vLife
adding value to life

Advantage of k-Nearest Neighbor Method for developing QSAR models

By

Dr. Subhash Ajmani

VLife Sciences Technologies Pvt Ltd.

Objective

The main objective of the present study is develop a single predictive QSAR model of a diverse set molecules from a wide variety of 11 structural classes inhibiting HIV integrase enzyme using k-Nearest Neighbour method.

Another objective is to develop and compare two kNN-QSAR models with reported two QSAR models formed by dividing the overall set of molecules into two clusters (cluster 1 and cluster 2).

Method

In k-nearest neighbor algorithm, for classifying a new pattern (molecule), the system finds the K nearest neighbors among the training set, and uses the categories of the k-nearest neighbors to weight the category candidates [1]. The nearness is measured by an appropriate distance metric (e.g., a molecular similarity measure, calculated using descriptors of molecular structures). The standard k-NN method is implemented simply as follows:

(1) calculate Euclidean 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;

(3) classify object u with the group to which a majority of the K objects belongs.

Literature reports QSAR models of a diverse set of molecules from a wide variety of structural classes inhibiting HIV integrase enzyme [2]. The biological activity of molecules was expressed as IC50 for 3í-processing, which range from 0.1µm to greater than 300µm. For QSAR analysis the biological activity was converted to log (1/IC50) value. This work reports QSAR analysis on two clusters formed by dividing the overall 11 structural classes of molecules into two clusters using hierarchical cluster analysis. QSAR models were generated for each cluster using genetic algorithm based method (GFA) and had q^2 values of 0.71 and 0.74 and pred_r^2 values of 0.65 and 0.78. The study also reported a single QSAR model using whole set of diverse molecules (11 classes), however, the model was not statistically significant (q^2 = 0.418).

In present work we have attempted to develop a single predictive QSAR model using

k-NN principle. This model may be useful for initial screening of biological activity of library of molecules (in absence of knowledge about their binding site). Two k-NN QSAR models (Cluster 1 and Cluster 2) were also developed using two clusters of molecules as reported by Yuan and Parrill [2].

1 Generation of Molecular Descriptors

All chemical structures and their descriptors [3-9] i.e. molecular connectivity indices (MCI), electrotopological indices (EI), alignment independent (AI) descriptors and other 2D descriptors such as logP (partition coefficient), number of hydrogen bond donor and number of hydrogen bond acceptor etc. were calculated by using VLifeMDS software [10]. MCI descriptors are calculated on the basis of chemical graph theory. AI descriptors are calculated as discussed in Baumannís paper [11]. In all 368 molecular descriptors were calculated for the molecules.

2 Generation of Training and Test Set (Sphere exclusion algorithm)

Calculated molecular descriptors have been used for the development of QSAR models. The whole data set was divided into training and test sets using sphere exclusion algorithm as described by Golbraikh and Tropsha [12]. This algorithm allows constructing training sets covering all descriptor space areas occupied by representative points.

3 Simulated Annealing k-NN QSAR Algorithm

This method uses simulated annealing variable selection and k-NN principle to build QSAR model. Development of SA k-NN QSAR method was done using algorithm as described in the paper by Zheng and Tropsha [13]. The parameter settings used for simulated annealing in the present study are as follows:

Temperature range from 1000-10-6; Decrement of temperature by 10 %

Number of terms in model varies from 2-20 with step size of 2

Perturbation of 1 term (2 and 3 terms perturbation was also used but no statistically significant improvement was observed);

4 Stepwise k-NN QSAR Algorithm

This method uses stepwise forward variable selection and k-NN principle to build QSAR model. In each step this method optimize (i) the number of nearest neighbors (k) used to estimate the activity of each compound and (ii) select variables (stepwise) from the original pool of all molecular descriptors that are used to calculate similarities between molecules.

The stepwise k-NN QSAR method involves the following steps.

(1) A step-by-step search procedure that begins by addition of a single independent variable with optimal k value (optimizing k value from the given range of k values) and highest sum of weighted k-nearest neighbor cross validation (q2) and external validation (pred_r2) (as described below) value amongst all available descriptors to form a model.

(2) Later on, in each iteration of this method, an independent variable gets added along with optimization of k value (using given range of k values) and examining the fit of the model using q2 and pred_r2 until there are no more significant variables remaining outside the model.

5 Cross-Validation (q^2) using weighted k-Nearest Neighbor

Following procedure as described in reference [13] was applied for cross validation.

(1) Eliminate a compound in the training set and predict its biological activity on the basis of the k-NN principle, i.e., as the weighted average activity of k most similar molecules (k is set to 1 initially) (eq 1). The similarities are evaluated as Euclidean distances between molecules (eq 2) using only the subset of descriptors that corresponds to the current model.

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

$$y_i = \sum w_i y_i \qquad (1)$$

$$d_{i,j} = \left[\sum_{k=1}^{n\,\mathrm{var}} (X_{ik} - X_{jk})^2\right]^{1/2} \qquad (2)$$

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

(3) Calculate the cross-validated r2 (q2) value using eq 3, where yi and ŷi are the actual and predicted activities of the ith compound, respectively, and ymean is the average activity of all molecules in the training set. Both summations are over all molecules in the training set. The obtained q^2 value is indicative of the predictive power of the current k-NN QSAR model in predicting molecules in training set.

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

(4) Repeat steps 1-3 for k = 2, 3, 4, etc. Formally, the upper limit of k is the total number of molecules in the data set. The k value that leads to the highest q^2 value is chosen for the current k-NN QSAR model.

6 External Validation (pred_$r^{2)}$ using weighted k-Nearest Neighbor

Following procedure was applied for external validation.

(1) Predict biological activity of a compound in the test set on the basis of the k-NN principle, i.e., as the weighted average activity of k (that corresponds k value for highest q^2 value) most similar molecules in the training set (eq 1 shown in cross validation). The similarities are evaluated as Euclidean distances between molecules (eq 2 shown in cross validation) using only the subset of descriptors that corresponds to the current model (for highest q^2 value).

(2) Repeat step 1 for every compound in the test set.

(3) Calculate the predicted r^2 (pred_r^2) value using eq 4, where yi and ŷi are the actual and predicted activities of the ith compound in test set, respectivel y, 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 obtained pred_r^2 value is indicative of the predictive power of the current k-NN QSAR model for external test set.

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

7 Randomization Test

To evaluate statistical significance of QSAR model for an actual data set, we have employed a one-tail hypothesis testing [13, 14]. 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 obtained is based on calculated Z score [13, 14].

8 Evaluation of QSAR model

Generated QSAR models were evaluated by following statistical measures:

n	number of observation (molecules)
nvar	number of terms (descriptors)
k	number of nearest neighbor
q^2	cross validated r^2 (by leave one out method)
pred_r^2	predicted r^2 for external test set
Zscore_q^2	Z score calculated by q^2 in randomization test
best_ran_q^2	highest q^2 value in the randomization test
α	Statistical significance parameter obtained by randomization test

Results and Discussion

1 k-NN QSAR of whole data set

The diverse set of 167 was divided into training set of 83 and test set of 84 molecules using all calculated descriptors and a dissimilarity value of 1.5 using sphere exclusion algorithm [12].

The QSAR models were developed using two variable selection methods viz. stepwise forward and simulated annealing coupled with k-NN method. At first, k value was varied from 1 to 5 and all the calculated descriptors was

subjected to stepwise k-NN (SW k-NN) method and the resulting model is reported in table 1. Further, SA k-NN QSAR algorithm was applied (after setting all parameters as described in SA k-NN QSAR algorithm above) along with k value varying from 1 to 5 and subjecting all the calculated descriptors that resulted in model with optimal statistical parameters as reported in table 1.

Table 1: QSAR models developed by stepwise and simulated annealing k-NN methods for whole data set of HIV intergrase inhibitors

	SW k-NN Model_WholeData	**SA k-NN Model_WholeData**
N (train/test)	83/84	83/84
nvar	6	10
k	3	2
q^2	0.785	0.817
pred_r^2	0.738	0.749
Zscore_q^2	10.303	5.970
best_ran_q^2	-0.227	-0.259
α	10^{-12}	10^{-8}
Descriptors	TOPO2O3,TOPO2N3, kappa3, H-DonorCount,3ClusterCount, TOPO2Cl5	TOPOCCl7, TOPOCCl5, TOPOCS1, TOPO3N7, TOPONO5, TOPOOO0, TOPOCC7, TOPO3O4, TOPOCN3, StNE-index

In contrast with the reports by Yuan and Parrill, we could develop a single model using k-NN method with the help of 2D-descriptors used herein. The model obtained by SW k-NN is statistically comparable with SA k-NN model. The number of variables have reduced from ten to six in case of SW k-NN compared to SA k-NN.

2 k-NN QSAR of Cluster 1

The diverse set of 83 molecules (in Cluster 1 as reported in [2]) was divided into training set of 36 and test set of 47 molecules using all calculated descriptors and a dissimilarity value of 2.0 using sphere exclusion algorithm [12].

The QSAR models were developed using two variable selection methods viz. stepwise forward and simulated annealing (all parameters set as described in SA k-NN QSAR algorithm above) coupled with k-NN method by varying k value from 1 to 5 and using all the calculated descriptors. The QSAR models with the important descriptors and the associated statistical parameters are reported in table 2.

Table 2: QSAR models developed by stepwise and simulated annealing k-NN methods for cluster 1 (as reported by Yuan and Parrill [2]) of data set of HIV intergrase inhibitors.

	SW k-NN Model	**SA k-NN Model**
N (train/test)	36/47	36/47
nvar	8	10
k	3	4
q^2	0.753	0.729
pred_r^2	0.705	0.695
Zscore_q^2	6.227	6.009
best_ran_q^2	0.004	-0.141
α	10^{-8}	10^{-8}
Descriptors	kappa1, TOPOOO3, TOPO2N1, StNE-index, TOPOOO0, chiV3Cluster, SssNHE-index, SaaOE-index	5ChainCount, StNE-index, 3ClusterCount, TOPONO1, TOPOOO4, SssNHcount,chiV3Cluster, SddsN(nitro)count, SddsN(nitro)E-index, k2alpha

Using molecules of cluster 1 in the study of Yuan and Parrill [2], we could obtain better QSAR models using both SW and SA k-NN methods compared to those reported (q^2 = 0.71, pred_r^2 =0.65) by them. The number of descriptors used has reduced from ten (for SA k-NN) to eight (for SW k-NN) with improved statistical parameters.

Table 3: QSAR models developed by stepwise and simulated annealing k-NN methods for cluster 1 (as reported by Yuan and Parrill [2]) of data set of HIV intergrase inhibitors.

	SW k-NN Model	**SA k-NN Model**
N (train/test)	37/47	37/47
nvar	8	16
k	3	2
q^2	0.850	0.885
pred_r^2	0.844	0.840
Zscore_q^2	5.785	5.860
best_ran_q^2	0.092	-0.131
α	10^{-6}	10^{-8}
Descriptors	TOPO222, TOPO227, SssSE-index, TOPO3N3, SssOcount, TOPOCS2, SsCH3count, SsOHcount	TOPOOS4, TOPO2Br4, TOPO2N5, TOPOCS3, TOPOCC5, TOPO2S6, TOPOSS4, TOPO3N5, TOPOCCl5, TOPONO5, TOPO3N4, SssCH2 E- index, SdssS(sulfone) E-index, SssOcount, TOPOOS2, TOPONS2

Using molecules of cluster 2 [2], we could obtain better k-NN QSAR models using both SW and SA k-NN methods compared to those reported (q^2 = 0.74, pred_r^2 =0.78) by them. The number of descriptors used has reduced from sixteen (for SA k-NN) to eight (SW k-NN) with comparable statistical parameters.

4 Summary

The k-NN approach used in this study resulted in single predictive QSAR model for a chemically diverse set of HIV integrase inhibitors that could not be developed earlier using conventional QSAR, viz. GFA [2]. The reported models were developed using easy and fast to calculate 2D molecular descriptors and thus may be useful for initial virtual screening of library of molecules in the absence of knowledge about their binding site. In addition, this study led to two individual k-NN QSAR models for Cluster 1 and Cluster 2 which are better than the reported QSAR models.

In the present study the models were developed using stepwise and simulated annealing variable selection procedures coupled with k-NN and it is observed that the descriptors selected by both methods are mostly different suggesting that all the descriptors in generated models are important for the significance of QSAR model. However, if any descriptor is selected in both methods, it suggests that the nearness of that descriptor plays important role for prediction of activities and hence may be critically important while designing new molecules. It should be noted that ultimate effect of the distance of a descriptor (which it contributes to overall distance) is reflected in terms of its contribution towards significance of QSAR (q^2 and pred_r^2 value) in explaining the variation of activity.

5 References

[1] *C. D. Manning, H. Schutze, Foundations of Statistical Natural Language Processing [M]. Cambridge: MIT Press., 1999.*

[2] *H.Yuan, A. L. Parrill, QSAR studies of HIV-1 Integrase Inhibition. Bio. Med. Chem. 10 (2002) 4169 – 4183.*

[3] *L.B. Kier, L.H. Hall, The Nature of Structure-Activity Relationships and their Relation to Molecular Connectivity. Eur. J. Med. Chem. 12 (1977) 307.*

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

[5] *L.B. Kier, A Shape Index from Molecular Graphs. Quant. Struct-Act. Relat. 4 (1985) 109.*

[6] *L.H. Hall, B.K. Mohney, L.B. Kier, The Electrotopological State: Structure Information at the Atomic Level for Molecular Graphs. J. Chem. Inf. Comput. Sci. 31 (1991) 76.*

[7] *L.H. Hall, L.B. Kier, Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. J. Chem. Inf. Comput. Sci. 35 (1995) 1039-1045.*

[8] *R. Wang, Y. Fu, L. Lai, A new atom-additive method for calculating partition coefficients. J. Chem. Inf. Comput. Sci. 37 (1997) 615-621.*

[9] *S.A. Wildman, G.M. Crippen, Prediction of Physicochemical Parameters by Atomic Contributions. J. Chem. Inf. Comput. Sci. 39 (1999) 868-873.*

[10] *MDS 1.0, Molecular Design Suite, VLife Sciences Technologies, Pvt. Ltd. Pune, India, 2003. See www.vlifesciences.com*

[11] *K. Baumann, An Alignment-Independent Versatile Structure Descriptor for QSAR and QSPR Based on the Distribution of Molecular Features. J. Chem. Inf. Comput. Sci.. 42 (2002) 26-35.*

[12] *A. Golbraikh, A. Tropsha, QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. J. Chem. Inf. Comput. Sci. 43 (2003) 144-154.*

[13] *W. Zheng, A. Tropsha, Novel variable selection quantitative structure property relationship approach based on the k-nearest-neighbor principle. J. Chem. Inf. Comput. Sci. 40 (2000) 185-194.*

[14] *N. Gilbert, Statistics, W.B. Saunders, Co.; Philadelphia, PA, 1976.*