Research on Drug Molecular Computing Based on Support Vector Machine and Neural Network

Ruihan Xu*
Dominican High School, Milwaukee Wisconsin, US

*Corresponding author: 1800022820@pku.edu.cn

Abstract. With the continuous development of science, a large number of drugs, molecules and genes have been stored in various databases. The relationship between drug like molecules and molecular properties can be identified by computer science methods according to the diversity of compound library, drug molecular characteristics and the differences between different molecules, so as to achieve the purpose of drug screening. In this paper, 300 FBPase inhibitors were selected from previous literatures to convert IC$_{50}$ into pIC$_{50}$. The whole data set is divided into 236 molecules of training set and 64 molecules of test set. Two machine learning algorithms, namely Neural Network and Support Vector Machine, are selected to construct the prediction model with the training set as the learning object. And the reliability of the model is verified by the test set, and it is applied to the prediction of FBPase inhibitors. The prediction results show that the model is stable and reliable. In addition, the randomization test shows that the current model is not caused by chance correlation. The explanation of the selected molecular descriptors proves that the polarity of the molecule plays an important role in the inhibitory activity of FBPase. The model can provide some useful guidance for drug researchers and can screen ideal drug molecules before drug development.

Keywords: Drug Molecular Computing; Support Vector Machine; Neural Network.

1. Introduction
With the continuous development of science, a large number of drug, molecular and gene details have been stored in various databases. Chemical informatics is to identify the relationship between drug molecules and molecular properties through computer science methods according to the diversity of compound library, drug molecular characteristics and the differences between different molecules. Chemical informatics is a new branch based on multidisciplinary. Its main idea is to use computer technology to express, manage, analyze, simulate and spread chemical information, so as to realize the extraction, transformation and sharing of chemical information, and to reveal the essence and internal relationship of chemical information. Although chemical informatics is a new subject, it has been paid more attention and developed rapidly because of its close combination with computer science. Chemical informatics mainly studies how to identify drug like molecules, the relationship between molecular structure and biological properties, which includes the research tasks and contents of machine learning and computational chemistry. With the development of artificial intelligence
technology, chemical informatics method based on logic has become an important method of drug screening. Data mining method based on machine learning is a common method for drug target prediction [1]. The results of this study are to predict the potential targets of molecular compounds and drugs, and accelerate the speed and accuracy of drug screening. As an excellent combination of modern computer technology and traditional pharmaceutical field, molecular computing has attracted the attention of research institutions and pharmaceutical companies in recent years. At present, more than 20 million organic small molecules have been found. These discoveries not only deepen people's understanding of drugs, but also derive many kinds of databases based on these data. However, it is impractical to directly use such a large amount of data for traditional new drug research and development, and screening based on molecular weight may provide an effective way to solve this problem. With the development of structural biology, it will be more and more easy to obtain the information of active compounds with specific target structures, which greatly reduces the research objectives, so as to improve the screening rate of compounds and reduce the cycle and cost of new drug discovery.

Many scholars have done a lot of work in drug screening based on chemical informatics methods. There are many modeling methods of chemical informatics, especially the classification and regression methods related to QSAR. In addition, the comparative molecular field analysis method, comparative molecular similarity index analysis method, molecular docking method and molecular dynamics method are also briefly introduced. Lipinski et al. [2] established the first analysis world drug index database (WDI). Only scholars from various countries have established different types of databases for centralized utilization of resources and chemical informatics analysis. At the beginning of the 21st century, Li et al. [3] used two databases including drug data and non drug data, and used molecular fingerprint as the input feature description of Support Vector Machine (SVM), and constructed a drug like prediction model with higher prediction accuracy. Wale et al. [4] used ecfp4 molecular fingerprint descriptor to study drug target prediction. Four different methods, including Bayesian classification, class II Support Vector Machine, sorting based Support Vector Machine and cascade Support Vector Machine, were used to train the model, and the prediction situation of the models was compared. Niwa [5] used the probabilistic neural network method to study 799 compounds associated with seven categories of target annotation information in MDDR database. Among them, 60% were used as training set, 20% were used to improve model parameters, and the remaining 20% was used to evaluate the external test set of model prediction ability. The prediction results show that about 90% of the target information can be correctly classified. Researchers from the school of computer science and information, University of Dublin [6] used the deep learning method to show how the recursive neural network method can be applied to predict the molecular properties, that is, to predict the water solubility of drugs, to predict the action sites of molecules, and to play a role in gene expression data.

Through years of research and exploration, the positive rate of molecular screening can reach 5% to 30%, which is much higher than 0.01-0.1% of high-throughput screening. Therefore, the application of virtual screening method in new drug research has been paid more and more attention. In recent years, virtual screening technology has been successfully used to discover lead compounds. Cherezov, Rosenbaum and others cooperated to screen the lead like sub Library in zinc database with β 2-adrenoceptor as the target and verified the first 25 candidate compounds. The results showed that the binding affinity of six compounds was less than 4 μm, and the best molecular affinity was 9 nm. The inhibitory activity of five active molecules was even higher than that of standard compounds. Okimoto et al. used molecular docking technology and MM-PB / SA re scoring method based on ligand structure optimization to study CDK2, trypsin, ache and HIV PR, which increased the enrichment rate of molecular docking results by 1.6 to 4 times.

This paper mainly relies on machine learning and chemical informatics methods for drug screening. Firstly, we need to divide the selected data set into training set and test set. Two machine learning algorithms are selected to construct the prediction model with the training set as the learning object,
and test set is used to verify the reliability of the model. Finally, the aim of screening drugs correctly is achieved.

2. Data
Diabetes is one of the most prevalent diseases worldwide, and its incidence rate is increasing year by year. More than 90% of the patients are type two diabetes. The pathogenesis of type two diabetes is very complicated, which involves progressive insulin resistance and relative deficiency of insulin secretion, which often lead to hyperglycemia. In addition, type two diabetes is closely related to metabolic syndrome, including glucose intolerance, obesity, hypertension and dyslipidemia. There is also a report that metabolic syndrome can lead to a marked increase in the incidence rate of coronary artery disease, and even cause blindness. This makes diabetic patients suffer from psychological and physical problems. An effective drug to treat the disease. At present, there are many oral drugs for the treatment of diabetes on the market. These drugs generally reduce glucose level by increasing glucose metabolism, insulin secretion or insulin sensitivity. However, there are still safety issues with these drugs, such as sulfonphthalein drugs that can cause weight gain. However, dimethylbiglycerides are not suitable for diabetic patients with liver disease and nephropathy. Therefore, there is still a need for a new drug with less side effects to treat the disease.

Fructose-1,6-diphosphate (FBPase) is a highly regulated enzyme, which catalyzes the second to the last step of Glyconeogenesis. It can inhibit the gluconeogenesis from the corresponding substrates, thus avoiding the direct side effects caused by glycolysis, glycolysis and tricuronic acid cycle. Thus, FBPase has become a potential target for the treatment of type 2 diabetes mellitus. In addition, evidence from clinical studies suggests that FBPase inhibitors are very safe. At present, many inhibitors have been developed for this target. However, these drugs have not achieved the ideal oral utilization, so it is still necessary to develop more safe and effective drugs for FBPase.

A total of 300 FBPase inhibitors were selected from the papers published by Dang and his colleagues. IC50 was converted into pIC50 (in mol). The whole data set is divided into 236 molecules of training set and 64 molecular data sets of test set, which describe the molecular structure and biological activity of representative FBPase inhibitors.

Firstly, the data are preprocessed as follows: remove the descriptors whose number of zeros is greater than 85% and those with zero variance or near zero variance: one of the two descriptors with absolute correlation coefficient greater than 0.75 is removed. After these steps, 300 original descriptors were reduced to 268 for further study. The representative molecules of FBPase inhibitors are shown in Table 1.

| Chemical compound | Substituent | pIC50(mol) |
|-------------------|-------------|------------|
| 1                 | CF3CH2      | 7.24       |
| 2                 | neopentyl   | 7.92       |
| 3                 | cyclobutyl  | 7.72       |
| 4                 | cyclopenryl | 7.68       |
| 5                 | cyclopropyl-CH2- | 6.10 |
| 6                 | cyclobutyl-CH2- | 6.10 |
| 7                 | cyclopentyl-CH2- | 5.82 |
| 8                 | cyclohexyl-CH2- | 5.6  |

3. Methods

3.1 Machine learning
The essence of machine learning is to use mathematical model as the core tool, combined with the research results of cybernetics, cognitive psychology and other disciplines, and finally simulate human
perception, reasoning, learning, decision-making and other functions by computer system. Machine learning can also be used to screen chemical informatics drugs based on existing drug databases. The main steps include: database processing (data acquisition, feature extraction, data conversion), model training, model selection and model prediction. The main purpose of this paper is classification and supervised learning. The user gives the features and labels. The algorithm mines the rules and learns a pattern. According to this pattern, it predicts the label corresponding to the new feature. Drug functional groups in the database are used as features and labels, and supervised learning is carried out through Support Vector Machine and neural network to realize drug screening.

3.2 Support Vector Machine

Support Vector Machine is a popular classification method proposed by Vapnik [7]. Based on his strong mathematical background and successful experience, it has the ability to deal with nonlinear classification and high-dimensional data in Figure 1. Support Vector Machine classifier is a two-step process: first, the sample data vector is mapped to a high-dimensional feature space, so that the dimension of this feature space is obviously larger than the original feature space. Then find an optimal hyperplane, which can correctly separate the two types of data and maximize the data interval. The solution of this plane can be known by solving the following quadratic optimization problem:

$$f(x) = w\phi(x) + b$$

Where $w$ is the weight vector; $\phi(x)$ is the mapping function of mapping sample $x$ from the original space to the high-dimensional space; $b$ is the offset value. In order to solve $w$ and $b$, SVR uses the insensitive loss function $\varepsilon$ to carry out linear regression. The regression problem is transformed into a convex quadratic programming problem about variables $W$ and $B$. Since radial basis function (RBF) has high fitting and prediction accuracy, this paper chooses RBF as kernel function

$$K(x_i, x_j) = \exp(-\gamma (x_i - x_j))^2, \gamma > 0$$

Where $\gamma$ is the width of RBF nucleus.

![Figure 1. Support Vector Machine](image-url)
3.3 Neural Network

BP (back propagation) network is a multilayer feedforward network based on error back propagation in Figure 2, which is one of the most widely used neural network models [8]. The topological structure of BP neural network is composed of input layer, hide layer and output layer. The input node and output node are determined by the problem itself, and the number of hidden layer layers and hidden nodes need to be determined according to the specific implementation. Each layer of BP neural network is composed of unconnected parallel neurons, and the neurons between layers are fully interconnected. BP neural network can be divided into two processes: the first stage is the forward propagation process of working signal. The information to be input is transferred from the input layer to the hidden layer, and the output value of each cell is calculated by layer by layer processing. In the second stage, the error signal back propagation process. If the output value obtained in the output layer does not meet the expected requirements, the signal is returned, and the error between the output value and the expected output value is calculated recursively layer by layer, and the weight is continuously modified by gradient descent method until the total error reaches the minimum.

4. Results

The parameters that need to be adjusted include penalty factor C and kernel parameter. In the current model, ‘gridsearch’ method is used to get the optimal parameters. Finally, a 10 fold cross validation method is used to determine C is 10, which is based on the optimal parameters.

In the neural network, we choose the time when the cost of training data begins to decrease immediately rather than oscillate or increase as the estimation of η threshold. We don't need to be too precise and determine the magnitude. If the cost starts to drop in the first few rounds of training, you can gradually increase η. Until you find a value that causes the cost to oscillate or increase at the beginning of several rounds; on the contrary, if the cost function curve starts to oscillate or increase, try to reduce the magnitude until you find the setting that the cost decreases at the beginning of the round. Take half of the threshold to determine the learning rate. The reason for using training data here is that the main purpose of learning rate is to control the step size of gradient descent, and monitoring the training cost is the best method to detect excessive step size; the classification accuracy will still jitter or vibrate when the overall trend drops. If we stop at the beginning of the decline in accuracy, we will certainly miss the better option. A good solution is to terminate if the classification accuracy does not improve for a period of time. In the process of statistical learning, in addition to the quality of the selected data set, the appropriate descriptor subset is also very important for the establishment of the
prediction model: after the Support Vector Machine and BP neural network, the final 40 descriptors are selected. Based on the optimal parameters of Support Vector Machine and BP neural network, the root mean square error of Support Vector Machine model is 0.15 for training set and 0.24 for test set. The decision coefficient of training set is 0.88, and the correlation coefficient of cross validation is. The prediction ability of the model is tested by the external verification set, which gives the correlation coefficient of the test set and the correlation coefficient of the prediction. We know that BP neural network is suitable for the case of multiple variables, so we compare the prediction of FBPase inhibitors by Support Vector Machine and BP neural network. Table 2 shows the statistical results of these two methods, and Figure 3 shows the scatter plot of Support Vector Machine and BP neural network model. It can be seen that the execution power of SVM model is slightly better than that of RF model. The point distribution of training set and test set of this model is closer to straight line, which shows that the model shows high internal and external prediction ability.

Table 2. $R^2$ statistical results of Support Vector Machine and BP neural network

| Model | Training set | Testing set |
|-------|--------------|-------------|
| SVM   | 0.88         | 0.87        |
| BP    | 0.85         | 0.82        |

Figure 3. Comparison of real value and predicted value of two models

5. Conclusions

In the current work, Support Vector Machine and BP neural network model have been successfully proposed as a new chemical information and scientific method and applied to the prediction of FBPase inhibitors. The model has been successfully analyzed, which shows that the model is stable and reliable. In addition, the randomization test shows that the current model is not caused by chance correlation. The explanation of the selected molecular descriptors proves that the polarity of the molecule plays an important role in the inhibitory activity of FBPase. In this way, the Support Vector Machine and BP neural network model technology can be used to predict and screen novel FBPase inhibitors in the early stage of drug development. The model can provide some useful guidance for drug researchers, and can screen ideal drug molecules before drug development.
References

[1] Gasteiger J. The scope of chemoinformatics [M]//Handbook of Chemoinformatics: From Data to Knowledge. Weinheim: Wiley-VCH Verlag GmbH, 2003: 3–5.

[2] Lipinski C A, Lombardo F, Dominy B W, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings[J]. Advanced drug delivery reviews, 2012, 64 (Supplement): 4-17.

[3] Li Q, Bender A, Pei J, et al. A large descriptor set and a probabilistic kernel-based classifier significantly improve druglikeness classification[J]. Journal of chemical information and modeling, 2007, 47 (5): 1776-1786.

[4] Wale N, Karypis G. Target fishing for chemical compounds using target-ligand activity data and ranking based methods [J]. J Chem Inf Model, 2009, 49: 2190–2201.

[5] Niwa T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors [J]. J Med Chem, 2004, 47: 2645–2650.

[6] Jing Y, Bian Y, Hu Z, et al. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era [J]. AAPS J, 2018, 20(3): 58.

[7] Hsieh C H, Lu R H, Lee N H, et al. Novel solutions for an old disease: Diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks[J]. Surgery, 2010, 149(1):87-93.

[8] Aizenberg I, Moraga C. Multilayer Feedforward Neural Network Based on Multi-valued Neurons (MLMVN) and a Backpropagation Learning Algorithm[J]. Soft Computing, 2007, 11(2):169-183.