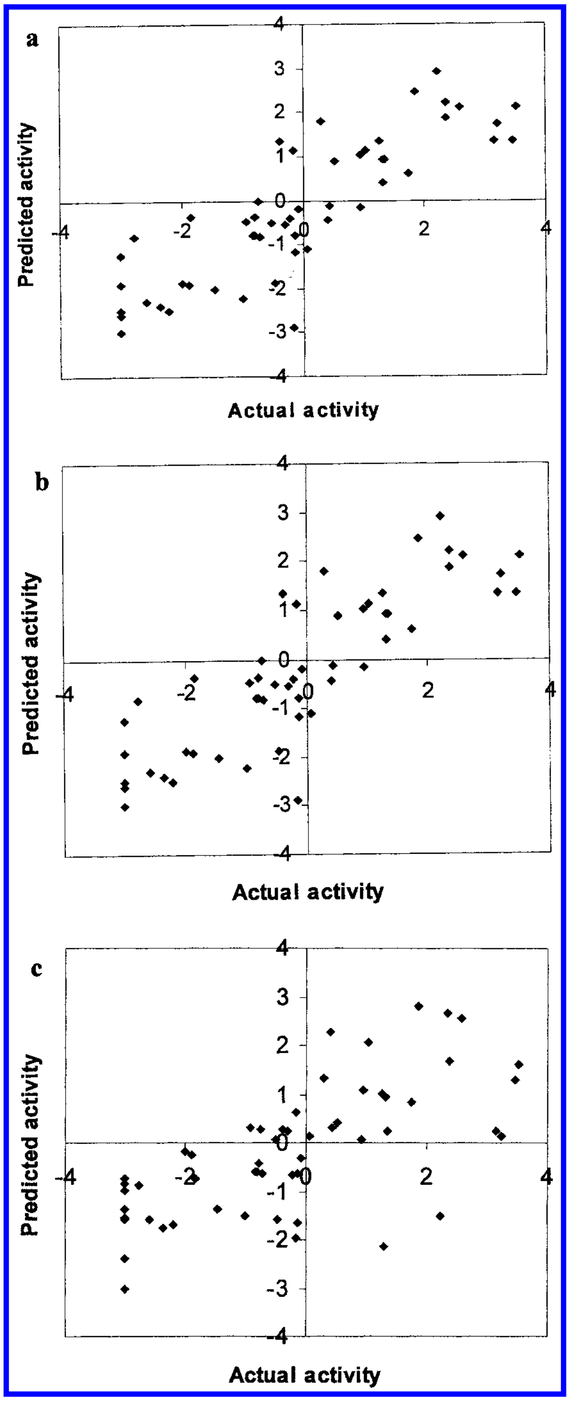

Figure 3. Plot of predicted vs actual activity obtained from 10-descriptor (a), 60-descriptor (b), and 120-descriptor (c) models for estrogen receptor ligands using MCI as molecular descriptors.

As discussed above, the robustness of the kNN-QSAR model (or, in fact, any QSAR model) should be subjected to a comparative analysis where the result for a real data set is compared to those for random data sets. The q^2 vs nvar relationships (cf. Figure 2) indicate that the kNN-QSAR models for estrogen receptor ligands give consistently higher q^2 values than those for random data sets. The statistical examination of the results has been performed by the means of hypothesis testing (see Computational Details). The q^2 values for ten 60-descriptor kNN-QSAR models obtained

Table 2. Standard One-Tail Hypothesis Testing for a 60-Descriptor kNN-QSAR Model for Estrogen Receptor Ligands Using MCI Descriptors

data sets	q^2	Z score	data sets	q^2	Z score
random 1	−0.132	−1.528	random 8	−0.044	−0.813
random 2	−0.028	−0.683	random 9	0.174	0.959
random 3	0.141	0.691	random 10	0.269	1.732
random 4	0.076	0.162	av of random 1−10	0.056	
random 5	−0.039	−0.772	std dev	0.123	
random 6	0.020	−0.293			
random 7	0.129	0.593	*Actual*	*0.63*	*4.677*

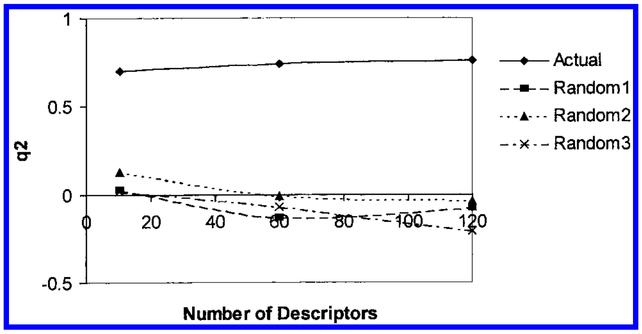

Figure 4. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for estrogen receptor ligands using AP as molecular descriptors. The results for both the actual data set and three random data sets are shown.

Table 3. Standard One-Tail Hypothesis Testing for a 60-Descriptor kNN-QSAR Model for Estrogen Receptor Ligands Using AP Descriptors

data sets	q^2	Z score	data sets	q^2	Z score
random 1	−0.129	−1.213	random 8	−0.060	−0.706
random 2	−0.008	−0.323	random 9	0.140	0.765
random 3	−0.066	−0.750	random 10	0.170	0.986
random 4	−0.192	−1.676	av of random 1−10	0.036	
random 5	−0.229	−1.948	std dev	0.136	
random 6	−0.081	−0.860			
random 7	0.096	0.441	*Actual*	*0.74*	*5.722*

for ten different random data sets are shown in Table 2. This table also contains the average q^2 value, the standard deviation of the q^2 values, and the Z score for the 60-descriptor kNN-QSAR model for the real data set. A Z score of 4.677 indicates that there is a probability of only about 10^{-5} that the kNN-QSAR model, constructed for the real estrogen data set, is random.

Atom Pairs as Molecular Descriptors. Figure 4 shows the q^2 vs nvar relationships when atom pairs were used to describe the molecular identities and calculate the molecular similarity. When the real activity values for estrogen receptor ligands were used in the kNN-QSAR analysis, the q^2 values were 0.70, 0.74, and 0.76 for the best 10-descriptor, 60-descriptor, and 120-descriptor models, respectively. The robustness of the kNN-QSAR models obtained using atom pair descriptors is apparent from Figure 4 and the results of standard hypothesis testing (Table 3). Overall, the q^2 values are very low for random data sets compared to those for the real data set. This suggests again that the kNN-QSAR model for the real data set is statistically distinguishable from those for random data sets. However, in the case of AP descriptors, the q^2 values for the real data set slightly increase with the increase in the number of descriptors.

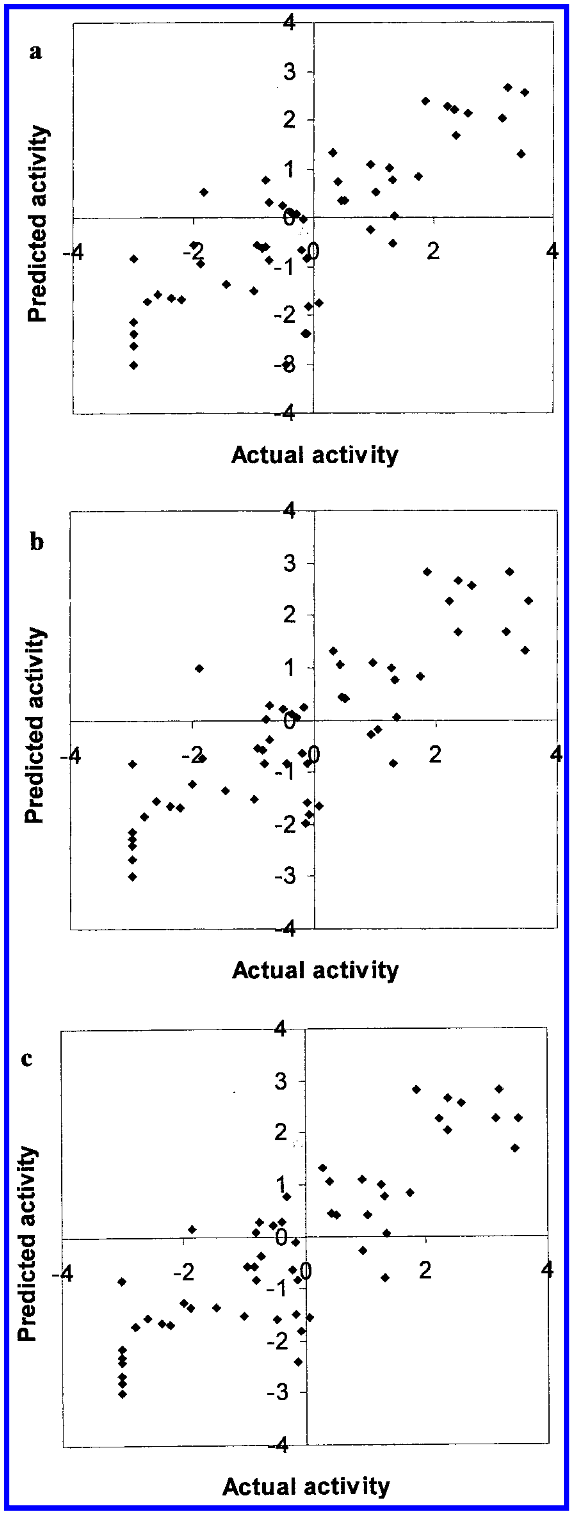

Figure 5. Plot of predicted vs actual activity obtained from 10-descriptor (a), 60-descriptor (b), and 120-descriptor (c) kNN-QSAR models for estrogen receptor ligands using AP as molecular descriptors.

The predicted vs actual activity as a result of the cross-validation analysis is shown in Figure 5a−c for 10-descriptor, 60-descriptor, and 120-descriptor models, respectively. It should be emphasized again that these figures show the results of prediction instead of fitting. These results demonstrate that the trend of the predicted values is similar to that of the real activity values. Table 2 lists the values of the actual activity, the predicted activity from all three models, and the associated absolute errors. Once again, the predictive power of kNN-QSAR models is better than that reported by Waller et al.[35] in terms of the predictions obtained

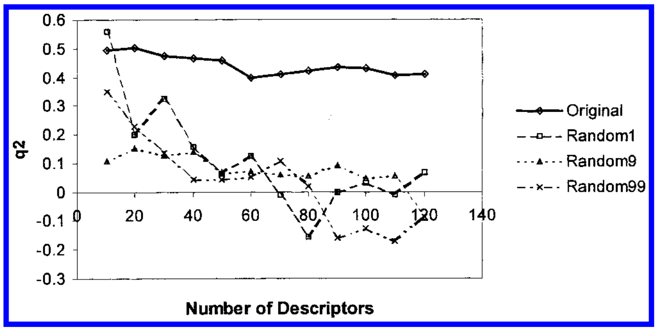

Figure 6. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for DHFR inhibitors using MCI as molecular descriptors. The results for both the actual data set and three random data sets are shown.

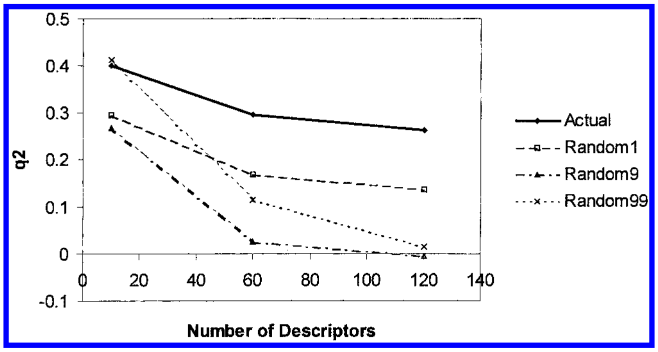

Figure 7. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for DHFR inhibitors using AP as molecular descriptors. The results for both the actual data set and three random data sets are shown.

in the cross-validation process for all models (cf. q^2 values of 0.70, 0.74, and 0.76 from this work vs 0.59 from Waller et al.).

Further statistical examination of the robustness of the kNN-QSAR models has been performed by the means of hypothesis testing. The q^2 values for the 60-descriptor models obtained for ten different random data sets are shown in Table 3. Also included in this table are the average q^2 value, the standard deviation of the q^2 values, and the Z score for the real 60-descriptor model. The value of this Z score (5.72) indicates that the probability of this kNN-QSAR model being random is less than 10^{-6}.

kNN-QSAR Analysis of Other Data Sets. To further demonstrate the effectiveness and generality of the kNN-QSAR method, three other data sets were analyzed. They included 23 DHFR inhibitors,[84] 60 AChE inhibitors,[44] and 14 5HT-receptor agonists.[85]

The compounds in each of the three data sets were represented by both MCI and AP descriptors. For each description method, the actual data were analyzed by kNN-QSAR using 10, 60, and 120 descriptors, respectively. To show the robustness of the derived models, three random data sets were generated in each case and also analyzed by kNN-QSAR method. For the DHFR data set, the q^2 vs nvar relationships obtained for both the actual data and random data sets are given in Figure 6 (using MCI descriptors) and Figure 7 (using AP descriptors), respectively. Similarly, for AChE inhibitors, the q^2 vs nvar values are plotted in Figure 8 and Figure 9 using MCI and AP descriptors, respectively, and those for the 5HT data set are given in Figure 10 and Figure 11 using MCI and AP descriptors, respectively.

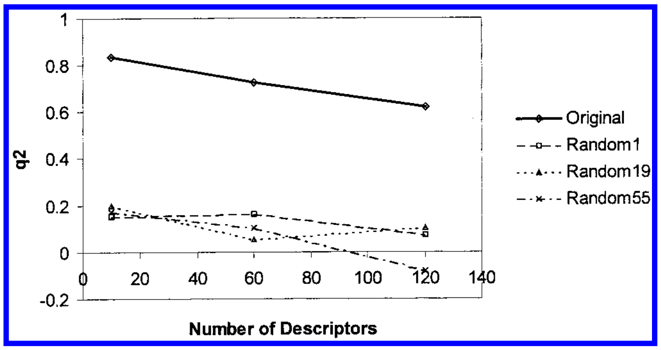

Figure 8. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for AChE inhibitors using MCI as molecular descriptors. The results for both the actual data set and three random data sets are shown.

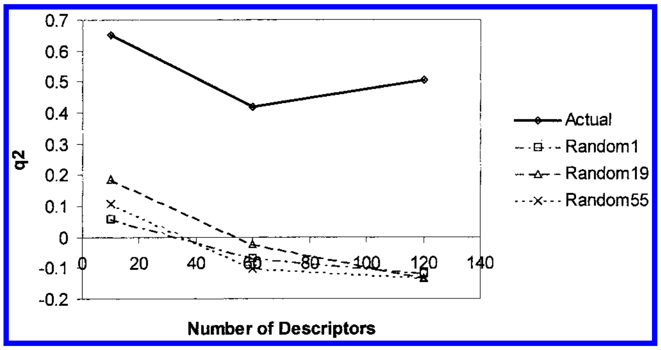

Figure 9. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for AChE inhibitors using AP as molecular descriptors. The results for both the actual data set and three random data sets are shown.

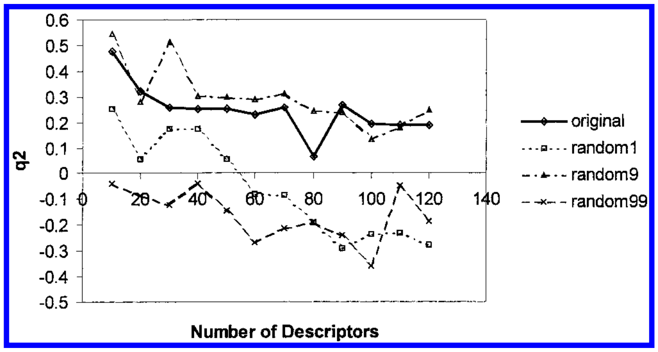

Figure 10. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for the 5HT data set using MCI as molecular descriptors. The results for both the actual data set and three random data sets are shown.

For the DHFR data set, none of the 10-descriptor models derived from both MCI and AP descriptors was any better than those generated from the random data. Therefore, the reliability of the 10-descriptor models is questionable. However, for both types of descriptors, 60-descriptor models and 120-descriptor models were clearly able to distinguish actual vs random data sets well. Furthermore, in this particular case, the MCI descriptor based models seem to work slightly better than those using AP descriptors in terms of the absolute values of q^2 (ca. 0.4 for MCI descriptors and 0.3 for AP descriptors).

In the case of AChE inhibitors, all models for the real data set were better than for the random data sets. It appears that in this case MCI description also works slightly better than AP description in terms of the q^2 values (0.6−0.82 for MCI description and 0.4−0.65 for AP description). Perhaps,

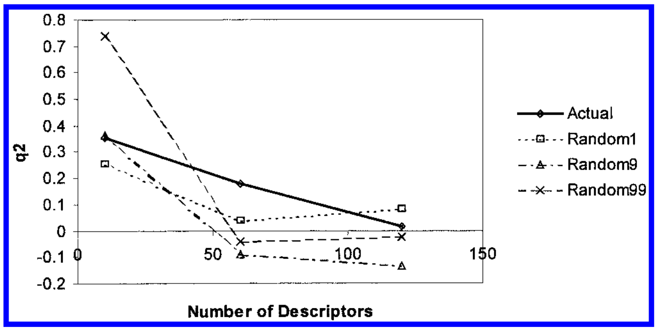

Figure 11. Plot of q^2 vs the number of descriptors (nvar) selected for the best kNN-QSAR model for the 5HT data set using AP as molecular descriptors. The results for both the actual data set and three random data sets are shown.

a more detailed classification of atom types in atom pair generation could help improve the quality of AP-based models. This point is under investigation.

The 5HT data set is unique: no reliable kNN-QSAR model could be developed, since the q^2 values obtained for the real data set were not better than those for random data sets. This may be due to the fact that the kNN-QSAR method was designed for analyzing relatively large data sets where a multitude of different classes of compounds are represented in the training set. Thus, the size of this data set may be insufficient for building a reliable model with the kNN-QSAR model since the 5HT data set contained only 14 compounds. In practical terms, this example has also demonstrated the importance of comparing the QSAR results obtained for an actual data set and those for random data sets to avoid spurious QSAR models.

CONCLUSIONS AND PROSPECTUS

We have developed a novel nonlinear QSAR technique that is based on the concept of molecular similarity and *K*-nearest neighbor principle. The philosophy of this method is straightforward and directly relies upon the active analogue similarity principle: since structurally similar compounds should have similar biological activities, then the activity of a compound can be predicted (or estimated) simply as the average of the activities of similar compounds. Other functional relationships can also be implemented such as those described by Hirst.[68]

The perception of structural similarity is relative and should always be considered in the context of a particular biological target. Since physicochemical characteristics of a receptor binding site vary from one target to another, the structural features that can best explain the observed biological similarities between compounds may be different for different biological sites of action. These critical structural features are defined in this work as topological pharmacophore (TP) for the underlying biological activity. Thus, the main task of building a kNN-QSAR model is to identify the best topological pharmacophore. This is achieved by the “bioactivity driven” variable selection, i.e., by selecting a subset of molecular descriptors that afford the most predictive kNN-QSAR model as judged by the cross-validated R^2. Since the number of all possible combinations of descriptors is huge, an exhaustive search of these combinations is not possible, and therefore stochastic optimization of variable selection should be applied. Because of the importance and difficulty of such optimization, much effort has gone into developing effective algorithms for finding good optima. The methods of simulated annealing,[82] genetic algorithms,[87] and taboo search[88] are three of the most popular techniques, inspired by the ideas from statistical mechanics, theory of evolutionary biology, and operations research, respectively. The generalized simulated annealing has been adapted in this paper for an efficient sampling of the combinatorial space. Other stochastic optimization algorithms will be investigated and implemented as well.

In most of the test cases, robust kNN-QSAR models have been obtained, which is supported by the results of the statistical hypothesis testing. One of the interesting aspects of the kNN-QSAR method is that it generates multiple QSAR models for a different (preselected) number of descriptors. Actually, this number can also be optimized in the course of the model development, along with the number of nearest neighbors (k) and the best choice of variables, which is the matter of further method development. Each model is both cross-validated and validated against random data sets with shuffled activity. Thus, multiple plausible QSAR models can be developed and validated, rather than a single best QSAR model. Although the concept of the method is simple, it works as well as, or even superior to, other known QSAR methods, e.g., CoMFA. While kNN-QSAR models do not provide any direct suggestions as to which modification should be made to the training set molecules to obtain more active compounds, it does provide an efficient way to predict activities of existing or newly designed compounds. The efficiency comes from the fact that no prealignment of molecules is needed during the search. Although molecular identity was characterized by MCI descriptors and 2D atom pairs descriptors in all the presented test cases, the kNN-QSAR technique is general enough to work with molecular field properties and many other descriptors, which is the subject of further investigations.

In the modern era of drug design by the means of combinatorial chemistry and high throughput screening, an unprecedented amount of experimental structure−activity relationship information has to be analyzed using adequate QSAR techniques. These techniques should be fast, automated, and applicable to large data sets of structurally diverse compounds where structure−activity correlations may not be described by conventional linear regression models. The kNN-QSAR method described in this paper can potentially satisfy these requirements. We shall continue to evaluate its actual range of applications, as larger data sets become available.

The authors gratefully acknowledge Professors L. B. Kier and L. H. Hall for providing the MOLCONN-X program, Drs. S. J. Cho and C. L. Waller for their interest in this work and helpful critical suggestions, and Tripos Assoc., Inc., for the software grant. W.Z. acknowledges the 1996 Award from the Chemical Structure Association Trust and the graduate assistantship from EPA/UNC Toxicology Research Program, Training Agreement No. T901915, with the Curriculum in Toxicology, UNC-CH.

REFERENCES AND NOTES

(1) Hansch, C.; Muir, R. M.; Fujita, T.; Maloney, P. P.; Geiger, E.; Streich, M. The Correlation of Biological Activity of Plant Growth Regulators

and Chloromycetin Derivatives with Hammett Constants and Partition Coefficients. *J. Am. Chem. Soc.* **1963**, *85*, 2817−2824.

(2) Fujita, T.; Iwasa, J.; Hansch, C. A New Substituent Constant, π, Derived from Partition Coefficients. *J. Am. Chem. Soc.* **1964**, *86*, 5175−5180.

(3) Hammett, L. P. Some Relations Between Reaction Rates and Equilibrium Constants. *Chem. Rev.* **1935**, *17*, 125−136.

(4) Hansch, C.; Leo, A. Exploring QSAR. In *Fundamentals and Applications in Chemistry and Biology*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1995.

(5) Hansch, C.; Leo, A.; Hoekman, D. In *Exploring QSAR: Hydrophobic, Electronic, and Steric Constants*; Heller, S. R., Ed.; American Chemical Society: Washington, DC, 1995.

(6) Verloop, A.; Hoogenstraaten, W.; Tipker, J. In *Drug Design*; Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol. VII, p 165.

(7) Randic, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.

(8) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Chemistry and Drug Research*; Academic Press: New York, 1976.

(9) Kier, L. B.; Hall, L. H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Chichester, England, 1986.

(10) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. In *Reviews in Computational Chemistry II*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: Cambridge, U.K., 1991; pp 367−422.

(11) Anker, L. S.; Jurs, P. C. Quantitative Structure−Retention Relationship Studies of Odor-Active Aliphatic Compounds with Oxygen-Containing Functional Groups. *Anal. Chem.* **1990**, *62*, 2676−2687.

(12) Jurs, P. C.; Ball, J. W.; Anker, L. S. Carbon-13 Nuclear Magnetic Resonance Spectrum Simulation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 272−278.

(13) Nelson, T. M.; Jurs, P. C. Prediction of Aqueous Solubility of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 601−609.

(14) Stanton, D. T.; Jurs, P. C. Computer−Assisted Study of the Relationship between Molecular Structure and Surface Tension of Organic Compounds. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 109−115.

(15) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

(16) Geladi, P.; Kowalski, B. R. Partial Least Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(17) Hellberg, S.; Sjostrom, M.; Skagerberg, B.; Wold, S. Peptide Quantitative Structure−Activity Relationships, a Multivariate Approach. *J. Med. Chem.* **1987**, *30*, 1126−1135.

(18) Hopfinger, A. J. A QSAR Investigation of Dihydrofolate Reductase Inhibition by Baker Triazines Based upon Molecular Shape Analysis. *J. Am. Chem. Soc.* **1980**, *102*, 7196−7206.

(19) Crippen, G. M. Distance Geometry Approach to Rationalizing Binding Data. *J. Med. Chem.* **1979**, *22*, 988−997.

(20) Crippen, G. M. Quantitative Structure−Activity Relationships by Distance Geometry: Systematic Analysis of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1980**, *23*, 599−606.

(21) Boulu, L. G.; Crippen, G. M. Voronoi Binding Site Models: Calculation of Binding Models and Influence of Drug Binding Data Accuracy. *J. Comput. Chem.* **1989**, *10*, 673−682.

(22) Holzbrabe, U.; Hopfinger, A. J.; Conformational Analysis, Molecular Shape Comparison, and Pharmacophore Identification of Different Allosteric Modulators of Muscarinic Receptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1018−1024.

(23) Rhyu, K.−B.; Patel, H. C.; Hopfinger, A. J. A 3D−QSAR Study of Anticoccidial Triazines Using Molecular Shape Analysis. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 771−778.

(24) Burke, B. J.; Dunn, W. J.; Hopfinger, A. J. Construction of a Molecular Shape Analysis-Three-Dimensional Quantitative Structure−Analysis Relationship for an Analog Series of Pyridobenzodiazepinone Inhibitors of Muscarinic 2 and 3 Receptors. *J. Med. Chem.* **1994**, *37*, 3775−3788.

(25) Hopfinger, A. J.; Burke, B. J.; Dunn, W. J. A Generalized Formalism of Three-Dimensional Quantitative Structure−Property Relationship Analysis for Flexible Molecules Using Tensor Representation. *J. Med. Chem.* **1994**, *37*, 3768−3774.

(26) Tokarski, J. S.; Hopfinger, A. J. Three-Dimensional Molecular Shape Analysis−Quantitative Structure−Activity Relationship of a Series of Cholecystokinin-A Receptor Antagonists. *J. Med. Chem.* **1994**, *37*, 3639−3654.

(27) Koehler, M. G.; Rowberg-Schaefer, K.; Hopfinger, A. J. A Molecular Shape Analysis and Quantitative Structure-Activity Relationship Investigation of Some Triazine-Antifolate Inhibitors of Leishmania Dihydrofolate Reductase. *Arch. Biochem. Biophys.* **1988**, *266* (1), 152− 161.

(28) Srivastava, S.; Crippen, G. M. Analysis of Cocaine Receptor Site Ligand Binding by Three-Dimensional Voronoi Site Modeling Approach. *J. Med. Chem.* **1993**, *36*, 3572−3579.

(29) Bradley, M. P.; Crippen, G. M. Voronoi Modeling: The Binding of Triazines and Pyrimidines to L. casei Dihydrofolate Reductase. *J. Med. Chem.* **1993**, *36*, 3171−3177.

(30) Smellie, A. S.; Crippen, G. M.; Richards, W. G. Fast Drug-Receptor Mapping by Site-Directed Distances: A Novel Method of Predicting New Pharmacological Leads. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 386−392.

(31) Boulu, L. G.; Crippen, G. M.; Barton, H. A. Voronoi Binding Site Model of a Polycyclic Aromatic Hydrocarbon Binding Protein. *J. Med. Chem.* **1990**, *33*, 771−775.

(32) Ghose, A. K.; Crippen, G. M.; Revankar, G. R. Analysis of the in Vitro Antiviral Activity of Certain Ribonucleosides against Parainfluenza Virus Using a Novel Computer Aided Receptor Modeling Procedure. *J. Med. Chem.* **1989**, *32*, 746−756.

(33) Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(34) Waller, C. L.; Juma, B. W.; Gray, L. E., Jr.; Kelce, W. R. Three-Dimensional Quantitative Structure−Activity Relationships for Androgen Receptor Ligands. *Toxicol. Appl. Pharmacol.* **1996**, *137* (2), 219−227.

(35) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H. K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, L. E., Jr. Ligand-Based Identification of Environmental Estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(36) Waller, C. L.; McKinney, J. D. Three-Dimensional Quantitative Structure−Activity Relationships of Dioxins and Dioxin-Like Compounds: Model Validation and Ah Receptor Characterization. *Chem. Res. Toxicol.* **1995**, *8*, 847−858.

(37) Waller, C. L.; Minor, D. L.; McKinney, J. D. Using Three-Dimensional Quantitative Structure−Activity Relationships to Examine Estrogen Receptor Binding Affinities of Polychlorinated Hydroxybiphenyls. *Environ. Health Perspect.* **1995**, *103*, 702−707.

(38) Oprea, T. I.; Waller, C. L.; Marshall, G. R. 3D-QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. III. Interpretation of CoMFA Results. *Drug Des. Discovery.* **1994**, *12* (1), 29−51.

(39) Oprea, T. I.; Waller, C. L.; Marshall, G. R. Three-Dimensional Quantitative Structure−Activity Relationship of Human Immunodeficiency Virus (I) Protease Inhibitors. 2. Predictive Power Using Limited Exploration of Alternate Binding Modes. *J. Med. Chem.* **1994**, *37*, 2206−2215.

(40) Waller, C. L.; Oprea, T. I.; Giolitti, A.; Marshall, G. R. Three-Dimensional QSAR of Human Immunodeficiency Virus (I) Protease Inhibitors. 1. A CoMFA Study Employing Experimentally-Determined Alignment Rules. *J. Med. Chem.* **1993**, *36*, 4152−4160.

(41) Waller, C. L.; Marshall, G. R. Three-Dimensional Quantitative Structure−Activity Relationship of Angiotesin-Converting Enzyme and Thermolysin Inhibitors. II. A Comparison of CoMFA Models Incorporating Molecular Orbital Fields and Desolvation Free Energies Based on Active-Analog and Complementary-Receptor-Field Alignment Rules. *J. Med. Chem.* **1993**, *36*, 2390−2403.

(42) Oprea, T. I.; Garcia, A. E. Three-Dimensional Quantitative Structure-Activity Relationships of Steroid Aromatase Inhibitors. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 186−200.

(43) Kim, K. H.; Martin, Y. C. Direct Prediction of Dissociation Constants (p*K*a's) of Clonidine-Like Imidazolines, 2-Substituted Imidazoles, and 1- Methyl-2-Substituted-Imidazoles from 3D Structures Using a Comparative Molecular Field Analysis (CoMFA) Approach. *J. Med. Chem.* **1991**, *34*, 2056−2060.

(44) Cho, S. J.; Garsia, M. L.; Bier, J.; Tropsha, A. Structure-Based Alignment and Comparative Molecular Field Analysis of Acetylcholinesterase Inhibitors. *J. Med. Chem.* **1996**, *39*, 5064−5071.

(45) Cho, S. J.; Tropsha, A.; Suffness, M.; Cheng, Y. C.; Lee, K. H. Antitumor Agents. 163. Three-Dimensional Quantitative Structure−Activity Relationship Study of 4′-*O*-Demethylepipodophyllotoxin Analogs Using the Modified CoMFA/q2-GRS Approach. *J. Med. Chem.* **1996**, *39*, 1383−1395.

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

(47) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.

(48) Kubinyi, H. Variable Selection in QSAR Studies. I. An Evolutionary Algorithm. *Quant. Struct.-Act. Relat.* **1994**, *13*, 285−294.

(49) Kubinyi, H. Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution. *Quant. Struct.-Act. Relat.* **1994**, *13*, 393−401.

(50) Luke, B. T. Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships and Quantitative Structure− Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.

(51) So, S. S.; Karplus, M. Evolutionary Optimization in Quantitative Structure−Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521−1530.

(52) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating Optimal Linear PLS Estimations (GOLPE): An Advanced Chemometric Tool for Handling 3D−QSAR Problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.

(53) Cho, S. J.; Tropsha, A. Cross-Validated R2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method To Achieve Consistent Results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(54) Kimura, T.; Hasegawa, K.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: GA-Based Region Selection for CoMFA Modeling, *J. Chem. Info. Comp. Sci.* **1998**, *38*, 276−282.

(55) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(56) So, S.-S.; Richards, W. G. Application of Neural Networks: Quantitative Structure−Activity Relationships of the Derivatives of 2,4-Diamino-5-(Substituted-Benzyl) Pyrimidines as DHFR Inhibitors. *J. Med. Chem.* **1992**, *35*, 3201−3207.

(57) Ajay. A Unified Framework for Using Neural Networks to Build QSARs. *J. Med. Chem.* **1993**, *36*, 3565−3571.

(58) Hirst, J. D.; King, R. D.; Sternberg, M. J. Quantitative Structure−Activity Relationships by Neural Networks and Inductive Logic Programming. I. The Inhibition of Dihydrofolate Reductase by Pyrimidines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 405−420.

(59) Hirst, J. D.; King, R. D.; Sternberg, M. J. Quantitative Structure−Activity Relationships by Neural Networks and Inductive Logic Programming. II. The Inhibition of Dihydrofolate Reductase by Triazines. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 421−432.

(60) Tetko, I. V.; Tanchuk, V. Yu.; Chentsova, N. P.; Antonenko, S. V.; Poda, G. I.; Kukhar, V. P.; Luik, A. I. HIV-1 Reverse Transcriptase Inhibitor Design Using Artificial Neural Networks. *J. Med. Chem.* **1994**, *37*, 2520−2526.

(61) Manallack, D. T.; Ellis, D. D.; Livingstone, D. J. Analysis of Linear and Nonlinear QSAR Data Using Neural Networks. *J. Med. Chem.* **1994**, *37*, 3758−3767.

(62) Maddalena, D. J.; Johnston, G. A. Prediction of Receptor Properties and Binding Affinity of Ligands to Benzodiazepine/GABA-A Receptors Using Artificial Neural Networks. *J. Med. Chem.* **1995**, *38*, 715−724.

(63) Bolis, G.; Pace, L.; Fabrocini, F. A Machine Learning Approach to Computer-Aided Molecular Design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 617−628.

(64) King, R. D.; Muggleton, S.; Lewis, R. A.; Sternberg, M. J. Drug Design by Machine Learning: The Use of Inductive Logic Programming to Model the Structure-Activity Relationships of Trimethoprim Analogues Binding to Dihydrofolate Reductase. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, (23), 11322−11326.

(65) King, R. D.; Muggleton, S.; Srinivasan, A.; Sternberg, M. J. Structure−Activity Relationships Derived by Machine Learning: The Use of Atoms and Their Bond Connectivities to Predict Mutagenicity by Inductive Logic Programming. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, (1), 438−442.

(66) Jain, A. N.; Dietterich, T. G.; Lathrop, R. H.; Chapman, D.; Critchlow, R. E., Jr.; Bauer, B. E.; Webster, T. A.; Lozano-Perez, T. A Shape-Based Machine Learning Tool for Drug Design. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 635−652.

(67) Peterson, K. L. Quantitative Structure−Activity Relationships in Carboquinones and Benzodiazepine Using Counter-Propagation Neural Networks. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 896−904.

(68) Hirst, J. D. Nonlinear Quantitative Structure−Activity Relationship for the Inhibition of Dihydrofolate Reductase by Pyrimidines. *J. Med. Chem.* **1996**, *39*, 3526−3532.

(69) Sharaf, M. A.; Illman, D. L.; Kowalski, B. R. *Chemometrics*; Wiley: New York, 1986.

(70) Hamamoto, Y.; Uchimura, S.; Tomita, S. A Bootstrap Technique for Nearest Neighbor Classifier Design. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 73−79.

(71) Djouadi, A.; Bouktache, E. A Fast Algorithm for the Nearest Neighbor Classifier. *IEEE Trans. Pattern Anal. Mach. Intell.* **1997**, *19*, 277−282.

(72) Strouf, O. *Chemical Pattern Recognition*; Research Studies Press: Chichester, England, **1986**.

(73) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting Conserved Water-Mediated and Polar Ligand Interactions in Proteins Using a K-Nearest-Neighbors Genetic Algorithm. *J. Mol. Biol.* **1997**, *265* (4), 445−464.

(74) Basak, S. C.; Grunwald, G. D. Tolerance Space and Molecular Similarity. SAR *QSAR Environ. Res.* **1995**, *3* (4), 265−277.

(75) Basak, S. C.; Bertelsen, S.; Grunwald, G. D. Use of Graph Theoretic Parameters in Risk Assessment of Chemicals. *Toxicol. Lett.* **1995**, *79*, 239−250.

(76) Basak, S. C.; Grunwald, G. D. Predicting Mutagenicity of Chemicals Using Topological and Quantum Chemical Parameters: A Similarity Based Study. *Chemosphere* **1995**, *31* (1), 2529−2546.

(77) Basak, S. C.; Grunwald, G. D. Estimation of Lipophilicity from Molecular Structural Similarity. *New J. Chem.* **1995**, *19* (2), 231.

(78) The program Sybyl is available from Tripos Associates, St. Louis, MO.

(79) Molconn-X version 2.0, Hall Associates Consulting, Quincy, MA.

(80) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(81) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087−1092.

(82) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671−680.

(83) Sun, L.; Xie, Y.; Song, X.; Wang, J.; Yu, R. Cluster Analysis By Simulated Annealing. *Comput. Chem.* **1994**, *18*, 103−108.

(84) Mabilia, M.; Pearlstein, R. A.; Hopfinger, A. J. Molecular Shape Analysis and Energetics-Based Intermolecular Modeling of Benzylpyrimidine Dihydrofolate Reductase Inhibitors. *Eur. J. Med. Chem.-Chem. Ther.* **1985**, *20*, 163.

(85) Agarwal, A.; Taylor, E. W. 3-D QSAR for Intrinsic Activity of 5-HT1A Receptor Ligands by the Method of Comparative Molecular Field Analysis. *J. Comput. Chem.* **1993**, *14*, 237−245.

(86) Gilbert, N. *Statistics*; W. B. Saunders, Co.: Philadelphia, PA, 1976.

(87) Forrest, S. Genetic algorithms: Principles of natural selection applied to computation. *Science* **1993**, *261*, 872−878.

(88) Cvijovic, D.; Klinowski, J. Taboo Search: An Approach to the Multiple Minima Problem. *Science* **1995**, *267*, 664−665.

CI980033M