Hindawi Publishing Corporation
BioMed Research International
Volume 2013, Article ID 798743, 7 pages
http://dx.doi.org/10.1155/2013/798743

Research Article

Predicting the DPP-IV Inhibitory Activity pIC_{50} Based on Their Physicochemical Properties

Tianhong Gu,[1] Xiaoyan Yang,[2] Minjie Li,[2] Milin Wu,[2] Qiang Su,[1] Wencong Lu,[2] and Yuhui Zhang[3]

[1] *School of Materials Science and Engineering, Shanghai University, 149 Yan-Chang Road, Shanghai 200072, China*
[2] *Department of Chemistry, College of Sciences, Shanghai University, 99 Shang-Da Road, Shanghai 200444, China*
[3] *Department of Neurosurgery, Changhai Hospital, Second Military Medical University, 168 Chang-Hai Road, Shanghai 200433, China*

Correspondence should be addressed to Wencong Lu; wclu@shu.edu.cn and Yuhui Zhang; gong_chang2008@126.com

Received 29 March 2013; Revised 10 May 2013; Accepted 28 May 2013

Academic Editor: Yudong Cai

The second development program developed in this work was introduced to obtain physicochemical properties of DPP-IV inhibitors. Based on the computation of molecular descriptors, a two-stage feature selection method called mRMR-BFS (minimum redundancy maximum relevance-backward feature selection) was adopted. Then, the support vector regression (SVR) was used in the establishment of the model to map DPP-IV inhibitors to their corresponding inhibitory activity possible. The squared correlation coefficient for the training set of LOOCV and the test set are 0.815 and 0.884, respectively. An online server for predicting inhibitory activity pIC_{50} of the DPP-IV inhibitors as described in this paper has been given in the introduction.

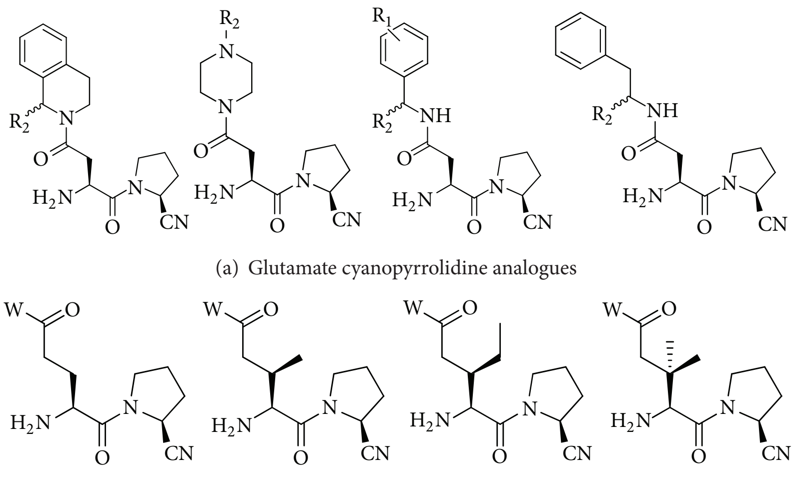
1. Introduction

The incretin hormones glucagon-like peptide-1 (GLP-1) and glucose-dependent insulinotropic polypeptide (GIP) are the endogenous peptides that stimulate glucose-dependent insulin secretion [1]. One of the important roles of dipeptidyl peptidase IV (DPP-IV) [2] is a rapid inactivation of the GLP-1 and GIP. Inhibition of DPP-4 increases the levels of endogenous intact circulating GLP-1 and GIP. Consequently, inhibitors of DPP-4 or gliptins have been recently regarded as a prospective approach for the treatment of type-2 diabetes mellitus.

In recent years, multiple small-molecule DPP-4 inhibitors have been reported [3, 4]. The development of a structurally diverse collection of DPP-4 inhibitors is a hot research [5–8]. Computational and various mathematical approaches have been widely employed in the quantitative structure-activity relationship (QSAR) analysis [9–13]. Using statistical methods, QSAR analyses were carried out on a dataset of 47 pyrrolidine analogs acting as DPP-IV inhibitors by Paliwal et al. [14]. Murugesan et al. used the comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) to analyze the structural requirements of a DPP-IV active site [15]. Gao et al. developed a novel 3D-QSAR model to assist rational design of novel, potent, and selective pyrrolopyrimidine DPP-4 inhibitors [16]. Moreover, several efforts by using computational and mathematical approaches have been made in investigating small molecules of DPP-4 inhibitors. In our previous studies [17], we have attempted to use the quantum chemistry method [18] to optimize a series of DPP-IV inhibitors, and a 2D-QSAR model has been built, which can predict the inhibitory activity of small molecule with satisfying results. However, it is time consuming to calculate the molecular descriptors adopted in 2D-QSAR model.

In view of this, here we will try to devise an effective method to correctly recognize the possible activity prediction of small molecules based on physical and chemical properties of the compounds.

According to the general development trend [19, 20] and the recent research progress [21–31], the following procedures should be considered to establish a powerful statistical predictor for a biological system: (i) a valid benchmark dataset is constructed or selected to train and test the predictor;



(a) Glutamate cyanopyrrolidine analogues

(b) (2S)-cyanopyrrolidines with glutamic acid derivatives

Figure 1: Molecular structure of cyanopyrrolidine amides as DPP-IV inhibitors.

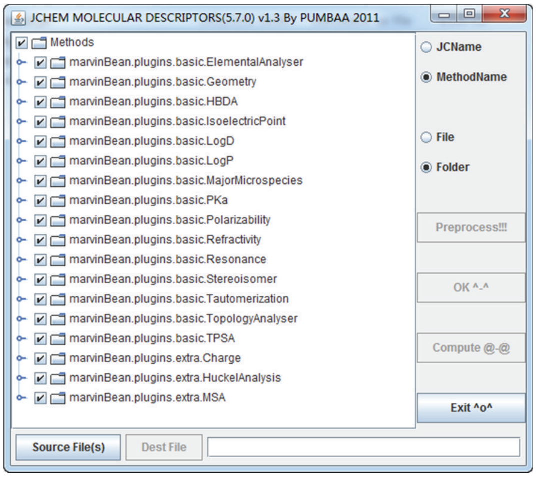
(ii) the samples are formulated with potent mathematical functions that are contributed to the prediction; (iii) a powerful algorithm is introduced or developed to operate the prediction; (iv) cross-validation tests are used to estimate the performance of the predictor; (v) a user-friendly online-server is established for the predictor that is accessible to the public. In this study, we attempt to describe how to deal with these steps for predicting the DPP-IV inhibitory activity pIC_{50} based on their physicochemical properties available via our program.

2. Materials and Methods

2.1. Data Preparation. The dataset used in the present work contains 48 pyrrolidine amides derivatives. In the current study, a diverse series of DPP-IV inhibitors with known IC_{50} values were collected from the papers [32, 33]. The detailed structures are documented in Supplementary Materials.(See Supplementary Material available at http://dx.doi.org/10.1155/2013/798743.) Figure 1 demonstrates the common structure of all of these analogues. All of the structures of compounds under investigation are based on the structure of Figure 1.

How to describe the molecules is an important problem in the establishment of the statistical model. In this study, the molecular descriptors for the 48 molecules were calculated by the second development software based on the calculator plugins, which is a product of ChemAxon [34]. ChemAxon is a company that provides chemical software development platforms and desktop applications for the biotechnology and pharmaceutical industries [35].

2.2. The Introduction of Procedure. Due to the use of Marvin Sketch graphic interface and JChem for Excel program, the calculations of small molecular descriptors are not very convenient. ChemAxon provides the calculation plugins of invoking function API, so our lab members have made a careful study and repeated experiments. The calculation results are compared with the ones of Gaussian 09 [18], JChem for Excel [34], HyperChem 7.5 [20, 36], and Dragon [37] programs calculation. By invoking the Calculator Plugins and using the Java language, we successfully developed a convenient and available customized batch calculation program (second development software) for the small molecular descriptors.


JCHEM MOLECULAR DESCRIPTORS(5.7.0) v1.3 By PUMBAA 2011

☑ Methods
- ☑ marvinBean.plugins.basic.ElementalAnalyser
- ☑ marvinBean.plugins.basic.Geometry
- ☑ marvinBean.plugins.basic.HBDA
- ☑ marvinBean.plugins.basic.IsoelectricPoint
- ☑ marvinBean.plugins.basic.LogD
- ☑ marvinBean.plugins.basic.LogP
- ☑ marvinBean.plugins.basic.MajorMicrospecies
- ☑ marvinBean.plugins.basic.PKa
- ☑ marvinBean.plugins.basic.Polarizability
- ☑ marvinBean.plugins.basic.Refractivity
- ☑ marvinBean.plugins.basic.Resonance
- ☑ marvinBean.plugins.basic.Stereoisomer
- ☑ marvinBean.plugins.basic.Tautomerization
- ☑ marvinBean.plugins.basic.TopologyAnalyser
- ☑ marvinBean.plugins.basic.TPSA
- ☑ marvinBean.plugins.extra.Charge
- ☑ marvinBean.plugins.extra.HuckelAnalysis
- ☑ marvinBean.plugins.extra.MSA

○ JCName
● MethodName

○ File
● Folder

Figure 2: The program interface for the computation of molecular descriptors.

This program contains a selection of tree box; the user can choose the visual way to the calculation of molecular descriptors (as shown in Figure 2, command-line version does not provide molecular descriptor selection). The molecule structures are constructed from Gauss View 5.0 package [38, 39] as MOL-format file. Command-line version of the program is operated commonly in Linux server, through the similar execution command as follows:

```
java-jar JChemCmd.jar Molecules Pathway Result.csv Method.xml
```

2.3. Model Validation

2.3.1. Dataset. The full dataset included training set (36 compounds) and test set (12 compounds). The whole samples were ranked by activity and were extracted every fourth sample for the generation of the test set.

2.3.2. Leave-One-Out Cross-Validation (LOOCV) and Predictive Validation. In this study, Leave-one-out cross-validation (LOOCV) [40, 41] was used to investigate the prediction quality of training set. In the cross-validation, each sample is used to test the model that is established by all of the other samples at the same time.

2.3.3. Fitting and Predictive Performances of Models. The fitting and predictive performances of model were measured by the squared correlation coefficient (q^2) and root mean square error (RMSE) for both the training set and the external test set. Here the performances of models can be estimated by q^2 and RMSE defined as follows, respectively:

$$q^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y_{\text{mean}})^2}, \qquad \text{RMSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N}}, \tag{1}$$

where y_i and $\hat{y}_i$ are the actual and predicted pIC_{50} values of i sample, respectively, and y_{mean} is the average pIC_{50} value of the entire samples. N is the numbers of the training set.

2.4. Methods. For the sake of the redundancy of some features, the selection of descriptors before establishing a suitable model is necessary. The selection of descriptors plays an important role in construction for the actual model. In this work, mRMR-BFS method (minimum redundancy maximum relevance-backward feature selection) [42, 43] was used for the selection of molecular descriptors. The support vector regression (SVR) model was established based on the feature selection results.

2.4.1. mRMR-BFS Algorithm. The mRMR (minimum-redundancy maximum-relevance) algorithm was introduced by Ding and Ping [44], which was used usually for feature selection. It sorts a feature based on score function which is maximum relevance to target and minimum redundancy to the already selected features. The score function is defined as follows:

$$\text{score}_j = I\left(f_j, c\right) - \frac{1}{m}\sum_{i=1}^{m} I\left(f_i, f_j\right), \tag{2}$$

where $f_j \subset S_n$, $f_i \subset S_m$, $S_m = S - S_n$, and S_m, S_n, and S are the feature sets. m and n are the feature numbers. The mutual information $I(x, y)$ is as follows:

$$I(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)} dx\, dy, \tag{3}$$

where $p(x, y)$, $p(x)$, and $p(y)$ are the probabilistic density functions.

More details about mRMR algorithm can be found in [44, 45].

To gain an even better performance of predictor and feature selection, backward feature selection (BFS) based on the result of mRMR is also used in this study. The most important 50 variables were obtained from the mRMR procedure. We initialize the BFS-selected feature set S_s with all features in S:

$$S_s = \left\{f'_1, f'_2, \ldots, f'_k\right\} \quad (1 \le k \le 50). \tag{4}$$

With the mRMR-selected feature subset S_s, the next BFS-selected feature set can be gained by the following steps.

(1) Suppose that the candidate feature set is $S_C = S_S - f_k$. Then an SVR model based on each S_C is established and evaluated by LOOCV method.

(2) The feature f which gets the lowest RMSE is selected when removed from S_S.

(3) The feature f is removed from S_s forming the next BFS-selected feature set.

2.4.2. SVM (Support Vector Machine). Vapnik and his coworkers developed the SVM algorithm, which is a supervised machine-learning method that is used for classification and regression analysis. Owing to embodying the structural risk minimization principle, the SVM exhibits a better whole performance. The SVM is suitable for the problems which are involved in the small sample set. In this work, SVM was applied to regression. The details of the algorithm can be found in reference [46]. The algorithm was performed by using the software package Weka 3.6.7 [47, 48].

3. Results and Discussion

3.1. Selection of Features. Firstly, mRMR method was applied to rank the total 75 features according to their mRMR scores. Secondly, we used the backward feature selection (BFS) algorithm based on SVR to search for the feature combinations. As different machine learning methods will lead to different results, several robust machine learning methods like the nearest-neighbor algorithm (NNA), support vector machine (SVM based on RBF kernel function), and Adaboost were employed to find an optimal feature subset with leave-one-out cross-validation, respectively. As a result, we adopted the SVM as the prediction engine based on the LOOCV in this study.

Table 1 lists an optimal subset attained by employing the above two-stage feature selection method, mRMR-BFS. The six features in optimal subset can be clustered into three categories (based on the category of Calculator Plugins [49]): elemental analysis, geometry, topology, and others. The geometry and topology factor are more important in this work. The geometry and topology factor are related to the size of the molecule as it indicates that the size of cyanopyrrolidine amides derivatives plays a main role in the inhibitory activity.

Table 1: Symbols for molecular descriptors involved in the model.

Molecular descriptor	Type	Description
OComposition	Elemental analysis functions	O Composition
MaximalProjectionArea	Geometry	Calculates the maximal projection area
MinimalProjectionArea	Geometry	Calculates the minimal projection area
BasicpKa	pKa	Constant denoting basic pKa
RingBondCount	Topology	Ring bond count
AliphaticRingCount	Topology	Aliphatic ring count

3.2. Results of Computation. In this work, q^2_{train}, $q^2_{\text{train-CV}}$, and q^2_{test} were used to present the squared correlation coefficients for the training set, cross-validation set, and external test set, respectively. Also $\text{RMSE}_{\text{train}}$, $\text{RMSE}_{\text{train-CV}}$, and $\text{RMSE}_{\text{test}}$ were adopted to present the root mean square errors for the training set, cross-validation set, and external test set, respectively.

The final model was built by the SVR based on the Gaussian kernel function (RBF) with the parameters C, ε, and γ that are 2.0, 0.05, and 1.0, respectively. The Gaussian kernel function (RBF) is given as follows:

$$K\left(x, x_i\right) = \exp\left(-\gamma\left\|x - x_i\right\|^2\right). \tag{5}$$

The model based on the above parameters with original data is given as follows:

$$\text{pIC}_{50} = 2.10 * \left[\sum_{i\subset \text{SV}} \beta_i \exp\left(-\left\|x - x_i\right\|^2\right) + 0.207\right] + 6.60, \tag{6}$$

where β_i is the Lagrange coefficient of support vectors.

The experimental versus predicted pIC_{50} values based on the SVR model for the training set and test set are shown in Figure 3. As a result, the values of q^2_{train}, $q^2_{\text{train-CV}}$, and q^2_{test} were 0.953, 0.815, and 0.884, respectively. And the values of $\text{RMSE}_{\text{train}}$, $\text{RMSE}_{\text{train-CV}}$, and $\text{RMSE}_{\text{test}}$ were 0.123, 0.247, and 0.193, respectively. Figure 3 illustrates that the regression straight line is appropriate not only for the fitting pIC_{50} values of the training set but also for the predicted pIC_{50} values of the external test set. Table 2 shows the experimental and the calculated values over the training set and the test set. From Figure 3 and Table 2, it can be concluded that the predicted values are in good agreement with the experimental ones. Figure 4 illustrates the dispersion plot of the residuals for the training and test sets. The predicted values are randomly dispersed around the zero-value line in Figure 4. It means that the model is appropriate for the data.

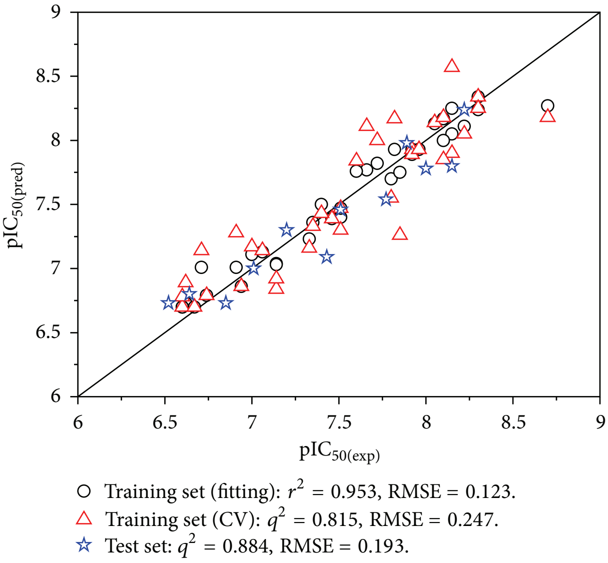


Figure 3: Predicted versus experimental pIC_{50} for the training (circles for fitting and triangle for CV, respectively) and test (stars) sets.

3.3. Analysis of the New Method. The secondary development program developed in this work was used to establish a robust model with $q^2_{\text{train}} = 0.953$, $q^2_{\text{train-CV}} = 0.815$, and $q^2_{\text{test}} = 0.884$, respectively. In order to validate the generalization and reliability of the descriptors obtained by using our secondary development program, the same training and test sets were also constructed and optimized at the HF/6 – 31G* level of theory with the Gaussian program; 1262 descriptors were computed by HyperChem 7.5 program [20], JChem for Excel package [34], and the Dragon program [37]. And a robust and reliable model was obtained with $q^2_{\text{train}} = 0.969$, $q^2_{\text{train-CV}} = 0.868$, and $q^2_{\text{test}} = 0.891$, respectively. The statistical comparisons were summarized in Table 3.

It is indicated that it takes less than 30 minutes for a molecule from the structure optimization to the computation of descriptors by using the second development program. In contrast, more than 36 hours were taken based on the Gaussian program. These results show that the computing speeds are greatly improved by using the secondary development program, while the statistical parameters of models are as good as those obtained with the Gaussian method. Therefore, the second development program is very helpful not only for saving the time of descriptor computation but also for providing the effective QSPR models online available in the future.

In a benchmark test, the support vector regression (SVR) was contrasted with the multiple linear regression (MLR) and the back propagation-artificial neural network (BP-ANN) on the $q^2_{\text{train-CV}}$. The statistical comparisons were shown in Table 4. From Table 4, SVR has a better generalization ability in our work.

Table 2: Experimental and predicted pIC_{50} for the training and test sets.

No.	$pIC_{50(exp)}$	$pIC_{50(Pred)}$	$pIC_{50(LOOCV)}$
1	7.00	7.11	7.17
2^T	7.20	7.30	—
3	7.35	7.36	7.33
4	7.33	7.23	7.16
5^T	7.01	7.00	—
6	7.14	7.04	6.92
7	7.14	7.03	6.84
8	6.71	7.01	7.14
9^T	6.64	6.80	—
10	7.06	7.13	7.14
11	6.91	7.01	7.28
12	6.62	6.73	6.89
13	6.60	6.70	6.78
14^T	6.85	6.73	—
15	6.67	6.70	6.70
16	6.60	6.70	6.70
17	6.94	6.86	6.86
18	6.74	6.79	6.79
19^T	6.52	6.73	—
20	8.70	8.27	8.18
21	8.30	8.34	8.34
22	7.46	7.39	7.39
23	7.40	7.50	7.43
24^T	8.22	8.24	—
25	8.15	8.25	8.57
26	8.30	8.24	8.25
27	8.05	8.13	8.14
28	8.22	8.11	8.05
29	8.15	8.05	7.90
30^T	8.00	7.78	—
31	7.66	7.77	8.11
32^T	8.15	7.80	—
33	7.82	7.93	8.17
34^T	7.77	7.54	—
35^T	7.51	7.46	—
36	8.10	8.00	7.85
37	7.72	7.82	8.00
38^T	7.43	7.09	—
39	7.96	7.93	7.93
40	8.10	8.17	8.18
41	7.51	7.40	7.30
42	7.92	7.89	7.89
43	7.51	7.47	7.47
44	7.92	7.93	7.93
45	7.80	7.70	7.55
46	7.60	7.76	7.84
47	7.85	7.75	7.26
48^T	7.89	7.98	—

^{T}indicates the test samples.

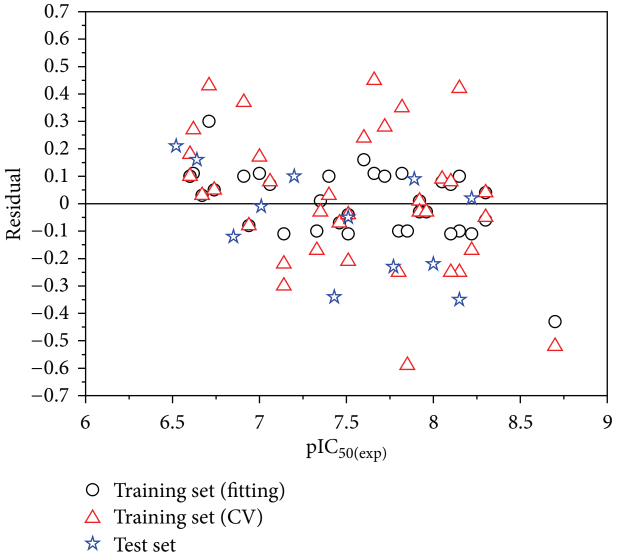


Figure 4: Dispersion plot of the residuals for the training and test sets.

Table 3: Comparative statistical parameters obtained by the secondary development program and the Gaussian program concerning the same compounds.

Program	q^2_{train}	$q^2_{train-CV}$	q^2_{test}
The secondary development program developed in this work	0.953	0.815	0.884
Gaussian, HyperChem 7.5, JChem for Excel package, Dragon	0.969	0.868	0.891

Table 4: $q^2_{train-CV}$ of different methods.

Method	SVR	BP-ANN	MLR
$q^2_{train-CV}$	0.815	0.761	0.721

3.4. The Online Web Server. Since user-friendly and publicly accessible online servers represent the trend for developing more useful models or predictors, we established a web server for predicting the DPP-IV inhibitory activity pIC_{50} at http://chemdata.shu.edu.cn:8080/QSARPrediction/index.jsp.

The web server allows users to upload the MOL-format file of a molecule, and the server will return the result of prediction according to the model of our mRMR-BFS-SVR method. In this course, the Calculator Plugins [49] of ChemAxon was invoked in the background program. The server developed has the most outstanding characteristic that users need to do nothing except for uploading the file of the unknown small molecule. Then they can get the predicted result after waiting for some time. It is a remarkable advance compared to our previous work [17, 20, 36].

4. Conclusions

In this paper, the secondary development program was proposed to bring an efficient and fast calculation means for molecular descriptors. The mRMR-BFS was adopted in the procedure of feature selection. The SVR was used to construct the model to map DPP-IV inhibitors to their corresponding inhibitory activity. The q^2_{train}, $q^2_{\text{train-CV}}$, and q^2_{test} of the model are 0.953, 0.815, and 0.884, respectively. These results are as good as those obtained with the Gaussian method. The web server, which provides a quick approach to predict the DPP-IV inhibitory activities pIC_{50} of unknown small molecules based on their MOL-format files, was established by using our secondary development program at http://chemdata.shu.edu.cn:8080/QSARPrediction/index.jsp. A user-friendly and rapid approach whose accuracy is approximate with the Gaussian method is proposed in this work.

Acknowledgments

This study was supported by the National Science Foundation of China (20973108, 20902056), the Shanghai Education Committee Project (11ZZ83), and the Leading Academic Discipline Project of Shanghai Municipal Education Commission, China (J50101). The authors also acknowledge ChemAxon for their excellent products.

References

[1] M. H. Kim and M. K. Lee, "The incretins and pancreatic beta-cells: use of glucagon-like peptide-1 and glucose-dependent insulinotropic polypeptide to cure type 2 diabetes mellitus," *Korean Diabetes Journal*, vol. 34, no. 1, pp. 2–9, 2010.

[2] A. Sarashina, S. Sesoko, M. Nakashima et al., "Linagliptin, a dipeptidyl peptidase-4 inhibitor in development for the treatment of type 2 diabetes mellitus: a phase I, randomized, double-blind, placebo-controlled trial of single and multiple escalating doses in healthy adult male japanese subjects," *Clinical Therapeutics*, vol. 32, no. 6, pp. 1188–1204, 2010.

[3] K. Augustyns, P. Van der Veken, and A. Haemers, "Inhibitors of proline-specific dipeptidyl peptidases: DPP IV inhibitors as a novel approach for the treatment of type 2 diabetes," *Expert Opinion on Therapeutic Patents*, vol. 15, no. 10, pp. 1387–1407, 2005.

[4] A. E. Weber, "Dipeptidyl peptidase IV inhibitors for the treatment of diabetes," *Journal of Medicinal Chemistry*, vol. 47, no. 17, pp. 4135–4141, 2004.

[5] S. D. Edmondson, A. Mastracchio, R. J. Mathvink et al., "(2S, 3S)-3-amino-4-(3,3-difluoropyrrolidin-1-yl)-N,N-dimethyl-4-oxo-2-(4-[1,2,4]triazolo[1,5-a]-pyridin-6-ylphenyl)butanamide: a selective α-amino amide dipeptidyl peptidase IV inhibitor for the treatment of type 2 diabetes," *Journal of Medicinal Chemistry*, vol. 49, no. 12, pp. 3614–3627, 2006.

[6] J. L. Duffy, B. A. Kirk, L. Wang et al., "4-Aminophenylalanine and 4-aminocyclohexylalanine derivatives as potent, selective, and orally bioavailable inhibitors of dipeptidyl peptidase IV," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 10, pp. 2879–2885, 2007.

[7] J. Xu, L. Wei, R. J. Mathvink et al., "Discovery of potent, selective, and orally bioavailable oxadiazole-based dipeptidyl peptidase IV inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 16, no. 20, pp. 5373–5377, 2006.

[8] J. Xu, L. Wei, R. Mathvink et al., "Discovery of potent, selective, and orally bioavailable pyridone-based dipeptidyl peptidase-4 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 16, no. 5, pp. 1346–1349, 2006.

[9] T. S. Garcia and K. M. Honório, "Two-dimensional quantitative structure-activity relationship studies on bioactive ligands of peroxisome proliferator-activated receptor δ," *Journal of the Brazilian Chemical Society*, vol. 22, no. 1, pp. 65–72, 2011.

[10] G. C. García, I. Luque Ruiz, and M. Á. Gómez-Nieto, "Analysis and study of molecule data sets using snowflake diagrams of weighted maximum common subgraph trees," *Journal of Chemical Information and Modeling*, vol. 51, no. 6, pp. 1216–1232, 2011.

[11] D. Jana, A. K. Halder, N. Adhikari, M. K. Maiti, C. Mondal, and T. Jha, "Chemometric modeling and pharmacophore mapping in coronary heart disease: 2-arylbenzoxazoles as cholesteryl ester transfer protein inhibitors," *MedChemComm*, vol. 2, no. 9, pp. 840–852, 2011.

[12] V. Kovalishyn, V. Tanchuk, L. Charochkina, I. Semenuta, and V. Prokopenko, "Predictive QSAR modeling of phosphodiesterase 4 inhibitors," *Journal of Molecular Graphics and Modelling*, vol. 32, pp. 32–38, 2012.

[13] B. Niu, Q. Su, X. C. Yuan, W. Lu, and J. Ding, "QSAR study on 5-lipoxygenase inhibitors based on support vector machine," *Medicinal Chemistry*, vol. 8, no. 6, pp. 1108–1116, 2012.

[14] S. Paliwal, D. Seth, D. Yadav, R. Yadav, and S. Paliwal, "Development of a robust QSAR model to predict the affinity of pyrrolidine analogs for dipeptidyl peptidase IV (DPP-IV)," *Journal of Enzyme Inhibition and Medicinal Chemistry*, vol. 26, no. 1, pp. 129–140, 2011.

[15] V. Murugesan, N. Sethi, Y. S. Prabhakar, and S. B. Katti, "CoMFA and CoMSIA of diverse pyrrolidine analogues as dipeptidyl peptidase IV inhibitors: active site requirements," *Molecular Diversity*, vol. 15, no. 2, pp. 457–466, 2011.

[16] Y. D. Gao, D. Feng, R. P. Sheridan et al., "Modeling assisted rational design of novel, potent, and selective pyrrolopyrimidine DPP-4 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 17, no. 14, pp. 3877–3879, 2007.

[17] X. Y. Yang, M. J. Li, Q. Su, M. Wu, T. Gu, and W. Lu, "QSAR studies on pyrrolidine amides derivatives as DPP-IV inhibitors for type 2 diabetes," *Medicinal Chemistry Research*, 2013.

[18] S. Peng, Z. Jian-Wei, Z. Peng, and X. Lin, "QSPR modeling of bioconcentration factor of nonionic compounds using Gaussian processes and theoretical descriptors derived from electrostatic potentials on molecular surface," *Chemosphere*, vol. 83, no. 8, pp. 1045–1052, 2011.

[19] T. Gu, W. Lu, X. Bao, and N. Chen, "Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors," *Solid State Sciences*, vol. 8, no. 2, pp. 129–136, 2006.

[20] J. Zhu, W. Lu, L. Liu, T. Gu, and B. Niu, "Classification of Src kinase inhibitors based on support vector machine," *QSAR and Combinatorial Science*, vol. 28, no. 6-7, pp. 719–727, 2009.

[21] V. Kovalishyn, J. Aires-de-Sousa, C. Ventura, R. Elvas Leitão, and F. Martins, "QSAR modeling of antitubercular activity of diverse organic compounds," *Chemometrics and Intelligent Laboratory Systems*, vol. 107, no. 1, pp. 69–74, 2011.

[22] L. Xing, R. Goulet, and K. Johnson, "Statistical analysis and compound selection of combinatorial libraries for soluble epoxide hydrolase," *Journal of Chemical Information and Modeling*, vol. 51, no. 7, pp. 1582–1592, 2011.

[23] S. Kar, O. Deeb, and K. Roy, "Development of classification and regression based QSAR models to predict rodent carcinogenic potency using oral slope factor," *Ecotoxicology and Environmental Safety*, vol. 82, pp. 85–95, 2012.

[24] B. Niu, X. C. Yuan, P. Roeper et al., "HIV-1 protease cleavage site prediction based on two-stage feature selection method," *Protein and Peptide Letters*, vol. 20, no. 3, pp. 290–298, 2013.

[25] B. Niu, Y. D. Cai, W. C. Lu, G. Z. Li, and K. C. Chou, "Predicting protein structural class with AdaBoost Learner," *Protein and Peptide Letters*, vol. 13, no. 5, pp. 489–492, 2006.

[26] B. Niu, Y. H. Jin, K. Y. Feng et al., "Predicting membrane protein types with bagging learner," *Protein and Peptide Letters*, vol. 15, no. 6, pp. 590–594, 2008.

[27] B. Niu, Y. H. Jin, K. Y. Feng, W. C. Lu, Y. D. Cai, and G. Z. Li, "Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins," *Molecular Diversity*, vol. 12, no. 1, pp. 41–45, 2008.

[28] B. Niu, Y. Jin, L. Lu et al., "Prediction of interaction between small molecule and enzyme using AdaBoost," *Molecular Diversity*, vol. 13, no. 3, pp. 313–320, 2009.

[29] B. Niu, Y. Jin, W. Lu, and G. Li, "Predicting toxic action mechanisms of phenols using AdaBoost Learner," *Chemometrics and Intelligent Laboratory Systems*, vol. 96, no. 1, pp. 43–48, 2009.

[30] B. Niu, L. Lu, L. Liu et al., "HIV-1 protease cleavage site prediction based on amino acid property," *Journal of Computational Chemistry*, vol. 30, no. 1, pp. 33–39, 2009.

[31] Q. Su, W. C. Lu, B. Niu, X. Liu, and T. H. Gu, "Classification of the toxicity of some organic compounds to tadpoles (Rana Temporaria) through integrating multiple classifiers," *Molecular Informatics*, vol. 30, no. 8, pp. 672–675, 2011.

[32] I. L. Lu, S. J. Lee, H. Tsu et al., "Glutamic acid analogues as potent dipeptidyl peptidase IV and 8 inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 15, no. 13, pp. 3271–3275, 2005.

[33] T. Y. Tsai, T. Hsu, C. T. Chen et al., "Rational design and synthesis of potent and long-lasting glutamic acid-based dipeptidyl peptidase IV inhibitors," *Bioorganic and Medicinal Chemistry Letters*, vol. 19, no. 7, pp. 1908–1912, 2009.

[34] L. Weber, "JChem base—chemAxon," *Chemistry World*, vol. 5, no. 10, pp. 65–66, 2008.

[35] 2013, http://www.chemaxon.com/.

[36] S. S. Yang, W. C. Lu, T. H. Gu, L. M. Yan, and G. Z. Li, "QSPR study of n-octanol/water partition coefficient of some aromatic compounds using support vector regression," *QSAR and Combinatorial Science*, vol. 28, no. 2, pp. 175–182, 2009.

[37] T. Todeschini, "Dragon 5.0: software for molecular descriptors," in *Milano Chemometrics and QSAR Research Group*, University of Milano-Bicocca, Milan, Italy, 2004.

[38] V. Mukherjee, K. Singh, N. P. Singh, and R. A. Yadav, "Quantum chemical determination of molecular geometries and interpretation of FTIR and Raman spectra for 2,4,5- and 3,4,5-trifluoro-benzonitriles," *Spectrochimica Acta A*, vol. 71, no. 4, pp. 1571–1580, 2008.

[39] Y. Chen, Z. Yi, S. J. Chen, J. S. Luo, Y. G. Yi, and Y. J. Tang, "Study of density functional theory for surface-enhanced raman spectra of p-aminothiophenol," *Spectroscopy and Spectral Analysis*, vol. 31, no. 11, pp. 2952–2955, 2011.

[40] T. Zhang, "A leave-one-out cross validation bound for kernel methods with applications in learning," *Computational Learning Theory Proceedings*, vol. 2111, pp. 427–443, 2001.

[41] J. Yuan, Y. M. Li, C. L. Liu, and X. F. Zha, "Leave-one-out cross-validation based model selection for manifold regularization," in *Advances in Neural Networks*, vol. 6063 of *Lecture Notes in Computer Science*, pp. 457–464, 2010.

[42] M. Kompany-Zareh, "An improved QSPR study of the toxicity of aliphatic carboxylic acids using genetic algorithm," *Medicinal Chemistry Research*, vol. 18, no. 2, pp. 143–157, 2009.

[43] M. Goodarzi, B. Dejaegher, and Y. Vander Heyden, "Feature selection methods in QSAR studies," *Journal of Aoac International*, vol. 95, no. 3, pp. 636–651, 2012.

[44] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," in *Proceedings of the IEEE Bioinformatics Conference*, pp. 185–205, August 2003.

[45] Z. He, J. Zhang, X. H. Shi et al., "Predicting drug-target interaction networks based on functional groups and biological features," *PLoS ONE*, vol. 5, no. 3, Article ID e9603, 2010.

[46] B. Üstün, W. J. Melssen, and L. M. C. Buydens, "Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel," *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 1, pp. 29–40, 2006.

[47] E. Frank, M. Hall, L. Trigg, G. Holmes, and I. H. Witten, "Data mining in bioinformatics using Weka," *Bioinformatics*, vol. 20, no. 15, pp. 2479–2481, 2004.

[48] L. Chen, L. Lu, K. Feng et al., "Multiple classifier integration for the prediction of protein structural classes," *Journal of Computational Chemistry*, vol. 30, no. 14, pp. 2248–2254, 2009.

[49] 2013, http://www.chemaxon.com/products/calculator-plugins/.