Integration of Systems Biology and Computer Aided Drug Design Technology in Action Mechanism of Traditional Chinese Medicine

Yu Dai*

School of Marxism, Yellow River Conservancy Technical Institute, Kaifeng, China
*Corresponding author(E-mail: daiyu_410@163.com)

Abstract
With the gradual deepening of people’s understanding for the complex pathogenesis of the disease, traditional Chinese medicine has received more and more attention due to its features of multi-component, multi-linkage, and overall regulation. After consulting literature, it was discovered that with the rapid development of systems biology, various system biology techniques have been widely used in the research and development of traditional Chinese medicine. At the same time, computer aided drug design (CADD) technology has also shown unique advantages in the screening of active ingredients, target discovery, toxicity prediction, and the study of the mechanism of prescription. Therefore, based on analyzing the current development mode of Chinese medicine and summarizing the role of computer aided drug design in Chinese medicine research, this paper uses machine learning algorithm to propose a new prediction model of Chinese medicine based on the combination of system biology technology and computer-aided drug design, which provides a powerful tool and means for the study of the complex theoretical system of Chinese medicine and the modernization of Chinese medicine.

Keyword: Systems Biology, Computer Aided Drug Design, New Medicine of Traditional Chinese Medicine, Research trend, Research and Development Model.

Integración de la Biología de Sistemas y la Tecnología de Diseño de Medicamentos Asistido por Computadora en el Mecanismo de Acción de la Medicina Tradicional China

Resumen
Con la profundización gradual de la comprensión de la gente sobre la compleja patogénesis de la enfermedad, la medicina tradicional china ha recibido cada vez más atención debido a sus características de multicomponentes, enlaces múltiples y regulación general. Después de consultar la literatura, se descubrió que con el rápido desarrollo de la biología de sistemas, varias técnicas de biología de sistemas se han utilizado ampliamente en la investigación y el desarrollo de la medicina tradicional china. Al mismo tiempo, la tecnología de diseño de medicamentos asistidos por computadora (CADD, por sus siglas en inglés) también ha demostrado ventajas únicas en la selección de ingredientes activos, el descubrimiento de objetivos, la predicción de toxicidad y el estudio del mecanismo de prescripción. Por lo tanto, basado en el análisis del modo de desarrollo actual de la medicina china y en el resumen del papel del diseño de medicamentos asistido por computadora en la investigación de la medicina china, este documento utiliza el algoritmo de aprendizaje automático para proponer un nuevo modelo de predicción de la medicina china basado en la combinación de la tecnología de biología de sistemas y diseño de drogas asistido por computadora, que proporciona una herramienta poderosa y un medio para el estudio del complejo sistema teórico de la medicina china y la modernización de la medicina China.

Palabras clave: Biología de Sistemas, Diseño de Medicamentos Asistido por Computadora, Nueva Medicina de la Medicina Tradicional China, Tendencia de Investigación, Modelo de Investigación y Desarrollo

1. Introduction
With the deepening of the research, people have become increasingly aware that most diseases are complex, multifactorial and systemic diseases, and the effect of drugs on the human body is a complex process of biological network regulation. The concept of traditional new drug research and development, which is dominated by "one drug, one target, one disease", is hard to meet the changing needs of modern diseases. At present, the research idea of regression analysis method and holistic method is obviously non-replicable, and it is difficult to become a general mode of innovation research of traditional Chinese medicine, especially the research of prescription [1-3]. Therefore, it is imperative to seek new research ideas and research means to
promote the modernization of traditional Chinese medicine. This paper summarizes the development of related technology in systems biology and its application in the field of TCM research and development and puts forward a new model of Chinese medicine research and development based on system biology and computer aided drug design [4-5].

2. Related Research Based on Computer Aided Drug Design Technology

2.1 The connotation of computer aided drug design technology

Computer aided drug design is the basis of drug molecular design and the tool of new drug research. Taking computer as a tool, it is the research results based on the biochemistry, enzymology, molecular biology, genetics and other life sciences. According to the potential drug design targets, including enzymes, receptors, ion channels and nucleic acids, which revealed in these basic studies, and referring the chemical structure characteristics of other endogenous ligands or natural products, the rational drug molecules are designed [6-7]. At the same time, CADD is also a Chinese medicine tool to study traditional Chinese medicine, explain the mechanism of traditional Chinese medicine and optimize the development of new prescription of traditional Chinese medicine, including molecular docking, reverse molecular docking / similarity search, substructure search, and other techniques, which are of great significance to the study of Chinese medicine and its prescription. CADD can not only reduce the blindness of drug screening, reduce the time of research and development, but also greatly save the research cost [8].

2.2. Classification of drug design

Since the concept of drug design was proposed, great progress has been made in its research and exploration. Both the basic theoretical knowledge and the application of practical methods have become more and more mature. Drug design can be basically divided into two categories: direct drug design and indirect drug design [9-10].

(1). Indirect drug design

The so-called indirect drug design is to select the required drug small molecules from the ligand structure database for the study of three-dimensional structure and some biological activities. The characteristic of this design model is that compared to direct drug design, and it reduces the scope of docking through previous research and greatly reduces the cost of drug docking. But at the same time, the value of accuracy and efficiency is sacrificed.

(2). Direct drug design

With the development of computer technology and the maturation of biology and chemistry, researchers can directly analyze the physical structure of target molecules (generally for protein, nucleic acid, etc.). By simulating the binding process of the compound, the combination of drug molecules and ligands is more stable. Intuitively, the matching process of target molecules and ligands is more direct and effective. The flow chart of
3. Integration of Systems Biology and Computer Aided Drug Design Technology in Action Mechanism of Traditional Chinese Medicine

3.1. Data acquisition

(1) Preparation of molecules in drug databases and non-drug databases

The database that selected for building the modeling are drug database MDDR and non-drug database ACD respectively. First, the subset of the molecular weight less than 600 is extracted from the MDDR, ACD and TCMCD databases, and then TCMCD and ACD subsets with similar molecular weight distribution were constructed with MDDR subsets with molecular weight less than 600. Finally, the molecular numbers of MDDR, ACD and TCMCD subsets with similar molecular weight distribution are 123927, 123927 and 33961 respectively.

(2) Preparation of training sets and data sets

First, 20,000 molecules were randomly selected from MDDR and ACD respectively to form the test set, and the remaining 103,927 molecules were used as the training set to build the model. How to select and generate training sets has always been a controversial issue in the construction of quantitative prediction of drug-like models. In the previous studies, there is no uniform standard for the training set used to construct the drug-like model. In order to systematically study the size of the training set and the influence of its composition on the prediction accuracy of the model, a series of prediction models are constructed based on different training sets. The size of the training set and the influence of its composition on the prediction accuracy of the model are studied in depth through the difference of prediction ability. A balanced training set is set up, that is, the number of drug molecules and non-drug molecules in the training set is equal. From MDDR and ACD, 10,000, 20,000 to 103,927 molecules were extracted successively, forming 10 set of balanced training sets. Different drug-like models were constructed and test set is carried out drug–like prediction by using balanced training sets containing different molecular numbers. In this way, we can systematically compare the influence of the size of the balanced training set on the model predictive ability.

Subsequently, 2000 drug molecules were randomly selected from the prepared test set, and then 40006000 to 18000 non-drug molecules were extracted and forming the non-equilibrium test sets with ratios of 1:2 to 1:9 respectively. Subsequently, 10000 drug molecules were extracted from data sets MDDR and ACD with molecular numbers of 103927 respectively, and 20,000, 30,000 to 90,000 non-drug molecules were extracted respectively, forming the non-equilibrium training sets with ratios of 1:2 to 1:9. Under each ratio, such as the ratio of 1:2, the training set based on the ratio of 1:2 is used to build drug-like models to predict the test set with ratio of 1:2. Through the prediction accuracy of the test set, the influence of the size of the non-equilibrium training set on the model construction can be compared and analyzed.

3.2. The calculation of molecular descriptors of traditional Chinese medicine in system biology

In this chapter, 21 descriptors of molecular properties widely used in drug-induced and ADME/T prediction (table 1) and molecular fingerprint descriptors characterizing molecular structure are calculated. All molecular property descriptors are calculated using Discovery Studio 2.529.

<table>
<thead>
<tr>
<th>The order number</th>
<th>Molecular parameters</th>
<th>Parameter meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A\log P$</td>
<td>Water partition coefficient</td>
</tr>
<tr>
<td>2</td>
<td>logD</td>
<td>Apparent partition coefficient at PH=7.4</td>
</tr>
<tr>
<td>3</td>
<td>logS</td>
<td>Water solubility of molecules</td>
</tr>
<tr>
<td>4</td>
<td>MW</td>
<td>Molecular weight</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>21</td>
<td>SASA</td>
<td>Molecular solvent accessibility surface</td>
</tr>
</tbody>
</table>

The molecular fingerprints used for model building are SciTegic extended connection fingerprints (ECFP, FCFP and LCFP) and Daylight-style path-based molecular fingerprints (EPFP, FPFP and LEFP). The English letters of these molecular fingerprints have their own meanings. The first letter F, E, L or F of the molecular fingerprint represents the functional role code. The second English letters (C or P) of molecular segment fingerprints represent the types of molecular fingerprints. C represents the molecular extended connection fingerprints, and P
represents path-based molecular fingerprints. The fourth letters of the molecular fingerprint indicate the maximum distance between atoms, such as the molecular fingerprints of the functional extension of the connection molecular fingerprints and the maximum distance of 6 between the atoms can be named FCFP 6. The molecular fingerprints with maximum distances of 4 and 6 between atoms are used here. Molecular fingerprinting descriptors were calculated by Discovery Studio 2.5.

3.3. Construction of algorithm model based on machine learning

The Bayesian model can process large amounts of data at the same time, and the learning process is fast and has good tolerance to the noise of the system. In addition, the Bayesian model only needs a training set consisting of a small number of samples to complete the construction of the model, which has been widely used in the related two element classification problems. Recursive segmentation model or decision tree model classifies samples according to certain hierarchical rules by simulating human learning and classification ability. The result of recursive segmentation is usually displayed through the graph of decision tree. Since the recursive segmentation model is more intuitive and easy to understand, it has been widely used to deal with related practical problems. In the construction of recursive segmentation model, the depth of decision tree has a great influence on the prediction accuracy of the model. In the study of this chapter, considering the number of elements in the training center is more, the minimum number of samples for each node is 30 when the decision tree model is built. The maximum number of nodes for each descriptor (that is, the maximum number of times descriptors appear as classification properties) is 20, adjust the depth of the decision tree from 5 to 15. Finally, the prediction accuracy of the decision tree classification model and the naive Bayesian classification model under the same conditions are compared and analyzed. For the Bayesian model and the recursive segmentation model, the parameters we used to measure the prediction ability are mainly as follows: True positive TP, false positive FP, true negative TN, false negative FN, sensitivity, specificity and Mathews correlation coefficient C:

\[ C = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FN)(TP + FP)(TN + FP)}} \]  

(1)

3.4. Experimental results and analysis

(1). Construction of Bayesian type drug model based on balance training

Based on the balance training set, a series of Bayesian models are built to evaluate the influence of the number of training concentrating molecule on the prediction accuracy of the model. First, for the training set with the least number of molecules, there are 10000 drug molecules and non-drug molecules. The prediction accuracy of the Bayesian model based on 21 physical and chemical properties (MP) and different molecular fingerprinting descriptors is shown in Table 2.

<table>
<thead>
<tr>
<th>Discorers</th>
<th>SE</th>
<th>SQ</th>
<th>Q-</th>
<th>Q+</th>
<th>C</th>
<th>SE</th>
<th>SQ</th>
<th>Q-</th>
<th>Q+</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>MP</td>
<td>0.690</td>
<td>0.586</td>
<td>0.625</td>
<td>0.654</td>
<td>0.277</td>
<td>0.692</td>
<td>0.577</td>
<td>0.621</td>
<td>0.652</td>
<td>0.271</td>
</tr>
<tr>
<td>MP+ECFP_4</td>
<td>0.874</td>
<td>0.833</td>
<td>0.840</td>
<td>0.869</td>
<td>0.708</td>
<td>0.864</td>
<td>0.822</td>
<td>0.829</td>
<td>0.858</td>
<td>0.687</td>
</tr>
<tr>
<td>MP+EPFP_4</td>
<td>0.801</td>
<td>0.678</td>
<td>0.713</td>
<td>0.773</td>
<td>0.482</td>
<td>0.802</td>
<td>0.673</td>
<td>0.710</td>
<td>0.772</td>
<td>0.479</td>
</tr>
<tr>
<td>MP+FCFP_4</td>
<td>0.824</td>
<td>0.789</td>
<td>0.795</td>
<td>0.818</td>
<td>0.163</td>
<td>0.815</td>
<td>0.785</td>
<td>0.791</td>
<td>0.810</td>
<td>0.601</td>
</tr>
<tr>
<td>MP+LPFP_4</td>
<td>0.754</td>
<td>0.710</td>
<td>0.722</td>
<td>0.742</td>
<td>0.464</td>
<td>0.752</td>
<td>0.714</td>
<td>0.724</td>
<td>0.742</td>
<td>0.466</td>
</tr>
<tr>
<td>MP+LCFP_4</td>
<td>0.878</td>
<td>0.838</td>
<td>0.844</td>
<td>0.873</td>
<td>0.716</td>
<td>0.865</td>
<td>0.832</td>
<td>0.837</td>
<td>0.860</td>
<td>0.697</td>
</tr>
<tr>
<td>MP+FPFP_4</td>
<td>0.835</td>
<td>0.834</td>
<td>0.834</td>
<td>0.834</td>
<td>0.669</td>
<td>0.823</td>
<td>0.824</td>
<td>0.824</td>
<td>0.823</td>
<td>0.647</td>
</tr>
<tr>
<td>MP+LPFP_6</td>
<td>0.873</td>
<td>0.876</td>
<td>0.875</td>
<td>0.873</td>
<td>0.748</td>
<td>0.858</td>
<td>0.864</td>
<td>0.864</td>
<td>0.859</td>
<td>0.722</td>
</tr>
<tr>
<td>MP+EPFP_6</td>
<td>0.782</td>
<td>0.759</td>
<td>0.764</td>
<td>0.777</td>
<td>0.541</td>
<td>0.779</td>
<td>0.759</td>
<td>0.764</td>
<td>0.774</td>
<td>0.538</td>
</tr>
<tr>
<td>MP+FCFP_6</td>
<td>0.853</td>
<td>0.844</td>
<td>0.846</td>
<td>0.852</td>
<td>0.698</td>
<td>0.843</td>
<td>0.837</td>
<td>0.838</td>
<td>0.842</td>
<td>0.680</td>
</tr>
<tr>
<td>MP+FPFP_6</td>
<td>0.751</td>
<td>0.764</td>
<td>0.761</td>
<td>0.754</td>
<td>0.514</td>
<td>0.747</td>
<td>0.764</td>
<td>0.760</td>
<td>0.751</td>
<td>0.511</td>
</tr>
<tr>
<td>MP+LCFP_6</td>
<td>0.893</td>
<td>0.860</td>
<td>0.864</td>
<td>0.889</td>
<td>0.753</td>
<td>0.877</td>
<td>0.850</td>
<td>0.854</td>
<td>0.874</td>
<td>0.728</td>
</tr>
<tr>
<td>MP+LPFP_6</td>
<td>0.824</td>
<td>0.788</td>
<td>0.795</td>
<td>0.818</td>
<td>0.613</td>
<td>0.815</td>
<td>0.785</td>
<td>0.791</td>
<td>0.810</td>
<td>0.601</td>
</tr>
</tbody>
</table>

The results show that the prediction accuracy of Bayesian model based on simple physicochemical properties is poor. The Mathews correlation coefficient C values of training set and test set are 0.277 and 0.271 respectively. Subsequently, after introducing molecular fingerprinting, the prediction ability of the model has been greatly improved. Among them, after introducing molecular fingerprint ECFP 4, ECFP 6, LCFP 4 and LCFP 6 based on the physical and chemical properties descriptors, the C value of the training set can reach more than 0.7. After predicting the balance test set based on the model constructed by different descriptors, we find that the four models (MP+ECFP_4, MP+LCFP_4, MP+ECFP_6, MP+LCFP_6) with the highest prediction
accuracy for the training set are also the best prediction accuracy for the prediction set. Among them, the model built based on MP+LCF_6 has the highest prediction accuracy for the test set. The sensitivity, specificity and Matthews correlation coefficients $C$ were 87.7%, 85.0% and 0.728, respectively. The prediction accuracy $C$ value of the model is shown in Figure 2.

![Figure 2 Prediction accuracy of the model](image)

Based on the MP+LCFP_6 molecular descriptor, the model is constructed under the equilibrium training set with a molecular number of 140000. The Bayesian scoring distribution chart of its training set and test set is shown in Figure 3. As shown in figure 3, the overall Bayesian score of drug molecules in the predicted results was higher than that of non-drug molecules. At the same time, the higher the score of Bayesian, the more likely the molecules are predicted to become drug like molecules, and vice versa. The constructed model predicts the optimal Bayesian score for separating the drug molecules and the non-drug molecules is -2.358. However, we observed that there was a significant overlap between the scores of drug and non-drug molecules within the range of -25 to 25 for Bayesian scores. As shown in figure 4, the area under the ROC curve of training set and test set, i.e., AUC value, is 0.984 and 0.967 respectively, indicating that the drug-like model constructed is very reliable.

![Figure 3a Bayesian scoring distribution map of training set](image)  ![Figure 3b Bayesian scoring distribution map of test set](image)
(2). Construction of Bayesian model based on non-equilibrium training set

In the process of constructing the model, the unbalance ratio of training set and test set is consistent. If the ratio of drugs to non-drug molecules in training set is 1:2, the ratio of drug concentration to non-drug molecules in test set should be 1:2. For the 9 sets of non-equilibrium training set and test set, the prediction results of Bayesian based on MP+LCFP 6 molecular descriptor are shown in Figure 5.

(3). An in-depth analysis of the misclassified compounds

A simple Bayesian classification model with the highest prediction accuracy is used to predict the molecules in the test set. It was found that 1814 molecules in the drug database MDDR were mistakenly predicted to be non-drug molecules, while 1885 non-drug molecules were mistakenly predicted to become drug molecules. According to the sorting of Bias's score, 20 compounds of MDDR were predicted to be the most likely non-drug molecules and 20 compounds of ACD were predicted to be the most likely drug molecules, as shown in Figure 4.9. As shown in Figure 6a, it is obvious that the compounds in the LVDR that are wrongly predicted to be non-drug molecules contain a substructure fragment (Fig. 6b) that has been predicted by the Bayesian model for the negative contribution to the efficacy of the drug-like.

We will use the number of molecules of 140000 balance training set (the number of drugs and drug molecules is 70000 respectively) to construct a multiple decision tree model based on the MP+LCFP 6 molecular descriptor in depth of the decision tree is range from 5 to 10. The prediction accuracy C value of the model is shown in figure 7. For the training set, as the depth of decision tree increases, its C value also increases. However, for the test set, as the depth of the decision tree increases from 5 to 13, its C value also increases. When the depth of decision tree reaches 13, the maximum value of C reaches 0.525.
Construction of a recursive segmentation model of medicinal properties

When the depth of the decision tree continues to increase, the value of C tends to decrease. Based on the above results, we know that when the depth of the decision tree is 13, the prediction accuracy of the decision tree model constructed is the best. In addition, under the same conditions (the same training set, prediction set and molecular descriptor), the Bayesian classification model has higher prediction accuracy than the recursive segmentation classification model.

Figure. 7 The change situation of C value of prediction accuracy of decision tree model changes with the depth of the decision tree

4. Conclusions

CADD is a bridge linking traditional Chinese medicine theory with modern science. The technology provides new ideas and technical support for the research of traditional Chinese medicine and its compound and plays an important role in the search of active substances and target screening of Chinese medicine. As a new
discipline, systematic biology has a unique theory of holistic view and system theory. In this paper, the machine-learning algorithm model is used to predict the drug-like model of Chinese medicine compounds. Experimental results verify the accuracy of the algorithm model. The research in this paper has promoted the deep integration of system biology and CADD in the mechanism of traditional Chinese medicine.

Acknowledgements
This work was supported by “the young backbone teacher training program” in Yellow River Conservancy Technical Institute.

References


