J Math Chem (2011) 49:796–805
DOI 10.1007/s10910-010-9776-1

ORIGINAL PAPER

QSAR study on cytotoxic activity (against KB cells) of some hederagenin diglycosides using support vector regression

Hua-Jun Luo · Jun-Zhi Wang · Kun Zou

Received: 27 February 2010 / Accepted: 29 October 2010 / Published online: 17 November 2010

Abstract Quantitative structure-activity relationship (QSAR) study on the cytotoxic activity (against KB cells) of 19 hederagenin diglycosides was performed by using spatial and electronic descriptors based on support vector regression (SVR) techniques. The predictive power of the models was verified with the leave one out cross validation (LOOCV) test and independent test methods. For the LOOCV test, the cross validation squared correlation coefficient Q^2 value for optimal SVR model was 0.8827. Compared with stepwise multiple linear regression (MLR) and back propagation artificial neural network (BPANN) models, the SVR model was the most powerful with a square of predictive correlation coefficient R^2_{pred} of 0.7285 for the test set, which indicates that the SVR model has better substantially predictive ability and be a useful and powerful tool to construct the QSAR model.

Keywords QSAR · Hederagenin diglycosides · Support vector regression

1 Introduction

Triterpene saponins are found in a wide variety of natural products possessing various biological and pharmacological activities, which consist of dammarane type, 1anostane type, oleanane type and ursane type triterpene structures linked to the sugar moiety normally including glucose, arabinulose, xylose, rhamnose, etc [1–3]. Many

H.-J. Luo (✉) · J.-Z. Wang · K. Zou
Hubei Key Laboratory of Natural Products Research and Development,
College of Chemistry & Life Science, China Three Gorges University,
Yichang, Hubei 443002, People's Republic of China
e-mail: luohuajun@21cn.com

K. Zou
e-mail: kzou@ctgu.edu.cn

hederagenin saponins are natural triterpenoids widely distributed in higher plants [4–7] and have been shown to possess cytotoxicity against various cancer cell lines and in vivo tumors [8–10]. Recently some hederagenin diglycosides were synthesized and their cytotoxic activities (against KB cells) were tested [11]. The nuclear structure of the KB cell was thus investigated by confocal microscopy, which revealed that the cytotoxic activity of hederagenin diglycosides was partially due to their interaction with the cell membrane. But the relationships of their structure and activity are still not well understood. The quantitative structure -activity relationship (QSAR) model is a powerful approach used to explain how structural features influence the biological activities. So the aim of this paper is to construct QSAR model of cytotoxic activity of hederagenin diglycosides that can be used to predict the activities from their molecular descriptors by stepwise multiple linear regression (MLR) and support vector regression (SVR) methods.

2 Materials and methods

2.1 Data set

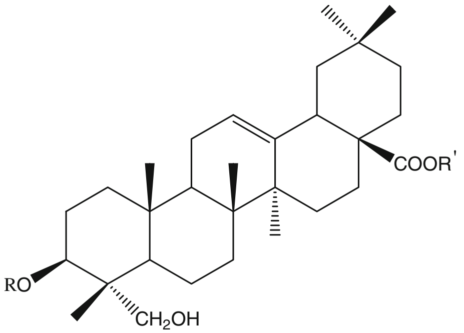
Cytotoxicity data for $\mathrm{IC}_{50}(\mu\mathrm{M})$, which is the concentration of hederagenin diglycosides causing 50% cell growth inhibition against KB cells, were obtained from the literature [11] and converted to negative logarithms of $\mathrm{IC}_{50}(\mathrm{pIC}_{50})$ in QSAR model. The molecular structures of the studied hederagenin diglycosides (**1-19**) are shown in Fig. 1 and the corresponding IC_{50} values are listed in Table 1.

2.2 Molecular descriptors

The molecular conformation was subjected to geometry optimization by PM3 semi-empirical quantum chemistry using HyperChem 8.0 (Evaluation version) [12]. Then the following molecular descriptors were collected: total energy (TE), binding energy (BE), heat of formation (HF), electronic energy (EE), nuclear energy (NE), dipole moment (DM), surface area (Area), volume, hydration energy (HE), LogP, refractivity (R), polarizability (Polar), and frontier orbital energies—$\mathrm{E}_{\mathrm{HOMO}}$ (energy of highest occupied molecular orbital), $\mathrm{E}_{\mathrm{LUMO}}$ (energy of lowest unoccupied molecular orbital) and the difference between $\mathrm{E}_{\mathrm{HOMO}}$ and $\mathrm{E}_{\mathrm{LUMO}}$ (ΔE) [13]. At last, Shadow Indices [14], Jurs Descriptors [15] and topological descriptors such as Wiener Index (WI) [16], connectivity Index $\left({}^{0}\mathrm{Xv}, {}^{1}\,\mathrm{Xv}\right)$ [17] were calculated.

2.3 Support vector regression

Support vector regression (SVR), originally proposed and developed by Vladimir Vapnik [18,19], is a relatively new nonlinear machine learning technique in the field of chemometrics and can handle higher dimensional data better even with a relatively low amount of training samples. In support vector regression, the basic idea is to map the data X into a higher dimensional feature space F via a nonlinear mapping Φ and then



R= α-L-Ara	R' = H	(**1**)			
R= α-L-Rha-(1→2)- α-L-Ara	R' = H	(**2**)	R= α-L-Rha-(1→3)- α-L-Ara	R' = CH_3	(**11**)
R= α-L-Rha-(1→3)- α-L-Ara	R' = H	(**3**)	R= α-L-Rha-(1→2)- β-L-Ara	R' = CH_3	(**12**)
R= β-D-Xyl-(1→2)- α-L-Ara	R' = H	(**4**)	R= β-D-Xyl-(1→2)- α-L-Ara	R' = CH_3	(**13**)
R= β-D-Xyl-(1→3)- α-L-Ara	R' = H	(**5**)	R= β-D-Xyl-(1→4)- α-L-Ara	R' = CH_3	(**14**)
R= β-D-Xyl-(1→4)- α-L-Ara	R' = H	(**6**)	R= β-D-Glc-(1→2)- α-L-Ara	R' = CH_3	(**15**)
R= β-D-Glc-(1→3)- α-L-Ara	R' = H	(**7**)	R= β-D-Glc-(1→3)- α-L-Ara	R' = CH_3	(**16**)
R= β-D-Glc-(1→4)- α-L-Ara	R' = H	(**8**)	R= β-D-Glc-(1→4)- α-L-Ara	R' = CH_3	(**17**)
R= α-L-Ara	R' = CH_3	(**9**)	R= β-D-Glc-(1→4)- β-D-Glc	R' = CH_3	(**18**)
R= α-L-Rha-(1→2)- α-L-Ara	R' = CH_3	(**10**)	R= β-D-Gal-(1→4)- β-D-Glc	R' = CH_3	(**19**)

Fig. 1 The molecular structures of the studied hederagenin diglycosides

to do linear regression in this space. Therefore, regression approximation addresses the problem of estimating a function based on a given data set $G = \{(x_i, d_i)\}_{i=1}^{l}$ (x_i is input vector, d_i is the desired value). SVR approximates the function in the following form:

$$y = \sum_{i=1}^{l} w_i \Phi_i(x) + b \tag{1}$$

where $\{\Phi_i(x)\}_{i=1}^{l}$ is the set of mappings of input features, $\{(w_i)\}_{i=1}^{l}$ and b are coefficients. They are estimated by minimizing the regularized risk function $R(C)$:

$$R(C) = C\frac{1}{N}\sum_{i=1}^{l} L_\varepsilon(d_i, y_i) + \frac{1}{2}\|w\|^2 \tag{2}$$

where

$$L_\varepsilon(d, y) = \left\{ \begin{array}{ll} |d - y| - \varepsilon \text{ for} & |d - y| \geq \varepsilon \\ 0 \qquad\qquad \text{otherwise} & \end{array} \right\} \tag{3}$$

and ε is a prescribed parameter in the insensitive loss function.

In Eq. (2), $\left[C(1/N)\sum L_\varepsilon(d_i, y_i)\right]$ is the so-called empirical error (risk) measured by ε-insensitive loss function $L_\varepsilon(d, y)$, which indicates that it does not penalize errors

Table 1 Experimental cytotoxic activity of hederagenin diglycosides (against KB cells)

No. compound	$IC_{50}(\mu M)$	pIC_{50}(exp.)
1	9.8	−0.9912
2	5.5	−0.7404
3	8.6	−0.9345
4	8.7	−0.9395
5	8.6	−0.9345
6	9.0	−0.9542
7	11.6	−1.0645
8	8.5	−0.9294
9	14.3	−1.1553
10	5.5	−0.7404
11	9.6	−0.9823
12	13.1	−1.1173
13	9.5	−0.9777
14	13.3	−1.1239
15	11.5	−1.0607
16	12.8	−1.1072
17	10.4	−1.0170
18	12.3	−1.0899
19	12.3	−1.0899

below ε. The second term, $\left[(1/2)\,\|w\|^2\right]$, is used as a measurement of function flatness. C is a regularized constant determining the tradeoff between the training error and the model flatness. The values of both ε and C have to be chosen by the user and the optimal values are usually data and problem dependent. Introduction of slack variables ξ leads Eq. (4) to the following constrained function Max:

$$\mathrm{Max} R\left(w, \xi^*\right) = \frac{1}{2}\,\|w\|^2 + C^* \sum_{i=1}^{l} \left(\xi_i + \xi_i^*\right) \tag{4}$$

$$\text{s.t.} \quad w\Phi(x_i) + b - d_i \le \varepsilon + \xi_i, \quad d_i - w\Phi(x_i) - b_i \le \varepsilon + \xi_i^*, \quad \xi_i, \xi_i^* \ge 0.$$

The minimization of Eq. (1) is a standard problem in optimization theory and it can be derived that the weight vector w equals the linear combination of the training data:

$$w = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) x_i \tag{5}$$

In this formula, α_i and α_i^* are Lagrange multipliers. Thus, decision function becomes the following form:

$$f(x) = \sum_{i=1}^{l} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + b \tag{6}$$

where $K(x_i, x)$ is the kernel function. The value is equal to the inner product of two vectors x_i and x_j in the feature space $\Phi(x)$. That is, $K(x_i, x_j) = \Phi(x_i)\Phi(x_j)$. The most used kernel functions include radial basis function (RBF) kernel, polynomial kernel and linear kernel. For the SVR calculations, a Matlab toolbox was used, developed by Gunn [20].

3 Results and discussion

3.1 Selection of the molecular descriptors

The experimental results in Table 1 show that the pIC_{50} values of the hederagenin diglycosides ranged from -1.1553 to -0.7404. To obtain a QSAR model with a more reliable predictive ability, external validation was used in model development. Compounds were classified into three clusters generated by K-means clustering method. The whole data set was divided into three clusters and the serial numbers of compounds in different clusters are listed in Table 2. So a test set of five compounds (**1, 7, 10, 11** and **15**) was selected. The other 14 compounds were used as the training set.

The molecular descriptors were selected using stepwise multiple linear regression method. In the stepwise MLR method, a multiple linear equation was built step by step. At each step all variables were assessed and evaluated to determine important descriptors. As the stepping process was terminated, the model with four important descriptors was obtained as follows:

$$\begin{aligned}\mathrm{pIC}_{50} = {} & (109.4171 \pm 15.0016) + (0.0214 \pm 0.0056)\ \text{Shadow\_Xlength} \\ & + (2.4815 \pm 0.4128)\ \text{Shadow\_YZfrac} + (10.6775 \pm 1.4519)\ \mathrm{E}_{\mathrm{HOMO}} \\ & - (12.1122 \mp 1.4935)\ \mathrm{E}_{\mathrm{LUMO}}\end{aligned} \tag{7}$$

($R^2 = 0.9297$, $R^2_{adj} = 0.8984$, $S = 0.0362$, $F = 29.7540$, $P = 0.0001$, $Q^2 = 0.8325$, $n_{\text{training}} = 14$)

where Shadow_Xlength is the length of molecule in the X dimension, Shadow_YZfrac is the ratio of molecular shadow area in the YZ plane, $\mathrm{E}_{\mathrm{HOMO}}$ is the energy of highest occupied molecular orbital, $\mathrm{E}_{\mathrm{LUMO}}$ is the energy of lowest unoccupied

Table 2 Serial numbers of compounds under different clusters

Cluster no.	No. of compounds in cluster	Serial number of compounds
1	9	**1, 3, 4, 5, 6, 8, 11, 13, 17**
2	8	**7, 9, 12, 14, 15, 16, 18, 19**
3	2	**2, 10**

molecular orbital. The statistical quality of the regression equation was examined using parameters such as the squared correlation coefficient (R^2), the squared adjusted correlation coefficient $\left(R^2_{adj}\right)$, the standard error (S), the Fisher ratio at the 95% confidence level (F), the statistical p value (P), and the cross validated squared correlation coefficient obtained based on Leave One Out (LOO) method (Q^2). Equation (7) predicted 83.25% of the variance and explained 89.84% of the variance of cytotoxic activity. According to the coefficients in Eq. (7), Shadow_Xlength, Shadow_YZfrac, and $\mathrm{E_{HOMO}}$ had positive impact on the cytotoxic activity, whereas $\mathrm{E_{LUMO}}$ decreased cytotoxic activity.

3.2 Selection of the SVR model parameters

The resulting descriptors in Eq. (7) decided by stepwise MLR were used for SVR model. The performance of SVR model is related to variables as well as the combination of parameters used in the model. So some parameters in SVR (the type of kernel function, the regularization parameter C and ε-insensitive loss function) ought to be optimized. In this work, Root mean squared error (RMSE) in leave one out cross validation (LOOCV) of SVR was used as the criteria of optimal set. The RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (Y_{pred} - Y_{\exp})^2}{n}} \tag{8}$$

The changing tendency of RMSE in LOOCV of SVR with the regularization parameter C and ε under three kernel functions including polynomial, radial basis and linear kernel function are given in Figs. 2 and 3. In general, the smaller value of RMSE was obtained, the better set of parameters gained. So it was found that the linear kernel function was suitable to build the SVR model among the three kernel functions.

Figure 4 illustrates the parameters (C and ε) optimized in SVR with linear kernel function. Thus the optimal SVR model was obtained with the minimum value of RMSE (0.0385) with $C = 250$ and $\varepsilon = 0.01$ under the linear kernel function. Figure 5 shows plots of the predicted values employing LOOCV of optimal SVM versus experimental values for pIC_{50}. It can be concluded that the predicted results are in good agreement with the experimental ones.

To compare the predictive capacity of QSAR model further, external validation of the test set was used in model development. The predictive R^2 $\left(R^2_{pred}\right)$ values were calculated according to the following equation:

$$R^2_{pred} = 1 - \frac{\sum \left(Y_{pred(Test)} - Y_{(Test)}\right)^2}{\sum \left(Y_{(Test)} - \bar{Y}_{\text{training}}\right)^2} \tag{9}$$

Here $Y_{pred(Test)}$ and $Y_{(Test)}$ represent the predicted and experimental values of the test set compounds, respectively. $\bar{Y}_{\text{training}}$ is the mean activity value of the training set.

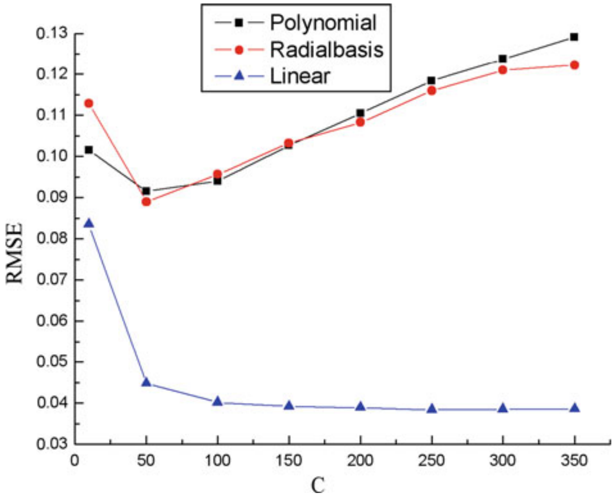


Fig. 2 RMSE in LOOCV versus the parameter C with $\varepsilon = 0.01$

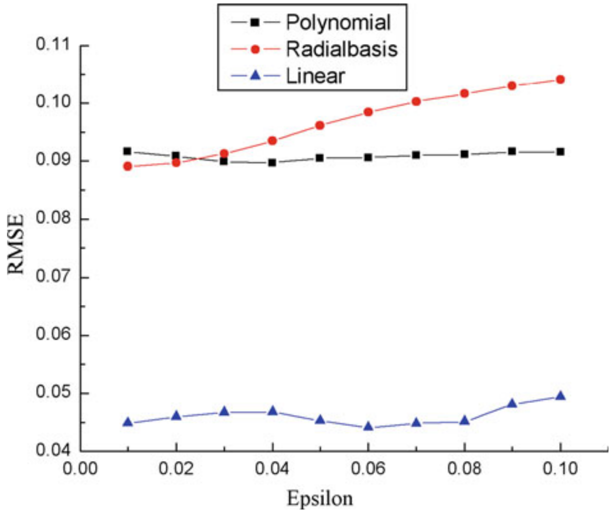


Fig. 3 RMSE in LOOCV versus the parameter ε with $C = 50$

The SVR model (optimal model as above) was compared with stepwise MLR and back propagation artificial neural network (BPANN). The parameters of BPANN model [21] with three layers used were as follows: the number of hidden nodes was three; the transformation function was sigmoid; the learning rate and momentum of each epoch were set to 0.10 and 0.20, respectively. The predicted activity values of QSAR models with some important descriptors were shown in Table 3 and the R^2_{pred} values of MLR, BPANN and SVR models were 0.6659, 0.6348 and 0.7285, respectively. Based on the above results, the quality of the SVR model outperforms those of

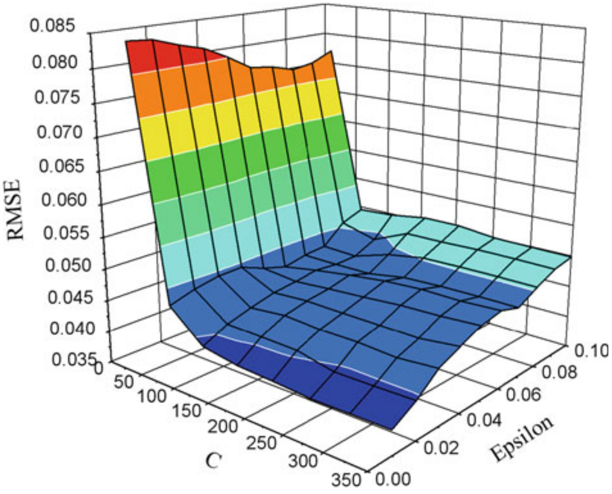


Fig. 4 RMSE in LOOCV versus C and ε with linear kernel function

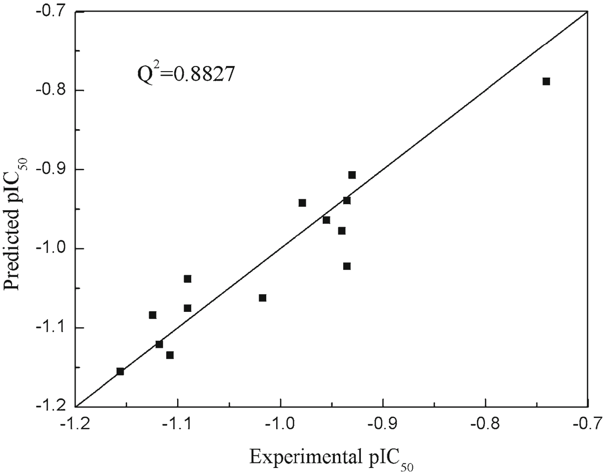


Fig. 5 Plot of experimental values versus predicted values for pIC_{50} using LOOCV of optimal SVR

MLR and BPANN models, which indicates that the SVR model has better substantially predictive ability.

4 Conclusions

The QSAR model based on stepwise MLR and SVR techniques for cytotoxic activity (against KB cells) of 19 hederagenin diglycosides was studied. The cytotoxic activity of hederagenin diglycosides depends strongly on spatial and electronic factors such as Shadow_Xlength, Shadow_YZfrac, E_{HOMO} and E_{LUMO}. The predictive power of the models was verified with the LOOCV test and independent test methods. For

Table 3 The predicted activity values of QSAR models with some important descriptors

Compound no.	Shadow_Xlength	Shadow_YZfrac	E_{HOMO}	E_{LUMO}	pIC_{50} (pred.)		
					MLR	BPANN	SVR
Training set							
2	20.8651	0.7270	−9.5540	0.8594	−0.7544	−0.7404	−0.7423
3	25.7028	0.7073	−9.4918	0.9347	−0.9484	−0.9345	−0.9366
4	20.2454	0.6819	−9.5085	0.9051	−0.9479	−0.9395	−0.9451
5	24.0200	0.6218	−9.4111	0.9885	−0.9860	−1.0381	−1.0149
6	23.7912	0.6964	−9.3929	1.0164	−0.9509	−0.9542	−0.9562
8	23.6076	0.7013	−9.3972	1.0096	−0.9054	−0.9294	−0.9116
9	20.1690	0.6718	−9.3839	1.0296	−1.1533	−1.1553	−1.1554
12	20.7023	0.7167	−9.4369	0.9920	−1.1405	−1.1173	−1.1195
13	20.2239	0.6861	−9.4840	0.9273	−0.9460	−0.9777	−0.9448
14	23.8491	0.6629	−9.3723	1.0383	−1.0765	−1.1238	−1.0881
16	25.8166	0.6644	−9.3866	1.0337	−1.1275	−1.1072	−1.1326
17	23.9242	0.6685	−9.3815	1.0294	−1.0524	−1.0170	−1.0625
18	24.9140	0.6515	−9.4588	0.9584	−1.0393	−1.0381	−1.0468
19	24.7172	0.6404	−9.4834	0.9377	−1.0823	−1.0381	−1.0878
Test set							
1	20.2841	0.6991	−9.4429	0.9661	−0.9433	−1.1334	−0.9426
7	24.7703	0.6516	−9.4220	0.9841	−0.9591	−1.0381	−0.9770
10	20.8166	0.7227	−9.5358	0.8812	−0.8360	−0.7404	−0.8220
11	25.6988	0.6996	−9.4892	0.9356	−0.9513	−1.0340	−0.9426
15	20.2532	0.7006	−9.4198	0.9921	−1.0085	−0.9894	−1.0054

the LOOCV test, the Q^2 value for optimal SVR was 0.8827. Compared with stepwise MLR and BPANN models, the SVR model was the most powerful with R^2_{pred} of 0.7285 for the test set. The SVR method has been shown to perform well for regression and be a useful and powerful tool to construct the QSAR model.

Acknowledgments This work was supported by grants from National Natural Science Foundation of China (No. 30760213, 30870254 and 30970296).

References

1. J.P. Vincken, L. Heng, A. de Groot, H. Gruppen, Phytochemistry **68**, 275–297 (2007)
2. S.G. Sparg, M.E. Light, J. van Staden, J. Ethnopharmacol. **94**, 219–243 (2004)
3. K. Hostettmann, A. Marston, *Saponins: Chemistry and Pharmacology of Natural Products* (Cambridge Univ. Press, Cambridge, 1995)
4. R. Higuchi, T. Kawasaki, Chem. Pharm. Bull. **24**, 1021–1032 (1976)
5. K. Hostettmann, Helv. Chim. Acta. **63**, 606–609 (1980)
6. N. Gopalsamy, J. Gueho, H.R. Julien, A.W. Owadally, K. Hostettmann, Phytochemistry **29**, 793–795 (1990)
7. H.J. Park, S.H. Kwon, J.H. Lee, K.H. Lee, K. Miyamoto, K.T. Lee, Plant. Med. **67**, 118–121 (2001)

8. C. Barthomeuf, E. Debiton, V. Mshvildadze, E. Kemertelidze, G. Balansard, Plant. Med. **68**, 672–675 (2002)
9. K.T. Lee, I.C. Sohn, H.J. Park, D.W. Kim, G.O. Jung, K.Y. Park, Plant. Med. **66**, 329–332 (2000)
10. S. Rooney, M.F. Ryan, Anticancer Res. **25**, 2199–2204 (2005)
11. M. Chwalek, N. Lalun, H. Bobichon, Biochimica Et Biophys. Acta. **1760**, 1418–1427 (2006)
12. HyperChem Release 8.0 Evaluation, Hypercube Inc. (2008), http://www.hyper.com
13. F. Jensen, *Introduction to Computational Chemistry* (Wiley, New York, 1999)
14. R.H. Rohrbaugh, P.C. Jurs, Analytica. Chimica. Acta. **199**, 99–109 (1987)
15. D.T. Stanton, P.C. Jurs, Anal. Chem. **62**, 2323–2329 (1990)
16. W.R. Muller, J. Comput. Chem. **8**, 170–173 (1987)
17. L.B. Kier, L.H. Hall, *Medicinal Chemistry, 14* (Academic Press, New York, 1976)
18. V. Vapnik, *Statistical Learning Theory* (Wiley, New York, 1998)
19. V. Vapnik, An overview of statistical learning theory. IEEE Trans. Neural Network. **10**, 988–999 (1999)
20. S.R. Gunn, *Support Vector Machines for Classification and Regression*. Technical report, Image speech and intelligent systems research group (University of Southampton, UK, 1997), http://www.isis.ecs.soton.ac.uk/isystems/kernel/
21. S. Haykin, *Neural Networks: A Comprehensive Foundation* (Macmillan, New York, 1999)