Implementing the Fisher's Discriminant Ratio in a *k*-Means Clustering Algorithm for Feature Selection and Data Set Trimming

Thy-Hou Lin,* Huang-Te Li, and Keng-Chang Tsai

Institute of Molecular Medicine & Department of Life Science, National Tsing Hua University,
Hsinchu, Taiwan 30013, R.O.C.

Received June 26, 2003

The Fisher's discriminant ratio has been used as a class separability criterion and implemented in a *k*-means clustering algorithm for performing simultaneous feature selection and data set trimming on a set of 221 HIV-1 protease inhibitors. The total number of molecular descriptors computed for each inhibitor is 43, and they are scaled to lie between 1 and 0 before being subjected to the feature selection process. Since the purpose is to select some of the most class sensitive descriptors, several feature evaluation indices such as the Shannon entropy, the linear regression of selected descriptors on the pK_i of selected inhibitors, and a stepwise variable selection program are used to filter them. While the Shannon entropy provides the information content for each descriptor computed, more class sensitive descriptors are searched by both the linear regression and stepwise variable selection procedures. The inhibitors are divided into several different numbers of classes. They are subsequently divided into five classes due to the fact that the best feature selection result is obtained by the division. Most of the good features selected are the topological descriptors, and they are correlated well with the pK_i values. The outliers or the inhibitors with less class-sensitive descriptor values computed for each selected descriptor are identified and gathered by the *k*-means clustering algorithm. These are the trimmed inhibitors, while the remaining ones are retained or selected. We find that 44% or 98 inhibitors can be retained when the number of good descriptors selected for clustering is three. The descriptor values of these selected inhibitors are far more class sensitive than the original ones as evidenced by substantial increasing in statistical significance when they are subjected to both the SYBYL CoMFA PLS and Cerius^2 PLS regression analyses.

INTRODUCTION

Feature selection is a procedure used to select the most important features out of a set of features so that their number is reduced and at the same time their class discriminatory information is retained as much as possible. The selection stage is crucial. The subsequent design of a classifier would lead to poor performance if one selected features with little discrimination power. On the other hand, if information- rich features are selected, the design of the classifier can be greatly simplified. For example, a problem common to the construction of a three-dimensional quantitative structure−activity relationship (3D-QSAR) model[1] such as using the partial least squares (PLS)[2] to derive a linear equation from the Comparative Molecular Field Analysis (CoMFA)[3] result is how to select the most informative variables and eliminate the background noise so as to increase the signal-to-noise ratio. There are several selection methods aimed at the selection of field variables that have been proposed for CoMFA modeling. Lindgren et al.[4] proposed interactive variable selection (IVS) as a chemometric technique. In their algorithm, variables are selected according to the weight value in the PLS model. Baroni et al.[5] proposed Generating Optimal Linear PLS Estimations (GOLPE) for 3D-QSAR studies. GOLPE uses fractional factorial designs to generate several PLS models with different combinations of variables, and the ones significantly contributing to the prediction are selected. Cho et al.[6] proposed cross-validated r^2 guided region selection (q^2-GRS) for CoMFA modeling. q^2-GRS divides the CoMFA box into many small boxes, and a separate CoMFA is performed for each box. Selection for further analysis is performed for boxes of greater q^2 values than a specified threshold. Recently, a genetic algorithm-based region selection method (GARGS) has been proposed by Hasegawa et al.[7] A combination of genetic algorithm and PLS is used in this method to select the subdivided CoMFA regions which give the best prediction result.

Feature selection is also used in searching correlations between molecular descriptors and chemical properties or biological activities of molecules of various types.[8] To yield a mathematically well-behaved relationship, the number of adjustable parameters must be kept small as compared to the number of compounds involved. The selected descriptors are often those found to obtain the best mathematical model after performing many trial-and-error runs.[9−11] In a more quantitative description, one needs to select descriptors leading to large between-class distance and small within-class variance in the descriptor vector space.[12] This means that descriptors should take distant values in different classes and closely located values in the same class. The descriptors are examined individually, and those with little discriminatory capability are discarded. Sometimes the application of a linear or nonlinear transformation to a descriptor vector is required to create a new one with better discriminatory properties.

* Corresponding author fax: 886-3-571-5934; e-mail: thlin@life.nthu.edu.tw.

10.1021/ci030295a CCC: $27.50 © 2004 American Chemical Society
Published on Web 11/12/2003

Table 1. 221 HIV-1 Protease Inhibitors Studied

class	inhibitor original ID number (pK_i, structure unit number)				
1	086(10.96, 3)	089(10.92, 3)	117(10.92, 4)	102(10.85, 3)	652(10.85, 6)
	088(10.80, 3)	115(10.80, 4)	069(10.70, 3)	083(10.70, 3)	087(10.70, 3)
	116(10.64, 4)	103(10.62, 3)	100(10.60, 3)	152(10.60, 5)	146(10.62, 6)
	085(10.57, 3)	638(10.57, 6)	28(10.52, 8)	146(10.48, 5)	145(10.43, 5)
	067(10.41, 3)	068(10.41, 3)	15(10.40, 8)	070(10.35, 3)	111(10.33, 4)
	110(10.28, 4)	079(10.20, 3)	096(10.19, 3)	071(10.18, 3)	109(10.16, 4)
	655(10.13, 5)	36(10.10, 8)	12(10.05,15)		
2	098(9.96, 3)	31(9.96, 8)	052(9.92, 1)	053(9.92, 1)	051(9.85, 1)
	18(9.85, 8)	097(9.74, 3)	32(9.70, 8)	43(9.70,14)	078(9.68, 3)
	090(9.61, 3)	084(9.59, 3)	25(9.57, 8)	054(9.55, 1)	059(9.55, 2)
	082(9.54, 3)	3(9.52,15)	9(9.52,15)	13(9.52,15)	066(9.48, 2)
	050(9.47, 1)	35(9.40, 8)	28(9.40,14)	2(9.40,15)	114(9.39, 4)
	049(9.38, 1)	075(9.37, 3)	073(9.24, 3)	056(9.22, 2)	10(9.22,15)
	11(9,22,15)	102(9.15, 4)	24(9.15, 7)	107(9.07, 4)	037(9.05, 1)
	48(9.05,14)	064(9.03, 2)	065(9.00, 2)	22(9.00, 7)	
3	055(8.96, 2)	23(8.92, 7)	6(8.92,15)	7(8.92,15)	023(8.89, 1)
	004(8.85, 1)	035(8.85, 1)	039(8.85, 1)	057(8.85, 2)	5(8.85,15)
	058(8.82, 2)	005(8.80, 1)	046(8.80, 1)	105(8.72, 4)	022(8.68, 1)
	060(8.64, 2)	41(8.64,14)	048(8.55, 1)	027(8.52, 1)	034(8.52, 1)
	019(8.47, 1)	063(8.44, 2)	8(8.42,15)	024(8.37, 1)	006(8.34, 1)
	038(8.28, 1)	038(8.28, 1)	061(8.28, 2)	042(8.24, 1)	283550(8.24,12)
	283143(8.22,12)	283490(8.22,12)	282664(8.19,10)	21(8.19, 7)	062(8.16, 2)
	013(8.15, 1)	041(8.15, 1)	282664(8.15,10)	018(8.14, 1)	283005(8.12,10)
	283366(8.11,12)	2(8.11, 9)	003(8.10, 1)	20(8.10, 7)	282981(8.08,10)
	282456(8.07,10)	6(8.04, 9)	283209(8.02,12)	283265(8.02,12)	029(8.01, 1)
	19(8.01, 7)	282540(8.00,10)			
4	282453 (7.99,10)	283055(7.99,10)	283010(7.98,10)	282822(7.97,10)	282350(7.92,10)
	012(7.92, 1)	283568(7.92,12)	282939(7.89,10)	283263(7.89,12)	282351(7.87,10)
	282835(7.85,10)	282714(7.82,10)	282756(7.82,10)	282796(7.82,10)	283374(7.82,12)
	283353(7.80,12)	17(7.77,13)	282730(7.77,11)	282558(7.77,10)	282916(7.74,10)
	282828(7.68,10)	021(7.66, 1)	043(7.66, 1)	282779(7.64,10)	282826(7.64,10)
	18(7.64,13)	282547(7.62,10)	282915(7.62,10)	283239(7.59,12)	040(7.57, 1)
	282529(7.57,10)	282713(7.57,10)	282978(7.57,10)	283489(7.55,11)	014(7.52, 1)
	282632(7.52,10)	282944(7.48,10)	283522(7.48,12)	033(7.47, 1)	282423(7.47,10)
	283441(7.47,12)	016(7.44, 1)	283052(7.44,10)	025(7.43, 1)	282349(7,38,10)
	282390(7.35,10)	011(7.31, 1)	044(7.29, 1)	020(7.22, 1)	282479(7.21,10)
	14(7.21,13)	283336(7.10,12)	42(7.09,14)	031(7.07, 1)	030(7.05, 1)
	283356(7.04,12)	002(7.00, 1)			
5	015(6.96,10)	283245(6.89,12)	282749(6.86,10)	282364(6.86,11)	37(6.86,14)
	028(6.84, 1)	047(6.80, 1)	282396(6.80,10)	282969(6.74,10)	15(6.70,13)
	283364(6.64,12)	036(6.62, 1)	007(6.59, 1)	283051(6.30,10)	39(6.30,14)
	282834(6.19,10)	40(6.18,14)	283406(6.17,12)	282389(6.13,10)	283520(6.11,12)
	008(6.10, 1)	283573(6.03,12)	282823(5.96,10)	009(5.96, 1)	283337(5.92,12)
	282807(5.85,10)	282808(5.77,10)	045(5.73, 1)	283567(5.72,12)	47(5.70,14)
	282967(5.64,10)	282658(5.60,10)	283497(5.60,12)	282832(5.57,10)	283186(5.43,11)
	026(5.40, 1)	283481(5.30,14)	38(5.24,12)	001(5.24, 1)	46(5.15,14)

In this work, we have used the Fisher's discriminant ratio[13] to perform feature selection on some 43 descriptors computed for a group of 221 human immunodeficiency virus type-1 (HIV-1) protease inhibitors.[14−21] The 43 descriptors computed included those of topologic, geometric, electronic, and 3D-QSAR features. A *k*-means clustering algorithm[22] is then employed to cluster compounds within each class for each descriptor selected. There are two clusters generated for each class, and the cluster with smaller descriptor values is identified. This is the bad cluster where some outliers are probably gathered inside. An inhibitor is considered as an outlier and needed to be discarded if it is identified in all the bad clusters for all the descriptors selected. This is a concomitant feature selection and data set trimming process since both the bad features (descriptors) and inhibitors with less class-sensitive descriptor values computed are eliminated during the process. Two feature selection criteria, namely, the stepwise variable selection program MREG described by Jurs[23] and the direct fitting of the selected descriptors to a linear relationship with the measured activity in pK_i are also computed to monitor the progress of the process. The process gives gradual reduction in both the number of descriptors and inhibitors while enhancing the significance of both CoMFA PLS[24] and Cerius^2 PLS[25] regression statistics on the retained ones. It is also found that the pK_i values of the retained inhibitors correlate well with some finally selected topologic descriptors.

METHODS

The original number of structures constructed for the HIV-1 protease inhibitors was 345, and the construction procedures were detailed previously.[26] These structures were screened using the SYBYL MOPAC program,[24] and the given default settings, i.e., structures that were too large to be fitted into the program, were abandoned. The number of structures finally screened was 221, and both the original designations and basic structure units corresponding to each of them were given in Table 1 and Figure 1. The alignment for these structures was based on the coordinates of some atoms in the three basic structures described previously.[26] The treatment by SYBYL MOPAC[24] gave the following seven descriptors: E_{HOMO} (energy of highest occupied molecular orbital), E_{LUMO} (energy of lowest unoccupied molecular orbital), μ (dipole moment), H_f (heat of formation),

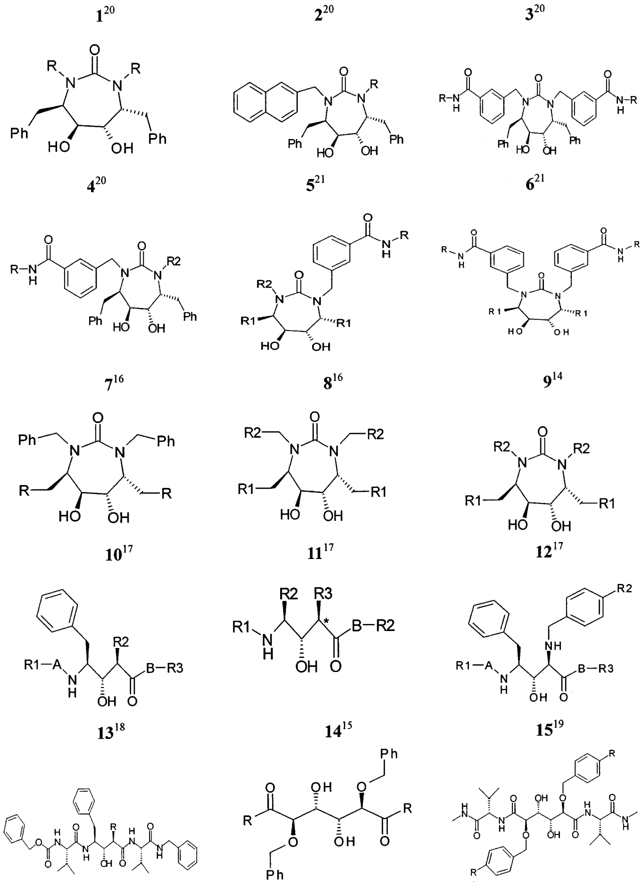

Figure 1. The basic structure units used to construct the 221 HIV-1 protease inhibitors[14−21] shown in Table 1.

E_i (ionization potential), E_{rep} (nucleus-nucleus repulse energy), and E_e (electronic energy). The 16 conventional descriptors described by Xu and Stevenson[27] were computed as described previously.[26]

There were six Hosoya topological indices[28] namely, Hosoya Z_1 (number of chemical bonds in a structure), Hosoya Z_2 (number of pair of disjoint chemical bonds in a structure), Hosoya Z_3 (number of triplet of disjoint chemical bonds in a structure), Hosoya Z_4 (number of quartet of disjoint chemical bonds in a structure), Hosoya Z_5 (number of quintet of disjoint chemical bonds in a structure), and Hosoya Z_6 (number of hexad of disjoint chemical bonds in a structure), plus the molecular connectivity index of path 1 ($^1\chi$)[29] computed for each structure. To compute these topological indices, each structure was treated as a graph[30] in which atoms were nodes and chemical bonds were edges. An adjacency matrix[31] was constructed first to identify each pair of connected nodes in a graph. This adjacency matrix was used to construct an adjacency matrix for edges which was then used to construct a connection table[31] for all the edges of the graph. Using the connection table,[31] a walk on the graph was initiated from one toward the other ends of the graph to count each type of Hosoya topological indices.[28] The other topological descriptor $^1\chi$[29] was computed as follows

$$^1\chi = \sum_{\substack{all\\ edges}} \frac{1}{(mn)^{1/2}} \tag{1}$$

where m and n were the degrees of the adjacent nodes joined by each edge. The valence delta values[23] of atoms C, N, and O in different bonding environment were taken as the degrees. To compute the 3D-QSAR descriptors which includes one CoMFA[3] and seven CoMSIA (Comparative Molecular Similarity Indices Analysis)[32] ones, the structures were aligned against the structure of the most active inhibitor among the set as described previously.[26] The five geometric descriptors, namely, MOLPROP_AREA, MOLPROP_PSA, MOLPROP_PV, MOLPROP_VOLUME, and MOL_WEIGHT, were computed using the SYBYL QSAR modules.[24]

To perform feature selection, the set of 221 inhibitors was divided into five classes according to their pK_i values. The number of structures in classes 1, 2, 3, 4, and 5 were 33, 39, 52, 57, and 40, respectively, as shown in Table 1. The Fisher's discriminant ratio[13] C defined as follows was used as a class separability criterion

$$C = \sum_{i}^{M} \sum_{j\neq i}^{M} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2} \tag{2}$$

where μ_i, σ_i and μ_j, σ_j were means and variances for classes i and j, respectively, and they were summed over all the classes M. To proceed the feature selection, all the descriptor values were scaled to lie within the same range, i.e., between 1 and 0. This was necessary since most of the original descriptors lie within a different dynamic range and those with larger values may have a larger influence in the cost function than those with small ones. The feature selection steps[13] were as follows: (i) compute C values for all the available descriptors and then rank them in descending order and choose the one i.e., $d_{n,1}$, with the largest C value computed; (ii) compute the cross-correlation coefficient[13] defined between the chosen $d_{n,1}$ and each of the remaining $M-1$ descriptors $r_{n,j}$ as follows

$$\rho_{ij} = \frac{\sum_n (d_{n,1} - \overline{d_{n,1}})(r_{n,j} - \overline{r_{n,j}})}{\sqrt{\sum_n (d_{n,1} - \overline{d_{n,1}})^2 \sum_n (r_{n,j} - \overline{r_{n,j}})^2}} \tag{3}$$

where the summation was conducted over all the inhibitors n and $\overline{d_{n,1}}$ or $\overline{r_{n,j}}$ were the average of $d_{n,1}$ or $r_{n,j}$ over all the inhibitors; (iii) select the second descriptor $d_{n,2}$ for which

$$d_{n,2} = \arg\max_j \{\alpha_1 C_j - \alpha_2 |\rho_{i,j}|\}, \text{ for all } j \neq 1 \tag{4}$$

where α_1 and α_2 were weighting factors that determine the relative importance we gave to the two terms. In words, for selection of the next descriptor, we took into account not only the class separability measure C but also the correlation with the already chosen descriptor. This then generalized for the kth step, i.e., select $d_{n,k}$, $k = 3,...l$, so that

$$d_{n,k} = \arg\max_j \left\{ \alpha_1 C_j - \frac{\alpha_2}{k-1} \sum_{m=1}^{k-1} \left|\rho_{i,j}\right| \right\}, \quad \text{for } j \neq m,\ m=1,2,....k\text{-}1 \tag{5}$$

That was, the average correlation with all previously selected descriptors was taken into account. Both the weighting parameters α_1 and α_2 were set as 1 during the selection process. A *k*-means clustering algorithm[22] was then employed to cluster descriptors within each class for each descriptor ranked and selected. Two or three clusters were created for each class first by setting the desired cluster number $k_{\mathrm{mm}} =$ 2 or 3. The cluster with the smallest descriptor values gathered was identified and regarded as the bad one. An inhibitor was considered as an outlier and needed to be abandoned if it was identified to be present in all the bad clusters generated for each descriptor selected.

To monitor the feature selection and data set trimming process, we have either used the stepwise variable selection program MREG described by Jurs[23] or directly fitted the selected descriptor values to each pK_i through a linear relationship by solving a normal equation as follows

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{6}$$

where β is a ($p \times 1$) vector of the regression coefficients, $\mathbf{X}'\mathbf{X}$ is a ($p \times p$) symmetric matrix, $\mathbf{X}'\mathbf{y}$ is a ($p \times 1$) column vector, $\mathbf{X}$ is a ($n \times p$) matrix of the levels of the regressor variables, and $\mathbf{y}$ is a ($n \times 1$) vector of the observations. The detail of the normal equation could be found elsewhere.[33] The normal equation was solved using the LU decomposition routines described by Press et al.[34] To proceed with the cross-validation process, two or three rows of input data were randomly omitted, and the model derived from the left data was used to predict the pK_i values of the omitted rows. The omitting-prediction cycle continued until the pK_i values of all the selected inhibitors have been predicted exactly once. The cross-validated r^2 (q^2) was computed as follows

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \tag{7}$$

where y_i were the actual pK_i values, $\hat{y}_i$ were the predicted pK_i values from the liner equation, and $\bar{y}$ was the average pK_i values. The summation in eq 5 was performed over all the selected inhibitors. The external validation process described by Golbraikh and Tropsha[35] was also employed to validate each model derived. Briefly, there were 20 inhibitors randomly selected for external validation from each selected and predicted set for computing the slope *a* and intercept *b* of a regression line of a plot of predicted versus actual pK_i values. The correlation coefficients R, R^2, $R_o{}^2$, and the slope *k* of an ideal regression line through the origin defined by Golbraikh and Tropsha[35] were also computed for the same set. The *F*-ratio[35] for the 20 inhibitors selected for external validation was then computed using $\hat{y}_i$ and y_i^r, the predicted pK_i values and the values computed from the linear equation $y^r = a\hat{y} + b$, respectively.

The Shannon entropy described by Godden and Bajorath[36] and the representation entropy H_R described as follows[37] were also computed for both the selected and discarded descriptors

$$H_R = -\sum_{j=1}^{l} \bar{\lambda}_j \log \bar{\lambda}_j \tag{8}$$

where $\bar{\lambda}_j$ was the normalized eigenvalues (eq 9) λ_j, $j = 1,...l$, of the $l \times l$ covariance

$$\bar{\lambda}_j = \frac{\lambda_j}{\sum_{j=1}^{l} \lambda_j} \tag{9}$$

matrix of a descriptor set of size l and $0 \leq \bar{\lambda}_j \leq 1$. For comparison, the Principal Component Analysis (PCA)-PLS module of the Cerius^2 package[25] was also used to perform the feature selection and data set trimming for all the 221 inhibitors.

RESULTS AND DISCUSSION

A large number of new HIV-1 protease inhibitors have been synthesized and assayed due to the growing problem of drug resistance.[14−21] The structure features of the 221 HIV-1 protease inhibitors studied are very diversified (Figure 1 and Table 1) since they includes some L-mannaric acid, C_2-symmetric $P_1/P_{1'}$, 2-heterosubstituted 4-amino-3-hydroxy-5-phenylpentanoic acid, 2-heterosubstituted statine, and cyclic urea derivatives.[14−21] The pK_i values of the set vary from 5.15 to 10.96 (Table 1). Based on the pK_i values, these inhibitors are initially divided into the following five classes: class 1 (pK_i 10.00−10.96), class 2 (pK_i 9.00−9.99), class 3 (pK_i 8.00−8.99), class 4 (pK_i 7.00−7.99), and class 5 (pK_i 5.00−6.99) (Table 1). The corresponding number of compounds in classes 1, 2, 3, 4, and 5 are 33, 39, 52, 57, and 40, respectively (Table 1). A CoMFA PLS[24] analysis on the 221 inhibitors aligned against the structure of the most active one (inhibitor 086 of Table 1) gives a q^2 of 0.606. The 43 molecular descriptors computed for the seven most active inhibitors are listed in Table 2. The descriptors are listed in the following order, namely, 16 conventional, 7 MOPAC, 6 Hosoya Z, 1 connectivity index, 8 QSAR, and 5 geometric descriptors. Since the values of these descriptors are widely varied, each of them is scaled to lie between 1 and 0. For each descriptor, the largest and the smallest values are identified first, and then the scaled ones are computed as the ratio between the difference of each original value with the smallest one and that of the two extremes. The feature values computed using the Fisher's discriminant ratio[13] and correlation coefficient[13] between them for all the 43 descriptors are shown in Table 3. Apparently, descriptors 27 (Hosoya Z_4), 29 (Hosoya Z_6), 26 (Hosoya Z_3), 25 (Hosoya Z_2), 32 (CoMSIA steric), and 38 (CoMSIA donor and acceptor) with corresponding feature values 788, 734, 677, 264, 190, and 162 computed were among the best or the most class sensitive descriptors identified (Table 3). Table 3 also lists the Shannon entropy[36] computed for each descriptor since it can serve as an indicator of the information content of each. The Shannon entropy[36] computed for the above six descriptors are 4.83, 2.55, 4.78, 3.79, 3.08, and 3.01, respectively. While the Shannon entropies[36] are not as class sensitive as the feature values computed (Table 3), they do indicate that the information content of some of the six best descriptors identified is high. The representation entropy[37] H_{R} is a more class sensitive criterion that one can use to evaluate the class separability of feature values computed. The function H_{R} attains a minimum value (zero)

Table 2. 43 Molecular Descriptors Computed for the Seven Most Active Inhibitors[a]

descriptor	086	089	117	102	652	088	115
1	60.00	60.00	46.00	54.00	54.00	62.00	46.00
2	11.00	11.00	10.00	13.00	13.00	11.00	10.00
3	2.00	2.00	5.00	8.00	8.00	2.00	5.00
4	2.00	2.00	0.00	0.00	0.00	4.00	0.00
5	2.00	2.00	2.00	2.00	2.00	2.00	2.00
6	4.00	4.00	7.00	10.00	10.00	4.00	7.00
7	3.00	3.00	3.00	5.00	7.00	3.00	4.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	6.00	6.00	4.00	6.00	6.00	6.00	4.00
10	9.00	9.00	8.00	9.00	9.00	9.00	8.00
11	26.00	26.00	22.00	26.00	26.00	24.00	22.00
12	10.00	10.00	7.00	8.00	8.00	12.00	7.00
13	58.00	58.00	60.00	62.00	62.00	56.00	60.00
14	19.00	19.00	17.00	23.00	23.00	19.00	18.00
15	13.00	13.00	10.00	11.00	11.00	13.00	10.00
16	1.00	1.00	2.00	3.00	1.00	1.00	2.00
17	−8.66	−7.94	−8.73	−8.88	−8.92	−8.31	−8.87
18	−4.86	−7.19	−4.89	−4.86	−4.86	−4.79	−5.24
19	14.18	3.92	8.35	7.30	7.69	14.62	7.32
20	2855.34	2661.15	2223.39	2636.22	2671.45	3034.84	2098.02
21	8.66	7.94	8.73	8.88	8.92	8.31	8.87
22	73011.12	75297.49	53391.01	65095.64	64885.31	76229.13	54255.99
23	−82034.95	−84820.21	−60456.63	−73490.46	−73278.61	−85501.63	−61393.43
24	68.00	68.00	54.00	60.00	60.00	70.00	54.00
25	124.00	124.00	100.00	110.00	110.00	128.00	100.00
26	180.00	180.00	156.00	168.00	168.00	184.00	156.00
27	265.00	265.00	240.00	243.00	243.00	269.00	240.00
28	376.00	376.00	401.00	468.00	468.00	480.00	420.00
29	496.00	496.00	427.00	456.00	456.00	504.00	427.00
30	27.94	27.45	21.72	24.29	23.71	28.52	21.64
31	4169.00	3813.00	2997.00	3621.00	3446.00	4352.00	2782.00
32	4.75	4.76	4.27	4.52	4.48	4.83	4.25
33	24.68	22.57	20.81	22.39	21.66	26.02	20.06
34	1.51	2.72	1.30	1.19	2.64	1.58	1.82
35	1.28	1.28	1.27	0.00	0.00	1.28	1.27
36	1.85	1.85	1.91	2.33	2.60	1.84	2.31
37	17.77	16.31	15.02	16.15	15.64	18.71	14.50
38	1.59	1.59	1.62	1.65	1.84	1.58	1.87
39	1200.00	1200.00	880.00	1000.00	1000.00	1300.00	873.00
40	270.00	269.00	275.00	313.00	309.00	254.00	317.00
41	175.00	206.00	134.00	139.00	133.00	179.00	161.00
42	2100.00	2100.00	1500.00	1800.00	1800.00	2100.00	1500.00
43	754.59	801.47	582.45	684.50	684.50	778.61	602.49

[a] 1.n-H(number of non-H atoms); 2.N&O(number of N and O atoms); 3.N&1-H(number of N atoms with at least one H atom); 4.MLipo(molecular lipophilicity,number of carbon atoms as the terminal group); 5.OH(number of hydroxyl group); 6.Hdon (number of H-bond donors); 7.Haccp(number of H-bond acceptor); 8.caps(number of caps); 9.2-deg(number of 2-degree chain atoms, acyclic atom connected with 2 non-H atoms); 10.3-deg(number of 3-degree chain atoms, acyclic atom connected with 3 non-H atoms); 11.2-dgc(number of 2-degree cyclic atoms, cyclic atom connected with 2 non-H atoms); 12.3-dgc(number of 3-degree cyclic atoms, cyclic atom connected with 3 non-H atoms); 13.MCD(molecular cyclized degree: number of ring atoms/total number of atoms); 14.n-Hpol(number of non-H polar bond, polarity); 15.n-Hrot(number of non-H rotating bonds, flexibility); 16.amide(number of amide linkage);17.E_{HOMO}; 18.E_{LUMO}; 19.μ; 20.$H_{f;}$ 21.E_i; 22.E_{rep}; 23.E_e; 24.Hosoya Z_1; 25.Hosoya Z_2; 26.Hosoya Z_3; 27.Hosoya Z_4; 28.Hosoya Z_5; 29.Hosoya Z_6; 30.${}^1\chi$; 31.CoMFA; 32.CoMSIA 1 (steric); 33.CoMSIA 2 (electrosteric); 34.CoMSIA 3 (hydrophobicity); 35.CoMSIA 4 (donor); 36.CoMSIA 5 (acceptor); 37.CoMSIA 6 (steric and electrosteric); 38.CoMSIA 7 (donor and acceptor); 39. MOLPROP_AREA (SYBYL); 40.MOLPROP_PSA(SYBYL); 41.MOLPROP_PV (SYBYL); 42.MOLPROP_VOLUME (SYBYL); 43.MOL_WEIGHT(SYBYL).

Table 3. Feature Values and Shannon Entropies[36] Computed for All the 43 Descriptors

27(1,[a]788,[b]4.83[c])
29 (2,734,2.55)
26 (3,677,4.78)
25 (4,264,3.79)
32 (5,190,3.08)
38 (6,162,3.01)
31 (7, 96,2.47)
15 (8, 95,2.13)
34 (9, 95,0.00)
24 (10,94,3.34)
9 (11,94,1.66)
28 (12,90,4.83)
33 (13,88,4.71)
22 (14,85,4.77)
37 (15,84,4.72)
11 (16,84,4.33)
23 (17,84,0.00)
36 (18,83,0.00)
13 (19,80,0.00)
12 (20,79,4.28)
40 (21,66,3.61)
43 (22,50,4.77)
14 (23,48,4.64)
1 (24,46,2.72)
4 (25,41,4.38)
8 (26,41,3.56)
3 (27,39,4.81)
21 (28,38,3.16)
19 (29,36,4.77)
30 (30,36,4.82)
17 (31,35,4.81)
20 (32,33,4.69)
5 (33,30,2.96)
42 (34,29,4.76)
16 (35,27,3.19)
7 (36,26,2.49)
10 (37,26,3.75)
35 (38,25,3.57)
2 (39,23,4.63)
39 (40,21,4.81)
41 (41,19,4.63)
6 (42,12,4.84)
18 (43,12,4.75)

[a] The rank of features according to the feature values computed. [b] The feature values computed for each descriptor using eqs 4 and 5. [c] The Shannon entropies[36] computed for each descriptor.

when all the eigenvalues except one are zero or, in other words, when all the information is present along a single coordinate direction. However, if all the eigenvalues are equal, i.e., information is equally distributed among all the

Table 4. Representation Entropies[37] Computed and the Linear Regression Equations Obtained for Inhibitors Selected Based on Different Numbers of Descriptors Selected

no. of descriptors selected	H_R (no. of inhibitors selected)	H_R (no. of inhibitors deleted)	linear regression equations obtained for inhibitors selected
1	- - - - (44)	- - - - (177)	$pK_i = 6.690 + 4.951*d1$
2	0.000(79)	0.000(142)	$pK_i = 5.422 + 8.107*d1 - 0.882*d2$
3	0.734(98)	0.069(123)	$pK_i = 5.810 + 12.491*d1 - 2.774*d2 - 3.546*d3$
4	1.175(161)	0.549(60)	$pK_i = 6.780 + 6.250*d1 - 0.934*d2 - 2.478*d3 + 1.507*d4$
5	1.539(171)	0.628(50)	$pK_i = 6.804 + 7.09*d1 - 0.665*d2 - 3.562*d3 + 1.473*d4 - 0.117*d5$
6	1.907(185)	1.033(36)	$pK_i = 6.819 + 6.269*d1 + 0.071*d2 - 3.753*d3 + 1.306*d4 - 1.181*d5 + 2.017*d6$

Table 5. Comparison of the Statistics Obtained from the Linear Regression Process, Cerius^2 PLS[25] Regression, and SYBYL CoMFA PLS[24] on the Cases of Different Numbers of Descriptors Selected

n.d.s.[a]	q^2	R	R^2	R_o^2	RMSE	F-ratio	k	Cerius^2 [b]	CoMFA[c]
1	0.354	0.995	0.990	0.999	0.0181	166.3	0.986	0.815(3)	0.277(6)
2	0.712	0.993	0.986	0.999	0.0248	683.8	0.984	0.847(3)	0.664(6)
3	0.727	0.989	0.979	0.998	0.0552	409.7	0.976	0.736(3)	0.710(6)
4	0.549	0.987	0.974	0.996	0.0830	96.6	0.969	0.698(3)	0.673(6)
5	0.536	0.984	0.968	0.998	0.0412	362.1	0.978	0.645(3)	0.612(6)
6	0.576	0.985	0.971	0.998	0.0307	588.5	0.981	0.658(3)	0.647(6)

[a] Number of descriptors (features) selected. [b] Values of q^2 obtained using the Cerius^2 PLS regression procedure, the number of principal components selected is parenthesized. [c] Values of q^2 obtained using the SYBYL CoMFA PLS procedure, the number of components used is parenthesized.

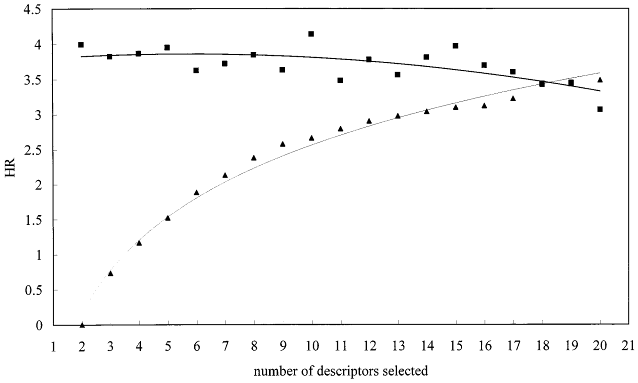

Figure 2. The representation entropies[37] H_R computed are plotted as a function of the number of descriptors selected. H_R computed for the inhibitors retained are represented as triangles while those computed for the discarded ones are represented as squares.

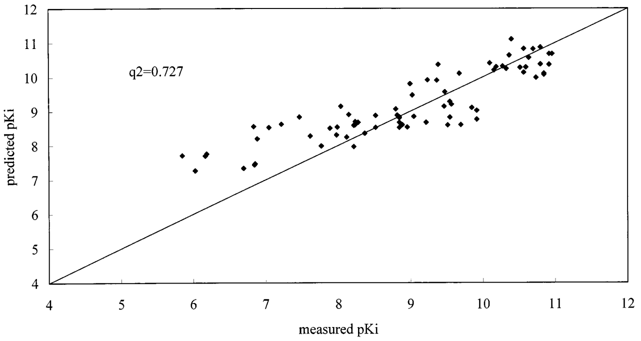

Figure 3. The predicted pK_i values are plotted against those of measured for the 98 inhibitors retained using the three best descriptors selected. The predicted pK_i values are computed using the linear regression equation obtained from solving eq 6 (Table 4).

features, H_R is maximum and so is the uncertainty involved in feature reduction. Since the proposed algorithm involves reduction of features of bad class sensitivity, it is expected that H_R computed for the class sensitive features selected will be lower than that for the class insensitive ones discarded. As shown in Figure 2, H_R computed is smaller for the class sensitive descriptors selected and is increasing as the number of descriptors selected is increased while that computed for the class insensitive ones discarded is much larger and is only slightly decreasing as the number of descriptors selected is increased. Figure 2 also indicates that H_R computed for selection of more than 18 or 19 descriptors will be similar to that for selection of none at all. A comparison for H_R computed for various numbers of descriptors selected is given in Table 4. The number of descriptors selected is from 1 to 6 which gives the corresponding number of inhibitors retained ranging from 44 to 185. The H_R computed is increasing from 0.0 (for two descriptors selected) to 1.907 (for six descriptors selected) as the number of good inhibitors retained is increased from 79 to 185 (Table 4). However, the H_R computed for the corresponding inhibitors discarded is also increased from 0.0 to 1.033 as they are decreased from 142 to 36 (Table 4). Note that for the two best descriptors (descriptors 27 and 29) selected, the H_R computed for both the retained (79 inhibitors) and deleted (142 inhibitors) sets are zero (Table 4) which implies that the class separability of the twos are equally well represented by the two descriptors selected.

The other criterion we use to judge the feature selection result is to fit the selected descriptors to the pK_i values of the inhibitors retained through a linear relationship. The rational for using such a criterion is that as the selected descriptors are more class sensitive they are deemed be more correlative with the pK_i values since the latter are also class sensitive parameters. The goodness of the fitting result is estimated through computation of the following parameters, namely, q^2, R, R^2, R_o^2, k, *RMSE* (residual mean square error), and the F-ratio as described by Golbraikh and Tropsha.[35] The corresponding linear relationships obtained for various number of descriptors selected are presented in Table 4. As shown in Table 5, q^2 values are increasing from 0.576 to 0.727 as the number of descriptors selected is reduced from 6 to 3. However, the parameter shows no sign of improving as the number of descriptors selected is reduced from 3 to 1 (Table 5). Figure 3 is a presentation of the predicted versus

Table 6. Stepwise Variable Selection Using the MREG Program Described by Jurs[23]

12[a](0.01[b])	11(0.13)	10(0.11)	9(0.03)	8(0.05)	7(0.03)	6(0.09)	5(0.17)	4(0.29)	3(0.49)	2(0.50)	1(0.51)
11(0.14)	10(0.36)	9(0.11)	8(0.08)	7(0.05)	6(0.13)	5(0.19)	4(0.34)	3(0.50)	2(0.51)	1(0.51)	12(0.58)
10(0.37)	9(0.36)	8(0.20)	7(0.08)	6(0.16)	5(0.37)	4(0.36)	3(0.50)	2(0.51)	1(0.51)	12(0.58)	11(0.58)
9(0.37)	8(0.37)	7(0.29)	6(0.16)	5(0.37)	4(0.44)	3(0.53)	2(0.52)	1(0.51)	12(0.59)	11(0.58)	10(0.58)
8(0.38)	7(0.41)	6(0.31)	5(0.38)	4(0.44)	3(0.54)	2(0.54)	1(0.52)	12(0.59)	11(0.59)	10(0.58)	9(0.58)
7(0.42)	6(0.42)	5(0.39)	4(0.45)	3(0.54)	2(0.55)	1(0.54)	12(0.59)	11(0.59)	10(0.59)	9(0.58)	8(0.60)
6(0.43)	5(0.47)	4(0.54)	3(0.54)	2(0.55)	1(0.55)	12(0.62)	11(0.59)	10(0.60)	9(0.59)	8(0.60)	7(0.61)
5(0.49)	4(0.55)	3(0.56)	2(0.55)	1(0.55)	12(0.62)	11(0.62)	10(0.61)	9(0.60)	8(0.61)	7(0.61)	6(0.62)
4(0.62)	3(0.57)	2(0.56)	1(0.55)	12(0.63)	11(0.62)	10(0.63)	9(0.61)	8(0.62)	7(0.62)	6(0.62)	5(0.63)
3(0.64)	2(0.57)	1(0.56)	12(0.63)	11(0.64)	10(0.63)	9(0.63)	8(0.63)	7(0.62)	6(0.64)	5(0.63)	4(0.64)
2(0.64)	1(0.57)	12(0.64)	11(0.64)	10(0.64)	9(0.63)	8(0.64)	7(0.63)	6(0.64)	5(0.64)	4(0.64)	3(0.64)
1(0.65)	12(0.65)	11(0.65)	10(0.65)	9(0.65)	8(0.65)	7(0.65)	6(0.65)	5(0.65)	4(0.65)	3(0.65)	2(0.65)

[a] The rank of the best 12 features (descriptors) selected (Table 3). [b] The value of r^2 computed by feeding each feature to the MREG program is parenthesized.

measured pK_i values for the case of three descriptor selected for which the value of q^2 computed is 0.727. It appears that most of the parameters presented in Table 5 fulfill the conditions given by Golbraikh and Tropsha[35] as the acceptable models namely, $q^2 > 0.5$, $R^2 > 0.6$, R_o^2closes to R^2, and the corresponding requirement $0.85 \leq k \leq 1.15$. To further validate the feature selection results, we have used the basic variable selection program MREG described by Jurs[23] to directly compare the significance of each feature selected. Program MREG is an interactive FORTRAN program to perform multiple linear regression. It steps forward adding one variable at a time, printing out the result of each step with the value of non-cross-validated r^2 (r^2) as an indicator of the goodness of the variables added so far. To proceed with the validation steps, we have chosen the best 12 descriptors selected (Table 3) and reversed their feature order for feeding them into the MREG program one by one. In other words, the 12th feature is fed into the program first then the 11th and 10th and until the last one or the first (best) one is fed. A new value of r^2 is computed at each feeding step, and it is recorded accordingly. Table 6 is a presentation of the validation result by the MREG program. Apparently, the validation result given by the MREG program agrees with ours since the feature order given by the MREG program is very similar to that by our method as indicated in Table 6. For example, the r^2 values computed by the MREG program for the best six features (Tables 3 and 4) can be also ranked as $1 > 2 > 3 > 4 > 5 > 6$ as shown in the first row of Table 6.

A direct Cerius^2 PLS[25] linear regression for each number of descriptors selected on pK_i of the inhibitors retained is also performed and presented in Table 5. This further indicates the feature selection result is effective since the PLS statistics expressed in q^2 is increasing from 0.658 to 0.815 as the number of inhibitors retained is reduced from 185 to 44 (Tables 4 and 5). Moreover, there is also an obvious improving in the CoMFA PLS[24] statistics (expressed in q^2) as the number of inhibitors retained is decreased from 185 to 98 (Tables 4 and 5) when the same sets of retained inhibitors are subjected to the SYBYL CoMFA[24] analysis.

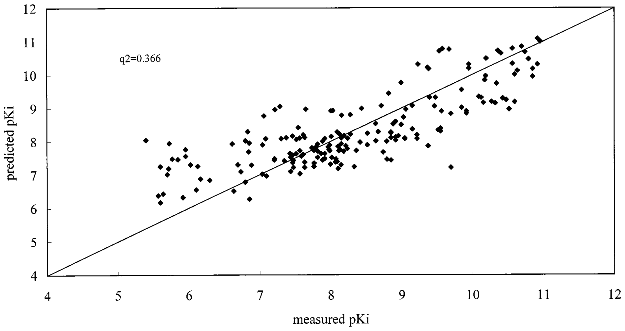

Figure 4. The predicted pK_i values are plotted against those of measured for the 188 inhibitors selected using the Cerius^2 PCA[25] program. The number of best features or principal components selected is three and the predicted pK_i values are computed from the following Cerius^2 PLS[25] equation: Predicted p$K_i = -0.0038225 * \mathrm{PC1} + 0.296482 * \mathrm{PC2} + 0.117598 * \mathrm{PC3} + 8.29382$.

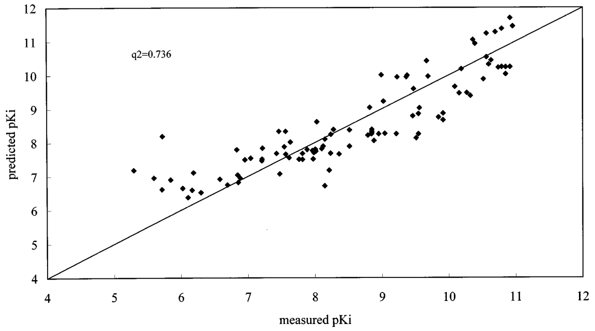

Figure 5. The predicted pK_i values are plotted against those of measured for the 98 inhibitors retained using the three best descriptors selected. The predicted pK_i values are computed using the linear equation shown below obtained from the direct Cerius^2 PLS[25] linear regression for all the 43 descriptors on the measured pK_i values: Predicted p$K_i = -0.00075263 * \mathrm{d1} - 0.012453 * \mathrm{d2} - 0.019508 * \mathrm{d3} - 0.038129 * \mathrm{d4} - 0.029575 * \mathrm{d5} - 0.026137 * \mathrm{d6} - 0.0044759 * \mathrm{d7} + 0.036015 * \mathrm{d8} - 0.022044 * \mathrm{d9} + 0.074083 * \mathrm{d10} - 0.010036 * \mathrm{d11} + 0.012923 * \mathrm{d12} - 0.0067395 * \mathrm{d13} - 0.0069779 * \mathrm{d14} + 0.022707 * \mathrm{d15} - 0.086987 * \mathrm{d16} - 0.034516 * \mathrm{d17} - 0.056364 * \mathrm{d18} - 0.0074647 * \mathrm{d19} + 8.5658\mathrm{e}{-}05 * \mathrm{d20} - 0.015766 * \mathrm{d21} + 2.8035\mathrm{e}{-}06 * \mathrm{d22} - 2.143\mathrm{e}{-}06 * \mathrm{d23} + 0.0044121 * \mathrm{d24} + 0.0099563 * \mathrm{d25} + 0.015174 * \mathrm{d26} + 0.009 * \mathrm{d27} - 0.0015909 * \mathrm{d28} + 0.0030409 * \mathrm{d29} + 0.00093125 * \mathrm{d30} - 6.6371\mathrm{e}{-}05 * \mathrm{d31} + 0.27043 * \mathrm{d32} - 0.012517 * \mathrm{d33} + 0.018872 * \mathrm{d34} + 0.187964 * \mathrm{d35} - 0.034779 * \mathrm{d36} - 0.015641 * \mathrm{d37} + 0.013782 * \mathrm{d38} - 0.00037612 * \mathrm{d39} + 0.00065056 * \mathrm{d40} - 0.001212 * \mathrm{d41} - 4.0779\mathrm{e}{-}05 * \mathrm{d42} - 0.00022202 * \mathrm{d43} + 3.39154$.

The Cerius^2 PCA[25] module is also capable of performing feature selection and data set trimming. The result by such an analysis for all the 221 inhibitors is presented in Figure 4. The number of best features or principal components selected is three, and the number of inhibitors retained is 188 (Figure 4). However, the selected principal components are poorly regressed on the pK_i of the retained inhibitors since the corresponding q^2 computed is only 0.366 (Figure 4). If we consider only the information content i.e., the Shannon entropy,[36] in the feature selection process, the best

Table 7. Feature Values, Number of Inhibitors Discarded, and Shannon Entropies[36] Computed for the Inhibitors Being Divided into Several Classes

no. of classes	class range (pK_i range)	the best four descriptors (feature values) selected	no. of inhibitors discarded	Shannon entroy of the best four descriptors selected

7	31(10.13−10.96)
	31(9.22−10.09)
	31(8.52−9.22)
	31(8.00−8.37)
	31(7.57−7.99)
	31(6.85−7.57)
	35(5.15−6.84)
		29(1726)
		27(1723)
		26(1222)
		15(511)
			124	4.844, 4.748, 4.626, 2.489

6	33(10.00−10.96)
	33(9.10−9.99)
	48(8.10−9.00)
	46(7.50−8.00)
	33(6.60−7.40)
	28(5.00−6.50)
		27(1050)
		29(946)
		26(934)
		25(392)
			123	4.748, 4.844, 4.626, 4.813

5	33(10.00−10.96)
	39(9.00−10.00)
	52(8.00−9.00)
	57(7.00−8.00)
	40(5.00−7.00)
		27(789)
		29(734)
		26(677)
		25(264)
			123	4.748, 4.844, 4.626, 4.813

4	55(9.40−10.95)
	55(8.15−9.40)
	55(7.47−8.14)
	56(5.00−7.44)
		26(245)
		27(238)
		29(200)
		25(111)
			133	4.626, 4.748, 4.844, 4.813

3	74(8.92−10.96)
	74(7.64−8.92)
	73(5.00−9.64)
		26(121)
		27(117)
		29(100)
		25(44)
			133	4.626, 4.748, 4.844, 4.813

2	110(8.15−10.96)
	111(5.0−8.14)
		27(31)
		26(28)
		29(23)
		31(11)
			149	4.748, 4.626, 4.844, 3.748

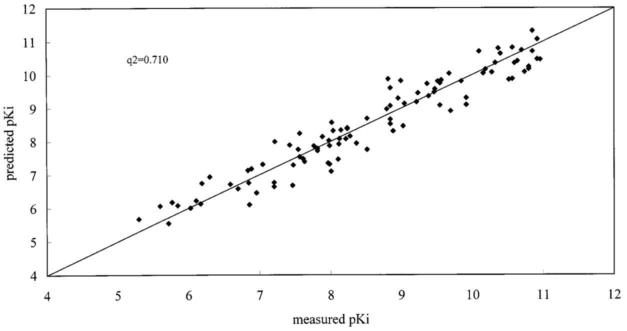

Figure 6. The predicted pK_i values are plotted against those measured for the 98 inhibitors retained using the three best descriptors selected. The predicted pK_i values are computed by the SYBYL CoMFA PLS[24] program for the 98 inhibitors retained.

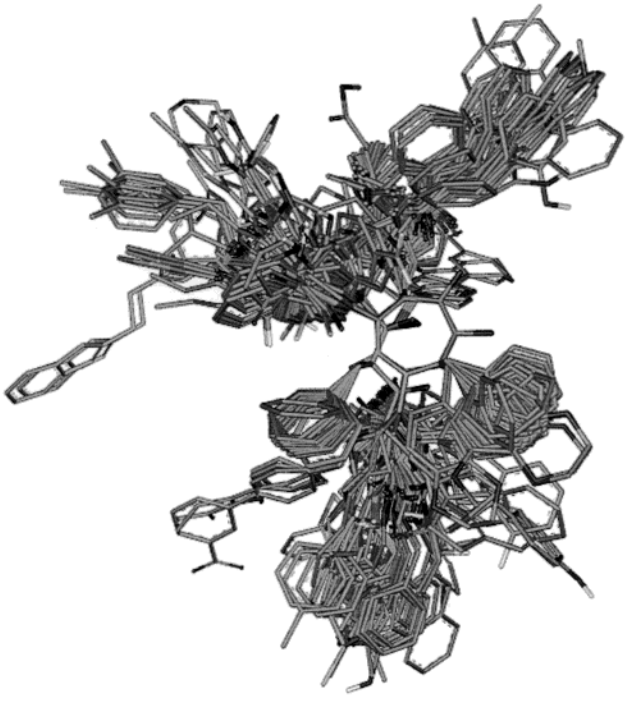

Figure 7. Superposition of the structures of the 98 selected and aligned inhibitors against that of inhibitor 086 (Table 1), the most active one studied in the series.

three descriptors selected are 29 (Hosoya Z_6), 18 (E_{LUMO}), and 20 (H_f), and the corresponding Shannon entropies[36] computed are 4.84, 4.83, and 4.83, respectively. The number of inhibitors retained using these three descriptors is 150. The regression of these descriptors on the pK_i of retained inhibitors gives a q^2 and correlation coefficient R of 0.382 and 0.639, respectively. This shows that feature selection on a class sensitive set could be successful only by using some class sensitive evaluation criteria. To examine the effect of difference in number of classes divided on the feature selection, we have varied the number of classes from two to seven and performed the feature selection on each of them. As shown in Table 7, there is no difference in the feature selection result for the inhibitors being divided into five or six classes since for them both the best four descriptors selected and the number of inhibitors retained are the same. There is also no difference in the feature selection result for the inhibitors being divided either into four or three classes (Table 7). However, the feature values computed for the best four descriptors selected for these twos are smaller than those computed for the five or six classes divided. The information content of the best descriptor selected for the latter (three

Table 8. Three Best Descriptors Selected Namely, 27 (Hosoya Z_4), 29 (Hosoya Z_6), and 26 (Hosoya Z_3) and the pK_i Values for the 98 Inhibitors Retained

inhibitor no.	measured pK_i	Hosoya Z_4	Hosoya Z_6	Hosoya Z_3	inhibitor no.	measured pK_i	Hosoya Z_4	Hosoya Z_6	Hosoya Z_3
Class 1									
086	10.9586	265	496	180	085	10.5686	263	494	178
089	10.9208	265	496	180	638	10.5686	243	456	168
117	10.9208	240	427	156	28	10.5229	249	510	166
102	10.8539	243	456	168	15	10.3979	277	588	180
652	10.8539	243	456	168	083	10.3665	257	484	172
088	10.7959	269	504	184	111	10.3279	231	399	149
115	10.7959	240	427	156	110	10.2757	234	404	152
068	10.7447	231	436	152	096	10.1938	251	472	174
087	10.6990	263	494	178	109	10.1611	233	405	153
116	10.6383	245	434	157	36	10.0969	247	478	164
100	10.6020	251	472	174					
Class 2									
052	9.9208	189	328	130	066	9.4815	217	406	144
053	9.9208	193	326	130	050	9.4685	197	348	134
051	9.8538	199	348	134	114	9.3872	232	408	146
43	9.7000	183	252	140	075	9.3726	229	432	152
078	9.6778	233	440	152	073	9.2373	227	426	148
25	9.5686	203	342	142	056	9.2218	181	297	126
054	9.5528	183	298	124	037	9.0506	183	298	124
059	9.5528	204	359	140	064	9.0315	215	407	144
13	9.5200	189	270	148	065	9.0000	225	431	149
Class 3									
055	8.9586	176	285	124	038	8.2840	181	300	126
023	8.8861	169	236	114	042	8.2441	181	300	126
035	8.8539	181	300	126	283550	8.2403	184	264	143
039	8.8539	183	298	124	283143	8.2197	185	239	148
057	8.8539	176	285	124	283490	8.2197	166	224	135
5	8.8500	189	258	144	041	8.1549	183	298	124
058	8.8239	188	311	130	283005	8.1198	171	224	139
046	8.7959	193	326	130	2	8.1079	165	262	114
027	8.5229	173	264	118	6	8.0362	197	366	134
034	8.5229	183	298	124	283265	8.0200	196	288	151
024	8.3665	167	232	122	029	8.0132	173	264	118
Class 4									
282453	7.9901	169	217	125	282978	7.5699	183	234	135
283055	7.9901	177	225	142	040	7.5686	181	300	126
283010	7.9800	171	222	138	283489	7.5500	175	221	139
282822	7.9698	163	204	132	283522	7.4800	169	238	128
283263	7.8901	181	244	147	033	7.4685	185	296	126
283374	7.8202	172	222	138	025	7.4318	173	264	118
282756	7.8199	167	216	136	020	7.2218	169	236	114
282730	7.7701	164	220	135	14	7.2147	163	221	132
282826	7.6400	178	226	142	282479	7.2100	162	216	132
282547	7.6200	169	214	133	030	7.0458	173	264	118
Class 5									
015	6.9586	153	189	112	283406	6.1700	149	184	118
283245	6.8900	165	232	123	283520	6.1100	136	185	103
37	6.8600	139	166	110	283573	6.0300	131	160	107
282364	6.8500	138	171	110	282807	5.8500	148	195	118
028	6.8386	173	264	118	282808	5.7700	148	195	118
15	6.6990	137	173	110	283567	5.7200	127	156	103
007	6.5850	135	173	100	283497	5.6000	159	198	126
283051	6.3000	145	192	109	38	5.3000	155	186	122
282834	6.1900	148	184	116					

and four classes) is also smaller than that for the former (five and six classes) (Table 7). Both the cases of two or seven classes divided are also ignored since descriptors of rather low information content such as 15 or 31 are selected for them as the best four ones (Table 7).

Based on the linear regression and SYBYL CoMFA PLS[24] results (Table 5), the best feature selection and data set trimming result for the inhibitors being divided into five classes is due to the case of three descriptors or 98 (Table 4) inhibitors selected or retained. The original designations of these inhibitors, their pK_i values, and the best three descriptors selected (Tables 2, 3), namely, 27 (Hosoya Z_4), 29 (Hosoya Z_6), and 26 (Hosoya Z_3) are listed in Table 8. Apparently, all the three descriptors are not only class sensitive but also correlated well with the pK_i measured. In other words, each of these descriptors may be used to represent the activity of the set of inhibitors selected. A Cerius^2 PLS[25] linear regression for all the 43 descriptors on the pK_i of the 98 inhibitors selected (Figure 5) gives a q^2 of 0.736 which is more significant in statistics than that presented in Figure 4 using the PCA module of the same program for feature selection and data set trimming. The 98

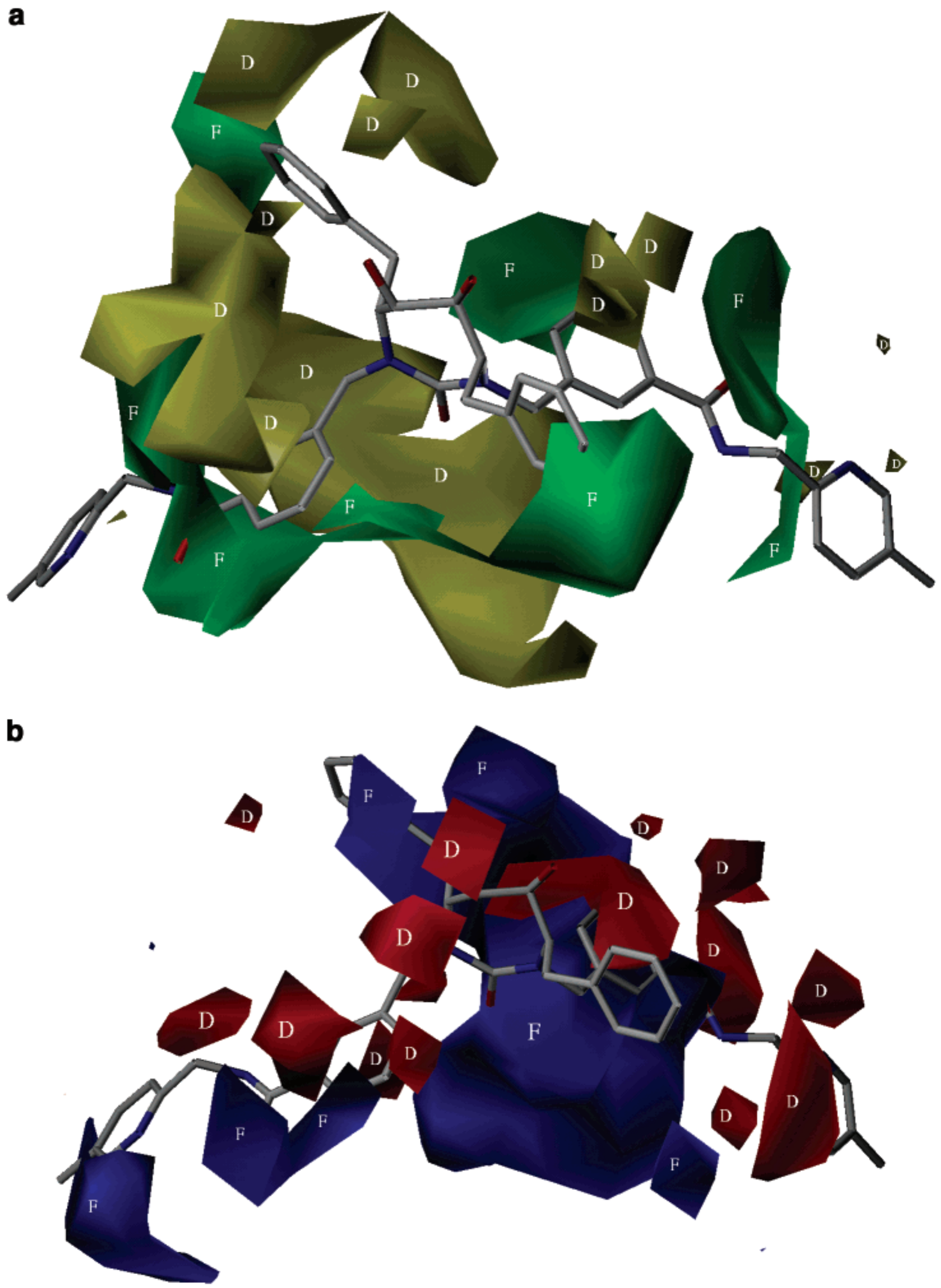

Figure 8. a. The CoMFA steric map obtained by the SYBYL CoMFA PLS[24] analysis on the 98 selected and aligned inhibitors. The favor regions for steric interaction are expressed with green contours (labeled as F regions) while for disfavored regions are expressed with yellow contours (labeled as D regions). b. The CoMFA electrostatic map obtained by the SYBYL CoMFA PLS[24] analysis on the 98 selected and aligned inhibitors. The favor regions for positive charges are expressed with blue contours (labeled as F regions) while favor regions for negative charges (or disfavor regions for positive charges) are expressed with red contours (labeled as D regions).

inhibitors selected are also subjected to the SYBYL CoMFA PLS[24] analysis, and a q^2 of 0.710 is obtained (Figure 6) which is better in statistics than that obtained for the original 221 inhibitors subjected to the same analysis. The aligned structures of the 98 selected inhibitors against that of the target one which is inhibitor 086, the most active one studied in the series (Table 1), are superimposed on each other and depicted in Figure 7. This alignment shows that most of the cyclic urea derivatives[26] (Figure 1 and Table 1) are well aligned on each other, while those of other linear chain derivatives[26] are not. The corresponding CoMFA steric and electrostatic maps obtained from the SYBYL CoMFA PLS[24] analysis on the 98 selected and aligned inhibitors are presented in Figure 8 (parts a and b, respectively). The favor regions for steric interaction and worthy of further exploration are expressed in green contours (Figure 8a). There are about six green contours (labeled as F regions) around the target structure and their distributions are as follows: two on each of the phenyl-pyridinyl side chains and one on each of the phenyl side chains of the cyclic urea (Figure 8a). This correlates with the fact that increasing the group bulkiness along the two phenyl-pyridinyl side chains enhancing the activity of the inhibitors.[20,26] The favor regions for electrostatic interaction are expressed in blue contours (labeled as F regions) for contribution from positive charges, while those expressed in red contours (labeled as D regions) are for contribution from negative charges (Figure 8b). These blue or red contours identified are coincident with the positions where the positively or negatively charged groups of the target structure are located (Figure 8b).

We have randomly altered the order of input for all the 43 descriptors and found that the feature selection result is unchanged. The correlation parameter α_2 has also been varied during the feature selection process to examine if there is any significant correlation between the descriptors. No difference in the feature selection result for α_2 being set as 1 or 0 is found, suggesting that no severe correlation is present in the set of descriptors computed. We have also applied the feature selection algorithm to the 43 unscaled descriptors and found that the best six descriptors selected are 32 (42.50,3.08), 38 (19.13,3.01), 35 (7.23,3.57), 36 (4.50,0.00), 8 (2.42,3.56), and 17 (1.87,4.81), where the corresponding feature values and information contents are parenthesized first and next, respectively. Although the descriptors identified are four QSAR, one conventional and one MOPAC ones, some of their corresponding information contents are low. This justifies the need for scaling all the descriptor values before performing the feature selection on them. On the other hand, the weakness of our algorithm is also revealed since it is not invariant under a transformation of the variables. Fortunately, the data set we dealt with is simple, and no complex decision boundaries[38] are present between classes. The same feature selection result is obtained when the data set is divided into two different numbers of classes. Since class separability is the most important criterion used to evaluate the feature selection result, we have tried a more complicated class separability measure such as the one-dimensional divergence[39] on the data set and found that it is not as effective as the Fisher's discriminant ratio[13] used.

CONCLUSION

In this report, we have presented an alternative way to improve the statistical significance of some 3D QSAR studies using the SYBYL CoMFA PLS[24] or Cerius^2 PLS[25] regression on a set of 221 HIV-1 protease inhibitors. Instead of focusing on the selection of field variables, we aim at the trimming of the data set by rejecting inhibitors that are unable to match the class separability criteria established for them using some descriptors selected. A random deletion of some inhibitors from the original set would not improve the value of q^2 significantly. The descriptors selected are those bearing good class sensitivity. It is interested to find that most of these descriptors are the topological ones. While the *k*-means clustering scheme performs well in combination with the feature selection process, other classification techniques may be also effective in providing the trimming in the process. However, a more complicated classification rule such as the Bayesian one requires some a priori knowledge about the distribution of classes. All CoMFA experience suggests that the only action capable of changing the sign of a q^2 value is realigning some or all of the molecules. Our method will be useful when all the biological data are sound and there are either no structurally obvious ways to align the molecules or those ways fail to produce a satisfactory q^2.

ACKNOWLEDGMENT

This work is supported in part by a grant from the National Science Council, ROC (NSC91-2313-B007-001). The SYBYL CoMFA PLS and Cerius^2 PLS studies are conducted at the National Center for High Performance Computing, ROC.

Supporting Information Available: Procedures for using the Cerius2 program to perform the PCA for feature selection and data set trimming. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Green, S. M.; Marshall, G..R. 3D-QSAR: A Current Perspective. *Trends Pharmcol. Sci.* **1995**, *16*, 285−291.

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

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

(4) Lingren, F.; Geladi, P.; Rannar, S.; Wold, S. Interative Variable Selection (IVS) for PLS Part I: Theory and Algorithm *J. Chemom.* **1994**, *8*, 349−363.

(5) Baron, M.; Constantino, 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.

(6) Cho, J. S.; Tropsha, A. Cross-Validated R^2-Guided Region Selection for Comparative Molecular Field Analysis: A Simple Method to Achieve Consistent Results. *J. Med. Chem*. **1995**, *38*, 1060−1066.

(7) Hasegawa, K.; Kimura, T.; Funatsu, K. GA Strategy for Variable Selection in QSAR Studies: Application of GA-Based Region Selection to a 3D-QSAR Study of Acetylcholinesterase Inhibitors. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 112−120.

(8) Dunn, W. J., III; Wold, S. A Structure-Carcinogenicity Study of 4-Nitroquinoline-1-Oxides Using the SIMCA Method of Pattern Recognition. *J. Med. Chem.* **1978**, *21*, 1001−1007.

(9) Hodes, L.; Hazard, G. F.; Geran, R. I.; Richman, S. A. Statistical Heuristic Method for Automated Selection of Drugs for Screening. *J. Med. Chem.* **1977**, *20*, 469−475.

(10) Shemetulskis, N. E.; Dunbar, J. B., Jr.; Dunbar, B. W.; Moreland, D. Chemical Database Clustering and Analysis. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 407−416.

(11) Menard, P. R.; Lewis, R. A.; Mason, J. R. Rational Screening Set Design and Compound Selection: Cascaded Clustering. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 497−505.

(12) Jain, A.; Zongker, D. Feature Selection: Evaluation, Application, and Small Sample Performance. *IEEE Trans. Patter. Anal. Mech. Intellig.* **1997**, *19*, 153−158.

(13) Theodoridis, S.; Koutroumbas, K. *Pattern Recognition*; Academic Press: New York, 1998.

(14) Hultén, J.; Bonham, N. M.; Nillroth, U.; Hansson, T.; Zuccarello, G.; Bouzide, A.; Åqvist, J.; Classon, B.; Danielson, U. H.; Karlén, A.; Kvarnström, I.; Samuelsson, B.; Hallberg, A. Cyclic HIV-1 Protease Inhibitors Derived from Mannitol: Synthesis, Inhibitory Potencies, and Computational Predictions of Binding Affinities. *J. Med. Chem.* **1997**, *40*, 885−897.

(15) Alterman, M.; Björsne, M.; Mühlman, A.; Classon, B.; Kvarnström, I.; Danielson, H.; Markgren, P. O.; Nillroth, U.; Unge, T.; Hallberg, A.; Samuelsson, B. Design and Synthesis of New Potent C2−Symmetric HIV-1 Protease Inhibitors. Use of L-Mannaric Acid as a Peptidomimetic Scaffold. *J. Med. Chem.* **1998**, *41*, 3782−3792.

(16) Nugiel, D. A.; Jacobs, K.; Cornelius, L.; Chang, C.; Jadhav, P. K.; Holler, E. R.; Klabe, R. M.; Bacheler, L. T.; Cordova, B.; Garber, S.; Reid, C.; Logue, K. A.; Gorey-Feret, L. J.; Lam, G. N.; Erickson-Viitanen, S.; Seitz, S. P. Improved P1/P1′Substituents for Cyclic Urea Based HIV-1 Protease Inhibitors: Synthesis, Structure−Activity relationship, and X-ray Crystal Structure Analysis. *J. Med. Chem.* **1997**, *40*, 1465−1474.

(17) Kroemer, R. T.; Ettmayer, P.; Hecht, P. 3D-Quantitative Structure−Activity Relationship of Human Immunodeficiency Virus type-1 Proteinase Inhibitors: Comparative Molecular Field Analysis of 2-Heterosubstituted Statine Derivatives-Implications for the Design of Novel Inhibitors. *J. Med. Chem.* **1995**, *38*, 4917−4928.

(18) Scholz, D.; Billich, A.; Charpiot, B.; Ettmayer, P.; Lehr, P.; Rosenwirth, B.; Schreiner, E.; Gstach, H. Inhibitors of HIV-1 Proteinase Containing 2-Heterosubstituted 4-Amino-3-Hydroxy-5- Phenylpentanoic Acid: Synthesis, Enzyme Inhibition, and Antiviral Activity. *J. Med. Chem.* **1994**, *37*, 3079−3089.

(19) Alterman, M.; Andersson, H. O.; Garg, N.; Ahlsén, G.; Lövgren, L.; Classon, C.; Danielson, U. H.; Kvarnström, I.; Vrang, L.; Unge, T.;

Samuelsson, B.; Hallberg, A. Design and Fast Synthesis of C−Terminal Duplicated Potent *C*2-Symmetric P1/P1−Modified HIV-1 Protease Inhibitors. *J. Med. Chem.* **1999**, *42*, 3835−3884.

(20) Debnath, A. K. Three-Dimensional Quantitative Structure−Activity Relationship Study on Cyclic Urea Derivatives as HIV-1 Protease Inhibitors: Application of Comparative Molecular Field Analysis. *J. Med. Chem.* **1999**, *42*, 249−259.

(21) Prabhakar, K.; Jadhav, P. A.; Woerner, F. J.; Chang, C. H.; Garber, S. S.; Anton, E. D.; Bacheler, L. T. Cyclic Urea Amides: HIV-1 Protease Inhibitors with Low Nanomolar Potency Against both Wild-Type and Protease Inhibitor Resistant Mutants of HIV. *J. Med. Chem.* **1997**, *40*, 181−191.

(22) Tou, J. T.; Gonzalez, R. C. *Pattern Recognition Principles*; Addison-Wesley: Reading, MA, 1974.

(23) Jurs, P. C. *Computer Software Applications In Chemistry*; John Wiley & Sons: New York, 1996.

(24) SYBYL 6.8; The Tripos Associates: 1699 S. Hanley Rd., St. Louis, MO.

(25) Accelrys Inc., Cerius^2 Modeling Environment, Release 4.0, Accelrys Inc.: San Diego, 2002.

(26) Lin, T. H.; Wang, G..M.; Hsu, Y. H. Classification of Some Active HIV-1 Protease Inhibitors and Their Inactive Analogues Using Some Uncorrelated Three-Dimenstional Molecular Descriptors and a Fuzzy C-Means Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1490−1504.

(27) Xu, J.; Stevenson, J. Drug-Like Index: a New Approach to Measure Drug-Like Compounds and Their Diversity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1177−1187.

(28) Hosoya, H. Topological Index: a Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn.* **1971**, *44*, 2332−2339.

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

(30) Balaban, A. T. Applications of Graph Theory in Chemistry. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 334−343.

(31) Randic, M.; Wilkins, C. L. Graph Theoretical Approach to Recognition of Structural Similarity in Molecules. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 31−37.

(32) Klebe, G.; Abraham, V.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130−4146.

(33) Montgomery, D. C.; Peck, E. A. *Introduction to linear regression analysis*; John Wiley & Sons: New York, 1982.

(34) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numeriacl Recipes. The Art of Scientific Computing*; Cambridge University Press: New York, 1986.

(35) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Model*. **2002**, *20*, 269−276.

(36) Godden, J. W.; Bajorath, J. Differential Shannon Entropy as a Sensitive Measure of Differences in Database Variability of Molecular Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1060−1066.

(37) Devijver, P. A.; Kittler, J. *Pattern Recognition: A Statistical Approach*; Prentice Hall: Englewood Cliffs, 1982.

(38) Lee, C.; Landgrebe, D. A. Decision Boundary Feature Extraction for Neural Networks. *IEEE Trans. Neural Networks* **1997**, *8*, 75−83.

(39) Fukunaga, K. *Introduction to Statistical Pattern Recognition*, 2nd ed.; Academic Press: New York, 1990.

CI030295A