ISSN 2616-5910 Vol. 5, Issue 1: 179-184, DOI: 10.25236/AJALS.2024.050124

Screening of Molecular Descriptors Based on Random Forest Model and Correlation Analysis

Yifan Fu[*,#], Zinuo Hao[#], Peijie Wu

Saxo Fintech Business School, University of Sanya, Sanya, 572022, China
[*]Corresponding author: 15501938775@163.com
[#]These authors contributed equally.

Abstract: *Breast cancer is the most common and one of the most lethal diseases in the world, the demand for its treatment is imminent. In this study, we explored a method to screen compounds with potential anti-breast cancer potential by targeting estrogen receptor alpha subtype (ERα). Through the analysis of data, 729 kinds of molecular descriptor of 1974 compounds were studied and analyzed. Firstly, strict data preprocessing steps were adopted, including vacancy detection, outlier processing, and feature screening. Finally, 359 valid features were identified. Subsequently, two different dimensionality reduction methods, random forest and entropy method, were used to screen out the 30 main variables with the most significant effect on ERα activity under each method. The comparative analysis shows that the random forest method has advantages in representativeness and effect. Then, by Pearson correlation analysis and gradually eliminate highly correlated variables, eventually determines the effect on activity of ER alpha 20 largest molecular descriptor. These descriptors included SHsOH, XLogP, etc., and their biological activity contribution rates ranged from 0.01045 to 0.00597. This study is helpful for the treatment of breast cancer, and the method of screening compounds with anti-breast cancer potential could be helpful.*

Keywords: *Breast cancer, estrogen receptor alpha subtype, biological activity*

1. Introduction

Breast cancer is one of the most common and deadly diseases in the world, and the need for its treatment is increasingly urgent. According to the latest data released by the World Health Organization's (WHO) Institute for Research on Cancer (IARC)[1], there were more than 19 million new cancer cases worldwide in 2020, among which the incidence of breast cancer in women remains high, posing a serious threat to public health. Estrogen receptor alpha subtype (ERα) is a key target in the treatment of breast cancer, and its activity regulation directly affects the therapeutic effect[2]. Therefore, finding compounds that can antagonize ERα activity has become one of the key goals of current breast cancer treatment drug research and development[3]. This paper focuses on the key problem in the research and development of breast cancer treatment drugs: how to quickly and accurately select candidate compounds with good biological activities through effective compound screening methods? We constructed a quantitative structure-bioactivity relationship model by analyzing the molecular structure descriptors of a large number of compounds and their bioactivity data to predict and optimize the activity of new compounds[4].

This paper first details the key steps of data preprocessing, including vacancy value processing, outlier detection[5], and feature screening to ensure the quality and representativeness of the selected features. Subsequently, we used two different dimensionality reduction methods, random forest and entropy method, to screen out the main variables with significant effects on ERα activity, respectively[6]. The comparative analysis of the effects of these two methods identifies the advantages of random forest in selecting variables with representative and predictive performance[7]. Furthermore, using Pearson correlation analysis method and the strategy of gradually eliminating highly correlated variables, we finally identified the 20 most influential molecular descriptors for ERα activity. These descriptors not only include structural features such as SHsOH and XLogP[8], but also show their important contribution rate range in biological activities. The results of this study not only help to guide the optimal design of existing drugs[9], but also provide theoretical support and practical guidance for the search of new breast cancer treatment drugs, thus providing new ideas and strategies for the progress of breast cancer treatment[10].

ISSN 2616-5910 Vol. 5, Issue 1: 179-184, DOI: 10.25236/AJALS.2024.050124

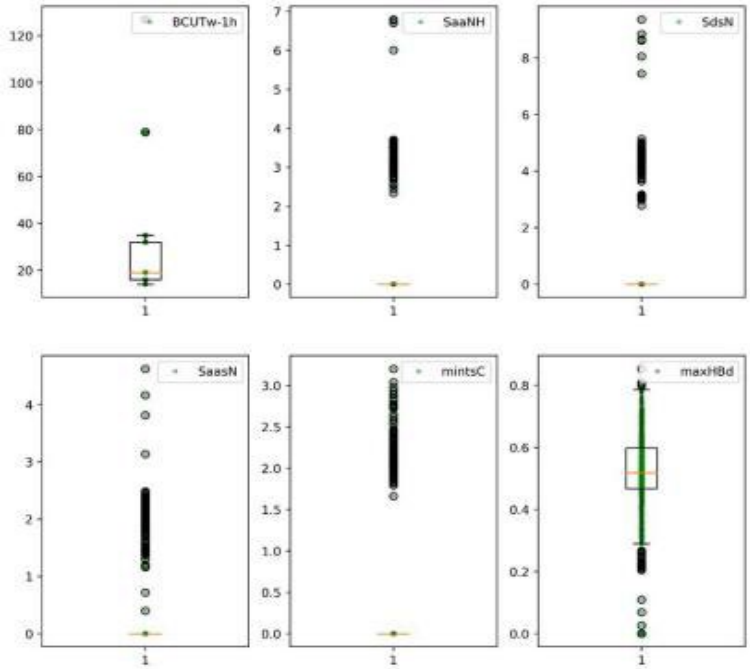
2. Modeling and analysis

It is necessary to select variables for 729 molecular descriptors of 1974 compounds, rank variables according to the importance of their impact on biological activity, and give the top 20 molecular descriptors (i.e., variables) with the most significant impact on biological activity. In this paper, the preprocessing of the above data includes the following processes: first, the vacancy value of the collected data is judged, and it is found that there is no vacancy; Then, the features with 0 ratio greater than 90% were eliminated, and 344 features were eliminated. Then, the Laida criterion was used to screen the data, and the features with more than 100 feature outliers were eliminated, and 26 features were eliminated. Finally, data with the number of characteristic outliers less than 100 are bounded.

After data preprocessing, dimension reduction is carried out. In this paper, random forest and entropy method are used to reduce dimension, and 30 groups of main variables are selected according to the contribution rate, and the correlograms of the 30 groups of variables under the two methods are compared, and it is found that the features extracted by random forest are more representative. Then Pearson correlation analysis was conducted on the 30 groups of variables selected by random forest, and the 10 groups of variables with strong correlation were eliminated one by one, and finally the top 20 characteristics were taken as the main variables.

In engineering practice, the data we get will have missing values, outliers, etc., which need to be pre-processed before use. On the basis of the attachment, this paper makes the following processing:

Firstly, the vacancy value of the collected data is judged, and there is no vacancy value after screening and screening. Then, the Laida criterion is used to screen the data, and the features whose eigenvalues are not within the range of $\mu \pm 3\sigma$ and the number of outliers is greater than 100 are eliminated, a total of 26 features are eliminated. On this basis, the features whose eigenvalues are not within the range of $\mu \pm 3\sigma$ and the number of outliers is less than 100 are processed by the maximum amplitude limiting method. The specific method of amplitude limiting is as follows: use $\mu+3\sigma$ to replace the abnormal value greater than $\mu+3\sigma$, use $\mu-3\sigma$ to replace the abnormal value less than $\mu-3\sigma$, so as to limit the abnormal data in the attachment; Finally, the data were screened, and 344 features were eliminated with the proportion of 0 values greater than 90%. The remaining 359 features were pretreated. In this paper, six out of the 26 features are selected to make boxplots for explanation. Figure 1 shows the boxplots of abnormal data. The data distribution is too scattered, and the box is very flat, even there is only one line left.



(Data source: https://www.sinomed.ac.cn/index.jsp)

Figure 1: Box plot of abnormal characteristics

ISSN 2616-5910 Vol. 5, Issue 1: 179-184, DOI: 10.25236/AJALS.2024.050124

3. Data dimensionality reduction

Pearson correlation, also known as product-difference correlation (or product-moment correlation), is a method for calculating straight-line correlation proposed by British statistician Pearson in the 20th century.

Suppose there are two variables X and Y, and Pearson's correlation coefficient is used to measure the degree of linear correlation between X and Y, with a value between -1 and 1. This linear correlation is intuitively expressed as follows: as X increases, whether Y increases or decreases simultaneously; When the two are distributed on a straight line, the Pearson correlation coefficient is equal to 1 or -1; The Pearson correlation coefficient is 0 when there is no linear relationship between the two variables.

Then the Pearson correlation coefficient between the two variables can be calculated by Formula (1) below:

$$\mathrm{P}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sigma Y\sigma Y} \tag{1}$$

Entropy is a physical concept in thermodynamics and is a measure of the degree of disorder (or disorder) in a system. Higher entropy means that the system is more chaotic and carries less information, while lower entropy means that the system is more orderly and carries more information. According to the characteristics of entropy, the entropy value can be calculated to represent the disorder degree of a random system, and the entropy value can also be used to judge the dispersion degree of an index. The general steps of the entropy method can be expressed as follows: data standardization, data preprocessing, calculating entropy and calculating weights.

We use the entropy method of dimension reduction to calculate the weights of the features in the attachments "Molecular_Descriptor.xlsx" and "ERα_activity.xlsx", and the results are shown in Figure 2 below:

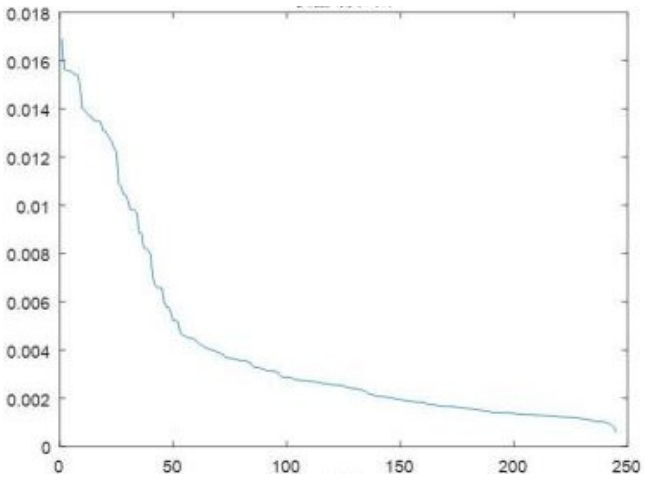


Figure 2: Weight calculation results of entropy method

As can be seen from the figure, the weight of the feature value gradually decreases, and the weight refers to the frequency of each number in the weighted average. Therefore, among the 729 molecular descriptors of 1974 compounds, the weight of the first 30 features is relatively large, indicating that it has a great influence on the biological activity and is representative to a certain extent. python language is used to implement the first 30 feature selection of entropy method.

In dimension reduction by entropy method, the top 30 molecular descriptors (i.e., variables) with significant influence on biological activities have strong positive correlation, which is generally in green (green represents positive correlation, and the correlation among themselves is 1), while the representation of negative correlation is relatively weak, which is not very representative. In addition, entropy method lacks horizontal ratio between pairs of features, and the weight of each feature changes with the change of samples, and the weight depends on the sample, so it is limited in application. Therefore, random forest dimension reduction method is used to make up for this shortcomer.

Random Forest (RF) is a newly emerging, highly flexible, tree-based machine learning algorithm, which uses the power of multiple trees to make decisions. Each tree in the forest is not the same, each tree

ISSN 2616-5910 Vol. 5, Issue 1: 179-184, DOI: 10.25236/AJALS.2024.050124

is randomly created, each node in each tree is a random subset of the features to be selected, and the output of all trees is integrated into the final output of the forest.

Random forest focuses on the word "random", which includes two aspects:

Random selection of training data. That is, the data used to train a single tree should be random and have the same amount of data as the original data set from all the data. This requirement is to prevent the consistency of the data used for training of each tree from causing each tree to be the same, thus losing the significance of building multiple trees.

Random selection of features to be selected. So that each tree in the forest can be different from each other, indicating the diversity of the system, thus improving the classification performance. The features in the attachment "Molecular_Descriptor.xlsx" and "ERα_activity.xlsx" were processed by random forest dimensionality reduction method, and the variables were ranked according to their importance (contribution rate) to biological activities. The percentage of contribution rate of each feature after ranking is shown in Figure 3 below:

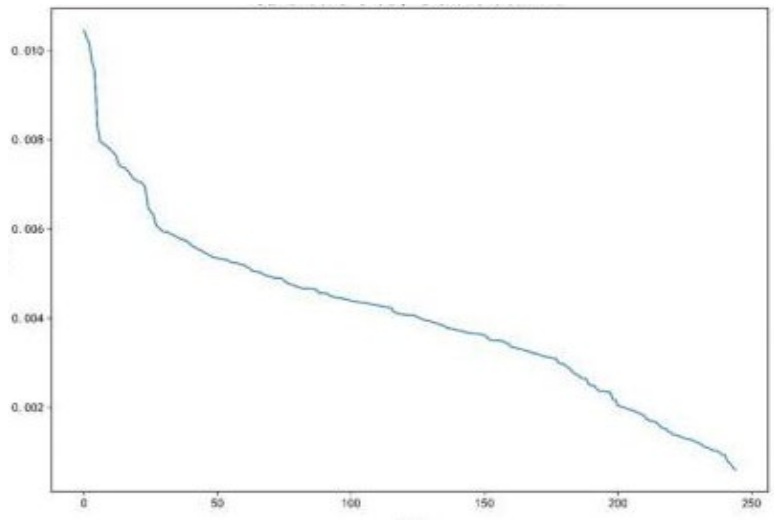


Figure 3: Percentage of feature contribution rate after sorting

As shown in Figure 3, the feature contribution rate of the features given in the attachment "Molecular_Descriptor.xlsx" and "ERα_activity.xlsx" is calculated and sorted. Among the 729 molecular descriptors of 1974 compounds, the top 30 features have a large contribution rate to biological activity. The latter feature contribution rate is less than 0.006, indicating a small impact factor.

Python language is used to select and reduce the first 30 features of the random forest, and the results of the RF dimension reduction method show that the correlation distribution is relatively uniform. The overall positive correlation is partial for the entropy method, which is independent of the dependent variable pIC50 in the process of selecting variables. Based on these two points, it can be seen that the 30 groups of features selected by the entropy method are not representative, while the random forest feature selection avoids these two shortcomings, so the 30 groups of features selected by the random forest are more representative. On this basis, the correlation analysis of 30 groups of variables of random forest was carried out, and 10 pairs of features with correlation greater than 0.85 in the 30 groups of features were removed one by one, and finally 20 groups of features were selected as the main variables.

It is concluded that the overall positive correlation of the entropy method is partial, and the method has nothing to do with the dependent variable pIC50 in the process of selecting variables. Based on these two points, it can be seen that the 30 groups of features selected by entropy method are not representative, while the random forest feature selection avoids these two shortcomings, so the 30 features selected by random forest are more representative. On this basis, this paper conducts correlation analysis on the 30 groups of variables of random forest, removes one of the 10 pairs of features whose correlation is greater than 0.85 among the 30 features one by one, and finally selects 20 groups of features as the main variables.

Among the 729 molecular descriptors of 1974 compounds, the top 20 molecular descriptors (i.e., variables) with the most significant impact on biological activities and their contribution rates are shown in Table 1 below:

ISSN 2616-5910 Vol. 5, Issue 1: 179-184, DOI: 10.25236/AJALS.2024.050124

Table 1: The 20 molecular descriptors with the most significant effects on biological activity

Name of feature	Rate of contribution	Name of feature	Rate of contribution
SHsOH	0.01045	XLogP	0.00711
LipoaffinityIndex	0.01016	maxssO	0.00708
MDEC-23	0.00953	minHBa	0.00705
BCUTc-1h	0.00833	TopoPSA	0.00702
hmin	0.00777	nHBAcc	0.00692
ATSc1	0.00771	gmin	0.00639
ATSc3	0.00739	WTPT-4	0.00632
minsOH	0.00738	ATSp5	0.00608
ATSc5	0.00733	BCUTp-1h	0.00604
WTPT-5	0.00726	maxaaCH	0.00597

Among the 20 molecular descriptors (i.e., variables) that have the most significant impact on biological activities, the ratio of quantity to cumulative contribution rate is shown in Figure 4 below.

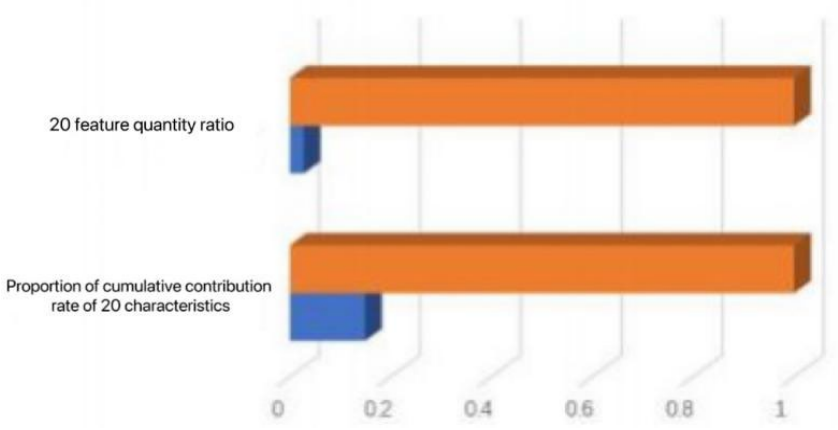


Figure 4: Relationship between contribution rate and quantity proportion

4. Conclusion

This study focused on the screening and bioactivity analysis of estrogen receptor alpha subtype (ERα) as an important target through in-depth discussion of the key issues in the research and development of breast cancer therapeutics. First, we implemented rigorous data preprocessing steps, including vacancy value processing, outlier detection, and feature screening, to ensure that the selected molecular descriptors are of high quality and representative. Subsequently, using two different dimensionality reduction methods, random forest and entropy method, we screened out the key variables that have a significant impact on ERα activity and clarified the advantages of random forest in terms of representativity and predictive ability. Furthermore, through Pearson correlation analysis and stepwise elimination of highly correlated variables, we finally identified the 20 molecular descriptors with the most significant effects on ERα activity. The detailed analysis of these descriptors showed that they not only reflected the structural characteristics of the compounds, but also showed an important role in the contribution rate of biological activity. The results of this study provide theoretical support and practical guidance for the optimal design of current breast cancer treatment drugs, and provide new methods and strategies for finding novel drug candidates. Future studies can further explore the pharmacokinetic properties and safety of these candidate compounds to accelerate their application and promotion in clinical treatment.

References

[1] ShoU Y Q. Optimization modeling of anti-breast cancer candidate drugs [C]// Chinese Society of Toxicology. Proceedings of the 10th National Toxicology Congress of the Chinese Toxicological Society. Qingdao University; 2023:1. DOI: 10.26914 / Arthur c. nkihy. 2023.012614.
[2] Wu S , Yang S , Luo L ,et al.A prediction model of insulation strength for gaseous medium considering the effect of external electric field[J].Journal of Molecular Modeling, 2024, 30(12). DOI:10.1007/s00894-024-06199-2.

ISSN 2616-5910 Vol. 5, Issue 1: 179-184, DOI: 10.25236/AJALS.2024.050124

[3] Liu R Y. Mathematical modeling of adaptive cancer therapy based on dynamic optimization [D]. Wuhan university, 2022. DOI: 10.27379 /, dc nki. Gwhdu. 2022.000328.
[4] Luo Ding. Molecular modeling of coronavirus major protease inhibitors and BRD4 inhibitors [D]. Shaanxi university of science and technology, 2022. DOI: 10.27290 /, dc nki. GXBQC. 2022.000366.
[5] Mr Chirac. Combination of drugs based on deep learning collaborative prediction research [D]. Shanghai ocean university, 2021. The DOI: 10.27314 /, dc nki. Gsscu. 2021.000805.
[6] LU X D. Theoretical study on CYP4F12 and anticancer drug Plocabulin based on DFT and MD simulation [D]. Tianjin medical university, 2021. DOI: 10.27366 /, dc nki. Gtyku. 2021.000305.
[7] YU Z. AEBoost: a method for predicting the sensitivity of anticancer drugs [D]. Shanghai normal university, 2021. DOI: 10.27312 /, dc nki. Gshsu. 2021.002335.
[8] Feng Yi. Ace inhibitors antitumor agents of QSAR study [D]. Shaanxi university of science and technology, 2021. The DOI: 10.27290 /, dc nki. GXBQC. 2021.000219.
[9] GAO H J. Exploring the synergistic effect of combined drugs by multi-scale drug model [D]. Southwest University, 2018. (In Chinese with English abstract)
[10] Huang, X. Research on computer-aided anticancer drug design and protein homology modeling [D]. Lanzhou University, 2011.