The Effect of Variable Selection on the Non-linear Modelling of Oestrogen Receptor Binding

ARTICLE *in* QSAR & COMBINATORIAL SCIENCE · JUNE 2006

Impact Factor: 1.55 · DOI: 10.1002/qsar.200510153

Taravat Ghafourian

University of Sussex

64 PUBLICATIONS **855** CITATIONS

The Effect of Variable Selection on the Non-linear Modelling of Oestrogen Receptor Binding

Taravat Ghafourian[a, b, c]* **and Mark T. D. Cronin**[a]

[a] School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England
[b] School of Pharmacy, Tabriz University of Medical Sciences, Daneshgah Street, Tabriz 51664, Iran
[c] Present address: Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham Maritime, Kent ME4 4TB, England,
Tel: +441634883846, Fax: +44163883927, E-mail: t.ghafourian@kent.ac.uk

Keywords: Variable selection, Oestrogen receptor binding, Recursive partitioning, Neural networks, Counter propagation, Support vector machine

Received: 16 June 2005; Accepted: 28 July 2005

DOI: 10.1002/qsar.200510153

Abstract

Oestrogen Receptor Binding Affinity (RBA) is often used as a measure of the oestrogenicity of endocrine disrupting chemicals. Quantitative Structure–Activity Relationship (QSAR) modelling of the binding affinities has been performed by three-dimensional approaches such as Comparative Molecular Field Analysis (CoMFA). Such techniques are restricted, however, for chemically diverse sets of chemicals as the alignment of molecules is complex. The aim of the present study was to use non-linear methods to model the RBA to the oestrogen receptor of a large diverse set of chemicals. To this end, various variable selection methods were applied to a large group of descriptors. The methods included stepwise regression, partial least squares and recursive partitioning (Formal Inference Based Recursive Modelling, FIRM). The selected descriptors were used in Counter-Propagation Neural Networks (CPNNs) and Support Vector Machines (SVMs) and the models were compared in terms of the predictivity of the activities of an external validation set. The results showed that although there was a certain degree of similarities between the structural descriptors selected by different methods, the predictive power of the CPNN and SVM models varied. Although the variables selected by stepwise regression led to poor CPNN models they resulted in the best SVM model in terms of predictivity. The parameters selected by some of the FIRM methods were superior in CPNN.

1 Introduction

A number of chemicals released into the environment are believed to disrupt normal endocrine function in animals, thereby causing reproductive disorders and abnormalities in wildlife [1]. There are a large number of interactions that may bring about endocrine disruption. Of these, the most studied are those arising from chemicals which are believed to mimic the effects of natural oestrogens. These compounds are believed to exert their effect via interaction with the oEstrogen Receptor (ER). The binding affinities of structurally diverse steroidal and non-steroidal chemicals capable of interacting with the ER cover a broad range. They vary from those more potent than the endogenous ligand 17β-estradiol (E2) to ligands with only a relatively weak binding affinity. Whilst not binding strongly, these latter compounds are nonetheless important ER ligands because of their potential harmful effects to environmental species and man.

There have been numerous efforts in recent years to model the ER binding affinities of chemicals using Quantitative Structure–Activity Relationship (QSAR) techniques. These act as ligand-based identification methods

Abbreviations: Artificial Neural Networks, ANNs; sixth order chain molecular connectivity index, ${}^{6}\chi_{ch}$; Comparative Molecular Field Analysis, CoMFA; Counter-Propagation Neural Network, CPNN; Energy of the Highest Occupied Molecular Orbital, E_{HOMO}; molecular mechanical torsion Energy, E_{tor}; oEstrogen Receptor, ER; Formal Inference Based Recursive Modelling, FIRM; indicator variable for the presence of the phenol group, I_{phenol}; third-order Kappa alpha shape index, ${}^{3}\kappa_{a}$; Multiple Linear Regression, MLR; National Center for Toxicological Research, NCTR; Number of carbon atoms, N_C; Number of halogen atoms, $N_{halogen}$; Partial Least Squares, PLS; Principal Components Analysis, PCA; Quantitative–Activity Relationship, QSAR; Receptor Binding Affinity, RBA; Recursive Partitioning, RP; Self-Organising Map, SOM; Support Vector Machine, SVM; Variable Importance in the Projection, VIP; Wiener topological index, W.

for environmental oestrogens [2–6]. The most widely used approaches are three-dimensional in nature and include Comparative Molecular Field Analysis (CoMFA) [7, 8] which may also incorporate the ligand-receptor interaction energy as an additional parameter [9]. However, CoMFA has some inherent shortcomings. In particular, its performance is highly dependent on the alignment of molecules and this, together with its sensitivity to molecular conformations, makes it difficult to use with structurally diverse datasets [10].

In traditional QSAR studies various techniques have been employed for variable selection and model development. In a previous study we employed a variety of linear techniques including variable clustering, Partial Least Squares (PLS) analysis, hierarchical PLS and stepwise regression analysis to identify the molecular descriptors influencing the ER binding affinities of a large diverse set of chemicals [11]. The study resulted in a number of successful models with good predictive ability. These studies were, however, based on linear techniques. In order to alleviate the problems brought about by linearity, Artificial Neural Networks (ANNs) and Support Vector Machines (SVMs) are among the more commonly used tools to generate non-linear models in QSAR studies [12–15].

The increasing use of ANNs in QSAR studies suggests the very high complexity and non-linear character of various relationships between biological activity and structure [16]. There are a variety of neural network techniques, and many reviews of their application in QSAR exist [cf. 16–18]. Of these, Counter-Propagation Neural Networks (CPNNs) are widely used in QSAR. A CPNN is an extension of the Self-Organising Map (SOM) or Kohonen neural network and has been described previously in many articles and textbooks [cf. 19, 20]. SVMs are also non-linear algorithms that can be used for classification and regression problems within the QSAR field (14–15, 21–22). These methods perform regression and classification tasks by constructing non-linear decision boundaries. Because of the nature of the feature space in which these boundaries are found, SVM can exhibit a large degree of flexibility in handling classification and regression tasks of varied complexities. There are a wide range of structural descriptors that can be used in QSAR modelling using CPNN or SVM methods. Due to the large numbers of molecular descriptors that are available through many commercially available software packages, variable selection has become a necessity in QSAR model development. This practice is essential to avoid overfitting to the training set data and the risk of chance correlation [23–25]. Various variable selection methods have been introduced. Formal Inference-Recursive based Modelling (FIRM) and stepwise regression are two examples of incremental methods [26]. While FIRM is a non-linear method (a form of decision tree) that selects variables for classification of data, stepwise regression selects variables that are relevant to the activity in a linear way. Multivariate methods such as PLS and Principal Components Analysis (PCA) also reduce the dimensionality by constructing orthogonal principle components. However, it has been shown that PLS can also benefit from variable selection, thus, strategies such as GOLPE and VIP have been implemented in order to increase the predictivity of the models [27, 28].

The aim of this study was to investigate the use of CPNN and SVM to predict the oestrogenic activities of a large diverse set of chemicals. To this end, various methods were employed for variable selection and the methods were compared in terms of their predictivity for an external validation set. The variable selection methods consisted of PLS and PCA, formal inference-based modelling and stepwise regression. Since quality of data (homogeneity and absence of influential outliers) and the representativity of the chemicals in the dataset are of vital conditions for applicability and validity of QSARs [29], a high-quality dataset is required for study. The dataset used in this study is oestrogen receptor binding activity data of the National Center for Toxicological Research (NCTR) compiled from a rat ER binding assay [30, 31]. It covers a broad range of synthetic and natural oestrogenic chemicals including steroids, phytoestrogens, diphenylmethanes, biphenyls, phenols and DES-like chemicals.

2 Materials and Methods

2.1 Oestrogen Receptor Binding Affinity Data

The oestrogen receptor binding affinity data were obtained from Blair *et al.* [30] and Branham *et al.* [31]. The ER binding activity is represented by the Relative Binding Affinity (RBA), where the RBA value for the endogenous ER ligand, 17β-estradiol (E2), was set to 100. This dataset contains chemicals that were selected to cover the structural diversity of ligands that bind to ER. These data have an activity distribution ranging over six orders of magnitude. Both factors (chemical diversity and a broad range of activity) are important for the prediction of the biological activity of oestrogens and for binding affinity in particular [32].

2.2 Structural Descriptors

A total of 175 structural descriptors were calculated by the TSAR 3D version 3.3 for Windows (Accelrys Ltd., Oxford), KOWWIN v1.66 (U.S. Environmental Protection Agency, Washington DC), and QSARis version 1.1 (SciVision – Academic Press, San Diego, CA) software packages. The descriptors included electronic parameters calculated by VAMP (using the AM1 Hamiltonian) in TSAR, atom and group counts, molecular weight and surface area and volume calculated by TSAR, the logarithm of the octanol–water partition coefficient (logP) calculated by KOWWIN, and topological, shape, and three-dimensional parameters calculated by QSARis. Molecular descriptors calculated by the software packages were checked and omitted if more than 98% of the values were the same. Highly intercorrelated descriptors were also discarded. The analyses were performed using a reduced dataset of 151 descriptors for the 131 compounds. The descriptors have been listed and explained in a recent publication [11].

2.3 Development of QSARs

QSAR models were developed using CPNN and Support Vector Machine SVM analyses. The CP-NN software package (National Institute of Chemistry, Hajdrihova 19, POB 3430, SI-1001 Ljubljana, Slovenia) was used for the CPNN analyses. A detailed description of this type of network is given in numerous articles and textbooks [cf. 33, 34]. The architecture of a CPNN allows one to combine an unsupervised mapping technique (Kohonen mapping) with a supervised learning strategy. A CPNN is built up from two layers of neurons arranged in two-dimensional rectangular matrices. The input (Kohonen) layer receives the input variables, *e.g.*, the molecular descriptors in a QSAR. The output layer, which has the same topological arrangement of neurons as the input layer, receives the target values (in this case the ER binding affinities) during the learning process. During the learning the object is directed to its winning neuron (a neuron with weights the most similar to the input vector) and the weights are corrected. Then the position of winning neuron is transferred from the input to the output layer and the weights of the output layer are corrected according to the target. Consequently, after the training has finished, the output neurons, arranged in output layer as points in a two-dimensional map, represent a response surface. Each point in the response surface offers one value as a possible prediction of target values. The number of different prediction values is limited by the number of neurons of an individual network. In this study, the number of neurons and number of epochs (iteration cycles) were optimised for the best prediction of the logRBA of chemicals in the test set. Therefore, network sizes ranging from 7×7 to 14×14 were examined. The numbers of epochs were between 80 and 220 with steps of 10, followed by 250 and 300 epochs. Other computational parameters of the networks were the maximal and minimal learning rates which were set at 0.5 and 0.01, non-thoroidal boundary conditions, triangular correction function of the neighbourhood (as recommended in references [35, 36]).

The LIBSVM software [37] was used to generate ν-SVMs with the error function given by Eq. 1:

$$\frac{1}{2}w^T w - C\left(\nu\varepsilon + \frac{1}{N}\sum_{i=1}^{N}\left(\xi_i + \xi_i^0\right)\right) \tag{1}$$

which is minimised subject to:

$$\begin{aligned}&(w^T\phi(x_i) + b) - y_i \leq \varepsilon + \xi_i\\ &y_i - (w^T\phi(x_i) + b_i) \leq \varepsilon + \xi_i^0\\ &\xi_i, \xi_i^0 \geq 0, \quad i = 1, ..., N, \quad \varepsilon \geq 0\end{aligned}$$

Here w is the vector of coefficients, b a constant and ξ_i are parameters for handling noisy data (inputs), the index i labels the N training cases, y_i and x_i are dependent and independent variables, respectively. The parameters C (the penalty parameter of the error term), ν and ε (tolerance of termination) were optimised in this study according to the generalisation performance of the model. The kernel function (ϕ) is used to transform data from the input to the feature space. The kernel used in this investigation was the Radial Basis Function (RBF) kernel defined by Eq. 2.

$$\phi = \exp\left(-\gamma|x_i - x_j|^2\right) \tag{2}$$

Here, γ is the kernel parameter that was optimised in this study.

In order to find the optimised combination of the parameters C, ε, ν and γ, a grid search was performed using 200 different combinations.

Chemicals were divided into three groups to form the training set, test set and validation set, with a ratio of 3:2:1. To achieve this, the compounds were sorted according to ascending logRBA values and from each group of six compounds, the first, third and the last was assigned to the training set, second and fifth to the test, and fourth to the validation set. Thus, the numbers of chemicals in the respective groups were 65, 44 and 22. The CPNNs and SVRs were trained using the chemicals in the training set and evaluated using the correlation coefficient between observed logRBA values and those predicted for the test set. The selected model was then used for the prediction of logRBA values of validation set.

Various sets of molecular descriptors were utilised in the CPNN and SVR analyses. These were selected by the following techniques:

2.3.1 Descriptors Selected by Stepwise Regression Analysis

Stepwise regression analysis was performed using the MINITAB statistical software (version 13.1, MINITAB Inc. State College PA 16801-3008, USA). In order to minimise the risk of chance correlations the maximum p-value for a parameter to be included in the model was set at 0.05 and the maximum number of parameters allowed in the Multiple Linear Regression (MLR) was restricted to eight.

2.3.2 Descriptors Selected by Various Formal Inference-based Recursive Modelling (FIRM) Analyses

FIRM analysis was performed using the TSAR 3D software to identify those variables that might be related to oestrogen receptor binding affinity in a non-linear manner. Parameters were standardised between 0 and 1 prior to analysis. Various maximum numbers of bins to divide the data at each step of splitting were examined. The maximum bin numbers were 2, 3, 4, 5, 7, 10, 20, 50 and 131. The models were evaluated by correlating the logRBA values calculated by the models against the observed values. Descriptors from the four best models (F1-F4 descriptors) were used in CPNN and SVM analyses.

Furthermore, a total of 87 chemicals were selected by random sampling from among the total number of 131. The samplings were performed 15 times and FIRM was performed on the samples using the bin numbers (for the continuous data splitting) set at 3, 10 and 20. The following descriptors selected by the models were used in CPNN and SVR analyses:

1. Those that occurred more than twice in the models with 3 bins (F5 descriptors).
2. The selected variables that were classified according to an earlier clustering analysis performed on this dataset [11] and the descriptors with the highest occurrence from each variable cluster (F6 descriptors).

2.3.3 *The Scores from PLS and PCA Models Performed on the Training Set and the Predicted Scores for Test and Validation Sets*

PLS and PCA analyses were carried using the SIMCA-P version 9.0 (Umetrics AB, Sweden) statistical software. The number of significant principal components was determined using the increase in the cross-validated regression coefficient (Q^2). The number of chemicals excluded in the cross-validation rounds was set to 25%. The PLS model was originally constructed using all the parameters. The model was improved after variable reduction using the VIP (Variable Importance in the Projection) criterion implemented in the SIMCA-P software [38]. Accordingly, variables with VIP values less than 0.7 were deleted. From the new model a small number of variables with the smallest coefficients were deleted and VIP values for the resulting model were examined. The steps were performed several times until further reduction of variables did not improve the Q^2 value. All the compounds were used for model development. Subsequently, the final model was carried forward for the chemicals in the training set (65 chemicals) and the scores were calculated. This model was also used for the prediction of the scores for the test and the validation sets consisting of 44 and 22 chemicals respectively. In order to study whether the severe reduction in the number of compounds in the training set had affected the PLS model and the consequent prediction of the scores dramatically, a model was performed on the chemicals in training and test sets (109 chemicals) and the scores were predicted only for the validation set.

PCA was performed using all the molecular descriptors. The scores of components with eigenvalues higher than one were used as variables in the CPNN and SVR analyses.

3 Results

The NCTR dataset consists of oestrogen receptor binding affinity data for 131 compounds covering a very wide range of diverse molecular structures [39]. The dataset has been used in a number of previous QSAR studies using three-dimensional modelling and receptor binding techniques such as Comparative Molecular Field Analysis (CoMFA) [7, 40]. Table 1 lists the compounds in the dataset and the corresponding logRBA values. In a previous study, we applied a number of linear QSAR methods to the dataset and some successful models were obtained [11]. In this study, in order to explore non-linear relationships between logRBA and the structural parameters, a form of neural network (counter-propagation) analysis was used to model the dataset. The models, built using different sets of descriptors following different techniques to select variables, are described.

Table 1. Compounds used in the analyses, and the logRBA values.

Compound	logRBA
4-Cresol ^{t}	– 4.50
4-Ethylphenol ^{s}	– 4.17
Rutin ^{v}	– 4.09
2-Ethylphenol ^{t}	– 3.87
7-Hydroxyflavanone ^{s}	– 3.73
4-Chloro-2-methyl phenol ^{t}	– 3.67
Phenolphthalin ^{t}	– 3.67
2-Cholor-4-methyl phenol ^{s}	– 3.66
2,4′-Dichlorobiphenyl ^{t}	– 3.61
4-*tert*-Butylphenol ^{v}	– 3.61
2-*sec*-Butylphenol ^{s}	– 3.54
3-Phenylphenol ^{t}	– 3.44
4-(Benzyloxyl)phenol ^{t}	– 3.44
Methyl 4-hydroxybenzoate ^{s}	– 3.44
6-Hydroxyflavone ^{v}	– 3.41
4-Chloro-3-methylphenol ^{t}	– 3.38
4-*sec*-Butylphenol ^{s}	– 3.37
4-*tert*-Amylphenol ^{t}	– 3.26
Phenol red ^{t}	– 3.25
3,3′,5,5′-Tetrachloro-4,4′-biphenyldiol ^{s}	– 3.25
n-Propyl 4-hydroxybenzoate ^{t}	– 3.22
Ethyl 4-hydroxybenzoate ^{v}	– 3.22
1,3-Diphenyltetramethyl disiloxane ^{s}	– 3.16
Diphenolic acid ^{t}	– 3.13
Morin ^{t}	– 3.09
4,4′-Sulfonyldiphenol ^{s}	– 3.07
n-Butyl 4-hydroxybenzoate ^{v}	– 3.07
6-Hydroxyflavanone ^{t}	– 3.05
Baicalein ^{s}	–3.05
4-Hydroxybiphenyl ^{t}	– 3.04
Bis(4-hydroxyphenyl)methane ^{t}	– 3.02
Formononetin ^{s}	– 2.98
Dihydrotestosterone ^{v}	– 2.89
4-Heptyloxyphenol ^{t}	– 2.88

Table 1. (cont.)

Compound	logRBA
o,p'-DDT ^{s}	−2.85
Chalcone ^{t}	−2.82
3′-Hydroxyflavanone ^{t}	−2.78
Triphenylethylene ^{s}	−2.78
2-Chloro-4-biphenylol ^{v}	−2.77
Myricetin ^{t}	−2.75
Prunetin ^{t}	−2.74
Doisynoestrol ^{s}	−2.74
4-Phenethylphenol ^{t}	−2.69
3a-Androstanediol ^{s}	−2.67
4′-hydroxyflavanone ^{v}	−2.65
2,4-Hydroxybenzophenone ^{t}	−2.61
4-Hydroxychalcone ^{s}	−2.55
Benzyl 4-hydroxybenzoate ^{t}	−2.54
4,4′-Dihydoxy-benzophenone ^{t}	−2.46
2,2′-Methylenebis(4-chlorophenol) ^{s}	−2.45
4′-Hydroxychalcone ^{v}	−2.43
Biochanin A ^{t}	−2.37
Fisetin ^{t}	−2.35
3′,4′,7-Trihydroxy isoflavone ^{s}	−2.35
4-*n*-Octylphenol ^{t}	−2.31
p-Cumyl phenol ^{s}	−2.30
3-Deoxy-estrone ^{v}	−2.20
4-Chloro-4′-biphenylol ^{t}	−2.18
Naringenin ^{s}	−2.13
Bisphenol A ^{t}	−2.11
Heptyl *p*-hydroxybenzoate ^{t}	−2.09
Kepone ^{s}	−1.89
Phenolphthalein ^{v}	−1.87
4-*t*-Octylphenol ^{t}	−1.82
2-Ethylhexyl-4-hydroxybenzoate ^{s}	−1.74
4-Dodecylphenol ^{t}	−1.73
3-Methyl-estriol ^{t}	−1.65
Daidzein ^{s}	−1.65
Kaempferol ^{v}	−1.61
Apigenin ^{t}	−1.55
Nonylphenol ^{s}	−1.53
Nordihydroguaiaretic acid ^{t}	−1.51
Aurin ^{t}	−1.50
16b-OH-16-Methyl-3-methyl-estradiol ^{s}	−1.48
4,4′-(1,2-Ethanediyl)bisphenol ^{t}	−1.44
2,5-Dichloro-4′-biphenylol ^{v}	−1.44
2′,4,4′-Trihydroxychalcone ^{s}	−1.26
Diethylstilbestrol dimethyl ether ^{t}	−1.25
Phloretin ^{t}	−1.16
Bisphenol B ^{s}	−1.07
3-b-Androstanediol ^{v}	−0.92
Monohydroxymethoxychlor ^{t}	−0.89
Equol ^{t}	−0.82
4′,6-Dihydroxyflavone ^{s}	−0.82
b-Zearalenol ^{t}	−0.69
2,2′,4,4′-Tetrahydroxybenzil ^{s}	−0.68
Norethynodrel ^{v}	−0.67
2,3,4,5-Tetrachloro-4′-biphenylol ^{t}	−0.64
Monohydroxymethoxychlor olefin ^{s}	−0.63
HPTE ^{t}	−0.60
4,4′-Dihydroxystilbene ^{t}	−0.55
Genistein ^{s}	−0.36
3,6,4′-Trihydroxyflavone ^{v}	−0.35
3-Deoxyestradiol ^{t}	−0.30
3-Hydroxy-estra-1,3,5(10)-trien-16-one ^{s}	−0.29
b-Zearalanol ^{t}	−0.19
6a-OH-Estradiol ^{t}	−0.15
Clomiphene ^{s}	−0.14
Nafoxidine ^{v}	−0.14
4-Ethyl-7-OH-3-(p-methoxyphenyl)coumarin ^{t}	−0.05
Coumestrol ^{s}	−0.05
a,a-Dimethyl-b-ethyl allenolic acid ^{t}	−0.02
Toremifene ^{t}	0.14
Tamoxifen ^{s}	0.21
Zearalanone ^{v}	0.32
Mestranol ^{t}	0.35
Dihydroxymethoxychlor olefin ^{s}	0.42
17a-Estradiol ^{t}	0.49
3-(*p*-Phenol)-4-(*p*-tolyl)-hexane ^{t}	0.60
Estrone ^{s}	0.86
2,6-Dimethyl hexestrol ^{v}	0.95
Monomethyl ether hexestrol ^{t}	0.97
Estriol ^{s}	0.99
Moxestrol ^{t}	1.14
17-Deoxyestradiol ^{t}	1.14
Dimethylstilbestrol ^{s}	1.16
ICI 164384 ^{v}	1.16
Droloxifene ^{t}	1.18
3,3′-Dihydroxyhexestrol ^{s}	1.19
Monomethyl ether diethylstilbestrol ^{t}	1.31
2-OH-Estradiol ^{t}	1.47
a-Zearalanol ^{s}	1.48
ICI 182780 ^{t}	1.57
Dienestrol ^{v}	1.57
Zearalenol ^{s}	1.63
4-OH-Estradiol ^{t}	1.82
17b-Estradiol (E2) ^{t}	2.00
4-OH-Tamoxifen ^{s}	2.24
Ethynylestradiol ^{v}	2.28
Hexestrol ^{t}	2.48
Diethylstilbestrol (DES) ^{s}	2.60

^{t}: training set, ^{s}: test set, ^{v}: validation set.

3.1 Results of Variable Selection Procedures

Variables selected by stepwise regression, FIRM and PLS analyses are listed in Table 2. Note that for PCA all of the 151 molecular descriptors were included in the model.

The variables selected by stepwise regression analysis (MLR descriptors) were $^{3}\kappa_{a}$, $^{6}\chi_{ch}$, N_C, W, $N_{halogen}$, I_{phenol}, E_{HOMO} and E_{tor}, all defined at Table 2.

The descriptors selected by FIRM analyses were obtained from the analyses performed on the whole dataset using different numbers of bins. Four sets of descriptors selected by the best four FIRM models, judging by the cross-validated correlation coefficient between observed and calculated values, are described in Table 2 as the F1 through F4 descriptors. They were obtained with 3, 7, 5 and 2 bins respectively, with the leave-one-out cross-vali-

Table 2. Molecular descriptors selected using different methods[a].

Descriptors	Method	Parameters
MLR	Stepwise regression	${}^{3}\kappa_{a}$, ${}^{6}\chi_{ch}$, E_{HOMO}, E_{tor}, I_{phenol}, N_C, $N_{halogen}$, W
F1	FIRM (3 bins)	${}^{5}\chi^{v}{}_{p}$, $\Delta^{10}\chi_p$, $\Delta^{3}\chi^{v}{}_{p}$, E_{angle}, I_{phenol}, N_{atom}, N_{rotat}, N_{viol}, Q
F2	FIRM (7 bins)	$\Delta^{5}\chi_p$, J, N_C, N_{phenol}, SsOH, SaaCH
F3	FIRM (5 bins)	${}^{0}\chi^{v}$, ${}^{10}\chi^{v}{}_{p}$, IS2, N_{phenol}, Q, SA1, W
F4	FIRM (2 bins)	$\Delta^{5}\chi_p$, $\Delta^{9}\chi^{v}{}_{p}$, E_{HOMO}, I_{phenol}, IS2, knotpv, logP, N_{atom}, nelem, N_{rotat}, μ1
F5	FIRM (descriptors repeated in more than 3 models on random samples with 3 bins)	$\Delta^{9}\chi^{v}{}_{p}$, E_{angle}, E_{tor}, I_{phenol}, N_{atom}, N_C, Q, Qsv, SssCH2
F6	FIRM (one descriptor from each cluster)	${}^{6}\chi_{ch}$, α_{SPC}, $\Delta^{5}\chi_p$, E_{angle}, E_{HOMO}, E_{tor}, I_{phenol}, MaxNeg, N_C, N_N, N_{OH}, Q, Qsv, SssCH2
F7	FIRM	$\Delta^{9}\chi^{v}{}_{p}$, E_{tor}, I_{phenol}, N_{atom}, N_C, Q, Qsv
PLS1 and PLS2	PLS	${}^{0}\chi$, ${}^{0}\chi^{v}$, ${}^{4}\chi_p$, α1, E_{tor}, I_{phenol}, IS2, IS3, logP, N_{Atoms}, N_C, N_H, N_{viol}, nvx, Q, SA1, SaasC, SssO, SsssCH, W

[a] The parameters are defined as:

$\Delta^{5}\chi_p$, $\Delta^{10}\chi_p$, $\Delta^{3}\chi^{v}{}_{p}$, and $\Delta^{9}\chi^{v}{}_{p}$:	5th and 10th order difference connectivity indices and 3rd and 9th order valence corrected difference connectivity indices respectively
${}^{0}\chi$, and ${}^{4}\chi_p$:	zero and fourth order path connectivity indices
${}^{0}\chi^{v}$, and ${}^{10}\chi^{v}{}_{p}$, and ${}^{5}\chi^{v}{}_{p}$:	zero, fifth and tenth order valence corrected path connectivity indices
${}^{3}\kappa_a$:	3rd order molecular shape index
${}^{6}\chi_{ch}$:	sixth order chain molecular connectivity index
E_{angle}:	energy of angle calculated by COSMIC Force Field
E_{HOMO}:	energy of the highest occupied molecular orbital
E_{tor}:	energy of torsion calculated by COSMIC Force Field
I_{phenol}:	indicator variable for the presence of phenol group
IS2:	size principle moment of inertia
IS3:	size principle moment of inertia
J:	Balaban topological index
knotpv:	difference between valence corrected connectivity indices of third order cluster and fourth order path/cluster
logP:	logarithm of partition coefficient
MaxNeg:	largest negative charge over the atoms in a molecule
N_{atom}:	total number of atoms
N_C:	number of carbon atoms
nelem:	number of elements
$N_{H:}$	number of hydrogen atoms
$N_{halogen}$,	number of halogen atoms
N_N:	number of nitrogen atoms
N_{OH}:	number of hydroxyl groups
N_{phenol}:	number of phenol groups
N_{rotat}:	number of rotateable bonds
N_{viol}:	number of violations from Lipinski's rule of five
nvx:	number of graph vertices
Q:	magnitude of the principle quadruple moment
Qsv:	dipolar descriptor
SA1:	solvent-accessible surface area
SaaCH:	atom-type electrotopological index for hydroxyl group
SaasC:	atom-type electrotopological index for substituted aromatic carbon atoms
SsOH:	atom-type electrotopological index for aromatic carbon atoms connected to a hydrogen
SssCH2:	atom-type electrotopological index for $-CH_2-$groups
SssO:	atom-type electrotopological index for −O−groups
SsssCH:	atom-type electrotopological index for −CH−groups
W:	Wiener topological index
α_{SPC}:	specific polarizability of a molecule equal to Polarisability/Volume
α1:	polarisability calculated by AM1 Hamiltonian
μ1:	dipole moment

dated r^2 ranging from 0.792 to 0.751. Two other sets of descriptors were selected by FIRM analyses (the F5 and F6 descriptors in Table 2) performed on 15 randomly selected samples containing 87 chemicals. A total of 54 parameters appeared in the 45 FIRM analyses performed using 3, 10 and 20 as the maximum number of bins for splitting the data. These have been summarised in Table 3 together with the number of occurrences of each descriptor. Some of the variables in Table 3 stand out as appearing in the majority of FIRM models. For example, α_1, N_{atoms}, N_C, Q and I_{phenol} have each appeared in more than nine models. F5 descriptors are the descriptors selected more than twice in the FIRM analyses with 3 bins. There were nine descriptors in this category: N_{atom}, N_C, $\Delta^9\chi^v{}_p$, E_{angle}, Q, I_{phenol}, E_{tor}, Qsv and SssCH2 (see Table 2 for definition of the parameters). F6 descriptors are the most frequently selected descriptors by all FIRM analyses from each variable cluster listed in Table 3 (This is according to previous variable clustering of this dataset [11]). F6 descriptors (14 in total) are also listed and defined in Table 2. In the F7 descriptors two of the descriptors (E_{angle} and SssCH2) that were not selected by any of the FIRM models with 10 or 20 bins were excluded from F5 descriptor set.

PLS was performed using the 151 original descriptors and logRBA to obtain the required descriptors. After variable reduction, the model was improved as indicated by Q^2. The final PLS model consisted of 3 significant components involving the 20 original descriptors listed and defined in Table 2. The scores of the PLS analysis performed on the chemicals in the training set and the scores predicted for the test and the validation set by this model were used as the PLS1 descriptors for further CPNN and SVM analyses. Furthermore, because in this strategy only 65 chemicals are used for PLS model generation and this might reduce the quality of the predicted variables (PLS scores) for the test and validation chemicals, a second strategy was examined as follows. The model was created on chemicals in both training and test sets. The scores of the PLS components were predicted for the validation set (PLS2 descriptors). Table 4 presents the coefficients and statistical parameters of the resulting PLS models. In Table 4, $R^2(X)$ and $R^2(Y)$ are cumulative sum of squares of all the X's and Y's explained by all extracted components respectively, Q^2 is cumulative variation of the Y's that can be predicted by the components in a leave-25%-out cross-validation where data were split into four groups and there were 100 iterations. The parameters r^2(pred) and rms in Table 4 are the squared correlation coefficient and root-mean-square differences between observed and predicted logRBA values calculated for the chemicals that have not been used in the model development, respectively. One outlier was observed, rutin, this is a chemical that is located in the validation set and is poorly predicted by both methods. Its deletion improves the correlations of observed and predicted logRBA values from 0.359 and 0.583 to 0.430 and 0.611 respectively.

Table 3. Parameters appearing in FIRM models for 15 randomly selected samples consisting of 87 compounds from the dataset.

Cluster	Parameter	Times occurred			
		F3 models	F10 models	F20 models	all models
C1	α_1	1	5	3	9
C1	α_2		1		1
C1	ET1	1			1
C1	IL1		1		1
C1	IL2		1		1
C1	IS3	1	1		2
C1	k3	2			2
C1	MW	1			1
C1	N_{atoms}	6	2	1	9
C1	N_C	3	8	7	18
C1	N_H	1			1
C1	Ovality	1			1
C1	Qs			1	1
C1	SA1	1	1	2	4
C1	Xv0	1	1		2
C1	Xv1		1		1
C2	Dxp10	2			2
C2	dxp3		1		1
C2	dxp5	2	1	2	5
C2	dxp6		1	1	2
C2	dxp8	2			2
C2	dxvp4	1			1
C2	dxvp8	1			1
C2	dxvp9	3			3
C2	J		1		1
C2	Rings_5			1	1
C2	rings_{6ali}	1			1
C2	rings_T	1			1
C2	Xp10			2	2
C2	xvp5	1		1	2
C2	xvp6		1		1
C2	xvp8		1		1
C2	xvp9	1			1
C3	E_{angle}	4			4
C3	xvc3		1		1
C4	xch6		1		1
C5	N_{OH}	1		1	2
C5	N_{phenol}		1		1
C5	SHHBd		1		1
C5	SsOH			1	1
C8	MaxNeg	1		1	2
C9	α_{SPC}			2	2
C9	$N_{halogen}$			1	1
C9	SsCl			1	1
C10	Q	6	5	3	14
C12	E_{LUMO}	1			1
C12	Qsv	3	1	1	5
C12	Qv			1	1
C12	SsCH3	2		1	3
C16	N_N		1		1
C20	I_{phenol}	13	2	1	16
C21	E_{HOMO}	2			2
C22	E_{tor}	5			5
C26	SssCH2	5			5

Table 4. Description of the PLS models; r^2(pred) and rms are calculated for the prediction set (validation and test sets).

Model	Dataset	No. of compounds	No. of components	$R^2(X)$	$R^2(Y)$	Q^2	r^2(pred)	rms
1	All chemicals	131	3	0.737	0.665	0.610	–	–
2	Training set	65	3	0.708	0.773	0.638	0.359	1.834
3	Training and test sets	109	2	0.592	0.646	0.546	0.583	1.619

Table 5. Description of the optimised CPNN models using the variables from different variable selection methods in addition to the square of the correlation coefficient (r^2) and root-mean-square error (rms) for the relationship between observed and predicted logRBA for training, test and validation sets.

Model	No. of variables	No. of epochs	Network size	r^2			rms		
				training	test	validation	training	test	validation
CPNN-MLR	8	100	11	0.849	0.537	0.329	0.681	1.226	1.525
CPNN-F1	9	200	8	0.940	0.599	0.700	0.426	1.158	0.993
CPNN-F2	6	180	11	0.968	0.518	0.528	0.314	1.252	1.292
CPNN-F3	7	200	12	0.972	0.379	0.399	0.296	1.45	1.642
CPNN-F4	11	130	12	0.927	0.568	0.616	0.477	1.221	0.616
CPNN-F5	9	210	11	0.946	0.541	0.456	0.406	1.219	1.372
CPNN-F6	14	120	9	0.941	0.584	0.718	0.422	1.176	0.956
CPNN-F7	7	210	14	0.984	0.549	0.504	0.222	1.219	1.336
CPNN-PLS1	3	250	12	0.989	0.488	0.557	0.183	1.297	1.344
CPNN-PLS2	2	100	12	0.965	0.554	0.392	0.339	1.207	1.486
CPNN-PCA	8	90	9	0.865	0.558	0.517	0.664	1.202	1.287

3.2 Results of CPNN Analyses

The parameters selected by the different methods were applied in the CPNN analyses, where the network size and the number of epochs were optimised. The final models are presented in Table 5. Although the r^2 values between observed and calculated logRBA for the training set are satisfactory, they have dropped dramatically for the test and validation sets. The rms values also show a big difference between training and test/validation sets.

The optimised CPNN models using the descriptors selected by FIRM (CPNN-F1 through CPNN-F7) are in general better than the CPNN model using the descriptors selected by stepwise regression analyses (CPNN-MLR). The predictive power of CPNN-F6, in particular, with prediction r^2 of 0.718 for the validation set, is the highest in comparison with the CPNN models using the descriptors selected by MLR or other FIRM strategies.

In the CPNN-PLS1 model the PLS scores calculated for the training set and predicted for the test and validation sets (model 2 in Table 4) were used as variables. In CPNN-PLS2 model PLS scores calculated for the training and test sets and predicted for the validation set were used. The results (Table 5) show that CPNN-PLS1, with a prediction r^2 of 0.557 for validation set, is a more predictive model than is CPNN-PLS2. The model built using the PCA scores is of similar quality to CPNN-PLS1.

3.3 Results of SVR Analyses

SVR analyses using the descriptors selected by different methods were performed and four model parameters were optimised to minimise rms error for the test set. C is a regularisation parameter that controls the trade off between maximising the margin and minmising the training error. A large value of C (*e.g.*, 1000) will lead to a more stable learning process. If C is too large then the algorithm will overfit the training data. C values of 1 – 1000 were examined in this study. ν values were between 0.01 and 1. γ controls the amplitude of the RBF function and therefore it controls the generalisation ability of SVM. The values between 0.001 – 100 were examined. ε values varied between 0.0001 and 1. The optimised models are presented in Table 6. The SVR model using the descriptors selected by stepwise regression (SVR-MLR) is the most predictive SVR model with a prediction r^2 of 0.863 for validation set. The SVR models using the variables selected by FIRM (SVR-F1 through SVR-F7) lead to prediction r^2 values of 0.473 – 0.707 for the validation set. SVR-F7 is the best model among the SVR models using FIRM descriptors in terms of the predictivity for the validation set ($r^2 = 0.707$). SVR models using the scores of PLS and PCA analyses as the molecular descriptors (SVR-PLS1, SVR-PLS2 and SVR-PCA in Table 6) have modest predictive powers.

4 Discussion

In this study CPNN and SVR studies were performed to explore non-linear relationships between the oestrogen receptor binding affinity of a diverse set of compounds and their structural descriptors. Various variable selection techniques were examined for the selection of structural descriptors from a pool of 151. Stepwise regression is one of the most widely used linear variable selection procedures. This incremental method is a combination of backward elimination and forward selection methods. It is fast and efficient but can end up in local minima. In this study, a relatively high error rate (α-value of 0.15) was used to enter and remove variables into and from the model. This high α-value was selected to enable a higher number of combinations of variables to be examined by stepwise regression. While the stepwise analysis leads to a model with many variables, the number of variables in which all the parameters selected had p-values lower than 0.05 was restricted to eight. When the descriptors were used in CPNN analysis, they had only a poor predictive ability with the correlation coefficients between observed and predicted logRBA being 0.537 and 0.329 for test and validation sets respectively. Table 5 shows that this descriptor set, in comparison with the descriptors selected by other methods, yields the poorest CPNN model in terms of predictivity. However the use of MLR descriptors in ν-SVR analysis using RBF kernel leads to the best SVR model (Table 6). In other word, the CPNN procedure used in this study fails to generalise well with this set of descriptors.

Recursive Partitioning (RP) is a statistical technique that can be used to explain biological activity in terms of a large number of structural features with an incremental approach to variable selection. The RP technique used in this study (FIRM) can be effective in uncovering structure in data with hierarchical, non-linear, non-additive, or categorical variables and has proven useful in classifying pharmaceutical data by discrete or continuous descriptors [41–43]. In this study FIRM was used to identify the structural descriptors (variables) that are relevant to predict ER binding affinity. The CPNN modelling using the variables selected by different FIRM strategies showed that CPNN-F6, CPNN-F1 and CPNN-F4 are the best models in terms of predictive ability for the validation set with r^2 values of 0.718, 0.700 and 0.616 respectively (Table 5). Therefore it can be concluded that variables selected with the aid of information from variable clustering (used in CPNN-F6) work better in terms of external predictivity than the parameters selected by a single FIRM analysis (CPNN-F1 and CPNN-F4). The parameters in CPNN-F6 were selected by various FIRM analyses on 15 random samples of the dataset, followed by the selection of one descriptor from each group of variables determined by cluster analysis on variables [11]. The variables used in CPNN-F1 and CPNN-F4, on the other hand, were selected by FIRM analysis on the whole dataset with the maximum number of bins for splitting the data being 3 and 2 respectively. A similar conclusion can be made from SVR results in which SVR-F7 with an r^2 value of 0.707 for the validation set is the better than other SVR models using FIRM parameters. PLS has been the method of choice in modelling datasets where the number of descriptors is higher than the number of observations [44]. However, this study and other investigations [45] show that PLS models can also benefit from variable selection. The PLS model for all the compounds in this study consisted of 20 structural descriptors with good predictive ability ($Q^2 = 0.610$). However, when performed on the training set of 65 compounds, the predictions for the test and validation sets are poor (Table 3, model 2). The models CPNN-PLS1 and SVR-PLS1 are based on the scores of the PLS model calculated for the training and predicted for the test and validation sets (Table 3, model 2). Comparing the results of CPNN-PLS1 (Table 5) and SVR-PLS1 (Table 6) with that of the PLS (model 2, Table 3) shows a significant improvement when applying the neural network, with the prediction r^2 increasing from 0.359 (for test and validation sets in model 2) to 0.488 for the test and 0.557 for the validation sets in CPNN-PLS1. A moderate improvement in the PLS model (Table 3) is observed when using the scores in SVR analysis (Table 6) with r^2 values of 0.399 and 0.598 for test and validation sets. Generally, although only 3 variables (PLS scores) are used in the CPNN-PLS1 and SVM-PLS1 models, the high number of descriptors involved in the calculation of the PLS components leads to the good CPNN and SVR mod-

Table 6. Description of the optimised ν-SVR models using the descriptors from different variable selection methods in addition to the square of the correlation coefficient (r^2) and root-mean-square error (rms) for the relationship between observed and predicted logRBA for training, test and validation sets.

Model	No. of variables	ν	C	γ	ε	r^2			rms		
						training	test	validation	training	test	validation
SVR-MLR	8	0.1	1000	0.1	1	0.886	0.669	0.863	0.709	1.179	0.504
SVR-F1	9	0.1	1000	100	0.001	1.000	0.507	0.614	0.000	1.709	1.387
SVR-F2	6	0.1	100	10	1	0.927	0.581	0.581	0.221	1.372	1.441
SVR-F3	7	0.1	1000	100	0.001	1.000	0.376	0.495	0.000	2.030	1.643
SVR-F4	11	0.1	100	10	1	0.932	0.494	0.556	0.218	1.651	1.437
SVR-F5	9	0.5	50	0.1	0.1	0.729	0.510	0.473	0.827	1.626	1.902
SVR-F6	14	0.5	50	0.1	1	0.759	0.575	0.557	0.756	1.385	1.717
SVR-F7	7	0.1	100	1	1	0.847	0.538	0.707	0.473	1.568	1.032
SVR-PLS1	3	0.7	5	0.1	1	0.773	0.399	0.598	1.442	1.946	1.911
SVR-PLS2	2	0.1	100	0.05	0.1	0.704	0.542	0.591	1.404	1.485	1.934
SVR-PCA	8	0.1	100	10	1	0.921	0.425	0.594	0.242	1.856	1.319

els. CPNN-PLS2 and SVR-PLS2, on the other hand, are not as successful; despite the relatively good r^2 values for the test sets, lower r^2 values were observed for the validation sets (Tables 5 and 6). In other words, predictions for the test set, which is used to assess the efficiency of the neural network and SVR, is not a good measure for the (external) validation set. This is probably due to the fact that for the test set, the PLS scores are calculated, whereas

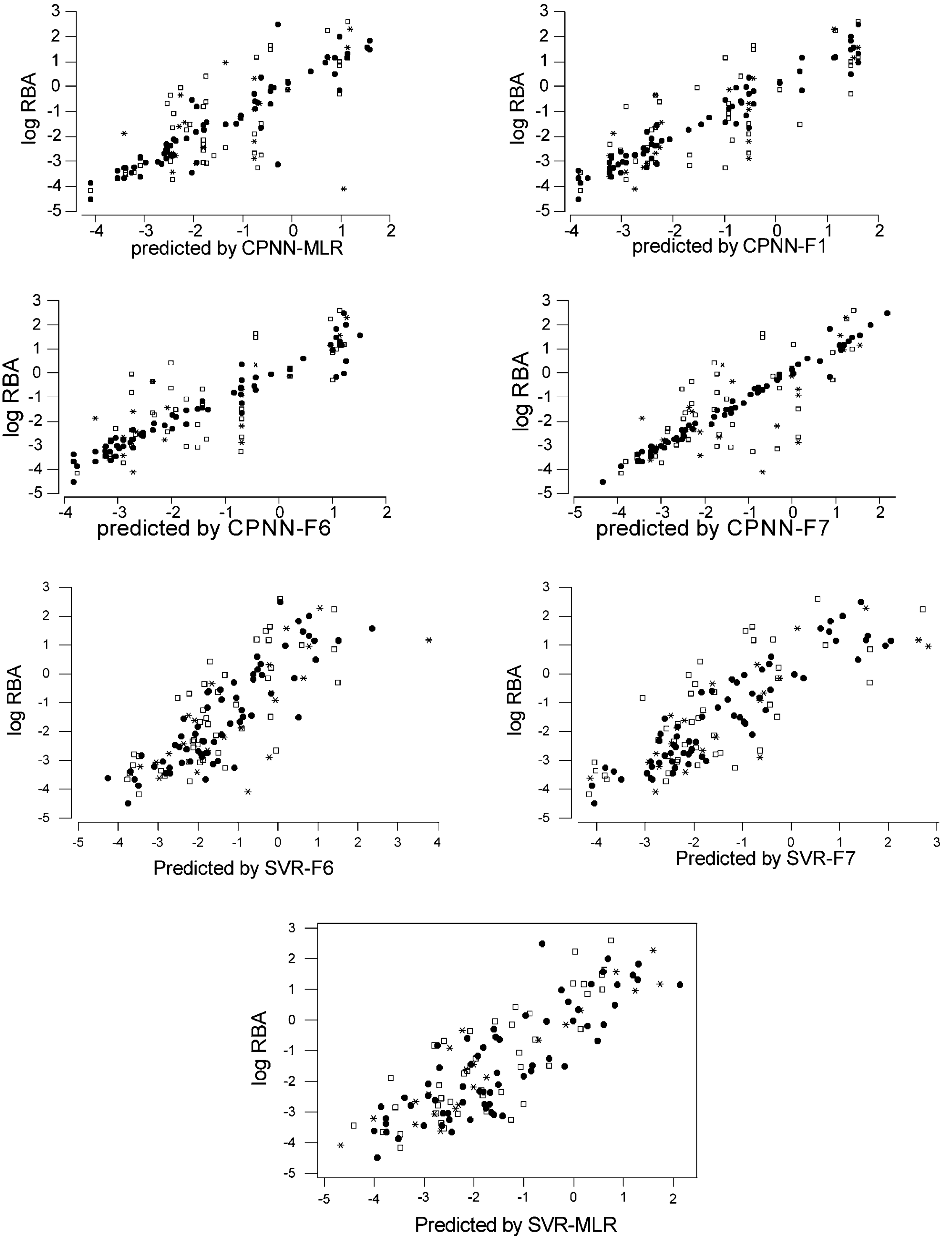

Figure 1. Plots of observed *versus* predicted logRBA values using some of the CPNN and SVR models; ●: chemicals in training set, □: chemicals in test set, *: chemicals in validation set.

they are predicted for the validation set, by the PLS model 3 (Table 3). There are also fewer CPNN parameters in CPNN-PLS2 in comparison with CPNN-PLS1. The unsupervised variable reduction by PCA, resulted in CPNN and SVR models with an intermediate predictivity in comparison with the models with PLS1 and PLS2 scores.

The summary of results in Table 5 shows that CPNN-F6, CPNN-F1 and CPNN-F4 have the best global predictive abilities among the CPNN models with the r^2 between observed and predicted logRBA values for the validation set ranging between 0.616 and 0.718. However, the SVR-MLR model (Table 6) outperforms the CPNN models. The physicochemical properties and structural descriptors used in the models are summarised and defined in Table 3. Some of the structural parameters are common to all the three best CPNN and SVR-MLR models. These are number of carbon, or all, atoms and the indicator variable for the presence of the phenol group. In addition E_{HOMO}, energy of angle, quadruple moment, number of rotatable bonds and topological parameter of $\Delta^5\chi_p$ are present in more than two out of the four models. The combination of the parameters selected by the stepwise regression analysis was a particularly poor predictor of logRBA when used in CPNN modelling. This is despite the fact that N_C, I_{phenol} and E_{HOMO} were present in the variables selected. Therefore although the variables work well together in a linear or SVM model, they fail to generalise in a CPNN analysis. Comparison of the r^2 values for training set with that of the test or validation sets in Table 5 and Table 6 shows that training r^2 values for CPNN models are very high but they drop dramatically for the test or validation sets. This shows the good learning abilities of the CPNN models with poor generalisation. Some of the SVR models, namely SVR-F1 through SVR-F4, also suffer from the same problem. However, for SVR-MLR the gap between r^2 values of training and that of test/validation sets is small.

Figure 1 shows the observed logRBA values against those predicted by some of the CPNN and SVR models. The plots show very good correlations for the training sets in CPNN models that might be interpreted as the CPNN models overfitting the training sets. It should be noted however, that reducing the number of epochs does not improve the fit for the test set, although it reduces the fit for the training set. The number of epochs examined in this study is within the range used in a number of other previous studies [35, 36, 46, 47]. An analysis of the outliers showed that some of the compounds are mis-predicted in most of the models. For example, the logRBA value of rutin (in validation set) is highly overestimated by 6 out of 10 CPNN models; only CPNN-F1, CPNN-F2, CPNN-F6 and CPNN-PLS1 predicted its value reasonably well. It is also an outlier from some of the SVR models (for example SVR-F6). The binding affinities of dihydrotestosterone (in validation set) and nordihydroguaiaretic acid (in test set) are also over-estimated by many models especially by CPNN-F7 and CPNN-PLS1. From the three selected CPNN models (CPNN-F1, CPNN-F4 and CPNN-F6), and two selected SVR models (SVR-MLR and SVR-F7) there is no single outstanding outlier.

5 Conclusion

Variable selection is an integral part of QSAR model generation. This is due to the increasing number of structural descriptors that are now available through specialised software packages. Variable reduction is necessary to avoid overfitting and the risk of chance correlations. This investigation has provided evidence that variable selection methods can greatly influence the QSAR models generated by SVM and CPNN. In this study, 151 structural descriptors were used for the QSAR modelling of oestrogen receptor binding affinity. Various supervised techniques were examined to select the variables for CPNNs and SVR analysis and the resulting models were compared in terms of predictivity for external validation. The results showed that some of the CPNN models are successful in modelling the oestrogen receptor binding affinities of a diverse set of chemicals. The neural network models obtained using the parameters selected by FIRM were superior to those obtained using the stepwise selected parameters. Neural network modelling using the PLS calculated scores for the training set could improve the prediction for the test and training set over the values predicted by the PLS model itself. While the variables selected by stepwise regression fail to generalise in CPNN analysis, they result in a very predictive SVR model.

Acknowledgement

This study was funded in part by the European Union EASYRING project (QLK4-2002-02286).

References

[1] T. Colborn, F. S. vom Saal, A. M. Soto, *Environ. Health Perspect.* **1993**, *101*, 378–384.

[2] T. W. Schultz, G. D. Sinks, M. T. D. Cronin, *Environ. Toxicol. Chem.* **2000**, *19*, 2637–2642.

[3] C. L. Waller, T. I. Opera, K. Chae, H.-K. Park, K. S. Korach, S. C. Laws, T. E. Wiese, W. R. Kelce, L. E. Gray, *Chem. Res. Toxicol.* **1996**, *9*, 1240–1248.

[4] A. G. Saliner, L. Amat, R. Carbo-Dorca, T. W. Schultz, M. T. D. Cronin, *J. Chem. Inf. Comp. Sci.* **2003**, *43*, 1166–1176.

[5] P. K. Schmieder, A. O. Aptula, E. J. Routledge, J. P. Sumpter, O. G. Mekenyan, *Environ. Toxicol. Chem.* **2000**, *19*, 1727–1740.

[6] S. Bradbury, V. Kamenska, P. Schmieder, G. Ankley, O. Mekenyan, *Toxicol. Sci.* **2000**, *58*, 253–269.

[7] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. Blair, W. Branham, D. M. Sheehan, *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186 – 195.

[8] W. Sippl, H.-D. Höltje, *J. Mol. Struct. (Theochem)* **2000**, *503*, 31 – 50.

[9] W. Sippl, *Bioorg. Med. Chem.* **2002**, *10*, 3741 – 3755.

[10] A. Berglund, M. C. De Rosa, S. Wold, *J. Computer-Aided Mol. Des.* **1997**, *11*, 601 – 612.

[11] T. Ghafourian, M. T. D. Cronin, *SAR QSAR Environ. Res.* **2005**, *16*, 171 – 190.

[12] D. Weekes, G. B. Fogel, *Biosystems* **2003**, *72*, 149 – 158.

[13] D. González-Arjona, G. López-Pérez, A. Gustavo González, *Talanta* **2002**, *56*, 79 – 90.

[14] D. C. Weaver, *Curr. Opin. Chem. Biol*. **2004**, *8*, 264 – 270.

[15] U. Norinder, *Neurocomputing* **2003**, *55*, 337 – 346.

[16] S. P. Niculescu, *J. Mol. Struct. (Theochem)* **2003**, *622*, 71 – 83.

[17] J. Devillers (Ed.), *Neural Networks in QSAR and Drug Design*, Academic Press, London, **1996.**

[18] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, Oxford, UK, **1995.**

[19] R. Hecht-Nielsen, *Appl. Opt.* **1987**, *26*, 4979 – 4984.

[20] J. Zupan, J. Gasteiger, *Neural Networks in Chemistry and Drug Design*, Wiley-VCH, Weinheim, Germany, **1999.**

[21] C. X. Xue, R. S. Zhang, H. X. Liu, X. J. Yao, M. C. Liu, Z. D. Hu, B. T. Fan, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693 – 1700.

[22] M. W. B. Trotter, S. B. Holden, *QSAR Comb. Chem.* **2003**, *22*, 533 – 548.

[23] D. M. Hawkins, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1 – 12.

[24] M. Clark, R. D. Cramer, *Quant. Struct.-Act. Relat*. **1993**, *12*, 137 – 145.

[25] D. C. Whitley, M. G. Ford, D. J. Livingstone, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160 – 1168.

[26] D. M. Hawkins, G. V. Kass, Automatic interaction detection, in: D. H. Hawkins (Ed.) *Topics in Applied Multivariate Analysis*, Cambridge University Press, Cambridge, UK, pp 269 – 302, **1982.**

[27] K. Tang, T. Li, *Chemom. Intl. Lab. Sys.* **2002**, 64, 55 – 64.

[28] M. Baroni, S. Clementi, G. Cruciani, G. Costantino, D. Riganelli, E. Oberrauch, *J. Chemom.* **1992**, 6, 347 – 356.

[29] L. Eriksson, J. Jaworska, A. P. Worth, M. T. D. Cronin, R. M. McDowell, *Environ. Health Perspect.* **2003**, *111*, 1361 – 1375.

[30] R. M. Blair, H. Fang, W. S. Branham, B. Hass, S. L. Dial, C. L. Moland, W. Tong, L. Shi, R. Perkins, D. M. Sheehan, *Toxicol. Sci.* **2000**, *54*, 138 – 153.

[31] W. S. Branham, S. L. Dial, C. L. Moland, B. Hass, R. Blair, H. Fang, L. Shi, W. Tong, R. Perkins, D. M. Sheehan, *J. Nutr.* **2002**, *132*, 658 – 664.

[32] T. W. Schultz, M. T. D. Cronin, *Environ. Toxicol. Chem.* **2003**, *22*, 599 – 607.

[33] J. Zupan, M. Novič, I. Ruisánchez, *Chemometr. Intell. Lab. Sys.* **1997**, *38*, 1 – 23.

[34] J. Dayhof, *Neural Network Architectures: An Introduction*, Van Nostrand-Reinhold, New York, **1990.**

[35] M. Novič, Z. Nikolovska-Cleska, T. Solmajer, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 990 – 998.

[36] I. Valkova, M. Vrako, S. C. Basak, *Anal. Chim. Acta* **2004**, *509*, 179 – 186.

[37] C.-C. Chang, C. -J Lin, LIBSVM : a library for support vector machines, **2001.** Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

[38] F. Lindgren, B. Hansen, W. Karcher, M. Sjöström, L. Eriksson, *J. Chemometrics*, **1996**, *10*, 521 – 532.

[39] W. Tong, H. Fang, H. Hong, Q. Xie, R. Perkins, D. M. Sheehan, Receptor-mediated toxicity: QSAR for oestrogen receptor binding and priority setting of potential oestrogenic endocrine disruptors, in: M. T. D. Cronin, D. J. Livingstone (Eds.), *Predicting Chemical Toxicology and Fate*, CRC Press, Boca Raton FL, **2004**, pp. 285 – 314.

[40] H. Fang, W. Tong, L. M. Shi, R. Blair, R. Perkins, W. Branham, B. S. Hass, Q. Xie, S. L. Dial, C. L. Moland, D. M. Sheehan, *Chem. Res. Toxicol.* **2001**, *14*, 280 – 294.

[41] D. M. Hawkins, S. S. Young, A. Rusinko, *Quant. Struct.-Act. Relat.* **1997**, *16*, 296 – 302.

[42] P. Blower, M. Fligner, J. Verducci, J. Bjoraker, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 393 – 404.

[43] J. W. Godden, J. R. Furr, J. Bajorath, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 182 – 188.

[44] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, *Multivariate and Megavariate Data Analysis-Principles and Applications,* Umea, Sweden: Umetrics AB, **2001.**

[45] Z. Daren, *Computers Chem.* **2001**, *25*, 197 – 204.

[46] M. Vracko, M. Novič, J. Zupan, *Anal. Chim. Acta* **1999**, *384*, 319 – 332.

[47] M. Vracko, D. Mills, S. C. Basak, *Environ. Toxicol. Pharmacol.* **2004**, *16*, 25 – 36.