Journal of Computer-Aided Molecular Design **17:** 155–171, 2003.

Classification of HIV protease inhibitors on the basis of their antiviral potency using radial basis function neural networks

S.J. Patankar & P.C. Jurs*

Department of Chemistry, 152 Davey Laboratory, Penn State University, University Park, PA 16802, USA

MS received 29 October 2002; accepted for publication 26 January 2003

Key words: Classification, GA, HIV protease inhibitors, *k*-NN, majority rules, QSAR, RBFNN classifier.

Summary

HIV protease inhibitors are being used as frontline therapy in the treatment of HIV patients. Multi-drug-resistant HIV mutant strains are emerging with the initial aggressive multi-drug treatment of HIV patients. This necessitates continued search for novel inhibitors of viral replication. These protease inhibitors may further be useful as pharmacological agents for inhibition of other viral replication. Classification models of HIV Protease inhibitors are developed using a data set of 123 compounds containing several heterocycles. Their inhibitory concentrations expressed as log (IC_{50}) ranged from -1.52 to 2.12 log units. The dataset was divided into active and inactive classes on the basis of their antiviral potency. Initially a two-class problem (active, inactive) is explored using *k*-nearest neighbor approach. In order to introduce non-linearity in the classifier different approaches were investigated. This led to the goal of a fast, simple, minimum user input, radial basis function neural network (RBFNN) classifier development. Then the same two-class problem was resolved using the (RBFNN) classifier. A genetic algorithm with RBFNN fitness evaluator was used to search for the optimum descriptor subsets. The application of majority rules was also tested for the RBFNN classification. The best six descriptor model found by the new cost function showed predictive ability in the high 80% range for an external prediction set.

Introduction

Advancements in the treatment of HIV infections have received considerable attention and coverage by the media in recent years. In particular, the HIV protease inhibitors have significantly changed the outcome of the disease. Peoples' perception of the HIV infection has changed in the past five years, and HIV infection is seen more as a chronic infection than the fatal killer of five years ago. The introduction of antiretroviral therapies has a dramatic impact on life span and life style. In fact the combination of protease inhibitors along with reverse transcriptase inhibitors has become a standard form of frontline therapy [1]. Two major problems encountered with the therapy are dosage regimen and HIV mutations [2] to multi-drug resistant strains due to aggressive initial therapy. Only half of patients starting the treatment actually maintain lower viral load [3]. The outlook is even grimmer for patients who are considered for salvage therapy [4,5]. This necessitates the continued search for inhibitors of viral replication. In addition, insight into preventing viral replication might also benefit development of inhibitors for other viral diseases such as Hepatitis C. Thus it is critical to continue research on protease inhibitors.

Viral replication can be interfered with at two points – one at the reverse transcriptase stage and the other at the protease stage. The current therapy includes the combination of both types of intervention. Nucleosidic and non-nucleosidic reverse transciptase inhibitors are available for intervention at the first stage. Several papers have been published on HIV protease inhibitors [6–9] that involve inhibition at the second stage.

In this paper we have used computer modeling to classify a structurally diverse set of HIV Protease inhibitors with known antiviral potencies. This data set consists of six moieties of various dihydropyranones.

*Correspondence: E-mail pcj@psu.edu

Classification of the actives/inactives was done on the basis of molecular structure alone. The predictive ability of all models developed is examined using external prediction sets. This modeling technique can be used even when complete knowledge of the binding site is unavailable. Models developed could be used to screen libraries of compounds to identify those likely to display activity as HIV Protease inhibitors.

Experimental

This study was performed on a combined data set taken from three papers by Hagen *et al.* [10–12] To maximize the likelihood of finding a relationship between chemical structure and HIV Protease inhibitory activity, only the data collected under the same experimental conditions is used. The combined data set included 129 compounds, out of which 123 compounds had reported IC_{50} values and 7 compounds had reported K_i values. The 123 compounds were variations of six different moieties, four of which are shown below.

The compounds were evaluated by Hagen *et al.* [10–12] for their *in vitro* activity against affinity purified HIV-1 protease [11] (His-Lys-Ala-Arg-Val-Leu-(p-NO_2-Phe)-Glu-Ala-Nle-Ser-NH_2 as the source of the substrate [12]. The Leu-(p-NO_2-Phe) bond was cleaved by the enzyme and the cleavage products were separated by HPLC. The absorbance was measured by UV spectrometry [11]. The in vitro activity was expressed as the nanomolar concentration of the test compound which inhibited enzyme activity by 50% (IC_{50}).

The molecular weight for the data set compounds varied from 427 to 592 amu with a mean of 553 amu. The log IC_{50} values ranged from -1.52 to 2.12 log (nM) units. All compounds contained at least one oxygen and sulfur, several contained multiple oxygens and sulfurs, and 90 compounds contained nitrogen. Overall, only 11 compounds contained halogens out of which nine compounds contained fluorine and two compounds contained chlorine. Additionally, all compounds were aromatics with anywhere from two to five aromatic rings. Structural information for all the compounds is available as supplementary material in Table 1.

In this paper we have investigated the classification of HIV protease inhibitors on the basis of their antiviral potencies. To generate the two-class models the data set was divided into active and inactive compounds on the basis of their activity. As there were no reported guidelines to use on this data set for the required subdivisions, compounds with log $IC_{50} < 0.85$ were considered active and compounds with log $IC_{50} \geq 0.85$ were considered inactive. A set of 12 compounds was randomly selected as a prediction set (PSET), and these compounds were not used anywhere in the model development process. The remaining 111 compounds comprised the training (TSET). Table 2 shows the distribution of compounds into classes for the data set. For the development of a neural network classifier, discussed later in the paper, an additional set of 12 compounds was selected and used as a cross-validation set (CVSET). This reduced the training set to 99 compounds for that portion of the work. However the same prediction set was used for the final model validations for all the models developed during this study.

Structure entry, descriptor generation, feature selection and model development and validation are four basic steps in the development of classification models. Several classification studies have been reported in the literature [13–16]. Numerous pattern recognition techniques are available, however in this paper we have investigated two approaches, *k*-nearest neighbor and radial basis function neural network approaches for classification.

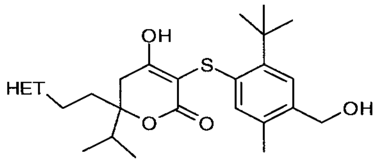
Structure entry

The compounds were sketched as 2-D representations using HyperChem [17], and optimized 3-D conformations were generated. Compound structures were stored as connection tables, which contain information about atom types, bond types, and bond connections. They were used to generate topological descriptors. At a later stage these structures were further refined to their lowest energy states using MOPAC [18], a semi empirical molecular modeling routine. A PM3 Hamiltonian was selected for geometry optimization [19]. These optimum 3-D conformations were used for generation of descriptors dependent on geometry. The compounds were also optimized using AM1 Hamiltonian before calculation of descriptors related to charges.

Descriptor generation

To relate molecular structure to HIV protease inhibitory activity, descriptors that accurately encode the structural features responsible for the observed activities are necessary. The ADAPT (**A**utomated **D**ata

Table 1. Structures, observed and predicted classes for HIV protease inhibitors



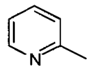
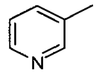
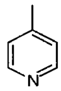
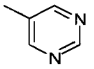
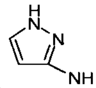
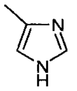
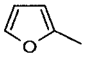
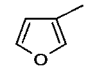
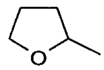
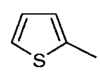
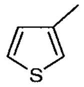
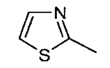
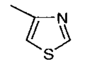
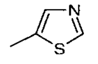
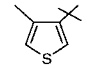
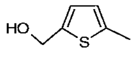
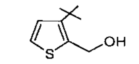
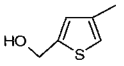
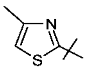
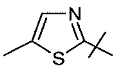
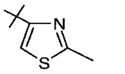
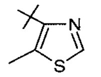
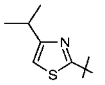
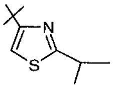
Comp	HET	IC_{50}[a]	k-NN Classifier		RBFNN Classifier	
#		nM	TRUE Class	Predicted Class	TRUE Class	Predicted Class
1[b]	Ph-NH_2	0,11	2	2	2	2
2[c]		1,2	2	2	2	2
3		0,94	2	2	2	1
4		1,5	2	2	2	2
5		17	2	2	2	2
6		0,28	2	2	2	2
7[b]		70	2	1	2	2
8[c]		0,78	2	2	2	2
9		0,34	2	2	2	1
10		5	2	2	2	2
11		0,35	2	2	2	2
12		0,17	2	2	2	1

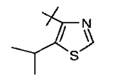
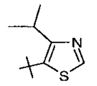
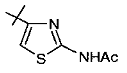
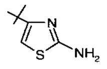
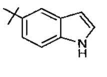
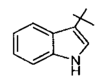
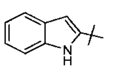
Table 1. Continued

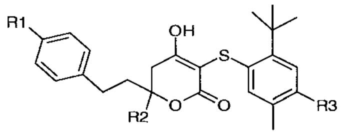
Comp	HET	$IC_{50}{}^{a}$	k-NN Classifier		RBFNN Classifier	
#		nM	TRUE Class	Predicted Class	TRUE Class	Predicted Class
13		0,57	2	2	2	1
14		0,13	2	2	2	2
15		0,24	2	2	2	2
16		0,41	2	2	2	2
17		0,58	2	2	2	2
18		14	2	1	2	2
19		0,19	2	2	2	2
20		0,45	2	2	2	2
21		1,9	2	2	2	2
22		13	2	1	2	2
23		1,3	2	2	2	2
24		14	2	1	2	2
25		1,6	2	2	2	2

Analysis and Pattern Recognition Toolkit) software package [20–21] was used to calculate about 125 molecular descriptors for each compound. As some of the compounds were beyond software limits, descriptor generation routines were modified to accommodate the larger molecules. This led to almost 50% reduction in the number of descriptors generated. The calculated descriptors encode the geometric, topological, electronic features and flexibility of the compounds.

Topological descriptors [22–25] are calculated on the basis of a two-dimensional sketch of the compound. A significant advantage of these type of descriptors is that geometry optimization of the structure is not needed. Topological descriptors calculated included connectivity indices, molecular distance edge descriptors, kappa indices, and flexibility indices. Connectivity indices and molecular distance edge descriptors encoded information about molecular size and branching while *k*appa indices provided information about molecular shape using the two-dimensional structure. The new flexibility indices routine provided collective information about the flexibility of the mo-

Table 1. Continued

Comp	HET	IC_{50}^{a}	k-NN Classifier		RBFNN Classifier	
#		nM	TRUE Class	Predicted Class	TRUE Class	Predicted Class
26		5,8	2	2	2	2
27		0,8	2	2	2	2
28		18	2	1	2	2
29		0,95	2	2	2	2
30		0,98	2	2	2	2
31		0,72	2	2	2	2
32		1,4	2	2	2	2



Comp	R1	R2	R3	IC_{50}^{a}	k-NN Classifier		RBFNN Classifier	
#				nM	TRUE Class	Predicted Class	TRUE Class	Predicted Class
33	4 - OH	isopropyl	NH_2	2,7	2	2	2	2
34	4 - OH	isopropyl	$NHSO_2Me$	4	2	2	2	2
35	4 - OH	isopropyl	$NHSO_2Ph$	1,8	2	2	2	2

lecule and mass equivalence of the rotatable as well as rigid atoms.

Geometric descriptors [26,27] such as volume, solvent accessible surface areas, and gravitational indices provide information about size and ability of the compounds to interact with solvents on a 3-D basis. However these could not be calculated as some of the compounds in the data set were beyond the software limit.

Electronic descriptors [28] are frequently computationally demanding. These descriptors encode the electronic environment of the compounds. These descriptors were calculated after geometry optimization of all the compounds. Our own as well as MOPAC routines were developed to accommodate the larger molecules in the data set. Calculated descriptors included dipole moment, polarizability, electronegativity, energy of the highest occupied molecular orbital, and the energy of the lowest unoccupied molecular orbital. However due to the software limit CPSA [28] descriptors could not be calculated.

Table 1. Continued

Comp #	R1	R2	R3	IC_{50}[a] nM	k-NN Classifier TRUE Class	k-NN Classifier Predicted Class	RBFNN Classifier TRUE Class	RBFNN Classifier Predicted Class
36[b]	4 - OH	isopropyl	$NHSO_2Ph$(4 - F)	4,1	2	2	2	2
37[c]	4 - OH	isopropyl	$NHSO_2Ph$(4 - Cl)	5,3	2	2	2	2
38	4 - OH	isopropyl	$NHSO_2Ph$(4 - CF_3)	18,3	1	1	1	2
39	4 - OH	isopropyl	$NHSO_2Ph$(3 - CN)	0,21	2	2	2	2
40[b]	4 - OH	isopropyl	$NHSO_2Ph$(4 - CN)	2,2	2	2	2	2
41[c]	4 - OH	isopropyl	$NHSO_2Ph$(2 - thiophene)	8,7	1	1	1	1
42	4 - OH	isopropyl	$NHSO_2Ph$(2 - pyridyl)	2,6	2	2	2	2
43	4 - OH	isopropyl	$NHSO_2Ph$(3 - pyridyl)	2,6	2	2	2	2
44	4 - OH	methyl	NH_2	13	1	1	1	2
45	4 - OH	methyl	$NHSO_2Me$	9,7	2	1	2	2
46	4 - OH	methyl	$NHSO_2Ph$	1,5	2	2	2	2
47[b]	4 - OH	methyl	$NHSO_2Ph$(4 - F)	3,7	2	2	2	2
48[c]	4 - OH	methyl	$NHSO_2Ph$(4 - Cl)	3,5	2	2	2	2
49	4 - OH	methyl	$NHSO_2Ph$(4 - CF_3)	12,3	1	1	1	2
50	4 - OH	methyl	$NHSO_2Ph$(4 - CN)	2,1	2	2	2	2
51	4 - OH	methyl	$NHSO_2Ph$(2 - thiophene)	6,3	2	2	2	2

Feature selection

The process of feature selection entails pruning the descriptor pool through subjective and objective means. Objective feature selection eliminates descriptors based solely on their values. In this process the dependent variable information (class label) is not utilized. To avoid chance correlation the descriptor pool was reduced to a reasonable level. In practice the ratio of descriptors for compounds to the number of TSET observations used was less than or equal to 0.6. This was carried out using 111 TSET observations. Any descriptor containing identical values for 90% or more of the TSET observations was eliminated. Additionally pairwise correlations were calculated for all descriptors. One of any two descriptors with a correlation above 0.9 was eliminated. These two steps reduced the all descriptor pool of 125 descriptors to 35 descriptors. These were acceptable levels as the ratio of descriptors to TSET observations was well below 0.6 for all cases investigated. Then models were developed on the basis of the reduced pool.

Table 1. Continued

Comp #	R1	R2	R3	IC$_{50}$^{a} nM	k-NN Classifier TRUE Class	k-NN Classifier Predicted Class	RBFNN Classifier TRUE Class	RBFNN Classifier Predicted Class
52	4 - OH	methyl	NHSO$_2$ (N-methylimidazole)	1,9	2	2	2	2
53	4 - OH	△	NHSO$_2$Ph	3,6	2	2	2	2
54	4 - OH	△	NHSO$_2$Ph	8,6	1	1	1	2
55	H	isopropyl	NH$_2$	3,6	2	2	2	2
56	H	isopropyl	NHSO$_2$Me	0,83	2	2	2	2
57	H	isopropyl	NHSO$_2$Ph(4 - CF$_3$)	27	1	1	1	2
58	H	isopropyl	NHSO$_2$Ph(3 - CN)	1,4	2	2	2	2
59	H	isopropyl	NHSO$_2$Ph(4 - CN)	9,1	2	1	2	2
60	H	isopropyl	NHSO$_2$Ph(2 - thiophene)	3,6	1	2	1	2
61	H	isopropyl	NHSO$_2$Ph(2 - pyridyl)	3	2	2	2	2
62	H	isopropyl	NHSO$_2$Ph(3 - pyridyl)	3,5	2	2	2	2
63	H	isopropyl	NHSO$_2$ (N-methylimidazole)	1,9	2	2	2	2
64	H	isopropyl	NHSO$_2$ [2-(5CF$_3$)pyridyl)]	1,04	1	2	1	2
65^{b}	4 - OH	isopropyl	OH	0,03	2	2	2	2
66^{c}	4 - OH	isopropyl	OSO$_2$Me	2,2	2	2	2	2
67	4 - OH	isopropyl	OSO$_2$Ph	10,4	2	1	2	2

Model development and validation

The reduced descriptor pool was screened using genetic algorithm [29,30] (GA) evolutionary optimization. The GA feature selection routines were written in-house. Different subsets of descriptors were evaluated to see if they could develop classifiers to determine the HIV protease inhibitory activity. Models were formed using *k*-nearest neighbor analysis (*k*-NN), and the newly developed radial basis function neural networks (RBFNN) classifier.

k-NN is a non-parametric classification technique [31,32] that takes an unknown input pattern and assigns it to the class of the majority among its *k*-nearest neighbors in the training set on the basis of Euclidian distance metric. It is effective when probabilities of the feature variables are unknown. Model sizes ranging from 3 to 10 descriptors were reviewed. For each model size GA was used to search the descriptor space for the subset of descriptors that produced the lowest cost function. The models were chosen on the basis of lowest cost function and the fewest number of descriptors. Once selected, the optimum model was

Table 1. Continued

Comp	R1	R2	R3	IC_{50}[a]	k-NN Classifier		RBFNN Classifier	
#				nM	TRUE Class	Predicted Class	TRUE Class	Predicted Class
68	4 - OH	isopropyl	OSO_2Ph (4-F)	9,6	1	1	1	2
69	4 - OH	isopropyl	OSO_2Ph (4-CN)	4,5	1	1	1	2
70[b]	4 - OH	isopropyl	OSO_2 (2-thiophene)	21,8	1	1	1	1
71[c]	4 - OH	isopropyl	OSO_2 (N-methylimidazole)	2,4	2	2	2	2
72	4 - OH	isopropyl	OSO_2 (2-pyridyl)	5	2	2	2	2
73	4 - OH	isopropyl	OSO_2 (3-pyridyl)	3,5	2	2	2	2
74[b]	4 - OH	methyl	OSO_2Ph	9,8	2	1	2	2
75[c]	4 - OH	methyl	OSO_2Ph (4-F)	16,9	2	1	2	2
76	4 - OH	methyl	OSO_2 (N-methylimidazole)	2	2	2	2	2
77	4 - OH	cyclohexyl	OSO_2Ph	131	1	1	1	2
78	4 - OH	cyclohexyl	OSO_2Ph (4-CN)	21	1	1	1	2
79	4 - OH	cyclohexyl	OSO_2Ph (4-F)	29	1	1	1	2
80	4 - OH	cyclohexyl	OSO_2 (N-methylimidazole)	8	1	1	1	2
81	4 - OH	n-propyl	OSO_2 (N-methylimidazole)	2,2	2	2	2	1
82	4 - OH	isobutyl	OSO_2 (N-methylimidazole)	3,6	2	2	2	2
83	4-H	isopropyl	OH	1,1	2	2	2	2

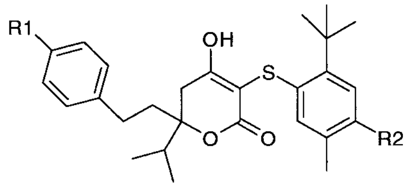
used to classify compounds in the PSET to verify the generalization ability of the model.

At this point two alternatives were evaluated in an effort to improve these linear classification methods. One was to transpose the non-linear data using an appropriate function to linear data and then use linear methods or hyperplanes such as SVM to classify the dataset [33–35]. The other was to explore networks capable of modeling non-linear phenomenon, as they would provide a potential advantage over the linear classifiers. This led to further investigation of radial basis function neural network as a (RBFNN) classifier.

Radial basis function (RBF) networks were introduced into the neural network world by Broomhead and Lowe [36]. The RBF model is based on locally tuned response of biological neurons. Attempts to speed up the training have been made [37]. Several applications in the field of optical character recognition and speech recognition have been reported [38–39]. However, the development of a RBFNN classifier is a complex problem [40].

Table 1. Continued

Comp #	R1	R2	R3	IC_{50}^{a} nM	k-NN Classifier TRUE Class	k-NN Classifier Predicted Class	RBFNN Classifier TRUE Class	RBFNN Classifier Predicted Class
84[b]	4-H	isopropyl	OSO_2 (N-methylimidazole)	3	2	2	2	2
85[c]	4-H	isopropyl	OSO_2 (2-pyridyl)	6,7	2	2	2	2
86	4-H	isopropyl	OSO_2 (3-pyridyl)	14	2	1	2	2
87	4-NH_2	isopropyl	OSO_2Ph (4-CN)	1,3	2	2	2	2
88	4-NH_2	isopropyl	OSO_2Ph(3-pyridyl)	4,1	2	2	2	2
89	4-NH_2	isopropyl	OSO_2 (N-methylimidazole)	2,2	2	2	2	2
90	4-NH_2	isopropyl	OSO_2Ph(2-thiophene)	18,5	1	1	1	2



Comp #	R1	R2	IC_{50}^{a} nM	k-NN Classifier TRUE Class	k-NN Classifier Predicted Class	RBFNN Classifier TRUE Class	RBFNN Classifier Predicted Class
91	4 - OH	$NHSO_2N(Me)_2$	3,6	2	2	2	2
		$NHSO_2NHEt$					
92[b]	4 - OH	$NHSO_2NHEt$	1,7	2	2	2	2
93[c]	4-H	$NHSO_2NHEt$	3,3	2	2	2	2
94	4 - OH	$OSO_2N(Me)_2$	3,4	2	2	2	2
95	4 - OH	OSO_2NHEt	4,9	2	2	2	2

The theoretical basis of the RBF approach lies in the field of interpolation of multivariate functions. The radial basis functions separate classes via hyperspheres. Given a set of N different points $\{x_i \in R^n \mid i = 1, \ldots. N\}$ and N real numbers $\{Y_i \in R \mid i = 1, \ldots. N\}$ construct a smooth function

$$F : R^n \longrightarrow R \qquad (1)$$

such that the interpolating conditions

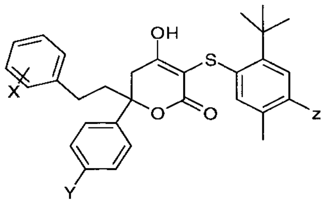
$$F(x_i) = y_i \quad i = 1, \ldots. N \qquad (2)$$

are satisfied. The points x_i are not assumed to satisfy any particular conditions. Choose a function $F(\bullet)$ to have the form

$$F(x) = \sum_{i=1}^{N} a_i \phi(\| x - x_i \|) \qquad (3)$$

Table 1. Continued

Comp #	R1	R2	$IC_{50}{}^a$ nM	k-NN Classifier TRUE Class	k-NN Classifier Predicted Class	RBFNN Classifier TRUE Class	RBFNN Classifier Predicted Class
96	4-NH_2	$OSO_2N(Me)_2$	3,9	2	2	2	2
97	4-NH_2	OSO_2NHEt	4,3	2	2	2	2
98	4-H	OSO_2NHEt	8,9	2	1	2	2
99	H	OSO_2Ph (4-CN)	0,7	2	2	2	2



Comp #	X	Y	Z	$IC_{50}{}^a$ nM	k-NN Classifier TRUE Class	k-NN Classifier Predicted Class	RBFNN Classifier TRUE Class	RBFNN Classifier Predicted Class
100	H	H	H	35	1	1	1	1
101	H	H	OH	33	1	1	1	2
102	H	H	$O(CH_2)_2OH$	6,8	1	2	1	2
103	H	H	CH_2OH	6,6	2	2	2	2
104	H	H	OMe	15	1	1	1	2
105	4 - OH	H	H	11	1	1	1	2
106	4-NH_2	H	H	24	1	1	1	1

where the function ϕ ($\bullet$) is a basis function that is continuous, $\|\bullet\|$ denotes the Euclidian norm, and vector x_i are the centers of the basis functions (usually assumed to be the data points themselves). Then, the interpolation conditions of equation 2 yield the following set of linear equations for the coefficients a

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1N} \\ A_{21} & A_{22} & \cdots & A_{2N} \\ \vdots & \vdots & & \vdots \\ A_{N1} & A_{N2} & \cdots & A_{NN} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}$$

or

$$\mathbf{A}a = y \qquad (4)$$

Where $a = (a_1, \ldots, a_N)$, $y = (y_1, \ldots, y_N)$ and $\mathbf{A}$ is an N × N matrix defined by

$$\mathbf{A} = (\phi(\| x_i - x_j \|)) \quad i, j = 1, \ldots\ldots, N$$

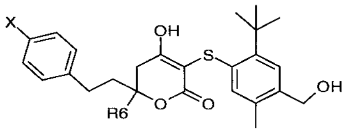
Provided the inverse of $\mathbf{A}$ exists, the solution a of the interpolation problem can be explicitly calculated and has a form of :

$$a = \mathbf{A}^{-1}y \qquad (5)$$

Several different radial basis functions as represented in the equations 6–10 were evaluated for the RBFNN classifier development where r represents the radius and σ represents the spread parameter. The cubic functions were not explored, as they are known to give rise

Table 1. Continued

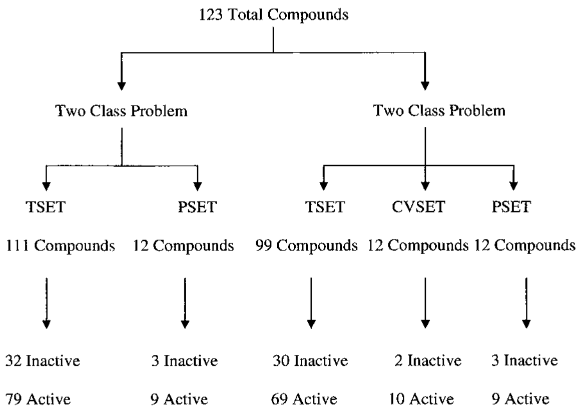
Comp #	X	Y	Z	IC_{50}[a] nM	k-NN Classifier		RBFNN Classifier	
					TRUE Class	Predicted Class	TRUE Class	Predicted Class
107	H	OH	H	40	1	1	1	2
108	H	NH_2	H	32	1	1	1	2
109	H	$O(CH_2)_2OH$	H	12	2	1	2	2
110	4 - OH	H	CH_2OH	1,7	2	2	2	2
111		H	CH_2OH	2,5	2	2	2	2
112	4-NH_2	H	CH_2OH	3,1	2	2	2	2
113[b]	3-NH_2	H	CH_2OH	4	2	2	2	2
114[c]	H	$O(CH_2)_2OH$	CH_2OH	1,4	2	2	2	2
115	H	$O(CH_2)_2OH$	$O(CH_2)_2OH$	6,4	2	2	2	2
116	H	$O(CH_2)_2OH$	OH	3,7	2	2	2	1
117	H	$O(CH_2)_2OH$	CH_2OCH_3	4,5	2	2	2	1
118	4 - OH	OH	CH_2OH	120	2	1	2	2



Comp #	X	R6	IC_{50}[a] nM	k-NN Classifier		RBFNN Classifier	
				TRUE Class	Predicted Class	TRUE Class	Predicted Class
119	OH	cyclohexyl	4,1	2	2	2	2

Comp #	X	R6	IC_{50}[a] nM	k-NN Classifier		RBFNN Classifier	
				TRUE Class	Predicted Class	TRUE Class	Predicted Class
120[b]	OH	isopropyl	3,6	2	2	2	2
121[c]	OH	methyl	4,3	1	2	1	1
122	NH_2	cyclohexyl	3,2	2	2	2	2
123	NH_2	isopropyl	2,7	2	2	2	2

Table 2. Distribution of compounds in TSET, CVSET and PSET



to a nonsingular linear system.

$$\phi(x) = e^{-r^2/2\sigma^2} \quad (6)$$

$$\phi(x) = r^2/\sigma^2 * \log r/\sigma \quad (7)$$

$$\phi(x) = r^2 * \ln r \quad (8)$$

$$\phi(x) = (r^2 + \sigma^2)^{-1/2} \quad (9)$$

$$\phi(x) = (r^2 + \sigma^2)^{1/2} \quad (10)$$

RBFNN classifier was developed in house to introduce nonlinearity into the classifiers. Usually RBFNN require the user to input multiple and complex parameters before you can start the optimizations process. Here an attempt is made to develop a classifier with minimum input from the user. A classifier is developed using the multi-quadratic radial basis function shown above in equation 10. A penalty term for the smaller radii (1/radii) was added which improved the overall classification rate. This classifier has radial basis function as the hidden layer in the three-layer, feed-forward computational neural network. This RBFNN uses a two-phase calculation, generating the centroids and spread parameters in the first phase, and generating weights and biases in the second phase. In the training phase the network parameters were determined and subsequently the output layer was adjusted. The output layer can be given as per equation 11.

$$y_k(x) = \sum_{m=1}^{n_h} a_{km}\phi_m(x) + b_k \quad (11)$$

Where y_k is the kth output unit for the input vector x, a_{km} is the weight connection between the kth output unit and the mth RBF unit, n_h is the number of RBF units, and b_k is the bias.

Some researchers have used each training set observation as a centroid and kept the width parameters value the same for every centroid. These networks do not generalize well, and large number of observations yield network size issues. Some papers have published work on RBFNNs using multiplayer perceptrons (MLPs) with backpropagation training [41]. In this paper a method is proposed to generate a self-configuring classifier. The initial centroids are formed using a *k*-nearest neighbor classifier. For this work the value of *k* was set at 3. As the order of presentation of the compounds to the network influences the formation of centroids, a committe of five random scrambles of the training set observations was used to generate the centroids while training the network. A singular value decomposition algorithm is used to compute weights and biases, as it handles large input matrices with high speed. Most algorithms have used a leave-N-out training, however here an external cross validation set is used for training. This external cross validation should give a better generalization ability to the network than the leave N out type of training.

A genetic algorithm coupled with RBFNN as a fitness evaluator is developed and used for performing feature selection. The cost function was computed for each descriptor subset. Instead of using a leave-N-out cross validation procedure, a separate CVSET was used to monitor the network performance. To improve the classification a penalty term was introduced for the smaller radii. This improved the classification by about 10%. The cost function given below gave the best results.

$$\begin{aligned} \mathrm{COST} &= T_{\mathrm{INC}} + 0.5 * |T_{\mathrm{INC}} - CV_{\mathrm{INC}}| \\ &\quad + 1/\mathrm{radii} \end{aligned} \quad (12)$$

Where the T $_{\mathrm{INC}}$ and CV_{INC} represent the average percent incorrectly classified for the TSET and CVSET respectively. This cost function also tested the ability of each model to generalize the external CVSET. Thus the models that could only predict the TSET compounds were avoided. In addition, models were avoided which reduced classification overall by increasing mis-classifications for a particular class, as this would affect data sets where there is an uneven distribution in the number of compounds in each class. Model sizes ranging from 3 to 10 descriptors were reviewed. For each model size GA was used to search the descriptor space for the subset of descriptors that produce the lowest cost function. Best models were chosen on the basis of lowest cost function and the fewest number of descriptors. Once selected the op-

Table 3. Six descriptors defining the optimal *k*-NN model

	Range		Average		
Descriptor[a]	Active	Inactive	Active	Inactive	Rel. Var[b]
FLEX2	196–84.1	166–84.1	115	117	4.80
EDIF1	18.0–13.6	18.0–13.6	14.5	14.8	0.003
FLEX5	14.3–1.69	9.41–1.76	3.31	3.74	1.59
MDE24	42.2–12.8	54.7–12.8	17.7	20.9	1.33
3SP3	2.00–00.0	2.00–00.0	0.81	0.59	0.313
ELOW1	11.6–3.45	12.4–3.45	5.38	5.83	1.11

[a] Explanations : FLEX2, mass of rotatable atom; EDIF1, difference between max. and min. E-state index; [42]FLEX5, distance weighted flexibility; MDE24, distance edge between S-Q carbons; [43]3SP3, tertiary SP3 carbons; ELOW1, distance between max. and min. E-state index. [42]

[b] Relative variance is the variance divided by the mean for a descriptor using all compounds for both the classes.

timum model was used to classify compounds in the CVSET to verify generalization ability of the model.

The computations for this work were performed on a DEC 3000 AXP Model 500 workstation. Those calculations involving HyperChem [24] were performed on a Pentium PC.

Results and discussion

k-NN classifier

GA was used to evaluate subsets with three to ten descriptors. The smallest descriptor subset that produced an acceptably low cost function was selected as optimal. These descriptor values for the TSET compounds were then used to develop the models. Once generated these models were used to classify all the compounds in the TSET and calculate the percent correct values. Then the compounds of the external prediction set were classified and the percent correct values were calculated for them.

Two-class problem

First, the problem of generation of binary models to differentiate actives from inactives was explored. Models were investigated using the reduced pool of 35 descriptors, using *k*-NN classification techniques. The optimal model formed from this reduced pool by *k*-NN classifier was a six-descriptor model. Of the 111 TSET compounds, 85.6% were classified correctly. The correct classification rate for the 12 PSET compounds was 83.3%, which clearly demonstrates that the optimum model is capable of classifying compounds not used in the model formation. The six descriptors selected in the best model are shown in Table 3. FLEX2 and FLEX5 describe the mass correlation of rotatable atoms and distance weightage of the flexibility index. EDIF and ELOW denote the difference in the electrotopological states of the molecule and lowest electrotopological state of the molecule respectively [42]. MDE24 describes the connection between the secondary and quaternary carbons [43]. 3SP3 descriptor is just a count of tertiary SP^3 carbons attached to three other carbon atoms. Table 3 also shows the range and average for the inactive and active classes for each of the six descriptors, as well as the relative variance for each descriptor. Attention should be drawn to the fact that the average descriptor value for the active compounds is usually smaller than the average descriptor value for the inactive compounds except for the 3SP3 descriptor. EDIF and 3SP3 descriptors have low relative variance values, but they still provide useful information to improve the model prediction rates.

Table 4a shows the confusion matrix for the TSET compounds using the six-descriptor model developed from all-descriptor reduced pool. Also shown are the correct classification rates for the classes using this model. As the inactive class is small, the prediction models seemed to misclassify more inactive compounds to increase the accuracy for the active and inactive classes.

The classification rate for the PSET compounds is close to that obtained for the TSET compounds, as is seen from Table 4b. Accuracy is good for the active and inactive classes. In an effort to keep the prediction

Table 4. Confusion matrix for the TSET and PSET compounds using the optimal six descriptor *k*-NN model

a. TSET Compounds

	Predicted class		
Actual Class	Inactive	Active	% Correct
Inactive	20	12	64.5
Active	4	75	94.9

b. PSET Compounds

	Predicted class		
Actual Class	Inactive	Active	% Correct
Inactive	2	1	66.6
Active	0	9	100.0

set completely external for all the work done on this data set, the distribution was not changed to make the PSET reasonably represented in the two classes. This added more difficulty to the problem at hand. It is also important to note that in spite of these problems the active compounds had a classification rate of 100%. In addition even if the percentage of inactive classified correctly is low it is important to note that only one compound was misclassified, and overall 11 out of 12 compounds were classified correctly.

RBFNN classifier

The newly developed RBFNN classifier with the most commonly used Gaussian radial basis function (equation 6) did not yield the best model among all the RBFNN models. The RBFNN classifier developed using the equation 10 yielded the best model. GA with the RBFNN as a fitness evaluator was used to evaluate subsets with three to ten descriptors. The smallest descriptor subset that produced an acceptably low cost function was selected as optimal. These descriptor values for the TSET compounds were then used to develop the models. Once generated these models were used to classify all the compounds in the TSET and calculate the percent correct values. Then the compounds of the external prediction set were classified and the percent correct values were calculated for them. This optimal model was a six-descriptor model. Of the 99 PSET compounds 75.9% were classified correctly. The classification rate for the CVSET was 81.7% and the 12 compound PSET was 83.3%. It is important to note that the CVSET in this case was part of the TSET in the *k*-NN classification results. Thus, a reasonable rate of classification was achieved.

Again the best model selected was based on six descriptors. The six topological descriptors selected in the best model are shown in Table 5. MDE24 describes the topological distances between secondary and quaternary carbons. EDIF1 denotes the difference between the maximum and minimum atomic electro-topological state value. Electro-topological state values provide information about inter-molecular interactions. The MOLC8 represents molecular connectivity for path clusters of four. NSB simply counts all the number of single bonds. This is not surprising based on the characterization of the data set. QNEG is the most negative charge on the atoms. which encodes the basicity. HOMO yields the information about the highest occupied molecular orbital and thus the reactivity of the molecule. The MOLC descriptor with flag eight gives information about degree of branching, which indirectly describes the stearic size of the molecule and also relates to the flexibility of the compounds. Flexibility of the compounds is an important factor influencing the binding at the HIV Protease site. Table 5 also shows the range and average for the inactive and active classes for each of the six descriptors, as well as the relative variance for each descriptor. In this case four of the six descriptors have the average descriptor value for the active compounds smaller than the average descriptor value for the inactive compounds. The fifth is very close and the sixth has the trend reversed. Again EDIF, MOLC and ELOW descriptors have very small relative variance values, but they still provide equally useful information to improve the model prediction rates.

As RBFNN classifier is sensitive to the order of presentation of TSET members in the first phase, the network was trained using five random scrambles of the TSET observations. If majority rules method was applied to the classification significant change in classification rates were observed. As seen from Table 6 by using the majority rules the TSET correct classification rate increased by 19.8%, the CVSET rate increased by 12.2%, however the PSET classification rate remained the same.

Table 7a shows the confusion matrix for the CVSET compounds using the six-descriptor model developed from all-descriptor reduced pool and majority rules method. Also shown are the correct classification rates for the classes using this model. The table demonstrates that the cost function found descriptor

Table 5. Six descriptors defining the optimal RBFNN model

	Range		Average		
Descriptor[a]	Active	Inactive	Active	Inactive	Rel. Var[b]
MDE23	32.7–6.75	53.5–6.80	13.7	18.9	1.87
EDIF1	32.7–13.6	18.0–13.6	14.5	14.8	0.03
MOLC8	4.71–2.84	4.74–3.00	4.01	3.80	0.04
NSB	28.0–18.0	31.0–17.0	24.4	23.2	0.18
QNEG	1.34–0.42	1.38–0.42	0.67	0.71	0.09
HOMO	−7.65–8.51	−7.66–8.73	−8.25	−8.53	0.002

[a]Explanations : MDE23, distance edge between S-T carbons; [43]EDIF1, difference between max. and min. E-state index; [42]MOLC8, molecular connectivity for a path cluster of four; NSB, number of single bonds; QNEG, most negative charge on the atom; HOMO, highest occupied molecular orbital.

[b]Relative variance is the variance divided by the mean for a descriptor using all compounds for both the classes.

Table 6. Effect of using the majority of compounds on classification rate

	TSET % Correct	CVSET % Correct	PSET % Correct
Initial	75.9%	81.7%	83.3%
Majority rule	90.9%	91.7%	83.3%
	(91/99)	(11/12)	(10/12)

Table 7. Confusion matrix for the CVSET and PSET compounds using the optimal Six descriptor RBFNN model

a. CVSET Compounds

	Predicted class		
Actual Class	Inactive	Active	% Correct
Inactive	2	0	100.0
Active	1	9	88.9

b. PSET Compounds

	Predicted class		
Actual Class	Inactive	Active	% Correct
Inactive	1	2	33.3
Active	0	9	100.0

subsets that are not biased for or against any one class for the active and inactive CVSET compounds. Even though the inactive class is small, the prediction models seemed to classify both inactive compounds and active compounds to similar extents. The percentage is lower for the active class but as only one of the actives is misclassified, and overall 11 out of 12 compounds are classified correctly. This model classifies the external prediction set well and therefore seems to generalize well.

The classification rate for the PSET compounds remains the same with the majority rules method. It is also same as the one obtained for the PSET compounds using *k*-NN classifier, as is clearly seen from Table 7b. It is also important to note that in spite of 83.3% results the active compounds had a classification rate of 100%. In addition overall 10 out of 12 compounds were classified correctly using majority rules. One of the misclassified compounds has failed every classification.

This model predicted 10 out of 12 PSET compounds correctly. This model classified all the active compounds correctly. Compound number seven was the inactive compound that was misclassified by all the models. Compound seven was the only compound in the data set that had an imidazole ring without any substituent as the heterocycle attached to the dihydropyranone ring, as seen from Table 1. So without further investigation into the reported activity results it could not be determined weather the problem is due to lack of exposure in the training phase or not.

To ensure that the results were not due to chance, Monte Carlo experiments were conducted in which models were generated after scrambling of class labels. These results were close to the random assignments. The result clearly demonstrated that the predictive ability of the six descriptor *k*-NN model and the six descriptor RBFNN model were very unlikely to have been due to chance.

Summary and conclusion

Computational models were developed for a small data set of pharmaceutically interesting heterocycles with substitution on six different moieties, using a topological, electronic and flexibility descriptor pool using different classification methods. In spite of a very small sized PSET (9 active, 3 inactive) no false negatives were found for PSET predictions using all optimal models. The two-class problem has over 80% classification rate was achieved using k-NN classifier. The flexibility descriptor proved useful in the current study. This model correctly classified 11 out of 12 PSET compounds correctly. In addition all the active PSET compounds (9 out of 9) were classified correctly. One of the two compounds that was misclassified failed all classification models.

The newly developed RBFNN classifier, which included penalty for the smaller radii, and GA routine with this RBFNN as fitness evaluator, yielded good results. Over 80% correct classification was achieved for the two-class problem, with 100% correct classification rate for the actives. When majority rules method was used the classification rate increased significantly for the TSET and the CVSET, however the PSET remained the same. The external cross validation set showed that the neural network generalized well.

Overall good classification accuracies were obtained for this data set using these descriptors and both the classification techniques. In addition to good classification rates the models had no false negatives. Flexibility was selected in about 40% of all top models in the GA search. These models were developed from an external prediction set, which was not exposed to during the process of model development and still yielded good classification rates for the two-class (active, inactive) problem. The newly developed RBFNN with penalty function for smaller radii needed only minimal inputs from the user, in addition to being fast.

References

1. Carpenter, C.C., Fischl, M.A., Hammer, S.M., Jacobson, D.M. and Katzenstein, D.A., J. Am. Med. Assoc., 280 (1998) 78.
2. Erickson, J.W., Gulnick, S.V. and Markowitz, M., AIDS, (London), 13 (Suppl. A), (1999) S189.
3. Valdez, H.N., Lederman, M.M., Wooley, I., Walker, C.J., Vernon, L.T., Heis, A. and Gripshover, B.M., Arch. Int. Med., 159 (1999) 1771.
4. Battegay, M., Harr, T., and Sponagel, L., Ann. Med., 31, 4, (1999) 253.
5. Deeks, S.G., Adv. Exp. Med. Biol., 458 (1999) 175.
6. Tummino, P.J., Vara Prasad, J.V.N., Fergusson, D., Nouhan, C., Graham, N., Domagala, J.M., Ellsworth, E., Gajda, C., Hagen, S.E., Lunney, E.A., Para, K.S., Tait, B.D., Pavolovsky, A., Erickson, J.W., Gracheck, S., McQuade, T.J., and Hupe, D.J., Bioorg. Med. Chem., 4 (1996) 1401.
7. Romaines, K.R. and Chrusciel, R.A., Curr. Med. Chem., 2 (1995) 825.
8. Thaisrivongs, S., Skulnik, H.I., Turner, S.R., Strohbach, J.W., Tommasi, R.A., Johnson, P.D., Aristoff, P.A., Judge, T.M., Gammill, R.B., Morris, J.K., Romaines, K.R., Chruscial, R.A., Hinshaw, R.R., Chong, K.T., Tarpley, W.G., Poppe, S.M., Slade, D.E., Lynn, J.C., Horng, M.M., Tomich, P.K., Seest, E.P., Dolak, L.A., Howe, W.J., Howard, G.M., Schwende, F.J., Schwende, F.J., Toth, L.N., Padbury, G.E., Wilson, G.J., Shiou, L., Zipp, G.L., Wilkinson, K.F., Rush, M.J., Koeplinger, K.A., Zhao, Z., Cole, S., Zaya, R.M., Kakuk, T.J., Janakiraman, M.N. and Watenpaugh, K.D., J. Med. Chem., 39 (1996) 4349.
9. Tait, B.D., Domagala, J.M., Ellsworth, E.L., Vara Prasad, J.V.N., Ferguson, D., Graham, N., Hupe, D., Nouhan, C., Tummino, P.J., Humblet, C., Lunney, E.A., Pavlovsky, A., Rubin, R.J., Baldwin, E.T., Bhat, T.N., Erickson, J.W., Gulnik, S. and Liu, B., J. Med. Chem., 40 (1997) 3781.
10. Hagen, S.E., Domagala, J., Gajda, C., Lovdahl, M., Tait, B.D., Wise, E., Holler, T., Hupe, D., Nouhan, C., Urumov, A., Zeikus, G., Zeikus, E., Lunney, E.A., Pavlovsky, A., Gracheck, S.J., Saunders, J., VanderRoest, S. and Brodfuehrer, J., J. Med. Chem., 44 (2001) 2319.
11. Boyer, F.E., Vara Prasad, J.V.N., Domagala, J.M., Ellsworth, E.L., Gajda, C., Hagen, S.E., Markoski, L.J., Tait, B.D., Lunney, E.A., Palvosky, C., Ferguson, D., Graham, N., Holler, T., Hupe, D., Nouhan, C., Tummino, P.J., Urumov, A., Zeikus, E., Zeikus, G., Gracheck, S.J., Sanders, J.M., Vander Roest, S., Brodfuehrer, J., Iyer, K., Sinz, M., Gulnik, S.V., and Erickson, J.W., J. Med. Chem., 43 (2000) 843.
12. Hagen, S.E., Vara Prasad, J.V.N., Boyer, F.E., Domagala, J.M., Ellsworth, E.L., Gajda, C., Hamilton, H.W., Markoski, L.J., Steinbaugh, B.A., Tait, B.D., Lunney, E.A., Tummino, P.J., Ferguson, D., Hupe, D., Nouhan, C., Gracheck, S.J., Sanders, J.M. and VanderRoest, S., J. Med. Chem., 40 (1997) 3707.
13. Kim, C.U., McGee, L.R., Krawczyk, S.H., Harwood, E., Harada, Y., Swaminathan, S., Bischofberger, N., Chen, M.S., Cherrington, J.M., Xiong, S.F., Griffin, L., Cundy, K.C., Lee, A., Yu, B., Gulnik, S. and Erickson, J. W., J. Med. Chem, 39 (1996) 3431.
14. Judge, T.M., Phillips, G., Morris, J.K., Lovasz, K.D., Romaines, K.R., Luke, G.P., Tulinsky, J., Tustin, J.M., Chruscie, Dolak, L.A., Mizsak, S.A., Watt, W., Morris, J., Vander Valde, S.L., Strohbach, J.W. and Gammill, R.B., J. Am. Chem. Soc., 119,15, (1997) 3627.
15. Baures, P.W., Org. Lett., 1, 2 (1999) 249.
16. Jadhav, P.K., Ala, P., Woerner, F.J., Chang, C.H., Garber, S.S., Anton, E.D. and Bacheler, L.T., J. Med. Chem., 40 (1997) 181.
17. Hypercube Inc. Waterloo, OH.
18. Stewart, J.P.P., MOPAC 6.0; Quantum Chemistry Program Exchange, Indiana University, Bloomsburg, IN, Program 455.
19. Stewart, J.P.P., J. Comput.-Aided Mol. Des., 4 (1990) 1.

20. Stuper, A.J., Brugger, W.E. and Jurs, P.C., Computer-Assisted Studies of Chemical Structure and Biological Function, Wiley-Interscience, New York, N.Y., 1979.
21. Jurs, P.C., Chou, T.J. and Yuan, M., In Computer -Assisted Drug Design, Olsen, E.C. and Christoffersen, R.E. (Eds.) American Chemical Society: Washington D.C.,1979, pp 103-129.
22. Kier, L.B. and Hall, L.H., J. Chem. Inf. Comput. Sci., 37 (1997) 548.
23. Madan, A.K., Gupta, S. and Singh, M., J. Chem. Inf. Comput. Sci., 39 (1999) 272.
24. Cao, C., Huaxue Tongbao, 22 (1996) 1238.
25. Balaban, A.T., Chem. Phys. Lett., 89 (1982) 399.
26. Bondi, A., J. Phys. Chem., 68 (1964) 441.
27. Stouch, T.R. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 26 (1986) 4.
28. Stanton, D.T. and Jurs, P.C., Anal. Chem., 62 (1990) 2323.
29. Luke, B.T., J. Chem. Inf. Comput. Sci., 34 (1994) 1279.
30. Kimura, T., Hasegawa, K. and Fanatsu, K., J. Chem. Inf. Comput. Sci., 38 (1998) 276.
31. Duda, R.O. and Hart, P.E., Pattern Classification and Scene Analysis, John Wiely & Sons, New York, 1973.
32. Dasarathy, B.V. Nearest Neighbour, NN Norm: NN Pattern Classification Techniques, IEEE Computer society Press, Los Alamitos, CA, 1991.
33. Boser, B.E., Guynon, I.M. and Vapnik, V.N., A Training Algorithm for Optimal Margin Classifiers. In Proceedings of 5th Annual Workshop on Computational Learning Theory, Haussler, D., (Eds.) ACM Press, 1992.
34. Osuna, E., Freund, R. and Girosi, F. Training Support Vector Machines: An application of Face Detection., (1997) 130.
35. Zhang, T. and Oles, F.J. Text Categorization Based on Regularized Linear Classification Methods., 4 (2001) 5.
36. Broomhead, D. and Lowe, D. Multivariable functional interpolation and adaptive Networks. Complex Systems, 2, 321-355.
37. Darken, C. and Moodey, J. Towards faster stochastic gradient search. Advances in neural information processing systems, 4 (1991) 1009.
38. Schwenker, F., Kestler, H.A. and Palm, G., Neural Networks, 14 (2001) 439.
39. Phillips, W.J., Tosuner, C. and Robertson, W. Speech Recognition Techniques Using RBF Networks. Proceedings of the IEEE WESCANEX95, Communications, Power, and Computing; IEEE: New York, 1995, 1, 185–190.
40. Bakken, G.A., PhD thesis, The Pennsylvania State University, 2001.
41. Magoulas, G.D., Vrahatis, M.N. and Androulakis, G.S, Neural Networks, 10 (1997) 69.
42. Kier, L.B. and Hall, L.H., J. Chem. Inf. Comput. Sci., 37 (1997) 548.