A Gene-disease Association Prediction Algorithm Based on Multi-source Data Fusion

Fei Wang

Information Management Center
Hohhot Vocational College
Hohhot 010010, China
E-mail: 200810143@hhvc.edu.cn

Received: August 30, 2021 **Accepted: February 24, 2022**

Published: March 31, 2022

Abstract: *Accurate gene-disease association prediction results are the basis for effective diagnosis and treatment of complex genetic diseases. However, existing studies related to this topic generally face problems in two aspects: large volume of original data and diverse data type, and data fusion difficulty. Therefore, this paper studied a gene-disease association prediction algorithm based on multi-source data fusion. At first, it processed the multi-dimensional gene phenotype data, analyzed the gene-disease associations of different phenotypes, and completed the selection of disease gene loci under multi-dimensional phenotypes. Then, this paper fused the multi-source data containing the gene expression data, gene sequence data, gene interaction data, and transcriptome sequencing data, and established the corresponding gene-disease association prediction model. At last, the effectiveness of the constructed prediction model was verified by experimental results. The research results obtained in this paper can improve the low utilization of gene datasets, restored the main features of the datasets to the greatest extent, reasonably processed the data noise, effectively enhanced the robustness of the model, and further improved the classification accuracy of the prediction of disease-causing genes.*

Keywords: *Gene-disease association prediction, Multi-dimensional phenotype, Multi-source data fusion.*

Introduction

Most of the known diseases in humans are related to genes [6, 11, 13, 20, 22, 27]. Until now, the pathogenic genes of nearly 1/4 of human genetic diseases have been found, but still, a lot of the pathogenic genes haven't been discovered yet [4, 7, 10, 14, 18, 23, 24]. Digging out valuable and accurate information of disease-causing genes from massive biological genetic data has now become an important task for biologists [2, 5, 15, 17, 21]. However, almost all studies on gene-disease association are based on experiments, and the high experimental costs, the non-repeatable feature of the experiments, and the low accuracy of experimental results have increased the difficulties of relevant research [1, 9, 16, 25]. Accurate gene-disease association prediction results are the basis for effective diagnosis and treatment of complex genetic diseases; therefore, innovative prediction methods are of great significance for the development of scientific research in the field of genomics.

Because the identification and association of genes and diseases need to conduct time-consuming and expensive experiments on the large number of potential candidate genes, Sikandar et al. [19] proposed a DisGeNET-based method for rapidly calculating gene-disease association data and identifying disease-related candidate genes, also, based on TP (true positive) rate, FP (false positive) rate, precision, recall rate, F1-measure, and ROC (receiver operating characteristic curve) curve, they evaluated the parameters and used 10-fold cross-validation to evaluate different calculation methods. Since the product proteins of genes

usually work together to achieve specific functions, they are widely used to predict disease genes by analyzing the relationship between the known disease genes and other genes in the network. Luo et al. [12] developed an integrated algorithm for predicting disease genes based on the clinical sample network, the algorithm can construct a single-sample network for each case sample of the study disease, and merge these single-sample networks into multiple fusion networks according to the clustering results of the samples, thereby realizing to use the central features extracted from fusion networks to train the logistic model, and use integrated strategies to predict the final probability of each gene related to the disease. In terms of small sample problems, the statistical methods and intelligent machine learning methods cannot obtain convergent gene set when sorting the biomarkers. Jiang et al. [8] designed a new generative adversarial network model, taking denoising autoencoder as generator and multilayer perceptron as discriminator, the predicted residuals are back-propagated to the decoder part of the DAE, thereby the capture probability distribution could be modified, moreover, in this study, they further designed a framework for predicting disease genes using RNA sequence data. Frasca et al. [3] proposed that unbalanced perception integration is a key requirement for improving the performance of gene priority sequencing method. In order to support the proposed viewpoint, in the paper, an integrated algorithm of imbalance perception was proposed for the research problem, and was compared with other latest integrated methods based on benchmark data. The identification of disease genes is a key step in revealing disease pathology and systematically analyzing polygenic diseases. Zhao and Lin [26] divided network-based disease gene prediction methods into three types: methods based on disease gene information, methods combining with phenotypic similarity, and methods integrating multiple results from multiple data sources into one final result.

After reviewing and summarizing existing research results in recent years, we found that although certain achievements have been made in the aspect of gene-disease association prediction, still, the prediction performance needs to be further improved. The data source of biological data is relatively wide, the data noise caused by human and equipment factors is unavoidable, and however, conventional data processing methods can hardly cope with this problem. At the same time, existing studies related to this topic generally face problems in two aspects: large volume of original data and diverse data type, and data fusion difficulty; moreover, the complex disease pathogenesis has also increased the difficulty of prediction. In view of these problems, this paper studied a gene-disease association prediction algorithm based on multi-source data fusion. In this paper, the second chapter processed the multi-dimensional gene phenotype data, analyzed the gene-disease associations of different phenotypes, and completed the selection of disease gene loci under multi-dimensional phenotypes. The third chapter fused the multi-source data containing the gene expression data, gene sequence data, gene interaction data, and transcriptome sequencing data, and established the corresponding gene-disease association prediction model. The fourth chapter employed experimental results to verify the effectiveness of the constructed prediction model. The research results obtained in this paper have improved the low utilization of gene datasets, restored the main features of the datasets to the greatest extent, reasonably processed the data noise, effectively enhanced the robustness of the model, and further improved the classification accuracy of the prediction of disease-causing genes.

Selection of disease gene loci under multi-dimensional phenotypes

With the continuous development of experimental technology, massive multi-dimensional gene phenotype data have been produced, and how to make full use of these data has become a critical work. Research shows that, usually, the effect of gene-disease association prediction using one-dimensional phenotype data is unsatisfactory, due to gene losses, the error rate is

high and the prediction result is unreliable, the method can only realize good prediction performance in terms of some certain types of disease.

To this end, this paper processed the data of multi-dimensional gene phenotypes, and analyzed the gene-disease associations between different phenotypes. Assuming: q represents the number of samples; A_1, A_2, …, A_w represent the target gene loci to be tested; B_1, B_2, …, B_s represent gene phenotypes; then, it's denoted as $A_i = (a_{1i}, a_{2i}, \ldots, a_{qi})^{\mathrm{T}}$ and $B_j = (b_{1j}, b_{2j}, \ldots, b_{qj})^{\mathrm{T}}$, wherein $i = 1, 2, \ldots, w$, and $j = 1, 2, \ldots, s$; and there's a corresponding relationship between the gene loci to be tested and the gene phenotypes:

$$O: \begin{bmatrix} a_{11} & \cdots & a_{1w} \\ \vdots & \ddots & \vdots \\ a_{q1} & \cdots & a_{qw} \end{bmatrix} \rightarrow \begin{bmatrix} b_{11} & \cdots & b_{1s} \\ \vdots & \ddots & \vdots \\ b_{q1} & \cdots & b_{qs} \end{bmatrix}. \tag{1}$$

Assuming: α_v represents the vector of the coefficient of association of the v-th phenotype, then, for each single phenotype B_v in the genetic relationship, there is:

$$\begin{bmatrix} b_{1v} \\ b_{2v} \\ \vdots \\ b_{(q-1)v} \\ b_{qv} \end{bmatrix} = \begin{bmatrix} g_v(a_{11}, a_{12}, \ldots, a_{1w}; \alpha_v) \\ g_v(a_{21}, a_{22}, \ldots, a_{2w}; \alpha_v) \\ \vdots \\ g_v(a_{(q-1)1}, \ldots, a_{(q-1)w}; \alpha_v) \\ g_v(a_{q1}, a_{q2}, \ldots, a_{qw}; \alpha_v) \end{bmatrix} + \begin{bmatrix} \sigma_{1v} \\ \sigma_{2v} \\ \vdots \\ \sigma_{(q-1)v} \\ \sigma_{qv} \end{bmatrix} \quad v = 1, 2, \ldots, s. \tag{2}$$

The genotype probabilities $WC(\sigma_{iv}) = 0$ and $WC(\sigma^2_{iv}) = \rho_{vv}{}^2$ are not related to each other. It is denoted as $A_i = (A_{i1}, A_{i2}, \ldots, A_{iw})$, then the regression equation satisfies $WC(B_v|A) = g_v(A;\alpha_v)$. For different gene phenotypes, the measurement results are different, however, since this paper adopted the data of multidimensional phenotypes of genes, the experiments assumed that the errors between different gene phenotypes could have certain correlations, that is, the different gene phenotypes can satisfy:

$$\sigma_{(i)} = \left(\sigma_{1i}, \ldots, \sigma_{qi}\right)^{\mathrm{T}}, \ \mathrm{cov}\left(\sigma_{(i)}, \sigma_{(j)}\right) = \rho_{ij} I \quad i, j = 1, 2, \ldots, s. \tag{3}$$

In the generalized additive model that the nonlinear relationship could be fitted, a single gene sample could be decomposed as shown in Formula 4:

$$b_{iv} = g_{v1}(a_{i1}) + g_{v2}(a_{i2}) + \ldots + g_{vv}(a_{iv}) = \sum_{j=1}^{n} g_{vj}(a_{ij}) + \sigma_{iv} \tag{4}$$

$$i = 1, 2, \ldots, q; \ v = 1, 2, \ldots, s.$$

The random errors are denoted as $\sigma_{iv} = (\sigma_{1v}, \ldots, \sigma_{qv})^{\mathrm{T}}$, then, the relationship between a single phenotype and multiple phenotypes of the target gene can be expressed as:

$$B_v = \sum_{v=1}^{5} g_{vj}\left(A_j\right)_+ \sigma_v, \quad v = 1, 2, \ldots, s$$
$$\left[b_1|b_2|\cdots|b_s\right] = \left[\sum_{j=1}^{w} g_{1_j}\left(A_j\right)\left|\sum_{j=1}^{w} g_{2j}\left(A_j\right)\right|\cdots\left|\sum_{j=1}^{w} g_{sj}\left(A_j\right)\right| + \left[\sigma_1|\sigma_2\right]\cdots|\sigma_s\right]. \tag{5}$$

When the phenotype of the target gene is uncertain, the genotype probability of the target gene locus to be tested can be analyzed. Assuming: H_{ij} represents the possible genotype of the j-th gene locus of the i-th gene sample, here, it's set that H_{ij} has three types 0, 1, and 2, that is, when H_{ij} = 0, it means genotype *cc*; when H_{ij} = 1, it means genotype *Cc*; when H_{ij} = 2, it means genotype *CC*; *EX*(*) represents the indicator function; r_{ij0}, r_{ij1}, and r_{ij2} respectively represent the probability that the corresponding gene locus to be tested takes genotype *cc*, *Cc*, and *CC*; then, calculation formula of the genotype probability of the gene locus to be tested is:

$$r_{ijl} = WC\left[EX\left(H_{ij} = l\right)\right]. \tag{6}$$

r_{ij0}, r_{ij1}, and r_{ij2} need to satisfy:

$$r_{ij1} + r_{ij2} + r_{ij3} = 1 \quad (i = 1, 2, 3, \ldots, q;\ j = 1, 2, \ldots, w). \tag{7}$$

The genotype probability of the j-th gene locus of the i-th gene sample can be described by (r_{ij0}, r_{ij1}, r_{ij2}). Assuming: e represents the three gene model types of gene loci; dominant gene, additive gene, and recessive gene respectively correspond to e = 0, e = 1, and e = 2; ω_{el} represents the influence coefficient of the l-th genotype under the e-th gene model, wherein, the additive genes are genes whose effects are cumulative and won't produce interactive effects (dominant and epistatic). This paper chose to characterize uncertain gene loci using their expected values:

$$A_{ij} = WC\left[EX\left(H_{ij}\right)\right] = \sum \omega_{eij} EX\left(H_{ij} = l\right) r_{ijl} = \omega_{eij}{}^{\mathrm{T}} EX_{ij}, \tag{8}$$

where $\omega_{eij} = (\omega_{e0}, \omega_{e1}, \omega_{e2})^{\mathrm{T}}$, $EX_{ij} = [EX/(H_{ij} = 0)r_{j0}, EX/(H_{ij} = 1)r_{ij1}, EX/(H_{ij} = 2)r_{ij2}]^{\mathrm{T}}$.

The influence coefficients of the gene models corresponding to dominant gene, additive gene, and recessive gene satisfy:

$$\left(\omega_0, \omega_1, \omega_2\right) = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{9}$$

The gene phenotype data have two types: discrete type, and continuous type. Assuming: b_i represents the result of the joint action of w gene loci to be tested of the i-th gene sample; A represents the designed matrix composed of the j-th gene locus of the i-th sample, and $A_i = (A_{il}, A_{i2}, \ldots, A_{iw})$; a_{ij} represents the genotype that is relatively determined; σ_i represents the i-th random error; $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_k)$ represents the coefficient vector. The regression equation $WC(B|A) = g(A;\alpha)$, it satisfies $WC(\sigma_i) = 0$ and $WC(\sigma_i^2) = \rho^2$ and the two are not correlated with each other. Eq. (10) gives the results presented by the gene model based on continuous gene phenotype data:

$$\begin{bmatrix} b_{1v} \\ b_{2v} \\ \vdots \\ b_{iv} \\ \vdots \\ b_{(q-1)v} \\ b_{qv} \end{bmatrix} = \begin{bmatrix} g_v\left(a_{11}, a_{12}, \ldots, a_{1w}; \alpha_v\right) \\ g_v\left(a_{21}, a_{22}, \ldots, a_{2w}; \alpha_v\right) \\ \vdots \\ g_v\left(\omega_{e11}^{\mathrm{T}} EX_{11}, \ldots, \omega_{eiv}^{\mathrm{T}} EX_{iv}; \alpha_v\right) \\ \vdots \\ g_v\left(A_{(q-1)1}, \ldots, A_{(q-1)v}; \alpha_v\right) \\ g_v\left(A_{q1}, A_{q2}, \ldots, A_{qv}; \alpha_v\right) \end{bmatrix} + \begin{bmatrix} \sigma_{1v} \\ \sigma_{2v} \\ \vdots \\ \sigma_{iv} \\ \vdots \\ \sigma_{(q-1)v} \\ \sigma_{qv} \end{bmatrix} \tag{10}$$

Based on $(B_i, A_i)^q{}_{i\text{-}1}$, the coefficient vector α could be estimated, and the probability distribution of each gene locus could be further obtained, as shown in Table 1.

Table 1. The probability distribution of each locus

Trait *B*	$\boldsymbol{b_1}$	$\boldsymbol{b_2}$	**…**	$\boldsymbol{b_q}$
$\boldsymbol{A_1}$	$(r_{110},\ r_{111},\ r_{112})$	$(r_{210},\ r_{211},\ r_{212})$	…	$(r_{q10},\ r_{q11},\ r_{q12})$
$\boldsymbol{A_2}$	$(r_{120},\ r_{121},\ r_{122})$	$(r_{220},\ r_{221},\ r_{222})$	…	$(r_{q20},\ r_{q21},\ r_{q22})$
⋮	⋮	⋮		⋮
$\boldsymbol{A_w}$	$(r_{1w0},\ r_{1w1},\ r_{1w2})$	$(r_{2w0},\ r_{2w1},\ r_{2w2})$	…	$(r_{qw0},\ r_{qw1},\ r_{qw2})$

If there are many target gene loci to be tested, the constructed genotype uncertainty model needs to be converted into the primary function form. At this time, the number of model parameters increases and the dimension of the designed matrix increases as well, this is because the existence of the cumulative effect will weaken the prediction effect.

In view of the above problems, this paper chose to shrink the insignificant coefficients to achieve the dimensionality reduction of the designed matrix. Least Absolute Shrinkage and Selection Operator (LASSO) regression is a shrinkage regression method, which could obtain a more refined model by building a penalty function, this paper applied this method to optimize the problem of disease gene locus selection:

$$\begin{aligned} &\min \sum_{i=1}^{q}\left(b_i - \sum_{j=1}^{w}\sum_{l=0}^{r}\alpha_{jl} Q_{lo}\left(a_{ij}^{*}; \theta_l\right)\right)^2 \\ &s.t. \sum_{j=1}^{w}\sum_{l=0}^{r}\left|\alpha_{jl}\right| \le D \end{aligned} \tag{11}$$

Further, the corresponding unconstrained penalty term could be constructed, and Eq. (12) gives its expression:

$$\min\left(\sum_{i=1}^{q}\frac{1}{2}\left(b_i-\sum_{j=1}^{w}\sum_{l=0}^{r}\alpha_{jl}Q_{lo}\left(a_{ij}^{*};\omega_l\right)\right)^2+\xi\sum_{j=1}^{w}\sum_{l=0}^{r}\left|\alpha_{jl}\right|\right)$$

$$\hat{\alpha}=\arg\min_{\alpha}\left(\sum_{i=1}^{q}\frac{1}{2}\left(b_i-\sum_{j=1}^{w}\sum_{l=0}^{r}\alpha_{jl}Q_{lo}\left(a_{ij}^{*};\omega_l\right)\right)^2+\xi\sum_{j=1}^{w}\sum_{l=0}^{r}\left|\alpha_{jl}\right|\right)$$

$$\frac{\partial}{\partial\alpha}\left(\sum_{i=1}^{q}\frac{1}{2}\left(b_i-\sum_{j=1}^{w}\sum_{l=0}^{r}\alpha_{jl}Q_{lo}\left(a_{ij}^{*};\omega_l\right)\right)^2+\xi\sum_{j=1}^{w}\sum_{l=0}^{r}\left|\alpha_{jl}\right|\right)=0 \tag{12}$$

$$\frac{\partial}{\partial\alpha}\left(\frac{1}{2}(B-Q\alpha)^{\mathrm{T}}(B-Q\alpha)+\xi 1|\alpha|\right)=0$$

$$q\xi\,\mathrm{sign}(\alpha)=Q^{\mathrm{T}}(B-Q\alpha)$$

$$M(\alpha)=\left\|B-Q\alpha\right\|_2+\sum_{i=1}^{w(r+1)}\left|\alpha_{jl}\right|=h(\alpha)+\sum_{il}f_i\left(\alpha_i\right)$$

Let:

$$M(\alpha)=\|\,B-Q\alpha\|_2+\sum_{i=1}^{w(r+1)}\left|\alpha_{jl}\right|=h(\alpha)+\sum f_i\left(\alpha_i\right). \tag{13}$$

For any α^l, there is:

$$\begin{aligned}&M\left(\alpha^l\right)-M\left(\alpha^{l-1}\right)=h\left(\alpha^l\right)+f\left(\alpha^l\right)-\left(h\left(\alpha^{l-1}\right)+f\left(\alpha^{l-1}\right)\right)\\&\geq\nabla h\left(\alpha^l\right)\left(\alpha^l-\alpha^{l-1}\right)+\sum f_i\left(\alpha_i^l\right)-f_i\left(\alpha_i^{l-1}\right)\\&=\sum\nabla_i h\left(\alpha^l\right)\left(\alpha_i^l-\alpha_i^{l-1}\right)+f_i\left(\alpha_i^l\right)-f_i\left(\alpha_i^{l-1}\right)\geq 0.\end{aligned} \tag{14}$$

Therefore, in order to obtain a numerical solution close to the optimal solution, the $M(\alpha)$ could be minimized based on the coordinate descent method to find a point α_0 that is close to the minimum value in all dimensional coordinates, the specific iteration process is elaborated as follows:

Step 1: Set the initial point according to Eq. (15):

$$\alpha^0=\left(\alpha_{01}^{(0)},\,\alpha_{11}^{(0)},\,\ldots,\,\alpha_{r1}^{(0)},\,\ldots,\,\alpha_{0w}^{(0)},\,\ldots,\,\alpha_{rw}^{(0)}\right). \tag{15}$$

Step 2: Starting from the l-th iteration, fix $\alpha_{0,1}^{(l)}$ parameter, and calculate the $\alpha_{0,1}$ when $M(\alpha)$ reaches the minimum value. Then, for the $wr + w - 1$ parameters after $\alpha_{0,1}^{(l)}$, repeat the above operation for $w(r + 1)$ times until $\alpha_i^{(l)}$:

$$\begin{aligned}
\alpha_{01}^{(l)} &= \underset{\alpha_{01}}{\operatorname{argmin}} M\left(\alpha_{01}, \alpha_{11}^{(l-1)}, \ldots, \alpha_{r1}^{(l-1)}, \ldots, \alpha_{0w}^{(l-1)}, \ldots, \alpha_{rw}^{(l-1)}\right) \\
\alpha_{11}^{(l)} &= \underset{\alpha_{11}}{\operatorname{argmin}} M\left(\alpha_{01}^{(l)}, \alpha_{11}, \ldots, \alpha_{r1}^{(l-1)}, \ldots, \alpha_{0w}^{(l-1)}, \ldots, \alpha_{rw}^{(l-1)}\right) \\
&\ldots \\
\alpha_{r1}^{(l)} &= \underset{\alpha_{r1}}{\operatorname{argmin}} M\left(\alpha_{01}^{(l)}, \alpha_{11}^{(l)}, \ldots, \alpha_{r1}, \ldots, \alpha_{0w}^{(l-1)}, \ldots, \alpha_{rw}^{(l-1)}\right) \\
&\ldots \\
\alpha_{rw}^{(l)} &= \underset{\alpha_{rw}}{\operatorname{argmin}} M\left(\alpha_{01}^{l}, \alpha_{11}^{(l)}, \ldots, \alpha_{r1}^{(l)}, \ldots, \alpha_{0w}^{(l)}, \ldots, \alpha_{rw}\right).
\end{aligned} \tag{16}$$

The gene-disease association prediction model based on multi-source data fusion

Using multiple types of data to form complementary information and jointly predict gene-disease associations can realize multiple screening of disease-causing genes. After analysis, if a gene exhibits a same trait in multiple types of data, then the analysis result is often reliable. Measuring gene-disease associations from the perspective of multi-source data will make the biological information of genes more reliable.

Studies discovered that, within a period of life activities, genes with similar phenotypes tend to have same functions or pathogenic characteristics. Genes with similar phenotypes in the gene interaction network also tend to form densely related functional modules, therefore, the various data of genes with similar phenotypes are not independent of each other. Thus, the works in the previous section, namely the selection of disease gene loci based on multi-dimensional phenotypes is very important for multi-source data fusion, and it plays an especially important role in the common annotation of multi-source and high-quality gene-disease association data.

The multi-source data used in this study mainly include gene expression data, gene sequence data, gene interaction data, and transcriptome sequencing data. They can provide comprehensive gene function information or disease-causing information; however, it should be noted that, the final results are affected by factors such as such as information noise, experimental technology, and the limitation of gene ontology databases, and this will cause some unknown and reliable information to be missed during the prediction of gene-disease associations.

Since the gene sample traits are characterized by phenotypic variables, it's assumed that, after the joint action of w gene loci to be tested of the i-th sample, the result of this sample shows that, when it has the disease, $b_i = 1$; when it hasn't the disease, $b_i = 0$. Based on logistic regression, this paper constructed the relationship between the genotypes of w gene loci to be tested and the sample trait b, let $b_1^* = \ln((r(b_1 = 1)/1 - r(b_1 = 1))$, then, there is:

$$\begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{q-1}^* \\ b_q^* \end{bmatrix} = \begin{bmatrix} g_1(a_{11}) & g_2(a_{12}) & \cdots & g_{w-1}(a_{1(w-1)}) & g_w(a_{1w}) \\ g_1(a_{21}) & g_2(a_{22}) & \cdots & g_{w-1}(a_{1(w-1)}) & g_w(a_{2w}) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ g_1(a_{(q-1)1}) & g_2(a_{(q-1)2}) & \cdots & g_{w-1}(a_{(q-1)(w-1)}) & g_w(a_{(q-1)w}) \\ g_1(a_{q1}) & g_2(a_{q2}) & \cdots & g_{w-1}(a_{q(w-1)}) & g_w(a_{qw}) \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{(w-1)} \\ \delta_w \end{bmatrix}. \tag{17}$$

Let $r_{ij} = \max\{r_{ij0}, r_{ij1}, r_{ij2}\}$, then the relationship between the genotypes of w gene loci to be tested and the sample trait b can be expressed as:

$$\begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{q-1}^* \\ b_q^* \end{bmatrix} = \begin{bmatrix} g_1\left(\max\left(r_{11}^1, r_{12}^1, r_{13}^1\right)\right) & g_2\left(\max\left(r_{11}^2, r_{12}^2, r_{13}^2\right)\right) & \cdots & g_w\left(\max\left(r_{11}^w, r_{12}^w, r_{13}^w\right)\right) \\ g_1\left(\max\left(r_{21}^1, r2_{12}^1, r_{23}^1\right)\right) & g_2\left(\max\left(r_{21}^2, r2_{12}^2, r_{23}^2\right)\right) & \cdots & g_w\left(\max\left(r_{21}^w, r2_{12}^w, r_{23}^w\right)\right) \\ \vdots & \vdots & \ddots & \vdots \\ g_1\left(\max\left(r_{q1}^1, r2_{q2}^1, r_{q3}^1\right)\right) & g_2\left(\max\left(r_{q1}^2, r2_{q2}^2, r_{q3}^2\right)\right) & \cdots & g_w\left(\max\left(r_{q1}^w, r2_{q2}^w, r_{q3}^w\right)\right) \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_{(w-1)} \\ \delta_w \end{bmatrix} \tag{18}$$

Based on above analysis, the genotype attribution of each gene locus to be tested could be obtained. According to the expected value characterization method described in the previous section, the corresponding genotype could be obtained, and its probability could be calculated:

$$QW\left(a_{ij}\right) = r_{ij1} + 2r_{ij2}\,. \tag{19}$$

To facilitate calculation, the $QW(a_{ij})$ in the above formula was replace by a_{ij}^*, at this time, the dimension of the genotype data was qr. This paper conducted B-spline processing on genotype data, assuming the number of primary functions is l, the corresponding genotype data dimension is equal to l^*qr, at this time, the relationship between the spline function and the spline set is given by Eq. (20):

$$\begin{cases} w_1\left(a_{i1}\right) = s_{11}\gamma_{1o}\left(a_{i1}\right) + \ldots + s_{1r}\gamma_{r,o}\left(a_{i1}\right) \\ w_2\left(a_{i2}\right) = s_{21}\gamma_{1o}\left(a_{i2}\right) + \ldots + s_{2r}\gamma_{r,o}\left(a_{i2}\right),\ i = 1, \ldots,\ q \\ \vdots \\ w_w\left(a_{iv}\right) = s_{w1}\gamma_{1d}\left(a_{iv}\right) + \ldots + s_{wr}\gamma_{r,o}\left(a_{iv}\right) \end{cases} \tag{20}$$

Eq. (21) gives the expression of the *logit* function in the *logistic* regression at this time:

$$\begin{bmatrix} b_1^* \\ b_2^* \\ \vdots \\ b_{q-1}^* \\ b_q^* \end{bmatrix} = \begin{bmatrix} w_1\left(a_{BS-1,1}^*\right) & w_2\left(a_{n1,2}^*\right) & \cdots & w_r\left(a_{n1,r}^*\right) \\ w_1\left(a_{n2,1}^*\right) & w_2\left(a_{n2,2}^*\right) & \cdots & w_r\left(a_{n2,r}^*\right) \\ \vdots & \vdots & \ddots & \vdots \\ w_q\left(a_{nq,1}^*\right) & w_q\left(a_{nq,2}^*\right) & \cdots & w_r\left(a_{nq,r}^*\right) \end{bmatrix}. \tag{21}$$

When taking two points on the dimensional coordinates, under cubic spline processing, there are 5 primary functions $\{\gamma_{13}, \gamma_{23}, \gamma_{33}, \gamma_{43}, \gamma_{53}\}$, and Eq. (22) gives the expression of γ_{i3}:

$$\gamma_{13}(a)=\begin{cases} n_1(a-a_1)^3, a\in[a_1,a_2] \\ n_1(a-a_1)^3+n_2(a-a_2)^3, a\in[a_2,a_3] \\ n_1(a-a_1)^3+n_2(a-a_2)^3+n_3(a-a_3)^3, a\in[a_3,a_4] \\ n_1(a-a_1)^3+n_2(a-a_2)^3+n_3(a-a_3)^3+n_4(a-a_4)^3, a\in[a_4,a_5] \\ 0, \text{ others} \end{cases} \tag{22}$$

Eq. (23) gives the expression of the constructed primary function:

$$\{\gamma_{13}(a), \gamma_{23}(a), \gamma_{33}(a), \gamma_{43}(a), \gamma_{53}(a)\}. \tag{23}$$

Then, in case of 5 primary functions, the cubic spline can be described by:

$$\begin{cases} w_1(a_{11})=\delta_{11}\gamma_{13}(a_{11})+\delta_{21}\gamma_{23}(a_{11})+\ldots+\delta_{51}\gamma_{53}(a_{11}) \\ w_2(a_{11})=\delta_{12}\gamma_{13}(a_{12})+\delta_{22}\gamma_{23}(a_{12})+\ldots+\delta_{52}\gamma_{53}(a_{12}) \end{cases} \tag{24}$$

Then, at this time, the model of the gene locus to be tested a_{ij} could be constructed as:

$$b_1^*=w_1(a_{11})+w_2(a_{12})+\ldots+w_r(a_{1r})=\sum_{i=1}^{5}\delta_{i1}\gamma_{i3}(a_{11})+\ldots+\delta_{ir}\gamma_{i3}(a_{1r}). \tag{25}$$

Similarly, b_2^* and b_3^* could be calculated.

Through above operations, the $a_{BS\text{-}ij}^*$ after B-spline processing and the logistic regression function of parameter δ can be obtained. For the possible genotypes of the j-th gene locus of the i-th gene sample, this paper introduced the B-spline curve to convert the logistic regression model into the form shown as:

$$\log itGV(B=1)=\sum_{j=1}^{r}w_j(a_j)\approx\sum_{j=1}^{r}\left[g_{j1}(a_j)\delta_{j1}+\ldots g_{jw}(a_j)\delta_{jw}\right]. \tag{26}$$

$g_{jr}(a_j)$ was replaced by $C_{jr}=QW\cdot g_{jr}(a_j)=g_{jr}(0)r_{j0}+g_{jr}(1)r_{j1}+g_{jr}(2)r_{j2}$, then there is:

$$\log itGV(B=1)=\sum_{j=1}^{r}\left[C_{j1}(a_j)\delta_{j,1}+\ldots+C_{jw}(a_j)\delta_{jw}\right]. \tag{27}$$

Since the relationship between the gene data and the phenotype that characterizes the whether the disease exists is non-linear, this paper used likelihood estimation to estimate the parameters of this non-linear relationship, let $r=r(B=1|\omega; a)$, Eq. (28) gives the expression of the *logstic* regression model at this time:

$$\ln\frac{r}{1-r}=\alpha_0+w_1(a_1)+\ldots+w_r(a_r). \tag{28}$$

Solve above formula to get:

$$\begin{aligned} & t^{\ln\frac{r}{1-r}} = t^{\alpha_0+w_1(a_1)+\ldots+w_r(a_r)} \\ & \Rightarrow \frac{r}{1-r} = t^{\alpha_0+w_1(a_1)+\ldots+w_r(a_r)} \\ & \Rightarrow \frac{r}{1-r} - 1 = t^{\alpha_0+w_1(a_1)+\ldots+w_r(a_r)} \\ & \Rightarrow r = \frac{t^{\alpha_0+w_1}(a_1)+\ldots+w_r(a_r)}{1+t^{\alpha_0+w_1(a_1)+\ldots+w_r(a_r)}} \end{aligned} \tag{29}$$

The probability function can be expressed as:

$$\begin{cases} r(b_i=1)=\theta_i \\ r(b_i=0)=1-\theta_i \end{cases} \tag{30}$$

Eq. (31) gives the combined written form of probability function b_i:

$$M(b_i) = \theta_i^{b_i}\left[1-\theta_i\right]^{1-b_i},\ (b_i = 0, 1;\ i = 1, \ldots, q) \tag{31}$$

The likelihood function b_1, b_2, …, b_q can be expressed as:

$$SR = \prod_{i=1}^{q} \theta^{b_i}\left(1-\theta^{1-b_i}\right) \tag{32}$$

Take the logarithm of the above formula:

$$\ln SR = \sum_{i=1}^{q}\left[b_i \ln\theta_i + (1-b_i)\ln(1-\theta_i)\right] = \sum_{i=1}^{q}\left[b_i \ln\frac{\theta_i}{(1-\theta_i)} + \ln(1-\theta_i)\right] \tag{33}$$

Combining Eq. (29) with above formula, then there is:

$$\therefore \ln SR = \sum_{i=1}^{q}\left(b_i\left(\alpha_0 + w_1(a_1)+\ldots+w_r(a_r)\right) - \ln\left(1+t^{\alpha_0+w_1(a_1)+\ldots+w_r(a_r)}\right)\right] \tag{34}$$

Based on the maximum likelihood estimation, a set of estimated values α_0, α_1, …, α_r was selected to realize the maximization of the regression model.

Experimental results and analysis

The gene datasets adopted in the experiments came from the association data of homologous genes and phenotypes of other species, and the source of disease features came from web pages about disease in Online Mendelian Inheritance in Man (OMIM). The disease similarity network had been used in the related datasets, using the parts of anatomy and disease in Medical Subject Headings (MeSH), relevant terms could be automatically extracted from OMIM entries. For each record, a feature vector was generated, the similarity between diseases was obtained by calculating the cosine angle after each eigenvector was normalized, and a network containing 5080 disease phenotypes, the MimMiner, was generated.

The simulation was realized in the R version 4.0.0 environment. At first, some population sample data were generated through simulation, after obtaining the three frequencies of a single genotype, genotypes that are relatively certain need to be filled in using the "dose method", and at last, the response variable was generated.

Due to the existence of error term, the simulation results have certain randomness. This study used 280 samples and 41 gene loci to simulate the association models of 3 sets of gene-disease data. The B-spline processing method of the genotype data was the cubic B-spline processing method. The normalized gene sample data were divided into several intervals: [0, 1/4], [1/4, 1/2], [1/2, 3/4], [3/4, 1]; and correspondingly, a total of 11 nodes {0, 0, 0, 0, 1/4, 1/2, 3/4, 1, 1, 1, 1} were set. Table 2 shows the genotype filling measurement for the missing genes, then, after genotype filling based on above table, the genome data were substituted into the model.

Table 2. Genotype filling

	Gene locus	A_1	A_2	A_3	A_4
Genotype	0	0.931	0.154	0.535	0.329
	1	0.032	0.103	0.218	0.423
	2	0.029	0.743	0.256	0.206
	SOFT CALL	0.101	1.588	0.721	0.840

For data sample 1, a preset normal distribution disturbance term was added in each model, and then the penalty coefficient was properly selected to minimize the mean square error.

Fig. 1 shows the mean square error of the prediction of phenotype B_1 under different penalty coefficients; and Fig. 2 shows the mean square error of the prediction of phenotype B_2 under different penalty coefficients in the independent cross-validation.

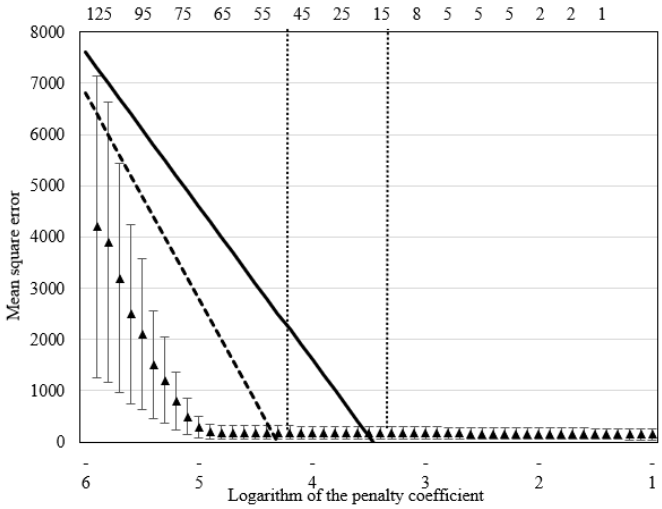

Fig. 1 The mean square error of the prediction of phenotype B_1 under different penalty coefficients

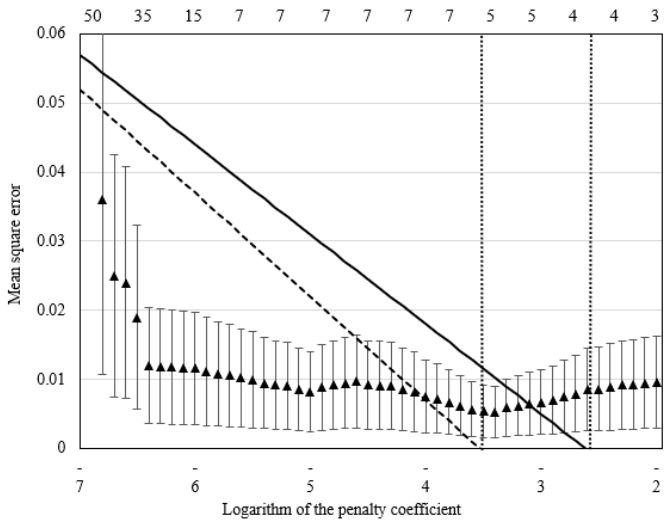

Fig. 2 The mean square error of the prediction of phenotype B_2 under different penalty coefficients in the independent cross-validation

After completing the filling of the missing genotypes, based on the model constructed in this paper, 298 primary functions were generated; the primary functions corresponding to the pathogenic genes were screened by LASSO regression, and the gene locus was determined. The simulation work, and the calculation of prediction accuracy, were repeated for 30 times, 50 times and 100 times. Table 3 lists the prediction accuracy of gene-disease association.

Table 3. Accuracy of gene-disease association prediction

Related phenotype		B_1	B_2	B_3	B_4	B_5
Accuracy	30 times	96.04%	96.66%	79.55%	69.56%	70.52%
	50 times	95.23%	96.76%	83.21%	83.53%	78.61%
	100 times	95.41%	96.37%	85.78%	85.81%	83.83%

According to the Table 3, after the gene samples were subject to above changes and screening operations, the probability that the corresponding disease-causing gene locus could be accurately associated and predicted was relatively ideal. After 100 times of simulations, under strong disturbance conditions, the accuracy of gene-disease association prediction still reached more than 83%, indicating that the model constructed in this paper can realize the screening of multiple gene-disease association loci in the multi-dimensional phenotype data.

According to the basic assumption of genotype uncertainty, this paper analyzed the multi-source sample data (Table 4). Based on the characteristics of the multi-source data, the Rstudio was employed to generate the final ultra-high-dimension gene data, and the screening of diseased and non-diseased related gene loci was realized through the screening of covariates. In order to obtain gene locus data of appropriate dimensions, the experiment conducted in this study selected 25 covariates, and gene data samples of 1,000, 1,500, and 2,000 dimensions; the risk allele frequency was set to 0.1. Then, values were assigned to δ_{jw}, δ_1 and δ_2 took the value of 1, and δ_w took the value of 0. In this way, this paper was able to derive the corresponding gene phenotype and the regression model that is closer to the real situation based on the binomial distribution of genotype probability, then, the variable selection was conducted using the Smoothly Clipped Absolute Deviation (SCAD) method, and whether the gene variables corresponding to δ_j under current coefficients can be screened out or not was verified.

Table 4. The labeling of multi-source sample data

Sample / Gene locus	1	2	…	1000
r_1	0	0	…	2
r_2	0	1	…	1
…	…	…	…	…
r_{30}	2	1	…	0

Table 5 shows the contribution of all gene loci to be tested of phenotype *B*.

Table 5. Contribution degree of gene loci to the disease

Phenotype	b_1	b_2	…	b_q
A_1	(0.88，0.19，0.02)	(0.23，0.75，0)	…	(0.74，0.07，0.15)
A_2	(0.65，0.31，0)	(0.12，0.88，0)	…	(0，0，1)
⋮	⋮	⋮		⋮
A_w	(0.94，0.05，0.01)	(0，1，0)	…	(0.42，0.53，0)

The calculated genotype probability and the corresponding genotype are given in Table 6 and Table 7, respectively.

Table 6. Calculation results of genotype probability

Phenotype *B*	b_1	…	b_q
A_1	0.90	…	0.81
A_2	0.63	…	1
⋮	⋮		⋮
A_w	0.96	…	0.54

Table 7. Genotype situation

Genotype / Genotype probability	*cc*	*Cc*	*CC*
A_1	0.91		
A_2	0.62		
A_w	0.96		
A_{q1}	0.81		
A_{q2}			1
A_{qw}		0.52	

After B-spline processing, the expected gene data $a_{BS\text{-}ij}^*$ could be obtained. The constructed matrix with a dimension of $5\times qr$ is given in Table 8.

Because the gene sample data were ultra-high-dimensional multi-source data that need to be reduced to a suitable dimension, this paper screened the variables under the condition of uncertain genotypes based on the SCAD method, and Table 9 gives the corresponding screening results. As can be seen from the table, the recognition accuracy of the SCAD method under different sample sizes could meet the requirement.

Table 8. Genotype data

	1	**2**	**3**	**…**	**1000**
W_1	0	0	1	…	0
W_2	0	0	0	…	0
W_3	0	0	0	…	0
W_4	0	0.512	0	…	0
W_5	0	0.476	0	…	0
W_7	0	0.08	0	…	0
…	…	…	…	…	…
W_{q-2}	0	0.513	0.574	…	0
W_{q-1}	1	0.496	0.451	…	0
W_q	0	0.02	0.05	…	0

Table 9. Screening results of gene loci of different samples under the condition of uncertain genotypes

Primary function	W_1	W_3	W_6	W_8
Corresponding sample gene locus	A_{q1}	A_{q1}	A_{q1}	A_{q2}
100	-2.078564	-3.853121	-4.272313	-9.321546
1000	0	0	-3.953213	-0.101213
2000	-1.653213	0	-5.231316	-12.863242
3000	-1.321312	-7.595653	-3.213545	-8.565646
Primary function	W_{11}	W_{143}	W_{144}	**…**
Corresponding sample gene locus	A_{q2}	A_{q21}	A_{q21}	…
100	-1.084532	0	0	0
1000	-0.6142351	-7.134652e-10	-9.654312e-14	0
2000	0	0	0	0
3000	-1.235464	0	0	0

According to above table, two gene loci corresponding to non-zero coefficients could be screened out. Compared with the 3 gene loci set by the real model, the screening accuracy was higher. The reason for the poor screening effect of the remaining gene locus was the too-small set value of the coefficient, and the coefficients of other gene loci in the table tended to be close to 0, which can be ignored.

This paper chose to optimize the constructed model through cross-validation. Fig. 3 shows the cross-validation numerical simulation results of the prediction of phenotype B_2. Through optimization processing, a hyperparameter value could be found so that the model can achieve the optimal generalization performance and realize variable screening. As for the specific real models, this paper determined the coefficients and assumed that the first three gene loci significantly correlated with the diseased and non-diseased gene phenotypes.

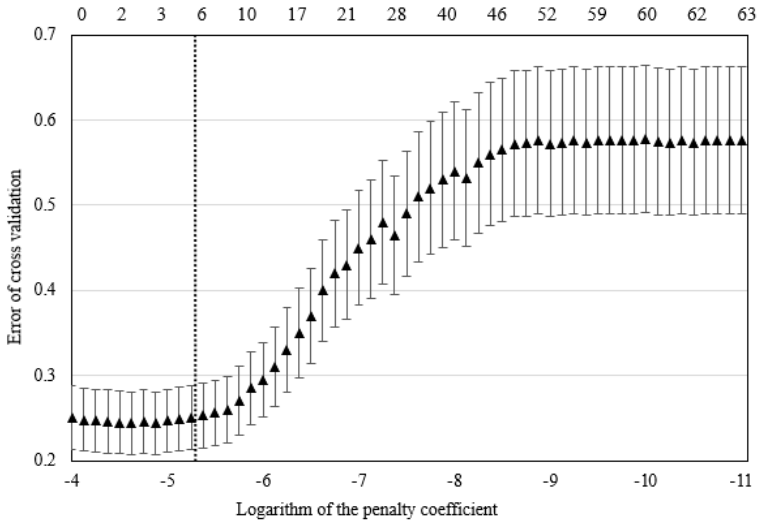

Fig. 3 Cross-validation numerical simulation of the prediction of phenotype B_2

The screened two gene loci corresponding to non-zero coefficients were subject to significance analysis, and the test results of the correlation coefficients are shown in Table 10.

Table 10. Test results of correlation coefficients of significance analysis

	Intercept term	A_1^*	A_2^*
Estimated value	1.2345	0.7812	1.4542
Standard error	0.4012	0.2742	0.2924
Z-value	3.2564	2.564	5.001
P-value	0.00214	0.01045	5.76e-06

According to the Table 10, gene locus A_1^* and gene locus A_2^* were significantly correlated with the intercept term, and the significance of A_2^* was higher. If the p-value obtained through the test is less than 0.05, then the coefficient can be judged to be significant, that is, the judgement of rejecting the original hypothesis could be made.

Conclusion

This paper studied a gene-disease association prediction algorithm based on multi-source data fusion. The innovation of this paper is to give an analysis on the gene-disease association between different phenotypes, and complete the selection of disease gene loci. At the same time, the paper also fused multi-source data including gene expression data, gene sequence data, gene interaction data and transcriptome sequencing data, and constructed the corresponding gene-disease association prediction model. Experimental results gave the mean square error of the prediction of phenotype B_1 under different penalty coefficients, and the mean square error of the prediction of phenotype B_2 under different penalty coefficients in independent cross-validation; also the prediction accuracy of gene-disease association was calculated, which had verified that the model constructed in this paper can realize the screening of multiple gene-disease association loci in multi-dimensional phenotype data. At last, this paper summarized the screening results of different samples under the condition of uncertain genotype, optimized the constructed model through cross-validation, gave test results of the correlation coefficients of the significance analysis, and the performance advantages of the optimized prediction model had been verified.

References

1. Al-Aamri A., K. Taha, Y. Al-Hammadi, M. Maalouf, D. Homouz (2019). Analyzing a Co-occurrence Gene-interaction Network to Identify Disease-gene Association, BMC Bioinformatics, 20(1), 1-15.
2. Chen X., Q. Huang, Y. Wang, J. Li, H. Liu, Y. Xie, Z. Li (2020). A Deep Learning Approach to Identify Association of Disease-gene Using Information of Disease Symptoms and Protein Sequences, Analytical Methods, 12(15), 2016-2026.
3. Frasca M., J. F. Fontaine, G. Valentini, M. Mesiti, M. Notaro, D. Malchiodi, M. A. Andrade-Navarro (2017). Disease-genes Must Guide Data Source Integration in the Gene Prioritization Process, International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, 60-69.
4. Grewal N., S. Singh, T. Chand (2016). Effect of Aggregation Operators on Network-based Disease Gene Prioritization: A Case Study on Blood Disorders, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(6), 1276-1287.
5. He M., C. Huang, B. Liu, Y. Wang, J. Li (2021). Factor Graph-aggregated Heterogeneous Network Embedding for Disease-gene Association Prediction, BMC Bioinformatics, 22(1), 1-15.
6. Hocker J. D., O. B. Poirion, F. Zhu, J. Buchanan, K. Zhang, J. Chiou, S. Preissl (2021). Cardiac Cell Type-specific Gene Regulatory Programs and Disease Risk Association, Science Advances, 7(20), eabf1444.
7. Jayanthi K., C. Mahesh, A. Arthi, K. T. Rajendran, B. Vijayalakshmi, N. R. Shanker (2021). Early Detection of Pediatric Cardiomyopathy Disease Using Window Based Correlation Method from Gene Micro Array Data, Journal of Sensors, Vol. 2021, Article ID 1055770, https://doi.org/10.1155/2021/1055770.
8. Jiang X., J. Zhao, W. Qian, W. Song, G. N. Lin, (2020). A Generative Adversarial Network Model for Disease Gene Prediction with RNA-Seq Data, IEEE Access, 8, 37352-37360.
9. Kambakam S., M. N. Ngaki, B. B. Sahu, D. R. Kandel, P. Singh, R. Sumit, M. K. Bhattacharyya (2021). Arabidopsis Non-host Resistance Pss30 Gene Enhances Broad-spectrum Disease Resistance in the Soybean Cultivar Williams 82, The Plant Journal, 107(5), 1432-1446.
10. Kawichai T., A. Suratanee, K. Plaimas (2021). Meta-path Based Gene Ontology Profiles for Predicting Drug-disease Associations, IEEE Access, 9, 41809-41820.
11. Korashy H. M., I. M. Attafi, K. S. Famulski, S. A. Bakheet, M. M. Hafez, A. M. Alsaad, A. R. M. Al-Ghadeer (2017). Gene Expression Profiling to Identify the Toxicities and Potentially Relevant Human Disease Outcomes Associated with Environmental Heavy Metal Exposure, Environmental Pollution, 221, 64-74.
12. Luo P., L. P. Tian, B. Chen, Q. Xiao, F. X. Wu (2020). Ensemble Disease Gene Prediction by Clinical Sample-based Networks, BMC Bioinformatics, 21(2), 1-12.
13. Mileva M., L. Dimitrova, M. Popova, V. Bankova, D. Krastev, H. Najdenski, Z. Zhelev, I. Aoki, R. Bakalova-Zheleva (2021). Redox-modulation, Suppression of "Oncogenic" Superoxide and Induction of Apoptosis in Burkitt's Lymphoma Cells Using *Geum urbanum* L. Extracts, Int J Bioautomation, 25(4), 315-330.
14. Narayan S., Z. Liew, J. M. Bronstein, B. Ritz (2017). Occupational Pesticide Use and Parkinson's Disease in the Parkinson Environment Gene (PEG) Study, Environment International, 107, 266-273.
15. Nguyen H. T., T. T. Phan, T. C. Dao, T. M. N. Phan, P. V. D. Ta, C. N. T. Nguyen, H. X. Huynh (2021). Gene Family Abundance Visualization Based on Feature Selection Combined Deep Learning to Improve Disease Diagnosis, Journal of Engineering & Technological Sciences, 53(1), 1-17.

16. Rath S. N., M. Patri (2020). Understanding miRNA Based Gene Regulation in Parkinson's Disease: An *in silico* Approach, Int J Bioautomation, 24(1), 15-28.
17. Ray S., S. M. M. Hossain, L. Khatun, A. Mukhopadhyay (2017). A Comprehensive Analysis on Preservation Patterns of Gene Co-expression Networks during Alzheimer's Disease Progression, BMC Bioinformatics, 18(1), 1-21.
18. Shi H., T. Xue, Y. Yang, C. Jiang, S. Huang, Q. Yang, X. Ye (2020). Microneedle-mediated Gene Delivery for the Treatment of Ischemic Myocardial Disease, Science Advances, 6(25), eaaz3621.
19. Sikandar M., R. Sohail, Y. Saeed, A. Zeb, M. Zareei, M. A. Khan, E. M. Mohamed (2020). Analysis for Disease Gene Association Using Machine Learning, IEEE Access, 8, 160616-160626.
20. Subbaiah K. C. V., O. Hedaya, J. Wu, F. Jiang, P. Yao (2019). Mammalian RNA Switches: Molecular Rheostats in Gene Regulation, Disease, and Medicine, Computational and Structural Biotechnology Journal, 17, 1326-1338.
21. Tsuchiya A., J. H. Kang, T. Mori, Y. Naritomi, S. Kushio, T. Niidome, Y. Katayama (2017). Efficient Delivery of Signal-responsive Gene Carriers for Disease-specific Gene Expression via Bubble Liposomes and Sonoporation, Colloids and Surfaces B: Biointerfaces, 160, 60-64.
22. Vasighizaker A., S. Jalili (2018). C-PUGP: A Cluster-based Positive Unlabeled Learning Method for Disease Gene Prediction and Prioritization, Computational Biology and Chemistry, 76, 23-31.
23. Verma P., A. Srivastava, C. V. Srikanth, A. Bajaj (2021). Nanoparticle-mediated Gene Therapy Strategies for Mitigating Inflammatory Bowel Disease, Biomaterials science, 9(5), 1481-1502.
24. Yotova G., S. Lazarova, V. Mihaylova, T. Venelinov (2021). Water Quality Assessment of Surface Waters and Wastewaters by Traditional and Ecotoxicological Indicators in Ogosta River, Bulgaria, Int J Bioautomation, 25(1), 25-40,
25. Zhang Z., K. L. Running, S. Seneviratne, A. R. Peters Haugrud, A. Szabo-Hever, G. Shi, J. D. Faris (2021). A Protein Kinase-major Sperm Protein Gene Hijacked by a Necrotrophic Fungal Pathogen Triggers Disease Susceptibility in Wheat, The Plant Journal, 106(3), 720-732.
26. Zhao J., L. M. Lin (2017). A Survey of Disease Gene Prediction Methods Based on Molecular Networks, J Univ Electron Sci Technol, 46, 755-765.
27. Zhou J., B. Q. Fu (2018). The Research on Gene-disease Association Based on Text-mining of PubMed, BMC Bioinformatics, 19(1), 1-8.

Assoc. Prof. Fei Wang
E-mail: 200810143@hhvc.edu.cn

Fei Wang has graduated from Institute of Information Engineering, Inner Mongolia University of Technology, and now serves as a Deputy Director and an Associate Professor of Computer Science at Information Management Center, Hohhot Vocational College. His research directions include computer science and technology, and computer network. Assoc. Prof. Fei Wang participated in one municipal scientific research project, published 5 papers, and co-edited one textbook.