MEDS Basic Medicine (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/medbm.2024.020108
ISSN 2616-2091 Vol. 2 Num. 1

Research on Efficient and Low-cost Drug-disease Association Prediction Method Based on Dual Attention in Heterogeneous Networks

Yujie Yang*, Yue Gao, Xiaohan Li, Wenhao Ding, Chengyang Gao

School of Software, Yunnan University, Kunming, 650500, China
**Corresponding author: cgyjhjy@163.com*

Keywords: Heterogeneous network, drug-disease association, attention mechanism, graph attention, SENet

Abstract: Drug development usually costs a high cost, so it is very important to establish an efficient, low-cost and accurate prediction method of drug-disease correlation. In this paper, a drug-disease prediction method based on dual attention in heterogeneous networks is proposed. First, the experimental data set is constructed through the biological database, then the node feature information in the heterogeneous network is extracted by the graph attention network, and the node feature information is filtered and enhanced by SENet. Finally, through the 10% discount cross verification evaluation, GASEDDA achieved an accuracy of 98.5%.

1. Introduction

The overall drug research and development can be summarized into three stages, the first is the drug discovery stage, the second is the preclinical research stage, and the third is the clinical research stage, which is difficult to develop. New drug development usually takes 10-15 years and an investment of 1.5 billion US dollars. The overall development process is time-consuming and risky. In the United States, more than 100 drugs are screened by the Food and Drug Administration (FDA) every year before they are approved for market, and eventually there are only about 20% on the market[1]. The number of new drugs approved around the world is declining year by year and the failure rate of new drug approval has been higher than 90% since the 1990s[2]. In order to solve the problems in the process of new drug development, relevant personnel try to develop new drugs through the method of drug repositioning[3].

Drug repositioning is the process of determining the potential indications of existing drugs and discovering new drug treatments for diseases. Drug repositioning is a new drug research and development strategy, and it is also considered as one of the best risk-benefit strategies in the existing drug research and development strategies[4]. It has attracted close attention all over the world. Compared with traditional drug development, drug repositioning has incomparable advantages, which can not only shorten the screening scope of drug development, but also save a lot of money and time. The traditional drug relocation methods generally include drug common biochemical characteristic analysis[5], drug prescription screenin[6], molecular activity similarity analysis[7] and

so on. With the continuous development of related research, the development and use of various biological databases, such as DrugBank[8], PubChem[9], SIDER[10], etc., provide a large number of opportunities for the development of drug relocation based on computing methods, so that computational drug relocation has a very broad development prospect and potential, and has been paid more and more attention by relevant researchers.

At present, researchers focus on identifying new drug targets by using drug chemical structure, pharmacology and genome properties. Some scholars have proposed to mine the potential indications of listed or unlisted drugs by directly predicting the relationship between drug diseases. The existing drug relocation methods are mainly divided into recommendation system-based methods, machine learning-based methods, deep learning-based methods and web-based methods. The method based on recommendation system mainly uses matrix decomposition to complete the task, but because of the problem of cold start, it is not suitable for the prediction of new drugs or new diseases. Machine learning-based methods are widely used, and then they rely heavily on input data that can reflect the characteristics of drug diseases, which is difficult to meet in practice. The method based on deep learning can make use of its strong learning ability to transform the original data features into abstract feature representation, which can perfectly solve the incompleteness of manual screening features. But they need a lot of training data to obtain high precision, that is to say, when the input drug-disease association network is too sparse, the method based on deep learning is easy to appear over-fitting. The web-based approach captures similar information from different types of biological networks as a feature of drugs and diseases. In this method, heterogeneous networks are usually introduced to represent different types of biological information, and their similarities are retained in different biological networks, so as to obtain unobserved associations between drugs and diseases.

Attention mechanism is widely used in a variety of deep learning tasks, such as natural language processing, image recognition and speech recognition, and has become one of the core technologies in the field of deep learning. When processing the information received by the outside world, the human brain will focus its attention on the key information of high value and interest, and the attention mechanism is inspired by the way the human brain processes information. It can be regarded as a combinatorial function, which highlights the influence of key inputs on output by calculating the probability distribution of attention. In bioinformatics, attention mechanism is also widely used, such as using layer attention mechanism to predict drug-disease association, integrating multiple biological relationships for drug-disease association, and so on.

In this paper, a graph attention heterogeneous network model based on SEnet is proposed to predict drug-disease association. The related information of drugs, diseases and genes is collected through the biological database, and a benchmark data set is constructed. Through the known drug-disease association, drug-gene association, disease-gene association and calculating drug similarity, disease similarity and gene similarity, we construct a heterogeneous network to predict drug-disease association. Based on this heterogeneous network, we extract information features from the similarity network through the graph attention mechanism, and then recalibrate through the channel features. Finally, an integrated embedded prediction module is used to predict the unobserved drug-disease association. According to the computer simulation experiment, the method proposed in this paper achieves 90.2% AUC score and 98.5% ACC score. Compared with other cash methods, the method proposed in this paper is better.

2. Construction of drug-disease heterogeneous network

2.1 Dataset

In order to effectively evaluate the model proposed in this paper, this paper collects relevant data through biological databases such as CTD[11] and DrugBank[8], and constructs a benchmark data

set, including 709 drugs, 5604 diseases, and 1513 proteins.

It contains 199214 drug-disease edge, drug-protein edge, disease-protein edge.

2.2 Heterogeneous network construction

2.2.1 Disease-disease similarity

The medical subject word identifier of disease can be described as a hierarchical directed acyclic graph DAGs.

In this paper, the DAG structure is used to calculate the semantic similarity of diseases. For disease d, DAG (d) = (N (d), E (d)), N (d) denotes the node set of disease d and all ancestors of d, and E (d) represents the relationship of all the relationships between diseases in N (d). The semantic contribution of a disease d ∈ N (d) to d can be expressed as follows:

$$D_d(d_t) = \begin{cases} 1, & d_t = d \\ \max\{\Delta * D_d(d_t') | d_t' \in children\ of\ d_t\}, & d_t \neq d \end{cases} \quad (1)$$

where Δ is the semantic attenuation factor, according to previous research, here we set it to 0.5, and the semantic contribution of disease d to itself has a value of 1. From Eq. 1, we know that the main contribution of disease d_t is determined based on the distance between disease d and disease d_t, and by summing up the contributions of all the ancestor nodes of disease d, we use Eq. 2 to obtain the semantic value of d_t.

$$DV(d) = \sum_{d_t \in d} D_d(d_t) \quad (2)$$

Combining Equation 1 and Equation 2, we can get the semantic similarity between disease d_i and disease d_j:

$$Sim^{di}(d_i, d_j) = \frac{\sum_{d_t \in (N(d_i) \cap N(d_j))} \left(D_{d_i}(d_t) + D_{d_j}(d_t)\right)}{DV(d_i) + DV(d_j)} \quad (3)$$

where the contribution of d_t to d_i and d_j is denoted as DV(d_i) and DV(d_j) respectively

2.2.2 Drug-drug similarity

Drugs are special chemicals used by human beings to prevent, treat, or diagnose diseases, or can regulate the function of the human body, improve the quality of life, and maintain good health. It usually has different characteristics of biological and chemical properties. We can convert drugs into many types of feature vectors by their characteristics and calculate drug similarity based on these features. In this paper, we download the drug SMILES sequences from DrugBank[8] and convert them into topological fingerprints of the drugs, and calculate the similarity between two drugs based on the fingerprint loci and Tanimoto similarity. Assuming drug dr_i and drug dr_j, the similarity between them can be calculated by Equation 4 and Equation 5

$$TM = \frac{dr_i dr_j}{dr_i^2 + dr_j^2 - dr_i dr_j} \quad (4)$$

$$Sim^{dr}(dr_i, dr_j) = 1 - \frac{\min(TM)}{\max(TM) - \min(TM)} \quad (5)$$

2.2.3 Gene-gene functional similarity

Calculating gene-gene functional similarity is the basic work of bioinformatics, which is an

important part of life science research. In this paper, we use GO to study the similarity between genes. The GO graph uses a directed acyclic graph to represent structured relationships between biological terms, as shown in Figure 1. A node in the graph represents a term, and in addition to the root node, each node has the possibility of multiple parent nodes and may have multiple children. The depth of a node indicates the shortest path between that node and the root node. The closer a node is to the root node indicates the more general the term semantics, and conversely, the further it is from the root node indicates the more explicit the term semantics. According to previous research, it is believed that the deeper the depth of the Lowest Common Ancestor (LCA) between two nodes, the more similar they are.

$$Sim^{ge}(g_i, g_j) = \frac{2H}{D_i + D_j + 2H} \tag{6}$$

where D_i and D_j denote the lengths of the shortest paths between g_i and g_j to the LCA, respectively, and H is the length of the shortest path between the LCA and the root node.

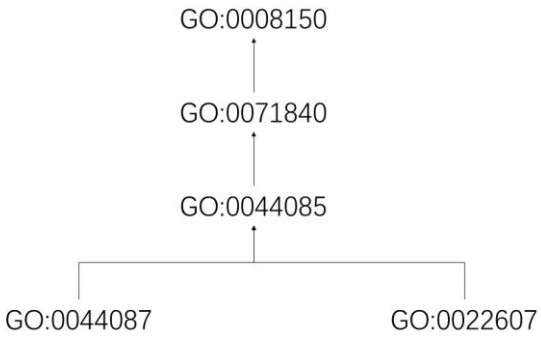

Figure 1: Relationships between some of the nodes in the biological process subgraph of the GO

2.2.4 Heterogeneous network construction

After obtaining drug-drug similarity, disease-disease similarity and gene-gene similarity, drug-disease associations we get from DrugBank[8], and drug-gene associations as well as disease-gene associations we get from CTD[11]. In order to better represent the heterogeneous networks, we use a parameterized form to represent the heterogeneous networks. Take drug-disease association as an example, the drug-disease association is represented as a kind of binary network $A^{didr} \in \{0,1\}^{N \times M}$, where N is the number of drugs and M is the number of diseases. If drug dr is associated with disease di, then $A_{di,dr} = 1$, otherwise $A_{di,dr} = 0$. Finally, the heterogeneous network model is constructed as shown in Equation 7:

$$H = \begin{bmatrix} Sim^{di} & A^{drdi} & A^{gedi} \\ A^{didr} & Sim^{dr} & A^{gedr} \\ A^{dige} & A^{drge} & Sim^{ge} \end{bmatrix} \tag{7}$$

3. Models and Methods

In this section, we will formally introduce the SEGAT method for drug-disease association. It includes GAT[12] node feature extraction; Senet feature aggregation and enhancement; and drug-disease association prediction. The workflow of SEGAT is shown in Figure 2.

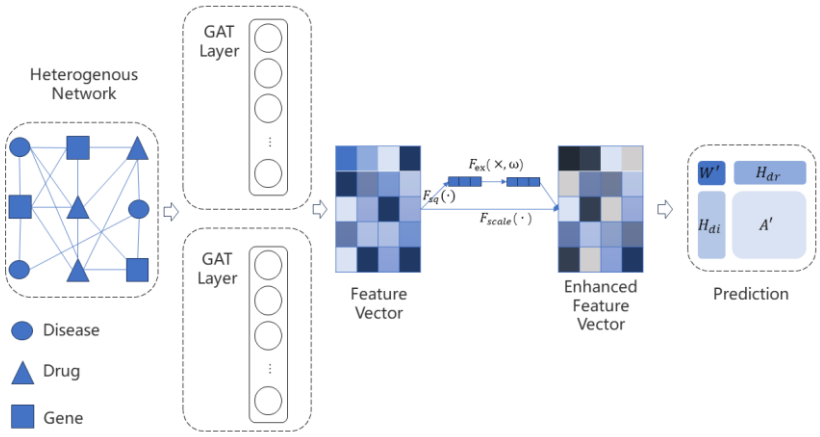

Figure 2: General model architecture

3.1 Graph attention network

Graph Attention Network GAT was proposed by Veličković[12] et al. in 2017 and the main idea is to apply the attention mechanism to graph structures. The core of GAT is the graph attention layer, which takes as input a set of node features $h = \{\vec{h}_1, \vec{h}_2, \dots, \vec{h}_N\}, \vec{h}_i \in \mathbb{R}^F$ as inputs, where N denotes the number of nodes and F denotes the node feature dimensions, and then outputs the new node feature F^' representations by going through the graph attention layer. After obtaining the new feature representation, a shared attention mechanism a: R^(F^')×R^(F^')→R is applied to obtain the attention coefficients:

$$e_{i,j} = a(W\vec{h}_i, W\vec{h}_j) \tag{8}$$

The attention coefficient $e_{i,j}$ indicates the importance of node j's features for node i. To make the attention coefficients of different nodes more interpretable, they are normalized using softmax.

$$a_{ij} = softmax(e_{i,j}) \tag{9}$$

After obtaining the normalized attention coefficients, the representation of the neighboring nodes is applied to the node and the features of the node are updated. After that, the features of the neighboring nodes are averaged by doing a weighting process and activated using a nonlinear function σ.

$$\vec{h}_i' = \sigma \sum_{j \in N_i} a_{ij} W \vec{h}_j \tag{10}$$

From this we can get the features output from the GAT layer.

3.2 Feature Enhancement

In this module, we hope to devise a way to enhance the model's focus on important features. We capture the contribution of each channel's features to the original signal through self-learning, and then enhance important features and suppress unimportant or even useless features based on the contribution of each channel to the original signal Chengdu. This method is also known as the principle of feature recalibration. The obtained $\vec{h}_i'$ is subjected to Squeeze operation to compress the features in spatial dimension.

$$a_{\text{spatial}} = F_{sq}(\vec{h}_i') \tag{11}$$

After learning the importance on the spatial dimension, we learn the importance of the channels through the Excitation operation to get the weights between different channels. The formula is shown below:

$$a_{channel} = F_{ex}(a_{\text{spatial}}, W) \tag{12}$$

Finally, the SEnet enhancement of the original features is accomplished by weighting the channel features through the Scale operation, which treats the output weights of Excitation as an important percentage of each channel.

$$h = F_{scale}(\vec{h}_i', a_{channel}) \tag{13}$$

3.3 Prediction

After graph attention and SEnet, we finally obtain the drug-disease embedding vector $\begin{vmatrix} H_{dr} \\ H_{di} \end{vmatrix}$. In the experiments of this paper, we use a bilinear inner product decoder to construct the association matrix between drug-disease.

$$H' = softmax(H_{dr} W' H_{di}) \tag{14}$$

where $W' \in R^{M \times N}$ is a trainable weight matrix for the association prediction score between drugs and diseases determined by the corresponding (i, j). Each element of $H'_{i,j}$ denotes the association score between drug i and disease j.

4. Experimentation

4.1 Evaluation Metrics

In order to objectively and effectively evaluate the accuracy of this experiment, we use ten times cross-validation to reduce the errors caused by data problems, and we use several indicators including AUC, AUPR, F1 score, and recall rate to evaluate the performance of the model as comprehensively as possible. The calculation formula is as follows:

$$Precision = \frac{TP}{TP+FP} \tag{15}$$

$$Recall = \frac{TP}{TP+FN} \tag{16}$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{17}$$

Where TP is true positive, indicating correctly predicted drug-disease relationships; FP is false positive, indicating incorrectly predicted drug-disease relationships; FN is false negative, indicating incorrectly predicted but actually labeled drug-disease relationships; and TN is positive negative, indicating correctly predicted unlabeled drug-disease relationships.

4.2 Analysis of experimental results

4.2.1 Comparison with other algorithms

We compare this experiment with other drug-disease association algorithms to demonstrate the effectiveness of our experiment. We conduct a comparative experiment between the method of this paper and the methods of Kang[13] and Chen[14] on the same dataset. As shown in Table 1, the experimental method used in this paper possesses an accuracy rate of 98.5%, which is far superior to

other methods, indicating that the method in this paper can more accurately make a better prediction of whether a drug can treat a disease.

Table 1: Comparison of method performance

Methodology	Acc	F1 score	recall rate
Kang [12]	0.866	0.3106	**0.408**
Chen[13]	0.3476	**0.3623**	0.3782
this text	**0.9850**	0.1989	0.265

4.2.2 Case Study

In order to validate the ability of this experiment in discovering new drug-disease associations, known drug-disease associations are obtained through biological databases such as Drugbank and CTD to train the model of this experiment, which is utilized to predict new drug-disease associations. We validated this through approved clinical trial studies and public literature (Table 2). For example, Etomidate[15] (Etomidate), a white powdery substance insoluble in water, is one of the commonly used drugs for induction of anesthesia and has been in clinical use for 30 years, with a rapid but short-lived action, fast sleep onset and awakening, and strong depressant effects on the central nervous system. Phenytoin[16] (Phenytoin) is mainly used for antiepileptic, antiarrhythmic, by highly selective inhibition of the cerebral cortex motor area. It is considered as the drug of choice for the treatment of grand mal and partial seizures.

Table 2: Top ten drug-disease associations predicted in this experiment

Drug	Disease	Evidence
Etomidate	Memory Disorders	PMID: 20180861
Phenytoin	Dermatomyositis	Literature[17]
Doxorubicin	Heart Failure	Literature[18]
Cefadroxil	Stevens-Johnson Syndrome	PMID: 25811541
Bicalutamide	Melanoma	Literature[19]
Nifedipine	Chemical and Drug Induced Liver Injury	Literature[20]
Rosiglitazone	Obesity	Literature[21]
Cimetidine	Hypertension	Literature[22]
Docetaxel	Chemical and Drug Induced Liver Injury	Literature[23]

In addition, to further test the validity of our model, we examined the top five disease candidates for Sulfasalazine and the top five drug candidates for HIV, a drug used to treat inflammatory bowel disease. Tables 3 and 4 show the results of our experiments, which can be confirmed according to some public literature and clinical medical studies.

Table 3: Top 5 Disease Candidates for Sulfasalazine

Drug	Disease	Evidence
Sulfasalazine	Pulmonary Fibrosis	Literature[24]
	Carcinoma, Hepatocellular	Literature[25]
	Pneumonia	Literature[26]
	Coma	PMID: 32850263
	Pruritus	PubMedID：6146502

Table 4: Top 5 Drug Candidates for HIV Infections

Disease	Drug	Evidence
HIV Infections	Ethanol	Literature[27]
	Maraviroc	OMIM:609423
	Dronabinol	PMID: 15308739
	Aplaviroc	PMID: 15644495
	Terameprocol	OMIM:609423

Based on Table 3 and Table 4, we can see that the present model can help to identify new drug-disease associations.

5. Conclusions

In this paper, we developed a GASEDDA model to discover drug-disease associations. The drug-disease associations were successfully predicted by combining the graph attention mechanism and SEnet attention mechanism through a heterogeneous network consisting of drug-drug similarity, disease-disease similarity and gene-gene similarity. From the results, it can be seen that the method of this model is superior to other drug-disease association prediction methods.

In future research, considering that biological network is a huge and interconnected large-scale network, we will consider adding more biological attribute networks, such as proteins, drug targets and so on. Secondly, although GAT is a powerful graph neural network method that can effectively extract the node information in the network, it loses the structural information in the network, and in the future, we hope to solve this problem by methods such as graph embedding.

References

[1] YAZDANIAN M, BRIGGS K, JANKOVSKY C, et al. The "High Solubility" Definition of the Current FDA Guidance on Biopharmaceutical Classification System May Be Too Strict for Acidic Drugs [J]. Pharmaceutical Res earch,, 2004, 21(2):293-299.

[2] HAY M, THOMAS D W, CRAIGHEAD J L, et al. Clinical development success rates for investigational drugs [J]. Nature biotechnology, 2014, 32(1):4051.

[3] GRABOWSKI, HENRY. Are the economics of pharmaceutical research and development changing? [J]. PharmacoEconomics, 2004, 22(2): 15-24.

[4] DUDLEY J T, DESHPANDE T, BUTTE A J. Exploiting drug-disease relationships for computational drug repositioning [J]. Briefings in Bioinformatics, 2011, 12(4): 303-311.

[5] QIAO Z. The research on drug repositioning algorithm by multiple information integration [J]. Hunan University, 2018.

[6] GLOECKNER C, GARNER A L, MERSHA F, et al. Repositioning of an existing drug for the neglected tropical disease Onchocerciasis [J]. Proceedings of the National Academy of Sciences, 2010, 107(8): 3424-3429.

[7] IORIO F, BOSOTTI R, SCACHERI E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses [J]. Proceedings of the National Academy of Sciences, 2010, 107(33): 14621-14626.

[8] WISHART D S, FEUNANG Y D, GUO A C, et al. DrugBank 5.0: a major update to the DrugBank database for 2018 [J]. Nucleic Acids Res, 2018, 46(D1): D1074-D1082.

[9] KIM S, CHEN J, CHENG T, et al. PubChem 2023 update [J]. Nucleic Acids Research, 2023, 51(D1): D1373-D1380.

[10] KUHN M, CAMPILLOS M, LETUNIC I, et al. A side effect resource to capture phenotypic effects of drugs [J]. Molecular Systems Biology, 2010, 6(1).

[11] DAVIS A P, GRONDIN C J, JOHNSON R J, et al. The Comparative Toxicogenomics Database: update 2019 [J]. Nucleic Acids Res, 2019, 47(D1): D948-D954.

[12] Xiu-Jie X U .The quality control in the research and development process of pharmaceutical[J].Heilongjiang Science, 2016.

[13] HONGYU K, QIN L, JIAO L, et al. Drug-Disease Association Prediction Based on Multi-Feature Fusion [J]. Chinese Journal of Biomedical Engineering, 2023.8 42 . 4

[14] HAO C, YUFANG Q, MING C. Collaborative filtering based on graph neural network for drug-disease association prediction [J]. Chinese Journal of Medical Physics, 2023.6, 40.6.

[15] CAVALCANTI B C, DO AMARAL VALENTE S á L G, DE ANDRADE NETO J B, et al. Etomidate is devoid of genotoxicty and mutagenicity in human lymphocytes and in the Salmonella typhimurium/microsomal activation test [J]. Toxicology in Vitro, 2020, 68.
[16] WANG Z, WANG X, WANG Z, et al. Potential herb-drug interaction risk of thymoquinone and phenytoin [J]. Chemico-Biological Interactions, 2022, 353.
[17] OKAMURA K, SUZUKI T, NOHARA K. Gestational arsenite exposure augments hepatic tumors of C3H mice by promoting senescence in F1 and F2 offspring via different pathways [J]. Toxicology and Applied Pharmacology, 2020, 408.
[18] KARHU S T, KINNUNEN S M, T öLLI M, et al. GATA4-targeted compound exhibits cardioprotective actions against doxorubicin-induced toxicity in vitro and in vivo: establishment of a chronic cardiotoxicity model using human iPSC-derived cardiomyocytes [J]. Archives of Toxicology, 2020, 94(6): 2113-2130.
[19] DASGUPTA P, KULKARNI P, BHAT N S, et al. Activation of the Erk/MAPK signaling pathway is a driver for cadmium induced prostate cancer [J]. Toxicology and Applied Pharmacology, 2020, 401.
[20] SHIMIZU Y, SASAKI T, YONEKAWA E, et al. Association of CYP1A1 and CYP1B1 inhibition in in vitro assays with drug-induced liver injury [J]. Chemical and Drug Induced Liver Injury, 2021, 46.
[21] ALFARHAN M W, AL-HUSSAINI H, KILARKAJE N. Role of PPAR-γ in diabetes-induced testicular dysfunction, oxidative DNA damage and repair in leptin receptor-deficient obese type 2 diabetic mice [J]. Chemico-Biological Interactions, 2022, 361.
[22] HUO C-J, YU X-J, SUN Y-J, et al. Irisin lowers blood pressure by activating the Nrf2 signaling pathway in the hypothalamic paraventricular nucleus of spontaneously hypertensive rats [J]. Toxicology and Applied Pharmacology, 2020, 394.
[23] LLEWELLYN H P, VAIDYA V S, WANG Z, et al. Evaluating the Sensitivity and Specificity of Promising Circulating Biomarkers to Diagnose Liver Injury in Humans [J]. Toxicological Sciences, 2021, 181(1): 23-34.
[24] LI G, XU Q, CHENG D, et al. Caveolin-1 and Its Functional Peptide CSP7 Affect Silica-Induced Pulmonary Fibrosis by Regulating Fibroblast Glutaminolysis [J]. Toxicological Sciences, 2022, 190(1): 41-53.
[25] YE L, ZHANG X, WANG P, et al. Low concentration triphenyl phosphate fuels proliferation and migration of hepatocellular carcinoma cells [J]. Environmental Toxicology, 2022, 37(10): 2445-2459.
[26] MALAVIYA R, A B, E A, et al. Pulmonary injury and oxidative stress in rats induced by inhaled sulfur mustard is ameliorated by anti-tumor necrosis factor-α antibody [J]. Toxicology and Applied Pharmacology 2021, 428.
[27] ANDERSON S M, NAIDOO R N, RAMKARAN P, et al. OGG1 Ser326Cys polymorphism, HIV, obesity and air pollution exposure influences adverse birth outcome susceptibility, within South African Women [J]. Reproductive Toxicology, 2018, 79: 8-15.