Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation

Subhash Ajmani,† Kamalakar Jadhav, and Sudhir A. Kulkarni*

VLife Sciences Technologies Private Limited, 1 Akshay, 50 Anand Park, Aundh, Pune 411 007, India

Received April 12, 2005

In this paper we report a novel three-dimensional QSAR approach, kNN-MFA, developed based on principles of the k-nearest neighbor method combined with various variable selection procedures. The kNN-MFA approach was used to generate models for three different data sets and predict the activity of test molecules through each of these models. The three data sets used were the standard steroid benchmark, an antiinflammatory and an anticancerous data set. The study resulted in kNN-MFA models having better statistical parameters than the reported CoMFA models for all the three data sets. It was also found that stochastic methods generate better models resulting in more accurate predictions as compared to stepwise forward selection procedures. Thus, kNN-MFA method represents a good alternative to CoMFA-like methods.

1. INTRODUCTION

Many different approaches to QSAR have been developed over the years. The rapid increase in three-dimensional structural information (3D) of bioorganic molecules, coupled with the development of fast methods for 3D structure alignment (e.g. active analogue approach), has led to the development of 3D structural descriptors and associated 3D QSAR methods. The most popular 3D QSAR methods are comparative molecular field analysis (CoMFA)[1] and comparative molecular similarity analysis (CoMSIA).[2] The CoMFA method involves generation of a common three-dimensional lattice around a set of molecules and calculation of the steric and electrostatic interaction energies at the lattice points. The interaction energies are numerically very high when a lattice point is very close to an atom and special care needs to be taken in order to avoid problems arising because of this. The CoMSIA method avoids these problems by using similarity function represented as Gaussian. This information around the molecule is converted into numerical data using the partial least squares (PLS) method that reduces the dimensionality of data by generating components. However, a major disadvantage is that PLS attempts to fit a linear curve among all the points in the data set. Further, the PLS method does not offer scope for improvement in results. It has been observed from several reports that the predictive ability of PLS method is rather poor due to fitting of a linear curve between the available points. In the case of the CoMSIA method, molecular similarity is evaluated and used instead of molecular field, followed by PLS analysis.

Recent trends in 2D/3D QSAR have focused on the development of procedures that allow selection of optimal variables from the available pool of descriptors of chemical structures i.e., ones that are most meaningful and statistically significant in terms of correlation with biological activity. This is accomplished by combining one of the stochastic search methods such as simulated annealing, genetic algorithms, or evolutionary algorithms with the correlation methods such as MLR, PLSR, or artificial neural networks.[3−8] Since the effectiveness and convergence of these algorithms are greatly affected by the choice of fitting function, several such functions have been used to improve their performance.[5,6] Since these techniques involve optimization of many parameters, the speed of the resulting analysis is relatively slow as compared to simple regression methods.

Variable selection methods have also been adopted for optimal region selection in 3D QSAR methods and shown to provide improved QSAR models as compared to the original CoMFA technique. For example, GOLPE[9] was developed using chemometric principles, and q^2-GRS was developed on the basis of independent analyses of small areas (or regions) of near-molecular space to address the issue of optimal region selection in CoMFA.[10]

These considerations provide an impetus for the development of fast, generally nonlinear, variable selection methods for performing molecular field analysis. We report here the development of a new method (kNN-MFA) that adopts a k-nearest neighbor principle for generating relationships of molecular fields with the experimentally reported activity. This method utilizes the active analogue principle that lies at the foundation of medicinal chemistry. The next section provides a detailed discussion of the kNN-MFA methodology. The methodology was tested on three different data sets. The chemical interpretation of results hence obtained is provided in section 3 followed by concluding remarks.

2. METHODOLOGY

Like many 3D QSAR methods,[1,2] k-nearest neighbor molecular field analysis (kNN-MFA) requires suitable alignment of given set of molecules. This is followed by generation of a common rectangular grid around the molecules. The steric and electrostatic interaction energies are computed at the lattice points of the grid using a methyl probe of charge +1. These interaction energy values are considered for relationship generation and utilized as descriptors to decide nearness between molecules. The term descriptor is

* Corresponding author phone/fax: +91-20-2588-6737; e-mail: sudhirk@vlifesciences.com.

† Current address: Centre for Molecular Design, Institute of Biomedical and Biomolecular Science, University of Portsmouth, King Henry 1 Street, Portsmouth PO1 2DY, U.K.

10.1021/ci0501286 CCC: $33.50
Published on Web 07/23/2005

utilized in the following discussion to indicate field values at the lattice points.

The optimal training and test sets were generated using the sphere exclusion algorithm.[11] This algorithm allows the construction of training sets covering descriptor space occupied by representative points. Once the training and test sets were generated, kNN methodology was applied to the descriptors generated over the grid.

2.1. k-Nearest Neighbor (kNN) Method. The kNN methodology relies on a simple distance learning approach whereby an unknown member is classified according to the majority of its k-nearest neighbors in the training set. The nearness is measured by an appropriate distance metric (e.g., a molecular similarity measure calculated using field interactions of molecular structures). The standard kNN method is implemented simply as follows:[12] (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; and (3) classify object u with the group to which the majority of the k objects belongs. An optimal k value is selected by optimization through the classification of a test set of samples or by leave-one out cross-validation. The variables and optimal k values were chosen using different variable selection methods as described below.

2.2. kNN-MFA with Simulated Annealing. Simulated annealing (SA) is the simulation of a physical process, 'annealing', which involves heating the system to a high temperature and then gradually cooling it down to a preset temperature (e.g., room temperature). During this process, the system samples possible configurations distributed according to the Boltzmann distribution so that at equilibrium, low energy states are the most populated.

The SA kNN-MFA method employs the kNN classification principle combined with the SA variable selection procedure. For each predefined number of variables (Vn) 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 molecule and (ii) the selection of variables from the original pool of all molecular descriptors that are used to calculate similarities between molecules (i.e., distances in Vn-dimensional descriptor space).

The implementation of SA kNN-MFA reported here is similar to that described in ref 13 and can be summarized as follows.

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

(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 kNN-MFA model.

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

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

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

$$\text{rnd} < e^{-(q^2\text{curr} - q^2\text{new})/T}$$

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

(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.[14] Thus, when a new solution is accepted or when a preset number of successive steps of generating trial solutions (20 steps) do not lead to a better result, the temperature is lowered by 10% (the default initial temperature is 1000 K). The calculations are terminated, when either the current temperature of simulations reaches 10^{-6} K or the ratio between the current temperature and the temperature corresponding to the best solution found equals 10^{-6}.

2.3. kNN-MFA with Stepwise (SW) Variable Selection. This method employs a stepwise variable selection procedure combined with kNN to optimize (i) the number of nearest neighbors (k) and (ii) the selection of variables from the original pool as described in simulated annealing. The step-by-step search procedure begins by developing a trial model with a single independent variable and adds independent variables, one step at a time, examining the fit of the model at each step (using weighted kNN cross-validation procedure described in section 2.5). The method continues until there are no more significant variables remaining outside the model.

2.4. kNN-MFA with Genetic Algorithm. Genetic algorithms (GA) first described by Holland[15] mimic natural evolution and selection. In biological systems, genetic information that determines the individuality of an organism is stored in chromosomes. Chromosomes are replicated and passed onto the next generation with selection criteria depending on fitness. Genetic information can however be altered through genetic operations such as mutation and crossover. In GAs, each "chromosome" is a set of genes, which constitutes a candidate solution to the discrimination problem. A population of "chromosomes" is used. The passage of each "chromosome" to the next generation is determined by its relative fitness, i.e., the closeness of its properties to those desired. Random combinations and/or changes of the transmitted "chromosomes" produce variations in the next generation of "offspring". Better the fitness (correspondence with desired properties), greater is the chance of that chromosome being selected for transmission. Optimal or near optimal solutions are obtained through evolution over many generations. There are four major components of GA: chromosome generation, fitness assessment, selection, and mutation.

This method employs a stochastic variable selection procedure, combined with kNN, to optimize (i) the number of nearest neighbors (k) and (ii) the selection of variables from the original pool as described in simulated annealing.

The implementation of GA based kNN-MFA involved the following steps:

(1) Generate the initial population of chromosomes (candidate solutions) by randomly selecting genes (descriptors) from the pool of available genes.

(2) Calculate pairwise Euclidean distances for all pair of molecules with respect to each chromosome.

(3) Calculate the fitness of each chromosome using a weighted kNN cross-validation procedure described in section 2.5.

(4) Select chromosomes for mating pool by roulette wheel selection.

(5) Apply uniform crossover and mutation operations on the mating pool chromosomes to create a new population of offspring.

(6) Calculate fitness of each offspring using a weighted kNN cross-validation procedure.

(7) Replace the least fit chromosomes in an initial population with the best offspring.

(8) Repeat steps 2−7 until the convergence criteria or the maximum number of generations is reached.

2.5. Cross-Validation Using Weighted k-Nearest Neighbor. The standard leave-one-out procedure was implemented as described in ref 13 and can be summarized as follows.

(1) A molecule in the training set was eliminated, and its biological activity was predicted as the weighted average activity of the *k* most similar molecules (eq 1). The similarities were evaluated as the inverse of Euclidean distances between molecules (eq 2) using only the subset of descriptors corresponding to the current trial solution.

$$w_i = \frac{\exp(-d_j)}{\sum_{\text{k-nearest neighbors}} \exp(-d_j)}$$

$$\hat{y}_i = \sum w_i y_i \tag{1}$$

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

(2) Step 1 was repeated until every molecule in the training set has been eliminated and its activity predicted once.

(3) The cross-validated r^2 (q^2) value was calculated using eq 3, where y_i and $\hat{y}_i$ are the actual and predicted activities of the *i*th molecule, respectively, and y_{mean} is the average activity of all molecules in the training set. Both summations are over all molecules in the training set. Since the calculation of the pairwise molecular similarities, and hence the predictions, were based upon the current trial solution, the q^2 obtained is indicative of the predictive power of the current kNN-MFA model.

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2} \tag{3}$$

(4) Steps 1−3 were repeated for $k = 2, 3, 4$, etc. Formally, the upper limit of *k* is the total number of molecules in the data set. However, the best value has been empirically found to lie between 1 and 5. The *k* value that led to the highest q^2 value was chosen for the current kNN-MFA model.

2.6. External Validation. The following procedure was applied for external validation.

(1) Predict the biological activity of a molecule in the test set as the weighted average activity of the *k* most similar molecules in the training set (eq 1). The similarities were evaluated as the inverse of Euclidean distances between molecules (eq 2) as calculated using the descriptors determined by the current model.

(2) Step 1 was repeated for every molecule in the test set.

(3) The predicted r^2 (pred_r^2) value was calculated using eq 4, where y_i and $\hat{y}_i$ are the actual and predicted activities of the *i*th molecule in test set, respectively, and y_{mean} is the average activity of all molecules in the training set. Both summations are over all molecules in the test set. The pred_r^2 value is indicative of the predictive power of the current kNN-MFA model for external test set.

$$\text{pred\_r}^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2} \tag{4}$$

2.7. Randomization Test. To evaluate the statistical significance of the QSAR model for an actual data set, we have employed a one-tail hypothesis testing.[13,16] 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 rearranging biological activities of the training set molecules. The significance of the models hence obtained was derived based on calculated Zscore.[13,16]

2.8. Evaluation of the QSAR Models. The QSAR models were evaluated using following statistical measures: *n*, number of observations (molecules); Vn, number of descriptors; *k*, number of nearest neighbors; q^2, cross-validated r^2 (by the leave-one-out method); pred_r^2, predicted r^2 for the external test set; Zscore, the Z score calculated by q^2 in the randomization test; best_ran_q^2, the highest q^2 value in the randomization test; and α, the statistical significance parameter obtained by the randomization test.

2.9. Data Sets. To assess the utility of approach, three data sets are considered for comparison of kNN-MFA models with CoMFA. The data sets used are (i) one of the most standard data sets for 3D QSAR i.e., steroids,[1] (ii) the data set of antiinflammatory selective COX-2 inhibitors,[17] and (iii) the data set of anticancer molecules.[18]

2.10. Alignment Rules. It should be noted that the original papers have already reported CoMFA results for these data sets. We have used the same set of alignment rules that have been reported earlier and generated the CoMFA as well as kNN-MFA models and compared them. It should be noted that we have not attempted different alignment rules for the chosen data sets since two of the data sets, viz. steroids and anticancer molecules have rigid templates, whereas for COX-2 inhibitors the alignment rule used was as described in the original paper.

3. RESULTS AND DISCUSSION

The importance and utility of the new three-dimensional QSAR method discussed herein has been established by applying it to known sets of molecules as described above. We hereby report the models, as generated by kNN-MFA in conjunction with simulated annealing (SA), genetic algorithm (GA), and stepwise (SW) forward variable selection methods and compare these models with the CoMFA models for these data sets. In the kNN-MFA method, several models were generated for the given or selected members of training and test sets, and the corresponding best models are reported herein. The method described above has been

Table 1. Comparison of the CoMFA and kNN-MFA Models for the Steroid Data Set Using the Reported Test Set[1]

parameter/molecule	actual	CoMFA	SA kNN-MFA	SW kNN-MFA	GA kNN-MFA
q^2		0.80	0.95	0.89	0.93
pred_r^2		0.56	0.78	0.58	0.85
Zscore			2.49	8.59	5.21
best_rand_q^2			0.23	−0.18	0.10
α			0.01	>0.001	>0.001
k/Vn			2/6	3/2	2/8
descriptors			S231, E219, E324, E166, S125, S132	S454, S167	E142, E166, E219, E252, E370, E379, E443, S393
test1	−7.512	−7.661	−7.639	−7.880	−7.548
test2	−7.553	−7.902	−7.685	−7.837	−7.760
test3	−6.779	−6.793	−7.312	−6.452	−6.577
test4	−7.200	−7.335	−6.535	−7.758	−7.739
test5	−6.144	−6.500	−6.162	−6.308	−5.838
test6	−6.247	−7.155	−6.180	−7.758	−6.280
test7	−7.120	−7.211	−7.793	−7.758	−7.534
test8	−6.817	−7.082	−7.583	−6.309	−7.551
test9	−7.688	−7.721	−7.691	−7.880	−7.683
test10	−5.797	−7.496	−5.335	−6.308	−6.278

Table 2. Comparison of the kNN-MFA Models for the Steroid Data Set Using the Optimal Test Set of 9 Molecules

parameter/molecule	actual	SA kNN-MFA	SW kNN-MFA	GA kNN-MFA
q^2		0.94	0.94	0.86
pred_r^2		0.87	0.72	0.86
Zscore		5.058	6.50	4.29
best_rand_q^2		−0.031	−0.17	0.20
α		>0.001	>0.001	>0.001
k/Vn		3/9	3/5	4/6
descriptors		S195, E96, E166, S203, E219, E403, E193, S245, E285	E453, E230, E52, E112, E298	E50, E96, E219, E230, E346, S393,
17OHpregnenlone	−5.00	− 5.241	−5.154	−5.113
cortisol	−7.881	− 7.516	−7.693	−7.636
cortisone	−6.892	− 6.416	−7.333	−6.746
deoxycorticosterone	−7.653	− 7.650	−7.539	−7.314
dihydrotestosterone	−5.919	− 6.349	−5.053	−6.551
hydroxyprog	−7.740	− 6.933	−6.595	−6.959
test5	−6.144	− 6.346	−6.709	−6.560
test7	−7.120	− 7.251	−7.278	−7.129
test8	−6.817	− 7.033	−7.258	−7.130

implemented in a software, VLife Molecular Design Suite (VLifeMDS),[19] which allows user to choose probe, grid size, and grid interval for the generation of descriptors. The variable selection methods along with the corresponding parameters are allowed to be chosen, and optimum models are generated by maximizing q^2.

(i) Steroid Data Set. First used by Cramer et al. to establish the usefulness of the CoMFA method, this data set has been used as a benchmark in many 3D QSAR studies.[20−23] The steric and electrostatic fields were calculated using the Tripos force field and Gasteiger−Marsili charges. The alignment was performed by the rigid body least-squares fitting of the 3, 5, 6, 13, 14, and 17 carbon atoms of each molecule to the corresponding atoms of deoxycortisol. A training set of 21 molecules, and a test set of 10 molecules was used as described in the original paper.[1] The q^2, pred_r^2, Vn, and *k* value of kNN-MFA with SA, GA, and SW variable selection methods were (0.95, 0.78, 6, 2), (0.93, 0.85, 8, 2), and (0.89, 0.58, 3, 2), respectively. The corresponding CoMFA results for an optimal number of 3 components are as follows: $q^2 = 0.80$ and pred_r$^2 = 0.56$. Thus, there is a significant improvement in the predictive ability of all the kNN-MFA methods as compared to CoMFA (Table 1). The kNN-MFA methods, in general, display higher prediction accuracies for activities for the molecules in the test set as compared to CoMFA. The points that are common in the SA and GA kNN-MFA models are E166 and E219, i.e., electrostatic interaction field at lattice points 166 and 219, implying that these points are indeed significant for the structure−activity relationship.

The results in Table 1 use the reported[1] training and test set thereby enabling comparison with CoMFA. However, as discussed previously, the sphere exclusion method can be used for choosing the optimal training and test set by using different dissimilarity values. The optimal training and test set resulted in a significant improvement in the statistical parameters of QSAR for SA kNN-MFA (pred_r$^2 = 0.87$) and SW kNN-MFA ($q^2 = 0.94$ and pred_r$^2 = 0.72$). The comparison of actual and predicted activities for the optimal test set is shown in Table 2.

(ii) Antiinflammatory Selective COX-2 Inhibitor Data Set. As described by Desiraju et al.,[17] a series of 1,5-diarylpyrazoles was used to build the CoMFA and kNN-MFA models. The steric fields were generated using the Tripos force field, and electrostatic fields were generated using MOPAC charges obtained from PM3 optimized structures of the molecules. The alignment of the molecules was done based on the common fragment of 1,5-diphenylpyrazole. As described in the original paper,[17] the training set consisted of 25 molecules, whereas the test set contained 5 molecules. The q^2, pred_r^2, Vn, and *k* value of kNN-MFA with SA, GA, and SW variable selection methods were (0.85,

Table 3. Comparison of the CoMFA and kNN-MFA Models for the Antiinflammatory Data Set Using the Reported Test Set[17]

parameter/molecule	actual	CoMFA	SA kNN-MFA	SW kNN-MFA	GA kNN-MFA
q^2		0.68	0.85	0.82	0.82
pred_r^2		0.68	0.81	0.89	0.90
Zscore			4.53	2.52	3.78
best_rand_q^2			−0.17	0.58	0.11
α			>0.001	0.01	0.001
k/Vn			2/10	2/7	2/6
descriptors			S404, E358, E352, S329, E216, E373, S495, E464, S257, S384	E373, S329, S494, E216, E146, S487, S404	E286, E464, S243, S323, S329, S495,
DE26	2.000	1.201	1.282	1.503	1.462
DE27	1.400	1.230	1.634	1.324	1.320
DE28	−0.420	−1.401	−1.001	−0.995	−0.999
DE29	2.330	1.834	1.879	2.099	2.283
DE30	1.300	0.813	1.634	1.628	1.411

Table 4. Comparison of the kNN-MFA Models for the Antiinflammatory Data Set Using the Optimal Test Set of 8 Molecules

parameter/molecule	actual	SA kNN-MFA	SW kNN-MFA	GA kNN-MFA
q^2		0.87	0.71	0.82
pred_r^2		0.90	0.88	0.90
Zscore		3.94	2.83	5.03
best_rand_q^2		0.14	0.02	−0.16
α		0.001	0.01	>0.001
k/Vn		2/6	2/3	2/6
descriptors		E465, S323, S329, S495, E216, E136	E463, S285, S474	E163, E375, S243, S329, S495, S533
DE02	1.570	1.396	1.000	1.397
DE05	0.920	1.122	0.879	1.174
DE08	1.250	1.237	1.201	1.247
DE10	0.960	0.638	0.974	0.873
DE14	2.050	2.205	1.617	2.390
DE23	1.570	1.253	1.239	1.248
DE24	1.680	1.251	1.237	1.250
DE26	2.000	1.395	1.547	1.396

Table 5. Comparison of the CoMFA and kNN-MFA Models for the Anticancer Data Set Using the Reported Test Set[18]

parameter/molecule	actual	CoMFA	SA kNN-MFA	SW kNN-MFA	GA kNN-MFA
q^2		0.69	0.96	0.82	0.86
pred_r^2		−0.72	0.72	0.65	0.70
Zscore			2.84	4.06	4.65
best_rand_q^2			0.02	0.00	−0.27
α			0.01	>0.001	>0.001
k/Vn			2/9	2/4	2/6
descriptors			E373, S419, S339, S475, S239, E346, S321, S437, S230	E538, E645, S645, E383	E331, E346, E485, E575, S359, S636
P-IIIc	0.850	0.980	0.719	0.690	0.690
P-IIIe	0.270	1.070	0.100	0.112	0.325
P-IIIf	1.050	0.591	0.739	1.024	0.777
P-IIIg	0.110	0.560	0.103	0.468	0.154
P-IIIh	0.770	1.016	0.740	0.729	0.745
P-IIIi	0.390	0.961	0.700	0.728	0.777

0.81, 10, 2), (0.82, 0.90, 6, 2), and (0.82, 0.89, 7, 2), respectively, as shown in Table 3. The corresponding CoMFA results for an optimal number of 6 components were as follows: $q^2 = 0.68$ and pred_r$^2 = 0.68$. The common descriptors among all kNN-MFA models are S329 and S495, whereas the points that are common among two of the methods are S404, E216, E373, and E464. This clearly indicates that the variables selected using different variable selection methods are indeed important for generating QSAR relationships.

The use of the sphere exclusion method for creating new training and test sets leads to significant improvement in SA kNN-MFA i.e., pred_r$^2 = 0.90$ with 6 descriptors, as compared to the training and test sets reported in the original paper (Table 4).

(iii) Anticancer Data Set. Suh et al.[18] have described the comparison of CoMFA, CoMSIA, and HQSAR models for a series of 1-N-substituted imidazoquinoline-4,9-dione derivatives. The molecules were optimized using the Tripos force field with Gasteiger−Huckel charges and a distance dependent dielectric constant. The alignment of the molecules was done based on the common substructure of imidazoquinoline-4,9-dione. A training set of 16 molecules and test set of 6 molecules was used as described by Suh et al.[18] Table 5 shows the comparison of statistical results and predicted activities of the test set using the kNN-MFA methods and the CoMFA method. The q^2, pred_r^2, Vn, and *k* value of kNN-MFA with SA, GA, and SW were (0.96, 0.72, 9, 2), (0.86, 0.70, 6, 2), and (0.82, 0.65, 4, 2), respectively. The corresponding CoMFA results for an

Table 6. Comparison of the kNN-MFA Models for the Anticancer Data Set Using the Optimal Test Set of 5 Molecules

parameter/molecule	actual	SA kNN-MFA	SW kNN-MFA	GA kNN-MFA
q^2		0.92	0.86	0.88
pred_r^2		0.84	0.57	0.84
Zscore		2.383	5.30	4.37
best_rand_q^2		0.18	−0.23	−0.20
α		0.01	>0.001	>0.001
k/Vn		2/10	2/9	2/9
descriptors		S150, E331, S645, E319, S142, S231, E580, E671, E346, E644	S635, E535, E517, S239, E547, E319, S241, E485, E331	E322, E331, E346, E381, E618, S132, S241, S247, S645
A04	1.000	1.032	1.047	1.008
A08	0.850	0.614	0.364	0.627
A09	0.680	0.624	0.304	0.646
A13	−0.290	−0.178	−0.150	−0.186
PIIIF	1.050	0.700	0.702	0.690

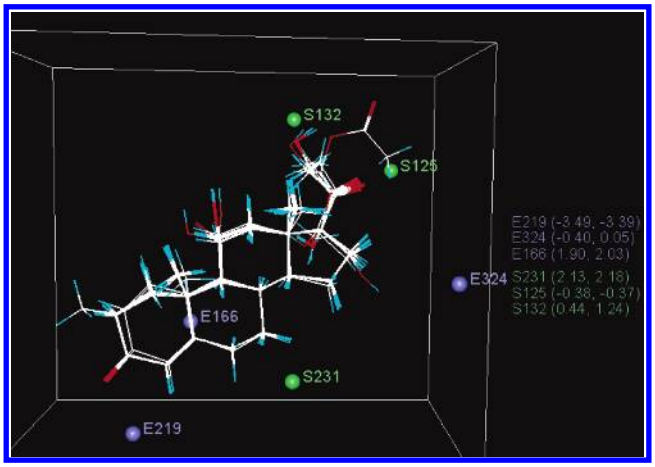

Figure 1. Distribution of chosen points in the SA kNN-MFA for the steroid data set with the reported test set molecules.

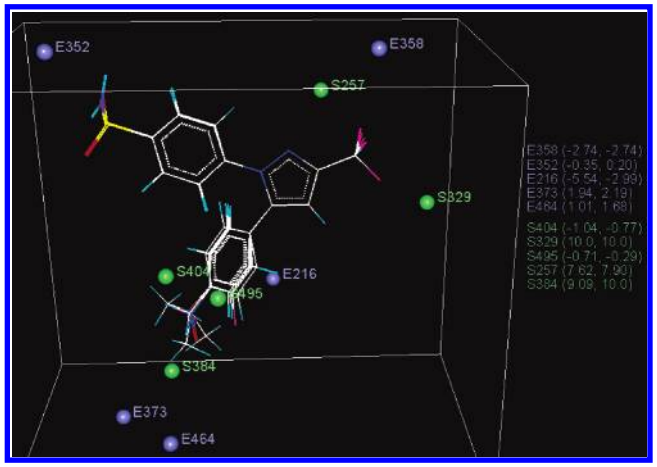

Figure 2. Distribution of chosen points in the SA kNN-MFA for the antiinflammatory data set with the reported test set molecules.

optimal number of 4 components were as follows: $q^2 = 0.69$ and pred_r$^2 = -0.72$. The predicted activities of the CoMFA model were rather inferior as compared to the kNN-MFA models. Although there are no common descriptors among these three methods, E346 gets selected with the simulated annealing and genetic algorithm variable selection methods. By using the sphere exclusion method to create a training set (17) and a test set (5) of molecules (cf. Table 6), SA and GA kNN-MFA models showed significant improvement in the pred_r^2 value as compared to the standard test set given in Table 5. Furthermore, these models show E331 as a common descriptor, whereas S645 and E346 were found common among SA and GA models. In addition, E319 and S241 were found common in two of the models.

For all the data sets, the sphere exclusion method generated several different training and test sets by varying dissimilarity value.[11] The kNN-MFA models with a reasonable number of test set members were chosen as the optimal one for the corresponding data set (cf. Tables 2, 4, and 6). For the steroid data set with the SA kNN method, 55 models were generated, and the optimal models with optimum q^2 and pred_r^2 for different training and test sets are shown in Figure 4. Similarly for the antiinflammatory and anticancer data sets, 121 and 77 models were generated, respectively. Figures 5 and 6 show the corresponding optimum q^2 and pred_r^2 for different training and test sets.

It is known that the CoMFA method provides significant value in terms of a new molecule design, when contours of the PLS coefficients are visualized for the set of molecules.

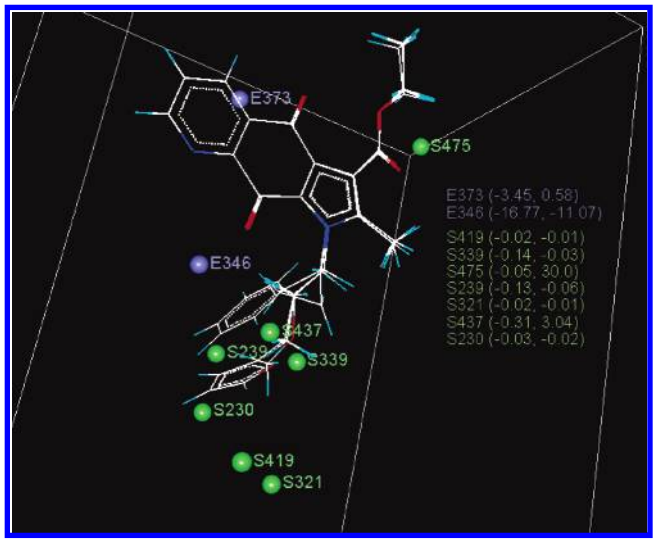

Figure 3. Distribution of chosen points in the SA kNN-MFA for the anticancer data set with the reported test set molecules.

Similarly, the kNN-MFA models provide direction for the design of new molecules in a rather convenient way. The points which contribute to the SA kNN-MFA models in all three data sets are displayed in Figures 1−3. The range of property values for the chosen points may aid in the design of new potent molecules (Figures 1−3). The range is based on the variation of the field values at the chosen points using the most active molecule and its nearest neighbor set. For the steroid data set with the reported[1] test set, there are six

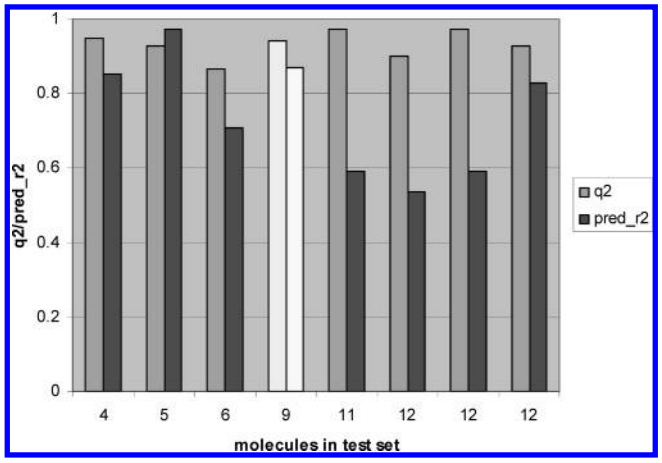

Figure 4. Optimal SA-kNN-MFA models for the steroid data set with different test sets and corresponding q^2 and pred_r^2 values. The model reported in Table 2 is shown in white.

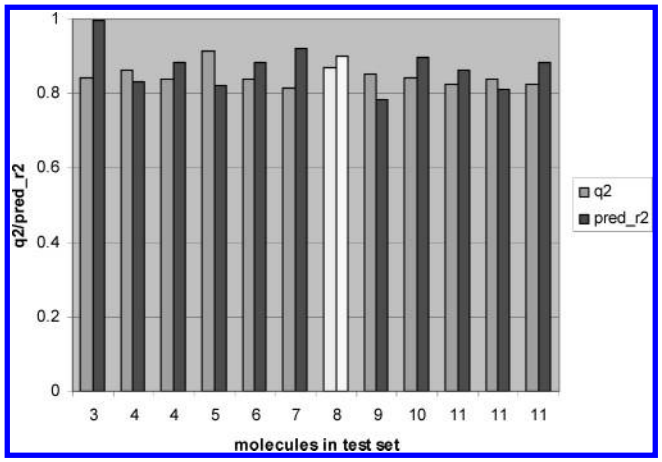

Figure 5. Optimal SA-kNN-MFA models for the antiinflammatory data set with different test sets and corresponding q^2 and pred_r^2 values. The model reported in Table 4 is shown in white.

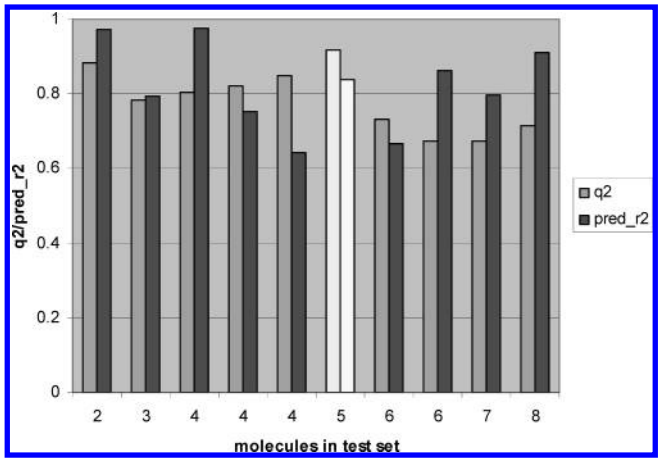

Figure 6. Optimal SA-kNN-MFA models for the anticancer data set with different test sets and corresponding q^2 and pred_r^2 values. The model reported in Table 6 is shown in white.

points for which descriptor distances are crucial for the SA kNN-MFA model; three points contribute through steric and three through electrostatic interactions. Some of the points are away from molecular framework and yet contribute to the model. The location and field values of these points can be used for the design of novel and better molecules. The distance of 10 points in the case of the antiinflammatory data set (Figure 2) and nine points in case of the anticancer data set (Figure 3) contribute to the SA kNN-MFA model. These points show regions that are important for variation in activity of these data sets.

Some of the underlying differences between kNN-MFA and conventional regression methods as well as limitations of this approach should also be noted. The kNN-MFA method does not assume a linear relationship between the dependent and independent variables and hence fitting a single linear equation is not required.

For predicting the activity of a molecule, regression methods use the following equation

$$\text{activity} = C_0 + C_1D_1 + C_2D_2 + ... + C_ND_N$$

where C_i's are coefficients and D_i's are descriptors.

In the case of the kNN-MFA method, the activity of a molecule is predicted using

$$\text{activity} = C_1A_1 + C_2A_2 + ... + C_kA_k$$

where C_i's are weights and A_i's are activities of the k-nearest neighbors in the training set. The nearest neighbors of any molecule are obtained from calculating the distance between the descriptors selected from various variable selection methods, described above.

Thus, kNN-MFA prediction uses an interpolative method, and hence predicted activities of new designed molecules will be within the range of activities of molecules in training set. Since the kNN method is based on distances of descriptors, their interpretation is quite difficult compared to the regression models. Although several models are generated by the kNN-MFA method, the time required for obtaining results is significantly more than for the CoMFA method. Typically SA and GA based variable selection methods require more time by a factor of 50−100 for any particular choice of training and test sets. However for SW variable selection method, the time required is comparable to CoMFA.

4. CONCLUDING REMARKS

A novel three-dimensional QSAR approach has been developed based on the principles of the k-nearest neighbor method. The method employs different variable selection procedures with either simulated annealing, stepwise forward, or genetic algorithm. For the three data sets reported in this study, it can be seen that stochastic methods generate better models with higher prediction accuracy as compared to the stepwise forward selection procedure. The regions shown by selected descriptors in the kNN-MFA models are similar to those obtained from the CoMFA models. The location and range of function values at the field points selected by the models provide clues for the design of new molecules. This method is expected to provide a good alternative for the generation of 3D QSAR models. The main disadvantage of the kNN based method as compared to CoMFA is the time required to generate models. As in case of CoMFA, this method too is highly dependent on the choice of structural alignment of molecules and does not address CoMFA's most limiting factor of working with a single alignment to model dynamic molecules. Hence, all the disadvantages and limitations resulting from inappropriate

choice of alignment in CoMFA are propagated in the kNN-MFA method as well. However, the method generates several models in a single run which gives a wider choice for model selection.

REFERENCES AND NOTES

(1) Cramer, R. D.; 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.

(2) Klebe, G.; Abraham, U.; Mietzner, T. *J. Med. Chem.* **1994**, *37*, 24−30.

(3) 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.

(4) 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.

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

(6) 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.

(7) 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.

(8) 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.

(9) 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.

(10) 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.

(11) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 144−154

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

(13) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the k-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

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

(15) Holland, J. *Adaptation in Natural and Artificial Systems*; University of Michigan Press: 1975.

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

(17) Desiraju, G. R.; Gopalakrishnan B.; Jetti, R. K. R.; Raveendra, D.; Sarma, J. A. R. P.; Subramanya, H. S. Three-Dimensional Quantitative Structural Activity Relationship (3D-QSAR) Studies of Some 1,5-Diarylpyrazoles: Analogue Based Design of Selective Cyclooxygenase-2 Inhibitors. *Molecules* **2000**, *5*, 945−955.

(18) Suh, M. E.; Park, S. Y.; Lee, H. J. Comparison of QSAR Methods (CoMFA, CoMSIA, HQSAR) of Anticancer 1-N-Substituted Imidazoquinoline-4,9-dione Derivatives. *Bull. Korean Chem. Soc.* **2002**, *23*, 417−422.

(19) VLifeMDS2.0; Molecular Design Suite, Vlife Sciences Technologies Pvt. Ltd., Pune, India, 2004 (www.vlifesciences.com).

(20) Jain, A. N.; Koile, K.; Chapman, D. Compass: predicting biological activities from molecular surface properties. Performance comparisons on a steroid benchmark. *J. Med. Chem.* **1994**, *37*, 2315−27.

(21) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-organizing molecular field analysis: a tool for structure−activity studies. *J. Med. Chem.* **1999**, *42*, 573−83.

(22) Stiefl, N.; Baumann, K. Mapping property distributions of molecular surfaces: algorithm and evaluation of a novel 3D quantitative structure−activity relationship technique. *J. Med. Chem.* **2003**, *46*, 1390−407.

(23) Polanski, J.; Gieleciak, R.; Magdziarz, T.; Bak, A. GRID formalism for the comparative molecular surface analysis: application to the CoMFA benchmark steroids, azo dyes, and HEPT derivatives. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1423−1435.

CI0501286