Mol Divers (2011) 15:877–887
DOI 10.1007/s11030-011-9321-6

FULL-LENGTH PAPER

# A classification study of human $\beta_3$-adrenergic receptor agonists using BCUT descriptors

**Ming Hao · Yan Li · Yonghua Wang · Shuwei Zhang**

Received: 13 March 2011 / Accepted: 17 May 2011 / Published online: 31 May 2011


**Abstract** Experimental $EC_{50}$s for 202 human $\beta_3$-AR agonists are used to develop classification models as a potential screening tool for a large library of target compounds before synthesis. A variable selection approach from random forests (VS-RF) is used to extract the structural information most relevant to the human $\beta_3$-AR activation properties of the collected data set. The obtained results indicate that the VS-RF method can be used for variable selection with smallest sets of non-redundant descriptors with highly predictive accuracy ($Q_{ex}\% = 96\%$ for the external prediction set). Thus, the proposed VS-RF models should be helpful for screening of potential human $\beta_3$-AR agonists before chemical synthesis in drug development.

**Keywords** Human $\beta_3$-adrenergic receptor agonists · Variable selection · Dragon descriptors · Random forest

## Introduction

Overactive bladder (OAB) is defined as urinary urgency with or without urgency incontinence, and the number of patients with OAB is estimated to be about 16% of adult population in the United States which is still steadily increasing worldwide [1]. The classical symptoms of OAB are urinary frequency and nocturia which seriously influence the life quality of the patients, and thus inducing an urgent requirement for corresponding therapeutic agents of the disease. Nowadays, the antimuscarinic agents that induce relaxation of the detrusor muscle have been widely used in the treatment of OAB [2]. These agents, however, have adverse effects such as dry mouth, constipation, and the potential for voiding difficulty in patients with poorly contractile bladders [2]. Consequently, drugs without these disadvantages would be a significant improvement over current therapy. Recently, it has been reported that $\beta_3$-AR, one of the three sub-types of $\beta$-AR (the other two are $\beta_1$- and $\beta_2$-ARs) which is a member of the G-protein coupled receptor (GPCR) family, is predominantly expressed in detrusor tissues in human, and its activation induces the relaxation of urinary bladder detrusor [3]. In addition, the concomitant activation of $\beta_1$- and $\beta_2$-ARs would cause undesirable side effects such as increased heart rate and/or muscle tremors. Therefore, $\beta_3$-AR agonists are expected to be new therapeutic candidates against OAB [4]. Besides, $\beta_3$-AR also mediates various pharmacological and physiological effects such as lipolysis in white adipocytes and thermogenesis in brown tissue adipocytes [5]. Thus, the $\beta_3$-AR activators are also recognized as potential drugs for the treatment of obesity and non-insulin-dependent diabetes.

Up to now, two generations of $\beta_3$-AR agonists have been developed. The first generation, some potent and selective rat $\beta_3$-AR agonists [6,7] such as BRL37344, CL316243, and FK175, has been reported to be effective anti-obesity and anti-diabetic agents in rodents. Unfortunately, these agonists discovered during the 1980s were found unsuccessful in the clinic either because of a lack of efficacy or due to an unfavorable cardiovascular side effect profile and/or poor pharmacokinetics [8]. Thus, a second generation of orally bioavailable human $\beta_3$-AR agonists with minimal side effects associated with the activation of $\beta_1$- and $\beta_2$-ARs has been a new target of recent

**Electronic supplementary material** The online version of this article (doi:10.1007/s11030-011-9321-6) contains supplementary material, which is available to authorized users.

M. Hao · Y. Li (✉) · S. Zhang
Department of Materials Science and Chemical Engineering, Dalian University of Technology, Dalian 116024, Liaoning, China
e-mail: yanli@dlut.edu.cn

Y. Wang
Center of Bioinformatics, Northwest A&F University, Yangling 71210, Shaanxi, China

research. Great efforts have been concentrated on the synthesis of selective $\beta_3$-AR agonists [6,9,8–11] such as LY377604, L796568, solabegron, and others; however, but these agonists are still not potent in terms of the pharmacokinetic properties. In light of these facts, Hattori et al. [12–17] have made continuous efforts and synthesized a large series of human $\beta_3$-AR agonists toward improving the potency and selectivity over human $\beta_1$- and $\beta_2$-ARs with good oral bioavailability, which makes it possible for later computational research.

As we know, in the laboratory, more accurate activity measurement generally requires more time and greater compound supply, which significantly slows down the process. Thus, the application of in silico methods such as quantitative structure–activity relationship (QSAR) to this issue could be considered at low cost before choosing a synthetic strategy [18] if the activity can be predicted only from the chemical structures. Nowadays, a dramatic increase in successful development of predictive, or computational models such as the prediction of P-glycoprotein substrates and inhibitors [19], androgenic and nonandrogenic compounds [20], PKC$\theta$ inhibitory activity [21], multidrug resistance reversal activity based on atom typing [22] and so forth can be seen. However, there is still, to our best knowledge, limited report of computational model to predict human $\beta_3$-AR antagonists up to date. Once Telvekar et al. [23] performed three-dimension QSAR study and pharmacophore mapping of a series of 80 biphenyl benzoic acid derivatives, the first two part of study reported by Hattori et al. [12,13], as selective human $\beta_3$-AR agonists. Using Pharmacophore Alignment and Scoring Engine (PHASE), they proposed a six-point common pharmacophore hypothesis with one acceptor, one negative charge, one positive charge, and three rings for pharmacophore-based alignment of molecules. And their subsequent comparative molecular field analysis (CoMFA) [24] and comparative molecular similarity index analysis (CoMSIA) [25] studies gave a predictive $r^2$ value of 0.664 and 0.867, respectively. They concluded that substitution of sterically favored groups at the biphenyl benzoic acid moiety with ether linkage and at methylene attached to amino and dibenzoic acid moieties increase the activity, and the phenyl ethanolamine moiety is responsible for providing pharmacophoric binding sites. While, in this study, we have enlarged the data set to 202 human $\beta_3$-AR agonists by collecting all six parts of Hattori's continuous study [12–17], with attempt to build high predictive classification models for rapid screening of potent $\beta_3$-AR agonists before synthesis.

Construction of a computational model often requires two basic elements. The first factor is the molecular descriptors that are used to capture the structural information of the molecules studied and correlate with the experimental observations. Since dragon software, a sophisticated program for calculation of molecular descriptors [26] developed by Milano Chemometrics and QSAR Research Group, has a quite good record of successful applications in various QSAR researches, presently we also employ it to calculate molecular descriptors based on only the two-dimension structures of the molecules expressed by SMILES notations, which require no specific orientation or through-space distances and thus alleviate the need for geometry optimization of the structures [27].

Another critical procedure is the choice of the data analysis approaches, since the assumptions that underpin a particular methodology must be shown to apply to the data under investigation. Often used classification methods include the simple but interpretable LDA and PLS, and nonlinear, relatively not being prone to interpretable but often having highly predictive methods such as SVM, RF, and so forth [20,28]. All these methods have a proven record of many successful applications in computational modeling. However, several of them suffer, sometimes, several limitations. Generally speaking, the largest limitation is the conditions where the number of the samples ($n$) is less than that of the descriptors ($p$). Under this circumstance, traditional statistical method like LDA can not be correctly applied unless a pre-selection of the descriptors is executed (e.g., by genetic algorithms [27,29] or genetic function approximation [30], etc). SVM, a nonlinear technique employed in classification problems, is also not robust to the presence of a large number of irrelevant descriptors [28]. Random forest (RF) has been reported as the combination of relatively high prediction accuracy and collections of desired features, which makes RF uniquely suited for modeling in cheminformatics [28] including prediction of quantitative or categorical biological activity of an unknown chemical based on a quantitative description of its molecular structure. RF can show excellent performance even when most predictive variables are noise, and be used when the number of variables is much larger than the number of observations, and returns measures of the variable importance. However, for approaching an ideal classification model (with high classification accuracy using less number of descriptors), a variable selection process is still required. To achieve the above object, in this study, a variable selection method by RF (described as "VS-RF" in this article) combined with the backward elimination using out-of-bag (OOB) error is selected to perform classification task for the current human $\beta_3$-AR agonists. Originally, this novel approach was proposed for gene selection and classification of microarray data, which has been proven as often yielding smaller sets of genes than alternative variable selection methods while retaining the predictive performance [31]. Although this method has been successfully applied in the field of gene selection and microarray data [31], there is still no record up to now of developing computational models for small molecular agonists. To extend the range of application, presently we examined the VS-RF combination method to classify the current dataset of $\beta_3$-AR agonists. For comparison, two other alternative methods (SVM and LDA) were also

performed on the basis of same selected descriptors within the same data sets. The obtained models, we hope, would be helpful for screening and identifying of potential $\beta_3$-AR agonists in a large library of target compounds before chemical synthesis.

## Material and experimental methods

### Data sets

A large, diverse dataset of 202 human $\beta_3$-AR agonists collected from articles [12–17] published by a same research group with $EC_{50}$ values ranging from 0.03 to larger than 100 nM were used as dataset in this study. Based on the inhibitory activity, the dataset is split into two classes, i.e., 102 low active (L) compounds with the range of $EC_{50}$ from 1.4 to >100 nM and 100 high active (H) ones with range from 0.03 to 1.3 nM. Table 1 depicts several representative compounds together with their classification labels. All information of the dataset with their diverse scaffolds of structures is provided in Table S1 (Supplementary Information).

### Descriptors calculation and pre-processing

In this study, the molecular structures of all agonists were built with the ISIS/Draw 2.3 program [32], and converted the SMILES format for calculation of their structural descriptors by Dragon soft [26]. These calculated descriptors have been reported using successfully in QSAR analysis [33]. Currently, 13 descriptor blocks for each molecule were calculated, including the constitutional and topological descriptors, walk and path counts, connectivity indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalues, topological charge and eigenvalue-based indices, functional group counts, atom-centered fragments, and molecular properties. Actually, originally Dragon calculated 929 molecular descriptors for each molecule. However, after excluding 263 constant or near-constant descriptors, we finally saved 666 ones which were further undertaken a preprocessing process (also called unsupervised selection of descriptors) as follows: (1) descriptors containing larger than 85% zero values were removed; (2) zero- and near zero-variance predictors were deleted; and (3) one of the two descriptors that have absolute correlations above 0.95 was omitted. After these steps, the number of original descriptors was reduced to 281 for further research.

### Split of the training and test sets

Rational division of an experimental SAR dataset into respective training and test sets for model development and validation is very important. The often used methods include random sampling (RS), Kennard–Stone (KS), $K$-mean clustering, self-organizing map (SOM), principal component analysis (PCA), etc. The basic rule should be that the points of the training set are distributed evenly within the whole area occupied by representative points, and the condition of closeness of the test set points to the training set ones is satisfied [34].

For the independent prediction set, we performed our selection on the basis of their distribution in the chemical space defined by PCA. In order to detect the homogeneities in the data set and identify possible outliers and clusters, PCA is performed within the calculated structure descriptors space for the whole data set. PCA is a useful multivariate statistical technique in which new variables (called principal components, PCs) are calculated as linear combinations of the old ones. These PCs are sorted by decreasing information content (i.e., decreasing variance) so that most of the information is preserved in the first few PCs. An important feature is that the obtained PCs are uncorrelated, and they can be used to derive scores which can be used to display most of the original variations in a smaller number of dimensions. These scores can also allow us to recognize groups of samples with similar behaviors.

### Statistical methods

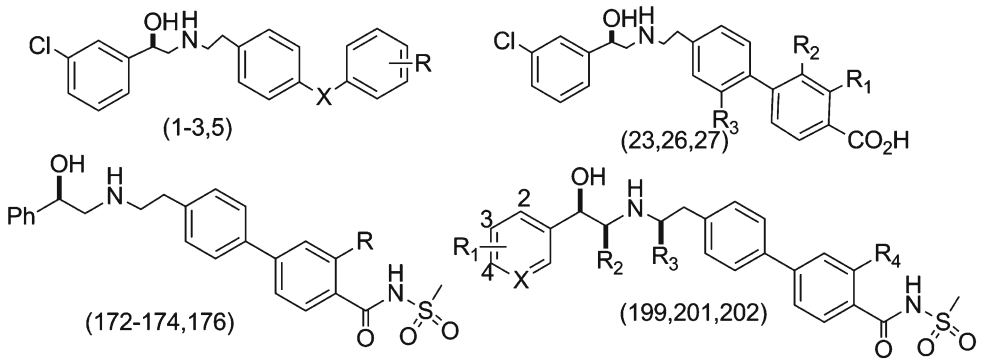
#### *VS-RF*

Random forest is a classification algorithm that uses an ensemble of unpruned decision trees, each of which is built on a bootstrap sample of the training data using a randomly selected subset of variables [35]. The random forest holds a number of appealing features making it well suited for performing classification task: (1) it is applicable when there are less observations than descriptors (predictors); (2) it performs embedded descriptor selection and it is relatively insensitive to the large number of irrelevant predictors; (3) it is based on the theory of ensemble learning that allows the algorithm to learn accurately both simple and complex classification function, and (4) it does not require much fine-tuning of parameters, and the default parameterization often leads to good performance [28,35].

Random forest has been successfully applied in the cancer microarray gene expression domain, but less in QSAR and QSPR (quantitative structure–property relationship) fields [21,28]. Thus, it should be of value to investigate whether RF can be applied to and obtains better statistical performance for the current dataset of human $\beta_3$-AR agonists. Here, only a brief introduction about RF is presented, since more details could be referred to the corresponding literatures [28,35]. In this study, the RF algorithm was employed using the R package randomForest [36].

Even though RF classifier is fairly insensitive to the number of irrelevant descriptors, we still applied following descriptor selection methods to further improve the classification

**Table 1** Representative compounds together with their classification labels for human $\beta_3$-AR agonists




| No. | Substituent | | | | | $EC_{50}$ (nM) | Class[a] | Ref.[b] |
|---|---|---|---|---|---|---|---|---|
| | X | | R | | | | | |
| 1[c] | O | | *p*-$OCH_2CO_2H$ | | | 48 | L | [12] |
| 2 | NH | | *p*-$OCH_2CO_2H$ | | | 12 | L | [12] |
| 3[c] | NMe | | *p*-$OCH_2CO_2H$ | | | 100 | L | [12] |
| 5 | S | | *p*-$OCH_2CO_2H$ | | | 85 | L | [12] |
| | $R_1$ | $R_2$ | | | $R_3$ | | | |
| 23 | H | F | | | H | 4.9 | L | [12] |
| 26[c] | H | H | | | Cl | 9.7 | L | [12] |
| 27 | H | H | | | OMe | 12 | L | [12] |
| | R | | | | | | | |
| 172 | | NH-*c*-Hex | | | | 0.14 | H | [16] |
| 173[c] | | *iso*-Bu | | | | 0.46 | H | [16] |
| 174 | | *c*-Pen | | | | 0.41 | H | [16] |
| 176 | | CH-*c*-Hex | | | | 0.54 | H | [16] |
| | $R_1$ | X | $R_2$ | $R_3$ | $R_4$ | | | |
| 199 | 3-OH | CH | H | H | O-*iso*-Pr | 0.044 | H | [17] |
| 201[c] | 4-OH | CH | $CH_3$ | H | O-*c*-Hex | 0.35 | H | [17] |
| 202 | 4-OMe | CH | $CH_3$ | H | O-c-Hex | 28 | L | [17] |

[a] H denotes high active compounds, L denotes low active compounds; [b] From the corresponding reference; [c] Test set

performance and increase the computational efficiency. To select the best set of descriptors, presently random forest-based backward elimination procedure was performed. The procedure involves iteratively fitting RFs and at each iteration building a new forest after discarding those descriptors with the smallest importance values; the selected set of descriptors is the one with the smallest OOB error rate. The variable selection procedure was achieved using the R package varSelRF [37]. We applied it with the recommended parameters: *ntree* = 2000, *fraction.dropped* = 0.2 (a parameter indicating fraction of descriptors with small importance values to be discarded during backward elimination procedure), *mtryFactor* = 1, *nodesize* = 1 and *c.sd* = 1 (a factor that multiplies the standard deviation of error for stopping iterations and choosing the best performing subset of descriptors). More details on the this algorithm can be referred to literature [31].

*Support vector machine (SVM)*

SVMs are a machine learning algorithm originally developed by Vapnik and co-workers [38]. SVM approach automatically controls the flexibility of the resulting classifier on the training data. By design of the algorithm, the deteriorating effect of the input dimensionality on the generalization ability is greatly suppressed. Due to its many attractive features and promising empirical performances, SVM is gaining increasing popularity in many fields, and thus was also performed in this study. The Gaussian RBF kernel was used in our experiment.

$$k(x, x') = \exp(-\sigma \left\| x - x' \right\|^2) \tag{1}$$

With this kernel, two parameters ($C$ and $\sigma$) were determined in the SVM model using 10-fold cross validation. And the R package kernlab [39] was used to develop the SVM classification model.

*Linear discriminant analysis (LDA)*

LDA is a multivariate statistical procedure that aims to split objects into two or more categories. The basic theory of LDA is to classify the dependents by dividing an $n$-dimensional descriptor space into two regions that are separated by a hyperplane defined by a linear discriminant function $\left(y = c + \sum b_i x_i\right)$. This is optimized by adjusting $c$ and $b_i$ to obtain a maximum separation of the two classes. In this study, the independent variables were the calculated molecular descriptors, and the discrimination property was $EC_{50}$ (represented by either high or low active compounds). LDA analysis was performed using the R package MASS [40].

Evaluation of the statistical performance

The performances of VS-RF, SVM, and LDA were measured using several statistics:

(1) Accuracy: the proportion of correctly classified instances:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

where true positives (TP) denote the correct classifications of positive examples (i.e., high active agonists in this study); true negatives (TN) are the correct classifications of negative examples (i.e., low active agonists here); false positives (FP) represent the incorrect classification of negative examples into the positive classes; and false negatives (FN) are the positive examples incorrectly classified into the negative classes.

(2) Sensitivity: the percentage of positive examples which are correctly classified;

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

(3) Specificity: the percentage of negative examples which are correctly classified;

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{4}$$

(4) Positive predictive value (PPV): the percentage of the examples predicted to be positive that are correct;

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

(5) Negative predictive value (NPV): the percentage of examples predicted to be negative that are correct;

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}} \tag{6}$$

In order to show how the classification system performs on the high active and low active compounds separately, the confusion matrix, which can transfer more information on the statistical results, for the best model is shown in the following form (Table 2).

In additional to above criteria, Matthews correlation coefficient (MCC) [41], which indicates the accuracy of real and estimated class, respectively, is also used to measure the prediction accuracies and can be given as follows:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FN} \times \text{FP}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \tag{7}$$

## Results and discussion

PCA of the dataset

As a kind of multivariate statistical technique that can be used for clustering, visualization, and abstraction tasks, PCA is suitable for data survey due to its visualization properties which have been successfully applied to dataset split [42]. In the current work, PCA gives three significant PCs, which together explain 54.68% of the variation in the data (with first PC 33.21%, second PC 12.67%, and third PC 8.80%, respectively). Figure 1 demonstrates the distribution of the molecules, where the blue asterisk denotes the training set and red hexagram stands for the test set, respectively. As seen from this figure, on one hand, representative points of the test set are close to those of the training one and on the other hand, the training and test sets uniformly occupy the

**Table 2** The format of confusion matrix for the best classification model

| | Observed | | Total | %Correct |
|---|---|---|---|---|
| | H | L | | |
| Predicted | | | | |
| H | TP | FP | TP + FP | PPV |
| L | FN | TN | FP + TN | NPV |
| Total | TP + FN | FP + TN | TP + FP + FN + TN | |
| %Correct | Sensitivity | Specificity | | Accuracy |

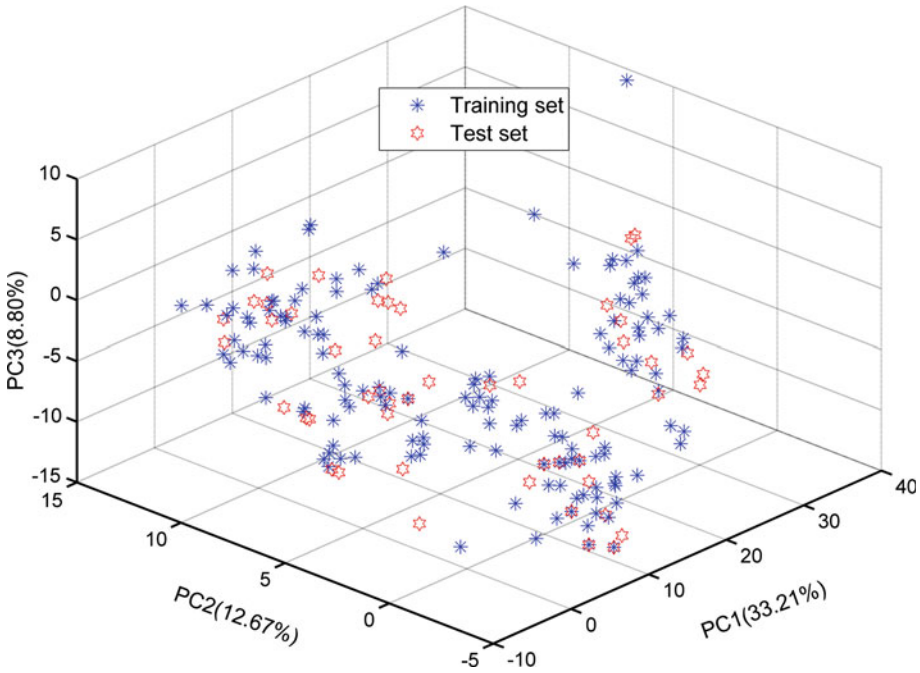



**Fig. 1** Principal components analysis of the structural descriptors for the human $\beta_3$-AR agonists. Asterisks and hexagrams denote the training and test set compounds, respectively

whole chemical space, which both indicate a rational selection of the training and test compounds in this study.

The training set was used for the development of the classification models with variable selection, and the independent prediction set was used for the assessment of the system. The training and independent test sets contain 152 (75 high active and 77 low active) and 50 (25 high active and 25 low active) compounds, respectively, with approximately one-fourth of the respective groups assigned in the independent prediction set. Table S1 (Supplementary Information) lists the structures of all the molecules.

## Selected descriptors using VS-RF

Apart from the quality of the data sets used, the selection of descriptors relevant to human $\beta_3$-AR activity is also important for optimization of the prediction system by reducing the noise and increasing computational efficiency in a statistical learning process. A VS-RF strategy has been developed successfully, with the final number of descriptors being reduced to 3 from the original 281 for further study. Since it is recommended that the number of compounds in the training set should be at least five times larger than that of the selected independent variables [43], the model developed by VS-RF obviously maintains the recommended ratio. Table 3 lists the selected descriptors together with their definitions, and Table S2 (Supplementary Information) lists their values.

## Performance evaluation

To estimate the performances of the statistical learning methods for prediction of the diverse set of $\beta_3$-AR agonists, it is useful to examine whether the accuracy from the three different statistical methods (VS-RF, SVM, and LDA) is at a similar level. Tables 4 and 5 summarize the detailed statistics, where the predicted results by the methods are presented in Table S3 (Supplementary Information).

A set of 152 agonists are chosen as a training set to derive the binary classification models and optimize the developed models, while the external test set including 50 compounds is only used to estimate the model performances. As a classification algorithm, random forest, generally, has only one parameter (i.e., $m_{\text{try}}$) that could be considered a tuning variable. Although it was shown in a previous report [28],

**Table 3** The selected 3 descriptors using VS + RF and their definition

| Descriptors | Definition | Class |
|---|---|---|
| BELm4 | Lowest eigenvalue $n$. 4 of Burden matrix/weighted by atomic masses | BCUT descriptors |
| BELp4 | Lowest eigenvalue $n$. 4 of Burden matrix/weighted by atomic polarizabilities | BCUT descriptors |
| BELv4 | lowest eigenvalue $n$. 4 of Burden matrix/weighted by atomic van der Waals volumes | BCUT descriptors |

**Table 4** Confusion matrices for the VS-RF using human $\beta_3$-AR agonists

| | Observed | | Total | %Correct |
|---|---|---|---|---|
| | H | L | | |
| *Training set* | | | | |
| Predicted | | | | |
| H | 73 | 1 | 74 | 98.65 |
| L | 2 | 76 | 78 | 97.44 |
| Total | 75 | 77 | 152 | |
| %Correct | 97.33 | 98.70 | | 98.03 |
| *Test set* | | | | |
| Predicted | | | | |
| H | 24 | 1 | 25 | 96.00 |
| L | 1 | 24 | 25 | 96.00 |
| Total | 25 | 25 | 50 | |
| %Correct | 96.00 | 96.00 | | 96.00 |

**Table 5** The prediction performance of high and low active compounds as human $\beta_3$-AR agonists from VS-RF, SVM, and LDA statistical methods for the external prediction set

| Model[a] | High active agonists | | | Low active agonists | | | $Q_{ex}$ (%) | MCC |
|---|---|---|---|---|---|---|---|---|
| | TP | FN | SE (%) | TN | FP | SP (%) | | |
| VS-RF | 24 | 1 | 96.00 | 24 | 1 | 96.00 | 96.00 | 0.92 |
| SVM | 24 | 1 | 96.00 | 22 | 3 | 88.00 | 92.00 | 0.84 |
| LDA | 23 | 2 | 92.00 | 24 | 1 | 96.00 | 94.00 | 0.88 |

[a] RF, $m_{try} = 1$; SVM, $C = 0.1$, $\sigma = 5.6$

although it has been shown that the performance of RF using a fixed set of descriptors is often relatively insensitive to the choice of $m_{try}$ specified as a function of number of descriptors ($p^{1/2}$ for classification), it is still necessary to attempt the variation of this value [20]. Since the number of final selected descriptors is three, the $m_{try}$ value is just tried from 1 to 3, the optimal one of which is determined by 10-fold cross-validation accuracy ($Q_{cv} = 0.80$). Ultimately, optimal RF results are obtained based on $m_{try} = 1$ and 500 trees in the forest. The confusion matrix for the optimal VS-RF model is given in Table 4. For the training set, the sensitivity is 97.33% (73 out of 75 high active compounds are correctly classified), and specificity is 98.70% (76 out of 77 low active compounds are correctly classified). Thus, both the sensitivity and specificity of the optimal VS-RF model show a perfect classification for the human $\beta_3$-AR agonists. Finally, an encouraging overall accuracy of 98.03% is obtained for the training set.

The validation of QSAR models is important because it assesses the model's reliability and prediction ability. Thus, both the external and internal validations are performed, where the internal validation is based on the training set data and the external one accomplished using a separate set of data (the test set) that is not used in the model development.

Here, the classification results in VS-RF for the test set is listed in Table 4. For the 50 tested compounds including 25 high active and 25 low active ones, the statistics for the set (sensitivity = 96%, specificity = 96, and overall $Q = 96\%$) exhibited good performances similar to the training set, proving the reliability and high predictive capacity of the proposed VS-RF model.

For comparison with the VS-RF, two other popular algorithms, SVM and LDA, are also performed using the same data set as VS-RF. Table 5 shows the detailed comparison results for the external prediction set. Based on the selected descriptors, the state-of-art implementation of SVM available in the R package kernlab is also employed for achieving classifications. Similar to other multivariate statistical models, the performance of SVM depends on the combination of several parameters including the capacity parameter $C$, the kernel type $K$, and its corresponding indices. $C$ is a regularization parameter which controls the tradeoff between maximizing the margin and minimizing the training error. In this study, the grid search technology was employed to obtain the optimum parameters ($C$ and $\sigma$) using the R package caret [44] on the basis of 10-fold cross validation. Here, the function sigest in the kernlab package was used to provide a good estimate of the $\sigma$ parameter, so that only the $C$ parameter was tuned. The final values used in the model are $C = 0.1$ and $\sigma = 5.6$ with the highest 10-fold cross-validation accuracy (0.77). Using the determined optimal parameters, the SVM obtains statistical results of 96, 88, and 92% for the sensitivity, specificity, and overall external accuracy ($Q_{ex}$) of the test set, respectively. And the MCC achieves 0.84 (Table 5).

Since the above depicted methods all based on a nonlinear technique, it is interesting to attempt a linear method to separate the categories of the studied compounds. Thus, LDA, a widely used classification technology, was also carried out to class the current dataset based on the selected three descriptors. As shown in Table 5, LDA gives relatively low sensitivity of 92%, specificity of 96% with a total accuracy of 94%. For the additionally statistical parameter, MCC, LDA gives 0.88 for the external prediction set.

## Comparison of different approaches

After the above discussion, it can be concluded that although the performances of all three methods are comparative, VS-RF still performs slightly non-significant better than the others in terms of internal ($Q_{cv} = 0.80$) and external ($Q_{ex} = 0.96$) prediction ability. Another popular machine learning method, SVM, though keeps a same high sensitivity as that of VS-RF indicating that both of them possess ability with high correct classification of the high active agonists, gives a relatively low specificity of 88% with the reduction in specificity of 8% compared to that of the VS-RF model, and the value of its MCC is also the lowest one (0.84) among the three models. In addition, by comparing with SVM, LDA decreases by 4% in the sensitivity, while increases by 8% in the specificity (as up to 96%) which is the highest one among the three models. Therefore, LDA also presents high prediction accuracy. In summary, by a comparison of the three models obtained in this study, the order of them in overall predictivity performance is VS-RF > LDA > SVM.

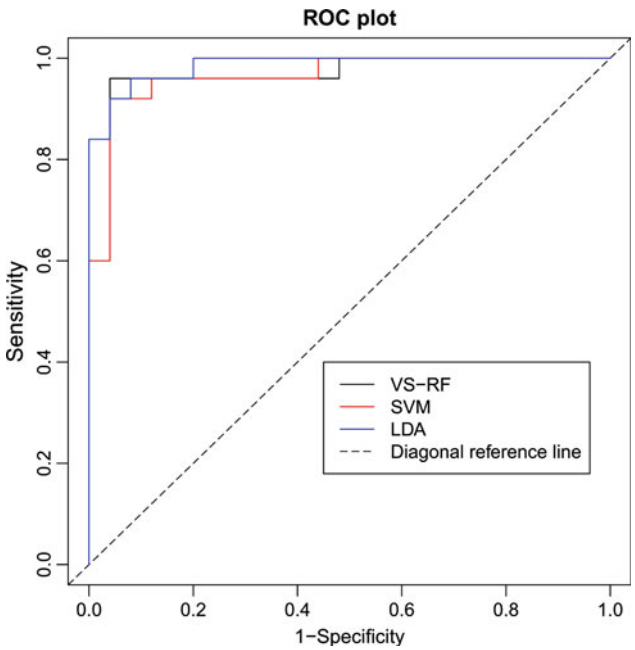



**Fig. 2** The ROC curves of VS-RF, SVM, and LDA for the prediction set

The area under the ROC curve (AUC) is also considered as an important criterion for measuring the performance of the model [22]. The ROC curve demonstrates the model's sensitivity, the ability to identify true positives, and specificity, the ability to avoid false negatives. The area under the ROC curve is a quantitative measure of the model performance. An AUC value of 1 indicates a theoretically perfect performance, while a value of 0.5 denotes no prediction ability. Clearly, the closer the AUC value is to 1, the better the model performance is. Figure 2 gives the ROC curves of VS-RF, SVM, and LDA for the prediction set. The computed AUC values of the three statistical methods are 0.96, 0.95 and 0.97, respectively, indicating good prediction ability and reliability of all three statistical models. However, when inspecting and comprising the three methods, it is noted that VS-RF seems to outperform the other two since, for the test set, there are only two misclassifications, while the SVM and LDA models present four and three wrongly classified compounds, respectively (Table 5). The reasons for the better performance of VS-RF may be that as an ensemble algorithm, the random forest approach, by constructing an ensemble out of all of these accurate classifiers, can "average" their votes and reduce the risk choosing the wrong classifier. Previous literature [28] also illustrates that random forest suited for modeling in QSAR field. However, it should also be kept in mind that no single technique can claim to be uniformly superior to any other. For example, the nonlinear algorithm of SVM has been reported outperformed LDA [45], while in this study, the performance of LDA is slightly better than that of SVM. Therefore, to aim directly at different

research systems, one should attempt multiple methods to find the optimal one to study further. In this study, we consider VS-RF as the best classifier for achieving current classification task.

### *Y*-Randomization check for VS-RF

*Y*-randomization, randomly scrambling the responses, is another validation approach that should be used in parallel with cross validation, and be applied to test the significance of the derived model [43]. In order to further investigate the validity of the generated VS-RF model here, we have repeated 100 *Y*-randomization checks and compared with the prediction statistics without such checks. In all random shuffles of the *Y* vector we tried, we noted that the values of sensitivity, specificity, and $Q_{ex}$ are all significantly reduced (more than 40%) for the prediction set. All these illustrate that the developed prediction model by VS-RF is not due to a chance correlation.

### Interpretation of the selected descriptors

Using feature selection, the most appropriate sets of molecular descriptors for predicting the low or high active $\beta_3$-AR agonists are extracted from the VS-RF models, the interpretation of which might provide some new insights into the physicochemical characteristics of human $\beta_3$-AR agonists.

It is very interesting to note that in this study, all the three parameters selected are BCUT descriptors (Table 3), who are the eigenvalues of a modified connectivity matrix, the Burden matrix [46]. The matrix is an *H* depleted molecular graph defined as follows: diagonal elements are atomic numbers of the elements ($Z_i$); off-diagonal elements ($B_{ij}$), representing bonded atoms *i* and *j* are equal to $\pi^* \times 10^{-1}$, where $\pi^*$ is the conventional bond order (i.e., 1, 2, 3, 1.5 for single, double, triple, and aromatic bonds, respectively); off-diagonal elements corresponding to the terminal bonds are increased by 0.01, and all other matrix elements are set to 0.001. The ordered sequence of the *n* smallest eigenvalues of *B* was proposed as a molecular descriptor based on the assumption that the lowest eigenvalues contain contributions from all the atoms and thus reflect the topology of the molecule. The BCUT descriptors are an extension of the Burden eigenvalues and consider three classes of matrices whose diagonal elements correspond to atomic charge related values, atomic polarizability related values, and atomic H bond abilities. A variety of definitions have been used for the off-diagonal terms, and both 2D and 3D approaches are considered. The highest and lowest eigenvalues of these matrices have been shown to be discriminating descriptors [47]. BELm4 is the lowest eigenvalue *n*. 4 of Burden matrix involving the atomic masses as weighting scheme. BELp4 is lowest eigenvalue *n*. 4 of Burden matrix/weighted by atomic polarizabilities, and BELv4 is lowest eignvalue *n*. 4 of Burden matrix/weighted by atomic van der Waals volumes.

In fact, the BCUT metrics have been successfully applied to QSAR studies. For example, Stanton [48] has found the BCUT metrics can provide unique information regarding the molecular structures and make significant contributions to resulting equations; In 2000, Pirard and Pickett [49] correctly classified the kinase inhibitors using the PLS discriminant analysis coupled with the BCUT descriptors, and thereafter the author [29] also presented an application of BCUT metrics and genetic algorithm in binary QSAR analysis with highly predictive models obtained. In addition, Ford et al. [50] applied both the LDA and a committee of neural networks using BCUT parameters as input variables to recognize compounds that act at biological targets belonging to protein kinases. Their results illustrated that BCUT metrics have utility in discriminating compounds that interact with particular gene families. Here, our highly predictive classification models further indicate that BCUT descriptors are useful in QSAR studies and should be extensively applied in the further study.

In summary, from the aforementioned discussion, it can be seen that the activity of these human $\beta_3$-AR agonists is mainly influenced by several factors including atomic masses, atomic polarizabilities, and van der Waals volumes. Our results are partly in agreement with the previous research [23]. For example, in their common pharmacophore hypothesis features, an acceptor, a negative charge and a positive charge are included, which are proven by our selected structure information (atomic polarizabilities). It also should be pointed out that though some structure features selected by our study cannot directly instruct the structural improvement of human $\beta_3$-AR agonists, in terms of developing a highly predictive classification model; however, the proposed VS-RF model in this study could implement this task (Tables 4 and 5).

## Conclusions

In this study, based on the up-to-date largest dataset to our best knowledge of 202 structurally diverse human $\beta_3$-AR agonists, a VS-RF classification model with good predictive performance (with an overall $Q = 96\%$ for the prediction set) has been built. By explanation of the selected descriptors, we conclude that atomic masses, atomic polarizabilities, and van der Waals volumes play a central role in the $\beta_3$-AR inhibition, which is supported by previous research [23]. Moreover, a comparison with other two statistical methods (i.e., SVM and LDA), the VS-RF model presents slightly non-significant better statistics both from the internal and external validations. Therefore, we hope that the proposed VS-RF method and the derived model would be of help for predictive tasks to

screen new and potent human $\beta_3$-AR agonists in early drug development.

**Acknowledgments** This study is financially supported by the National Natural Science Foundation of China (Grant No. 10801025). The authors thank the R Development Core Team for affording the free R2.10 software.

## References

1. Stewart W, Van Rooyen J, Cundiff G, Abrams P, Herzog A, Corey R, Hunt T, Wein A (2003) Prevalence and burden of overactive bladder in the United States. World J Urol 20:327–336. doi:10.1007/s00345-002-0301-4
2. Abrams P, Andersson K (2007) Muscarinic receptor antagonists for overactive bladder. BJU Int 100:987–1006. doi:10.1111/j.1464-410X.2007.07205.x
3. Emorine L, Marullo S, Briend-Sutren M, Patey G, Tate K, Delavier-Klutchko C, Strosberg A (1989) Molecular characterization of the human $\beta_3$-adrenergic receptor. Science 245:1118–1121. doi:10.1126/science.2570461
4. Yamaguchi O, Chapple C (2007) $\beta_3$-adrenoceptors in urinary bladder. Neurourol Urodyn 26:752–756. doi:10.1002/nau.20420
5. Arch J, Ainsworth A, Cawthorne M, Piercy V, Sennitt M, Thody V, Wilson C, Wilson S (1984) Atypical $\beta$-adrenoceptor on brown adipocytes as target for anti-obesity drugs. Nature 309:163–165. doi:10.1038/309163a0
6. Hu B, Jennings LL (2003) Orally bioavailable $\beta_3$-adrenergic receptor agonists as potential therapeutic agents for obesity- and type-II diabetes. Prog Med Chem 41:167–194. doi:10.1016/S0079-6468(02)41005-3
7. Uchida H, Shishido K, Nomiya M, Yamaguchi O (2005) Involvement of cyclic AMP-dependent and -independent mechanisms in the relaxation of rat detrusor muscle via $\beta$-adrenoceptors. Eur J Pharmacol 518:195–202. doi:10.1016/j.ejphar.2005.06.029
8. de Souza C, Burkey B (2001) $\beta_3$-adrenoceptor agonists as anti-diabetic and anti-obesity drugs in humans. Curr Pharm Des 7:1433–1449. doi:10.2174/1381612013397339
9. Mathvink RJ, Tolman JS, Chitty D, Candelore MR, Cascieri MA, Colwell LF, Deng L, Feeney WP, Forrest MJ, Hom GJ, MacIntyre DE, Miller RR, Stearns RA, Tota L, Wyvratt MJ, Fisher MH, Weber AE (2000) Discovery of a potent, orally bioavailable $\beta_3$ adrenergic receptor agonist, $(R)-N$-[4-[2-[[2-hydroxy-2-(3-pyridinyl)ethyl] amino] ethyl] phenyl]-4-[4-[4-(trifluoromethyl)phenyl]thiazol-2-yl]benzenesulfonamide. J Med Chem 43:3832–3836. doi:10.1021/jm000286i
10. Uehling DE, Shearer BG, Donaldson KH, Chao EY, Deaton DN, Adkison KK, Brown KK, Cariello NF, Faison WL, Lancaster ME, Lin J, Hart R, Milliken TO, Paulik MA, Sherman BW, Sugg EE, Cowan C (2006) Biarylaniline phenethanolamines as potent and selective $\beta_3$ adrenergic receptor agonists. J Med Chem 49:2758–2771. doi:10.1021/jm0509445
11. Shearer BG, Chao EY, Uehling DE, Deaton DN, Cowan C, Sherman BW, Milliken T, Faison W, Brown K, Adkison KK, Lee F (2007) Synthesis and evaluation of potent and selective $\beta_3$ adrenergic receptor agonists containing heterobiaryl carboxylic acids. Bioorg Med Chem Lett 17:4670–4677. doi:10.1016/j.bmcl.2007.05.069
12. Imanishi M, Tomishima Y, Itou S, Hamashima H, Nakajima Y, Washizuka K, Sakurai M, Matsui S, Imamura E, Ueshima K, Yamamoto T, Yamamoto N, Ishikawa H, Nakano K, Unami N, Hamada K, Matsumura Y, Takamura F, Hattori K (2008) Discovery of a novel series of biphenyl benzoic acid derivatives as potent and selective human $\beta_3$-adrenergic receptor agonists with good oral bioavailability. Part I. J Med Chem 51:1925–1944. doi:10.1021/jm701324c
13. Imanishi M, Itou S, Washizuka K, Hamashima H, Nakajima Y, Araki T, Tomishima Y, Sakurai M, Matsui S, Imamura E, Ueshima K, Yamamoto T, Yamamoto N, Ishikawa H, Nakano K, Unami N, Hamada K, Matsumura Y, Takamura F, Hattori K (2008) Discovery of a novel series of biphenyl benzoic acid derivatives as highly potent and selective human $\beta$3 adrenergic receptor agonists with good Oral bioavailability. Part II. J Med Chem 51:4002–4020. doi:10.1021/jm8000345
14. Imanishi M, Nakajima Y, Tomishima Y, Hamashima H, Washizuka K, Sakurai M, Matsui S, Imamura E, Ueshima K, Yamamoto T, Yamamoto N, Ishikawa H, Nakano K, Unami N, Hamada K, Matsumura Y, Takamura F, Hattori K (2008) Discovery of a novel series of benzoic acid derivatives as potent and selective human $\beta_3$ adrenergic receptor agonists with good oral bioavailability. 3. Phenylethanolaminotetraline (PEAT) skeleton containing biphenyl or biphenyl ether moiety. J Med Chem 51:4804–4822. doi:10.1021/jm800222k
15. Nakajima Y, Imanishi M, Itou S, Hamashima H, Tomishima Y, Washizuka K, Sakurai M, Matsui S, Imamura E, Ueshima K, Yamamoto T, Yamamoto N, Ishikawa H, Nakano K, Unami N, Hamada K, Hattori K (2008) Discovery of novel series of benzoic acid derivatives containing biphenyl ether moiety as potent and selective human $\beta_3$-adrenergic receptor agonists. Part IV. Bioorg Med Chem Lett 18:5037–5040. doi:10.1016/j.bmcl.2008.08.009
16. Hattori K, Toda S, Imanishi M, Itou S, Nakajima Y, Washizuka K, Araki T, Hamashima H, Tomishima Y, Sakurai M, Matsui S, Imamura E, Ueshima K, Yamamoto T, Yamamoto N, Ishikawa H, Nakano K, Unami N, Hamada K, Matsumura Y, Takamura F (2009) Discovery of highly potent and selective biphenylacylsulfonamide-based $\beta_3$-adrenergic receptor agonists and evaluation of physical properties as potential overactive bladder therapies. Part 5. J Med Chem 52:3063–3072. doi:10.1021/jm9000709
17. Hattori K, Orita M, Toda S, Imanishi M, Itou S, Nakajima Y, Tanabe D, Washizuka K, Araki T, Sakurai M, Matsui S, Imamura E, Ueshima K, Yamamoto T, Yamamoto N, Ishikawa H, Nakano K, Unami N, Hamada K, Matsumura Y, Takamura F (2009) Discovery of highly potent and selective biphenylacylsulfonamide-based $\beta_3$-adrenergic receptor agonists and molecular modeling based on the solved X-ray structure of the $\beta_2$-adrenergic receptor. Part 6. Bioorg Med Chem Lett 19:4679–4683. doi:10.1016/j.bmcl.2009.06.083
18. Sun X, Li Y, Liu X, Ding J, Wang Y, Shen H, Chang Y (2008) Classification of bioaccumulative and non-bioaccumulative chemicals using statistical learning approaches. Mol Divers 12:157–169. doi:10.1007/s11030-008-9092-x
19. Wang Y, Li Y, Yang S, Yang L (2005) Classification of substrates and inhibitors of P-glycoprotein using unsupervised machine learning approach. J Chem Inf Model 45:750–757. doi:10.1021/ci050041k
20. Li Y, Wang Y, Ding J, Wang Y, Chang Y, Zhang S (2009) In silico prediction of androgenic and nonandrogenic compounds using random forest. QSAR Comb Sci 28:396–405. doi:10.1002/qsar.200810100
21. Hao M, Li Y, Wang Y, Zhang S (2010) Prediction of PKC$\theta$ inhibitory activity using the random forest algorithm. Int J Mol Sci 11:3413–3433. doi:10.3390/ijms11093413
22. Sun H (2005) A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. J Med Chem 48:4031–4039. doi:10.1021/jm050180t
23. Telvekar V, Patel D, Jadhav N, Mishra S (2010) Three-dimensional QSAR and pharmacophore mapping of biphenyl benzoic acid

derivatives as selective human $\beta_3$-adrenergic receptor agonists. Med Chem Res 19:1174–1190. doi:10.1007/s00044-009-9261-1
24. Richard D, David E, Jeffrey D (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc 110:5959–5967. doi:10.1021/ja00226a005
25. Klebe G, Abraham U, Mietzner T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 37:4130–4146. doi:10.1021/jm00050a010
26. DRAGON, rel. 5.2 for Windows (2004) Talete srl, Milano, Italy
27. Kauffman GW, Jurs PC (2001) QSAR and $k$-nearest neighbor classification analysis of selective cyclooxygenase-2 inhibitors using topologically-based numerical descriptors. J Chem Inf Comput Sci 41:1553–1560. doi:10.1021/Ci010073h
28. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP (2003) Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 43:1947–1958. doi:10.1021/ci034160g
29. Gao H (2001) Application of BCUT metrics and genetic algorithm in binary QSAR analysis. J Chem Inf Comput Sci 41:402–407. doi:10.1021/Ci000306p
30. Roy K, Mitra I, Saha A (2009) Molecular shape analysis of antioxidant and squalene synthase inhibitory activities of aromatic tetrahydro-1,4-oxazine derivatives. Chem Biol Drug Des 74:507–516. doi:10.1111/j.1747-0285.2009.00888.x
31. Díaz-Uriarte R, Alvarezde Andrés S (2006) Gene selection and classification of microarray data using random forest. BMC bioinformatics 7:3. doi:10.1186/1471-2105-7-3
32. ISIS Draw 2.3, MDL Information Systems, Inc
33. Hemmateenejad B, Yazdani M (2009) QSPR models for half-wave reduction potential of steroids: a comparative study between feature selection and feature extraction from subsets of or entire set of descriptors. Anal Chim Acta 634:27–35. doi:10.1016/j.aca.2008.11.062
34. Golbraikh A, Tropsha A (2002) Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. J Comput Aided Mol Des 16:357–369. doi:10.1023/A:1020869118689
35. Breiman L (2001) Random forests. Mach Learn 45:5–32. doi:10.1023/A:1010933404324
36. randomForest: Breiman and Cutler's random forests for classification and regression. http://cran.r-project.org/web/packages/randomForest/index.html. Accessed on 27 May 2011
37. varSelRF: Variable selection using random forests. http://cran.r-project.org/web/packages/varSelRF/index.html. Accessed on 27 May 2011
38. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46:389–422. doi:10.1023/A:1012487302797
39. kernlab: Kernel-based machine learning lab. http://cran.r-project.org/web/packages/kernlab/index.html. Accessed on 27 May 2011
40. MASS: Main package of venables and Ripley's MASS. http://cran.r-project.org/web/packages/MASS/index.html. Accessed on 27 May 2011
41. Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451. doi:10.1016/0005-2795(75)90109-9
42. Ren Y, Liu H, Yao X, Liu M (2007) Prediction of ozone tropospheric degradation rate constants by projection pursuit regression. Anal Chim Acta 589:150–158. doi:10.1016/j.aca.2007.02.058
43. Eriksson L, Jaworska J, Worth A, Cronin M, McDowell R, Gramatica P (2003) Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 111:1361–1375. doi:10.1289/ehp.5758
44. caret: Classification and regression training. http://cran.r-project.org/web/packages/caret/index.html. Accessed on 27 May 2011
45. Ren Y, Liu H, Xue C, Yao X, Liu M, Fan B (2006) Classification study of skin sensitizers based on support vector machine and linear discriminant analysis. Anal Chim Acta 572:272–282. doi:10.1016/j.aca.2006.05.027
46. Burden F (1997) A chemically intuitive molecular index based on the eigenvalues of a modified adjacency matrix. Quant Struct Act Relatsh 16:309–314. doi:10.1002/qsar.19970160406
47. Mercader A, Duchowicz P, Fernández F, Castro E, Bennardi D, Autino J, Romanelli G (2008) QSAR prediction of inhibition of aldose reductase for flavonoids. Bioorg Med Chem 16:7470–7476. doi:10.1016/j.bmc.2008.06.004
48. Stanton DT (1999) Evaluation and use of BCUT descriptors in QSAR and QSPR studies. J Chem Inf Comput Sci 39:11–20. doi:10.1021/ci980102x
49. Pirard B, Pickett S (2000) Classification of kinase inhibitors using BCUT descriptors. J Chem Inf Comput Sci 40:1431–1440. doi:10.1021/ci000386x
50. Ford M, Pitt W, Whitley D (2004) Selecting compounds for focused screening using linear discriminant analysis and artificial neural networks. J Mol Graph Model 22:467–472. doi:10.1016/j.jmgm.2004.03.006