Construction of QSAR model between the ligand and γ-Aminobutyric acid type A receptor using support vector regression algorithm

Shu Cheng
School of Science, Jiangnan University
Wuxi, China
e-mail: 18351821170@163.com

Yanrui Ding
School of Science, Jiangnan University
Wuxi, China
e-mail: yr_ding@jiangnan.edu.cn

***Abstract*—Quantitative structure-activity relationship (QSAR) plays an important role in the prediction of biological activity based on machine learning. According to the characteristics of the binding interface between ligands and the γ -Aminobutyric acid type A (GABA$_A$) receptor, we used random forest feature selection and support vector regression (SVR) to establish three QSAR models. The best QSAR model features include docking ligand molecular descriptors and ligand-receptor interactions. We also used Leave-One-Out-Cross-Validation (LOOCV) to select the appropriate value C = 2, g = 0.0221. The result of cross validation (Q^2_{LOO}) is 0.8225, R^2 of test set is 0.8326, and MSE is 0.0910. In addition, we found that BELm2, BELe2, Mor08v, Mor29m, refRMS and intermol _ energy are key features, which helps to build QSAR model more accurately.**

Keywords-QSAR; GABA$_A$; docking; SVR; LOOCV

I. Introduction

Quantitative structure-activity relationship (QSAR) refers to the relationship between activity and structural characteristics of compounds [1]. It uses the method of mathematical statistics for regression analysis and mathematical model to express and generalize the regular pattern, so as to analyze the action mode of drugs and predict the biological activity of compounds.

There are many QSAR studies of drugs that interact with GABA$_A$ receptor based on machine learning. D.J. Maddalena *et al*. examined the QSAR between substituent constants and random noise and their binding affinities (log IC50) for benzodiazepine GABA$_A$ receptor preparations by multilinear regression (MLR) and back-propagation ANNs [2]. M. Goodarzi *et al*. introduced a new HGA-SVR hybrid method in QSAR field for the first time, and compared its statistical performance with partial least square, back-propagation artificial neural network and support vector machine. It is proved that HGA-SVR method is the best method for predicting the activity of the flavone derivatives binding to GABA$_A$ receptor [3]. A. M. Bianucci *et al*. proposed a new method of recursive neural network based on processing domain for QSAR analysis. By using this model, they can express and process the structure of the compound in the form of a marker tree, which can be used to predict the affinity of benzodiazepine / GABA$_A$ receptor [4].

According to the three-dimensional structure of GABA$_A$ receptor and the precise binding site [5], the molecular docking between candidate ligands and the GABA$_A$ receptor were carried out. Here, after determining the optimal conformation, the binding properties of the ligands to the GABA$_A$ receptor were calculated, and the QSAR model was established by using the support vector regression (SVR) algorithm. Finally, we determined the structure-activity relationship and the interaction mode between ligands and the GABA$_A$ receptor. In the construction of QSAR based on machine learning, we not only consider the molecular characteristics of the docking ligand itself, but also pay attention to the interaction characteristics of the binding interface between the ligand and the receptor.

II. Construction of data sets

A. Data preprocessing

First, we select 76 ligand molecules with known activity values (IC50) from ChEMBL and BindingDB database and convert the value of IC50 to pIC50 (-LogIC50). Then, we obtain the GABA$_A$ receptor from PDB database. Candidate ligands and the receptor protein have pretreated uniformly.

B. Molecular docking

We use Autodock4.2 [6] to batch dock 76 ligands with the GABA$_A$ receptor. After docking, the optimal conformation of the ligands and the characteristics of the interaction between ligands and the GABA$_A$ receptor are stored in the data set.

III. Calculation of features

First, for the docking ligands, the molecular descriptors are calculated by E-Dragon [7]. 20 kinds of descriptors are calculated including Constitutional descriptors, Information indices, Edge adjacency indices, Topological charge indices, Randic molecular profiles, RDF descriptors, Walk and path counts, WHIM descriptors, Charge descriptors, Functional group counts, Topological descriptors, 2D autocorrelations, BCUT descriptors, Connectivity indices, Eigenvalue-based indices, Geometrical descriptors, 3D-MoRSE descriptors, GETAWAY descriptors, Atom-centred fragments and Molecular properties. After pre-screening, 1286 molecular descriptors are left. In addition, for the characteristics of interaction between ligands and the GABA$_A$ receptor, we obtain 10 interaction characteristics as descriptors, including inhib _ constant (uM), binding _ energy, ligand _ efficiency, intermol _ energy, electrostatic _ energy, total _ internal, torsional _ energy, unbound _ energy, vdw _ hb _ desolv _ energy and refRMS.

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00060

IV. Feature selection

Because there are redundant and noisy features in 1286 calculated molecular descriptors that are not related to the QSAR model, we use Mean Decrease Impurity in random forest to select features. In this method, the features are sorted based on the importance score, and the optimal condition can be determined by using the impurity. Finally, 53 molecular descriptors are selected as features input of QSAR model. For interaction characteristics, the original 10 characteristics we obtained are retained as features input of QSAR model.

V. Construction process of SVR

epsilon-SVR and RBF in LIBSVM are used to build QSAR model and predict the pIC50 of ligand binding to the $GABA_A$ receptor.

A. Model 1

Of the 76 ligands described by 53 descriptors, 61 are randomly selected as training set and the rest as testing set. For training set $S = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$ with N(N = 61) ligands, assuming x is the set of training ligands, x_i is the i-th ligand ($i = 1,2,\cdots,\mathrm{N}$); y is the pIC50 experimental value of the ligand, y_i is the pIC50 experimental value of the i-th ligand. Put x_i mapping to a high dimensional feature space $\varphi(x_i)$, the modeling process is as follows:

The optimization goal of SVR is to find a regression plane, and make all molecular descriptors closest to this plane. The hyperplane is shown in (1).

$$f(x) = \omega^T x + b \tag{1}$$

ω is the normal vector perpendicular to the hyperplane and b is the deviation. We can calculate ω based on Lagrange multiplier. The Φ function maps x_i and x to a higher dimensional space, as shown in (2) and (3).

$$\omega = \sum_{i=1}^{N} \lambda_i y_i x_i \tag{2}$$

$$f(x) = \sum_{i=1}^{N} \lambda_i y_i \Phi^T(x_i)\Phi(x) + b \tag{3}$$

λ_i is Lagrange multiplier, and ω can't be expressed in high-dimensional feature space, so RBF kernel function $k(x_i, x)$ is introduced to replace $\Phi^T(x_i)\Phi(x)$, as shown in (4) and (5).

$$f(x) = \sum_{i=1}^{N} \lambda_i y_i k(x_i, x) + b \tag{4}$$

$$k(x_i, x) = exp(-gamma * |x1 - x|^2) \tag{5}$$

$gamma$ is the parameter g in kernel function, which has an important influence on the training of the model. Another main parameter to be adjusted in this study is penalty coefficient C, which affects the smoothness of regression plane. In order to find the optimal combination of C and g, we use Leave-One-Out-Cross-Validation(LOOCV) to select C and g,, and then apply them to the testing set to complete the prediction of the candidate ligand predicted pIC50.

The QSAR model is Model 1, and the remaining 15 testing set ligands are used to predict pIC50. The accuracy of the model is checked according to the deviation between the experimental pIC50 and the predicted pIC50.

B. Model 2

For 76 ligands described by 10 docking interaction features, the original 53 molecular descriptors are replaced by 10 docking interaction features using the above steps of SVR algorithm, and the constructed model is called Model 2.

C. Model 3

76 ligands are described by 63 descriptors, including 53 descriptors in Model 1 and 10 features in Model 2. Based on these 63 descriptors, the QSAR model constructed by SVR algorithm is called Model 3.

VI. Evaluation criterion

We use the LOOCV evaluation criterion (Q^2_{LOO}) to evaluate the prediction ability of QSAR model. In addition, the reliability of QSAR model is evaluated by the coefficient of determination (R^2) and mean square error (MSE).

$$Q^2{}_{LOO} = 1 - \frac{\sum_{i=1}^{N}(Y_{act} - Y_{LOO(pre)})^2}{\sum_{i=1}^{N}(Y_{act} - Y_{avg})^2} \tag{6}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_{act} - Y_{pre})^2 \tag{7}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_{act} - Y_{pre})^2}{\sum_{i=1}^{N}(Y_{act} - Y_{avg})^2} \tag{8}$$

In the above equation, N represents the total number of ligands, Y_{act} represents the experimental pIC50, Y_{pre} represents the predicted pIC50, Y_{avg} is the average of the experimental pIC50. $Y_{LOO(pre)}$ is the predicted value of cross validation. If Q^2_{LOO} is greater than 0.5, it indicates that the model has credibility, and the closer it is to 1, the stronger the prediction ability of the model.MSE is used to evaluate the proximity between the experimental pIC50 and the predicted value. The smaller the MSE is, the smaller the error of the prediction model of pIC50 is, indicating the higher the reliability of the model. R^2 indicates the fitting effect. The closer the distance 1 is, the stronger the ability of the square independent variable to explain the dependent variable is, the better the fitting of the model effect is.

VII. Results and Analysis

A. Results of LOOCV

The Q^2_{LOO} of the three models in Fig. 1 are 0.7966, 0.8074 and 0.8225 respectively, all of which are greater than 0.5. In addition, C and g selected by LOOCV in the three models are: C = 1.4142, 8, 2; g = 0.0313, 0.0321, 0.0221. From Figure 1, the predicted pIC50 in Model 3 is the most consistent with the experimental pIC50, so Model 3 has the highest internal prediction ability in the three models and we choose C = 2, g = 0.0221 finally.

B. Comparison of QSAR models

Fig. 2 shows the prediction effect of pIC50 value of ligands in the testing set. R^2 of the three models are 0.8254, 0.7937 and 0.8326, which show that the three models are reliable for the pIC50 of the ligands. By combining the two types of descriptors (Model 3), the ability to predict pIC50 is better than the other two models. In addition, the MSE of the three models are 0.1038, 0.0963 and 0.0910, which show that the error of Model 3 is the smallest when

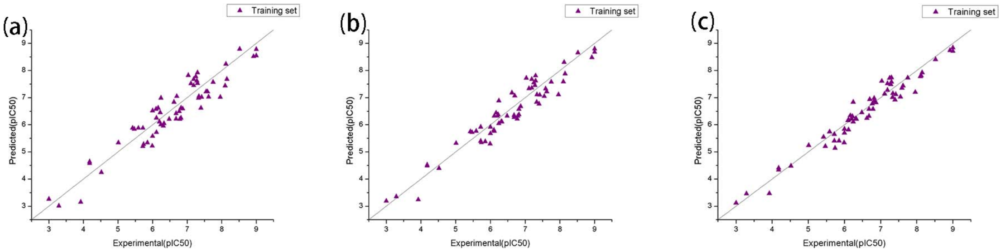

Figure.1 Correlation between predicted and experimental pIC50 by LOOCV. (a): Model 1; (b): Model 2; (c): Model 3.

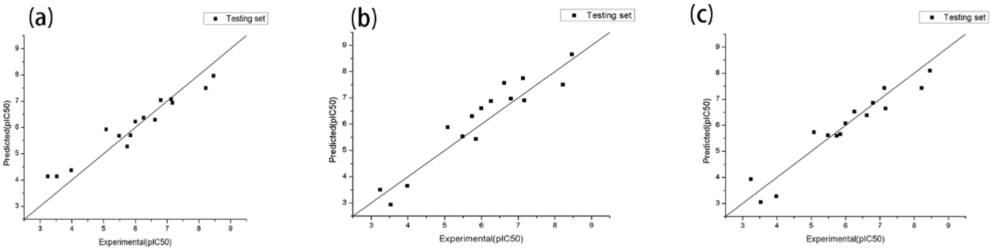

Figure.2 The predicted and experimental pIC50 of the testing set. (a): Model 1; (b): Model 2; (c): Model 3.

predicting pIC50. Therefore, the next result analysis is based on Model 3.

C. Selection of important features

In SVR model, ω= model.SVs'*model.sv _ coef. Among them, SVs are specific support vectors stored in the form of sparse matrix, sv _ coef is the coefficient of support vector, ω is the feature weight vector, that is, the proportion of each feature value in the SVR model established, and the top 10% weight histogram of 63 descriptors are shown in Tab. 1.

From Tab. 1, we can see that in the SVR model, the two most critical descriptors, BELm2 and BELe2, belong to BCUT descriptors. BCUT descriptors come from the eigenvalues of the adjacency matrix. BElm2 is a descriptor whose lowest eigenvalue is no.2 and weighted by atomic mass, while BELe2 is a descriptor weighted by Sanderson's electronegativity, which has the greatest impact on the activity of the GABA$_A$ receptor after binding to ligands. From these two descriptors, we can see that the diagonal elements of the whole matrix now correspond to the atomic mass and Sanderson electronegativity, and the lowest eigenvalue represents the topology of the whole molecule. The diagonal elements of BCUT descriptor correspond to atomic mass, van der Waals volume, Sanderson electronegativity and atomic polarizability [8]. We should understand this descriptor as a broader expression, not only limited to atomic mass and Sanderson electronegativity, but also the Mor08v descriptor in the third place is a 3D-MoRSE descriptor weighted by atomic van der Waals volume. At the same time, Mor29m is also weighted by atomic mass, which is the descriptor obtained by summing the atomic weights observed in signal 29.

refRMS is the rms difference between the current conformation coordinate and the reference structure, and its size mainly depends on which ligand conformation is taken as the reference. In this experiment, the original conformation of the ligand before docking of the receptor is chosen as the reference, so the establishment of QSAR model is closely related to the change of the conformation of the ligand before and after docking. We use semi flexible docking, and the conformation of ligands will change to some extent during docking. Intermolecular energy is

TABLE I. TOP 10% FEATURES

Rank	Feature Name	Category	Description
1	BELm2	BCUT descriptors	lowest eigenvalue n. 2 of Burden matrix / weighted by atomic masses
2	BELe2		lowest eigenvalue n. 2 of Burden matrix / weighted by atomic Sanderson electronegativities
3	Mor08v	3D-MoRSE descriptors	3D-MoRSE - signal 08 / weighted by atomic van der Waals volumes
4	Mor29m		3D-MoRSE - signal 29 / weighted by atomic masses
5	refRMS	Interaction force	The RMS of this conformation and the input conformation, and the input conformation is the initial conformation of the small molecule ligand
6	intermol_energy		Intermolecular energy is the binding energy minus the rotational free energy

binding _ energy minus torsional _ Energy and has a certain impact on the prediction of biological activity. In summary, it shows that Model 3 is reliable and has strong generalization ability. In addition, the top 10% important features selected by the model have certain credibility.

VIII. Conclusion

We used SVR to construct three QSAR models, and predicted pIC50 of the ligands after docking with the $GABA_A$ receptor. The reliability of the model constructed by combining the ligand molecular descriptor after docking with the ligand-receptor interaction characteristics is superior to the other two models. In addition, according to the optimal QSAR model, we determined and analyzed the first six important characteristics, which affect the ligand-receptor binding. The combination of these six characteristics is very important for predicting the pIC50 value after docking. It provides an important reference for predicting the bioactivity of different drugs combined with the same receptor in the future.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (21541006, 61772241).

References

[1] G. Gini, “QSAR: What Else?,” Methods Mol Biol, vol.1800, 2018, pp.79-105, doi:10.1007/978-1-4939-7899-1_3.

[2] D.J. Maddalena and G.A. Johnston, “Prediction of receptor properties and binding affinity of ligands to benzodiazepine/GABAA receptors using artificial neural networks,” J Med Chem, vol. 38(4), Feb. 1995, pp. 715-724, doi:10.1021/jm00004a017.

[3] M. Goodarzi, P.R. Duchowicz, C.H. Wu, F.M. Fernández and E.A. Castro, “New hybrid genetic based Support Vector Regression as QSAR approach for analyzing flavonoids-GABA(A) complexes,” J Chem Inf Model, vol. 49(6), Jun. 2009, pp.1475-1485, doi:10.1021/ci900075f.

[4] A. M. Bianucci, A. Micheli, A. Sperduti and A. Starita, “A novel approach to QSPR/QSAR based on neural networks for structures,” Soft Computing Approaches in Chemistry, vol. 120, 2003, pp. 265-296, doi: 10.1007/978-3-540-36213-5_10.

[5] S. Zhu, C.M. Noviello, J. Teng, R.M. Walsh, J.J. Kim and R.E. Hibbs, “Structure of a human synaptic GABAA receptor,” Nature, vol. 559(7712), Jul. 2018, pp.67-72, doi:10.1038/s41586-018-0255.

[6] G.M. Morris, R. Huey, W. Lindstrom, M.F. Sanner, R.K. Belew, D.S. Goodsell and A.J. Olson, “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem,” vol. 30(16), 2009, pp.2785-2791, doi:10.1002/jcc.21256.

[7] F. Grisoni, V. Consonni and R. Todeschini, “Impact of Molecular Descriptors on Computational Models,” Methods Mol Biol, vol. 1825, 2018, pp.171-209, doi:10.1007/978-1-4939-8639-2_5.

[8] R.S. Pearlman and K.M. Smith, “Metric Validation and the Receptor-Relevant Subspace Concept,” Journal of Chemical, Information & Modeling, vol. 39(1), 1999, pp.28-35, doi: 10.1021/ci980137x.