


A deep learning method for predicting metabolite–disease associations via graph neural network

Feiyue Sun, Jianqiang Sun and Qi Zhao 

Corresponding author. Qi Zhao, School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China.
E-mail: zhaoqi@lnu.edu.cn

Abstract

Metabolism is the process by which an organism continuously replaces old substances with new substances. It plays an important role in maintaining human life, body growth and reproduction. More and more researchers have shown that the concentrations of some metabolites in patients are different from those in healthy people. Traditional biological experiments can test some hypotheses and verify their relationships but usually take a considerable amount of time and money. Therefore, it is urgent to develop a new computational method to identify the relationships between metabolites and diseases. In this work, we present a new deep learning algorithm named as graph convolutional network with graph attention network (GCNAT) to predict the potential associations of disease-related metabolites. First, we construct a heterogeneous network based on known metabolite–disease associations, metabolite–metabolite similarities and disease–disease similarities. Metabolite and disease features are encoded and learned through the graph convolutional neural network. Then, a graph attention layer is used to combine the embeddings of multiple convolutional layers, and the corresponding attention coefficients are calculated to assign different weights to the embeddings of each layer. Further, the prediction result is obtained by decoding and scoring the final synthetic embeddings. Finally, GCNAT achieves a reliable area under the receiver operating characteristic curve of 0.95 and the precision-recall curve of 0.405, which are better than the results of existing five state-of-the-art predictive methods in 5-fold cross-validation, and the case studies show that the metabolite–disease correlations predicted by our method can be successfully demonstrated by relevant experiments. We hope that GCNAT could be a useful biomedical research tool for predicting potential metabolite–disease associations in the future.

Keywords: metabolite, disease, metabolite–disease associations, graph attention network, graph convolutional network

Introduction

During long-term evolution, biological organisms interact with the surrounding environment to absorb and excrete material and energy, which is called metabolism. It plays a vital role in the process of material and energy change as an important life activity of organisms. A growing number of biological and medical experiments show that some patients have different concentrations of certain metabolites than those in healthy people [1]. Deoxycholic acid is a secondary bile acid produced by the liver, which is recirculated through the liver, bile ducts, small intestine and portal vein, forming the entero-hepatic circuit. They are strongly toxic in their anionic form at physiological pH and thus require a carrier for transport across the membranes of the gut and liver tissues. When deoxycholic acid is present at sufficiently high levels, it can act as a hepatotoxin, a metabolic toxin and a tumor metabolite [2]. Liver toxins can cause damage to the liver or liver cells. When chronically high, it promotes tumor growth and survival [3]. In addition to being associated with liver disease, chronically high

levels of deoxycholic acid are also associated with several forms of cancer, including colon cancer [4], breast cancer [5] and many other gastrointestinal cancers [6]. Besides, the pathogenesis of cardiovascular and cerebrovascular diseases [7] and some immune diseases [8–10] have also been confirmed to be associated with metabolites. Metabolite-based disease diagnosis is an important judgment in medical diagnosis. For multifactorial diseases such as metabolic syndrome or degenerative diseases, there is also a need to identify disease-related metabolites as biomarkers for clinical research. Traditional biological experiments can propose hypotheses and test the relationships between metabolites and diseases, but these experiments are often time-consuming and labor-intensive. Therefore, it is necessary to construct effective computational models that complement biological experimental processes to predict the potential associations between metabolites and diseases.

During recent years, many kinds of research such as long non-coding RNA (lncRNA)–protein interaction [11], microRNA (miRNA)–lncRNA interaction [12, 13],

Feiyue Sun is a graduate student in the University of Science and Technology Liaoning. Her research interests include bioinformatics and deep learning.

Jianqiang Sun is an associate professor in Linyi University. His research interests include bioinformatics and deep learning.

Qi Zhao is a professor in the University of Science and Technology Liaoning. His research interests include bioinformatics, complex network and machine learning.

Received: April 25, 2022. Revised: June 4, 2022. Accepted: June 6, 2022

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

miRNA–disease associations prediction [14–16] and circular RNA (circRNA)–disease associations prediction [17, 18] have been carried out in bioinformatics. These studies have promoted the development of methods for predicting metabolite–disease associations to a certain extent. At present, the prediction of the relationships between metabolites and diseases mainly adopts optimization and complex network algorithms. In 2018, Hu *et al.* proposed the Random Walk for Metabolite–Disease Association prediction algorithm (RWRMDA) and applied the random walk algorithm to this field for the first time to predict the associations between metabolites and diseases [19]. However, when calculating metabolite similarity, they only considered disease similarity. Later, Lei *et al.* used a KATZ-based method to predict the correlations between metabolites and diseases [20]. This method considered the relevant information between metabolites and diseases but was not applicable to isolated metabolites or diseases in the network. To improve predictive accuracy, Lei *et al.* presented the ABC algorithm based on the espionage strategy to predict the relationships between metabolites and diseases [21]. In the same year, Zhang *et al.* developed a metabolite–disease associations prediction algorithm (LightGBM) based on optical gradient enhancement [22]. This method took results from statistics, graph theoretical measurements and matrix factorization, combined with principal component analysis to remove noise. As an improvement, Ma *et al.* used the hypergraph-based logistic matrix factorization approach to predict the associations between metabolites and diseases [23]. In 2020, Zhao *et al.* presented a graph deep learning-based method, named Deep-DRM, for identifying disease-related metabolites [24]. As far as we know, this is the first computational method using the deep learning method. It applied a graph convolutional network (GCN) to encode features separately. Then, the dimensionality of these features was reduced by principal component analysis. A deep neural network (DNN) was built to derive true metabolite–disease associations. Although there are several algorithms for predicting metabolite–disease relationships, they still have some shortcomings. Network algorithms use relatively small data sets for prediction and are not as effective as deep learning algorithms. Moreover, network algorithms may not be able to better deal with sparse data and data noise contained in the model. Furthermore, computational models with high precision and strong performance for the prediction of potential metabolite–disease relationships are still challenging.

Therefore, we develop a model built by GCN with graph attention network (GAT) (GCNAT) to infer the correlations between metabolites and diseases. First, we construct a heterogeneous network based on known associations information, the Gaussian kernel similarity network for metabolites and the integrated similarity network for diseases. Then, a graph convolutional neural network is used to learn feature embeddings for input graph-structured data. Moreover, the embeddings of

multiple convolutional layers are combined using the GAT to assign different weights to the embeddings of different convolutional layers. Furthermore, by scoring the final composite embeddings, we can employ our model to predict potential associations between metabolites and diseases. We assess the performance of GCNAT in the 5-fold cross-validation (5-fold CV) experiment and the results outperform five previous computational methods. Meanwhile, we conduct case studies based on GCNAT, and the results of the predicted top 10 metabolite–disease pairs can be confirmed in the relevant literature. These results demonstrate that GCNAT is an efficient and feasible model for predicting potential metabolite–disease associations.

Materials and methods

Data sets

The Human Metabolomics Database (HMDB, <https://hmdb.ca/>) is currently the most complete and comprehensive database of human metabolites and metabolism in the world, which contains rich information on small molecule metabolites found in humans [25]. We obtain the correlations between metabolites and diseases from HMDB, and the current version of this database is 5.0. The extracted data set includes diseases with ID of diseases (DOID) and related metabolites information. At last, 216 diseases, 2262 metabolites and 4357 associations between them are screened from such data set.

Metabolite–disease associations

To describe the correlations between metabolites and diseases more intuitively, an adjacency matrix $A (n_m \times n_d)$ is introduced. n_m and n_d are the number of metabolites and diseases, respectively. If disease i has been approved to be associated with metabolite j , then the value of $A(i, j) = 1$, otherwise $A(i, j)$ is 0.

Disease semantic similarity

For diseases, the relationship between them can be described as a directed acyclic graph (DAG). Each disease has one or more tree numbers indicating its position in the DAG. The DAG of disease d can be expressed as $DAG(d) = (d, T(d), E(d))$, where $T(d)$ represents d itself and all ancestor nodes of d and $E(d)$ represents the edge from all parent nodes to child nodes. Here, we choose the Medical Subject Headings (MeSH) descriptor of the disease to construct the DAG of the disease. The MeSH descriptor of the corresponding disease can be obtained from the MeSH database of the National Medical Library (<https://meshb.nlm.nih.gov/>). In the DAG, the semantic value $DV(d)$ of disease d can be defined as $DV(d) = \sum_{n \in N(d)} D_d(n)$, and the semantic contribution value $D_d(n)$ of disease n to disease d is defined as follows:

$$\begin{cases} D_d(n) = 1 & \text{if } n = d \\ D_d(n) = \max \{ \Delta * D_d(n') \mid n' \in \text{children of } d \} & \text{if } n \neq d \end{cases} \quad (1)$$

where Δ is the semantic contribution decay factor, generally defined as 0.5, which affects $D_d(n)$ between the parent node d and its child node n' . $D_d(n)$ is inversely proportional to the distance between diseases d and n [26].

We add the contribution values of all ancestors of the disease to get $D_d(n)$. Through observation, it is found that two diseases with larger DAG sharing parts may have a higher similarity score. Therefore, the semantic similarity score SD between the two diseases d_i and d_j is obtained as:

$$SD(d_i, d_j) = \frac{\sum_{t \in T(i) \cap T(j)} (D_i(t) + D_j(t))}{DV(i) + DV(j)} \quad (2)$$

Disease GIP kernel similarity

In the known metabolite–disease network, similar diseases may have similar patterns of associations. Taking advantage of this feature, we apply the Gaussian function to the topological association network between biological information nodes and use the kernel method to establish a kernel function from the feature vector to achieve the purpose of extracting association features. The Gaussian kernel similarity network GIPD between each disease pair d_i and d_j can be calculated as follows:

$$GIPD(d_i, d_j) = \exp(-\omega_d \|IP(d_i) - IP(d_j)\|^2), \quad (3)$$

where $IP(d_i)$ represents the binary interaction map of disease d_i . In Equation (3), ω_d is the bandwidth that controls the nuclear similarity of Gaussian interaction attributes. ω_d' is used to regularize the Gaussian interaction property kernel similarity bandwidth; here, we set its value as 1, and ω_d is computed as Equation (4):

$$\omega_d = \omega_d' / \left(\frac{1}{n_d} \sum_{i=1}^{n_d} IP(d_i)^2 \right). \quad (4)$$

From the past experiment [27], we learn that adding a logistic function to the obtained disease Gaussian kernel similarity can improve the prediction accuracy. NGD represents the improved disease GIP kernel similarity. The logistic function NGD is shown:

$$NGD(d_i, d_j) = \frac{1}{1 + e^{a \times GIPD(d_i, d_j) + b}}. \quad (5)$$

According to the experiments of related applications [27], it is known that $a = -15$, $b = \log(9999)$ in Equation (5). Through such processing, the association between nodes with greater correlation in the interaction network can be made to be closer to 1, and the association of smaller correlation can be made to be closer to 0.

Integrated similarity for diseases

Not every disease has a corresponding MeSH molecular descriptor, so there will be some sparse data in the

semantic similarity matrix of the diseases; to solve this problem, we combine SD and NGD to get the integrated similarity matrix NSD.

$$NSD(d_i, d_j) = \begin{cases} NGD(d_i, d_j) & \text{if } SD(i, j) = 0 \\ (1 - \mu) SD(d_i, d_j) + \mu GIPD(d_i, d_j) & \text{otherwise} \end{cases}, \quad (6)$$

where μ is a weighted parameter determines the proportion of the two types of disease similarities. It is set as 0.1 according to the previous ref. [20].

Metabolite GIP kernel similarity

We also construct the Gaussian kernel similarity network GIPM of the metabolites in a similar way:

$$GIPM(m_i, m_j) = \exp(-\omega_m \|IP(m_i) - IP(m_j)\|^2) \quad (7)$$

$$\omega_m = \omega_m' / \left(\frac{1}{n_m} \sum_{i=1}^{n_m} IP(m_i)^2 \right), \quad (8)$$

In Equation (7), $IP(m_i)$ represents the binary interaction map of metabolite m_i . In Equation (8), ω_m is the bandwidth that controls the nuclear similarity of Gaussian interaction attributes.

GCNAT

Figure 1 shows the workflow of GCNAT. This method is divided into three steps. First, we obtain metabolite–disease associations from HMDB, screen disease information with DOID and related metabolites and get MeSH descriptors of diseases from the National Library of Medicine. After processing, we use the integrated disease similarity matrix NSD, the metabolite Gaussian kernel similarity matrix GIPM, and the known metabolite–disease correlations to construct a heterogeneous network. We take such a heterogeneous network as the input to GCN and use GCN encoder to learn the features of both metabolite and disease. Then, we put the three convolutional layers of GCN into GAT, using the graph attention mechanism to assign different weights to different embeddings. Finally, we decode the resulting metabolite–disease embeddings to obtain a reconstructed metabolite–disease adjacency matrix and infer potential metabolite–disease relationships.

Graph encoding

GCN is a framework that uses deep learning to learn graph-structured data directly [28]. By mapping the nodes in the graph, the feature information of them and their neighbors can be integrated. This gives a regular expression for each node in the graph and feeds it into a convolutional network to generate a new representation of the nodes. It can effectively utilize graph structure information. The propagation rule of GCN is:

$$H^{(l+1)} = f(H^{(l)}, G) = \sigma(D^{-\frac{1}{2}} G D^{-\frac{1}{2}} H^{(l)} W^{(l)}), \quad (9)$$

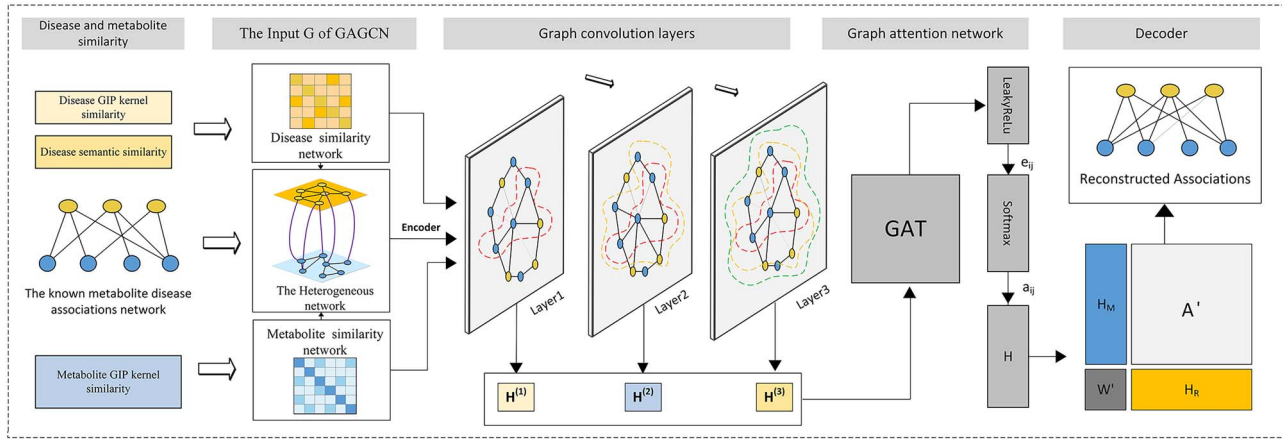


Figure 1. The workflow of GCNAT.

Among them, D is the degree matrix of the input graph, $H^{(l)}$ is the embeddings of the nodes in the l th layer, $W^{(l)}$ is a matrix of layer-specific trainable weights and $\sigma(\cdot)$ is a nonlinear activation function. Here, we use the nonlinear activation function Relu [28]. In our study, considering the contribution of similarity in the propagation process, we set the input graph G as:

$$G = \begin{bmatrix} \lambda \times \text{GIPM} & A \\ A^T & \lambda \times \text{NSD} \end{bmatrix}, \quad (10)$$

G represents the adjacency matrix composed of metabolite similarity GIPM, integrated disease similarity NSD and their correlations. A penalty factor λ is added into the similarity matrix of G to control the contribution value. Then, we initialize the embeddings as:

$$H^{(0)} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}, \quad (11)$$

According to Equation (11), the first layer of the network is:

$$H^{(1)} = \text{ReLU} \left(D^{-\frac{1}{2}} A D^{-\frac{1}{2}} H^{(0)} W^{(0)} \right), \quad (12)$$

$W^{(0)}$ is an input-to-hidden weight matrix. In the subsequent propagation, following the rule in Equation (12), L K -dimensional embeddings can be obtained after L iterations. Through the embeddings of different layers, different structural information of the heterogeneous network can be captured. Considering that the contributions of different embeddings on different layers are also different, we add a graph attention layer (GAT) to our model.

Graph attention network

The GAT is a new type of convolutional neural network that acts on graph-structured data [29]. In the

propagation process, a masked attention mechanism is introduced, and the attention is only allocated to the neighbor nodes set N_i of node i , so that j belongs to N_i . The hidden state of each node is calculated by paying attention to its neighbor nodes. The following are the core formulas of GAT:

$$Z_i^{(l)} = W^{(l)} H_i^{(l)} \quad (13)$$

$$e_{ij} = \text{LeakyReLU} \left(a^{(l)T} \left(Z_i^{(l)} \parallel Z_j^{(l)} \right) \right) \quad (14)$$

$$a_{ij}^{(l)} = \text{softmax}_j(e_{ij}) = \frac{\exp(e_{ij}^{(l)})}{\sum_{k \in N(i)} \exp(e_{ik}^{(l)})} \quad (15)$$

$$H_i^{(l+1)} = \sigma \left[\sum_{j \in N(i)} a_{ij}^{(l)} Z_j^{(l)} \right], \quad (16)$$

$W^{(l)}$ is a weight matrix in Equation (13), $H_i^{(l)}$ is the input of the GAT. In our model, three convolutional layers of GCN are used. To more fully represent the characteristics of the nodes, the nodes need to be linearly transformed. Applying $W^{(l)}$ to each node can obtain the linearly transformed matrix $Z_i^{(l)}$ of the node. $Z_j^{(l)}$ represents the linearly transformed matrix of the neighbor node j . Using the attention mechanism on each node, the attention weight e_{ij} between any two nodes can be calculated by Equation (14). Since a single-layer feedforward neural network is applied here, we use LeakyRelu as the activation function. Then, to make the attention coefficient easier to calculate and compare, we use Equation (15) to introduce Softmax to normalize the attention weights e_{ij} between i and adjacent nodes j to get the final attention coefficient a_{ij} .

Through this method, a_{ij} and its corresponding features are linearly combined, and the comprehensive representation of each node is obtained by summation, so that our model can pay more attention to important nodes and reduce the influence caused by edge noise, thus improving the predictive accuracy of GCNAT.

GCN can realize the encoding of node features, and different convolutional layers can capture different structural information of the heterogeneous network. After combining GAT, we get the final embeddings of metabolites and diseases as H_M and H_D , which are decoded by the Adam [28, 30]. The decoding method is shown in the following formula:

$$A' = \text{sigmoid}(H_M W' H_D^T). \quad (17)$$

Here, W' is the trainable weight matrix, and A' indicates the reconstructed metabolite–disease adjacency matrix.

Cross-entropy loss function

The loss function is used to measure the difference between the value predicted by the model and the true value, and this difference can be back-propagated to update the network parameters. The cross-entropy loss function is often adopted in the last step of classification problems, usually using logistic regression. The so-called logistic regression is a nonlinear activation function sigmoid added to the linear regression. When gradient descent is utilized to solve the problem in the process of backpropagation, if the squared error is applied, the gradient may disappear. Using the cross-entropy loss function is more convenient to obtain the optimal solution, when the parameters are updated. The mathematical expression of this function is:

$$L = - \sum_{i,j=1}^N y^{(ij)} \log \hat{y}^{(ij)} + (1 - y^{(ij)}) \log (1 - \hat{y}^{(ij)}). \quad (18)$$

Considering that the number of negative samples in the data set we use is much larger than that of positive samples, we add parameter γ to Equation (18). The absolute value of γ is obtained by dividing the negative sample y^- and the positive sample y^+ , and then a new weighted cross entropy loss function is obtained. The modified formula is as follows:

$$L = - \frac{1}{N_D \times N_M} \left(\gamma \times \sum_{(i,j) \in y^+} \log \hat{y}^{(ij)} + \sum_{(i,j) \in y^-} \log (1 - \hat{y}^{(ij)}) \right), \quad (19)$$

where (i, j) denotes the pair of disease d_i and metabolite m_j .

Adam optimizer

In the deep learning predictive model, the optimization algorithm is usually used to minimize the loss function. In our study, the weight matrices $W^{(l)}$ and W' are initialized by the Xavier method [31]. In addition, there are some parameters like bias in the model, which are

generally applied to calculate the output value and play an important role in training the neural network model.

Adam optimizer mainly adopts the gradient descent method to update the parameters of the model by finding the minimum value and controlling the variance to finally make the model converge and minimize the loss function [32]. In the optimization process, to prevent overfitting and solve the problem of slow training speed, we add two kinds of dropouts in the convolution layer [33], including node dropout and regular dropout, to randomly delete some neurons to achieve the purpose of training different small networks in different batches.

Results

Performance evaluation

Our model utilizes the metabolite–disease heterogeneous network as the input of GCN for prediction. In the model, the metabolite–disease heterogeneous network is based on an integrated similarity network of diseases, a Gaussian kernel similarity network of metabolites and known metabolite–disease correlations. After the model is built, we adopt a 5-fold CV method to evaluate the performance of GCNAT on the main data set, that is, the known metabolite–disease association pairs are randomly divided into five small sets of equal size. Take each of them as the testing set and the remaining four as the training set. A predictive model is established on the known associations in the training set, and the associations in the testing set are predicted to obtain the corresponding results. Then, as the thresholds in the model are changed, the true positive rate (TPR) and false positive rate (FPR) can be calculated as follows:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (20)$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}. \quad (21)$$

In the above formula, TP and TN represent the number of correctly predicted positive and negative samples. FN and FP indicate the number of falsely predicted negative and positive samples. According to TPR and FPR, the receiver operating characteristic (ROC) curve can be drawn and the corresponding area under the ROC curve (AUC) value can be calculated. In addition to AUC, we also calculate other metrics, including recall, specificity, the precision-recall curve (AUPR), Accuracy (ACC) and F1-score (F1). Owing to the imbalance of positive and negative samples in our data set, the ratio is roughly 1:100, and other metrics can provide more information [34]. The recall represents the probability that positive samples in data sets are correctly predicted as positive. Specificity is indicating the probability that a negative sample is predicted to be negative. AUPR represents the area of the PR curve, and the value of AUPR can represent the predictive effect of the model on an unbalanced data set. ACC is the proportion of correctly classified samples

Table 1. The evaluation metrics of GCNAT by six different penalty factors

Results						
λ	AUPR	AUC	F1	ACC	RE	SP
2	0.3711	0.9410	0.0959	0.9789	0.6651	0.9772
3	0.3540	0.9404	0.0949	0.9782	0.6402	0.9789
4	0.3500	0.9393	0.1012	0.9765	0.6301	0.9801
5	0.3597	0.9390	0.0950	0.9768	0.6595	0.9773
6	0.3590	0.9406	0.8930	0.9736	0.6885	0.9741
7	0.2886	0.8493	0.0811	0.9804	0.5300	0.9812

Table 2. The parameters of GCNAT

Structure	Parameters
Layer	Amount(L):3 Units(K):64 Activation function: ReLu Node dropout(δ): 0.5 Edge dropout(β): 0.5
Loss function	Cross-entropy loss function
Optimizer	Adam
Epoch	4000
Initial learning rate(α)	0.01
Penalty factor(λ)	2

to the total number of samples. Precision represents the accuracy of the predicted accurate results over all results. F1 is the harmonic mean of recall and precision. Their formulas are as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \quad (22)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (23)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (24)$$

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

$$\text{F1} = \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} \quad (26)$$

In our experiment, there are several super parameters, including embedded dimension k , number of layers L , initial learning rate l_s of optimizer and training epochs α , two dropouts (node dropout δ and edge dropout β) and penalty factors λ . We choose δ and β from {0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7}, while the value of α from {500, 1000, 2000, 4000}.

Considering that other parameters are related to the model training process, the penalty coefficient controls the contribution of similarities in the propagation process of GCN. Therefore, we choose six values of penalty factors λ and keep other coefficients unchanged. According to Table 1, GCNAT has higher AUC and AUPR scores when $\lambda = 2$. After fixing the penalty factor, we test combinations of other parameters, and the final parameters of GCNAT are shown in Table 2.

Comparison with previous methods

In this subsection, we compare GCNAT with five current methods for association prediction applied in the bioinformatics domain and test each method on the same data set to highlight the superiority of our approach.

- Random walk (RWR) [19] used a restart random walk approach to predict metabolite–disease associations.
- PageRank [35], also known as the webpage ranking algorithm, sorted the relevance by calculating the weight, to achieve the purpose of prediction.
- KATZ [20] is a network measure that predicted metabolite–disease correlations by measuring the number and length of walks between nodes.
- DNN [36], using a DNN structure, can improve the learning process with more training data. Here, we use the same similarity matrix and heterogeneous network as GCNAT to extract the topology of each metabolite–disease association, thereby training a DNN model and making correlation predictions.
- EKRRMDA [37] proposed a Kernel Ridge Regression algorithm to complete the prediction.

Based on the associations between known metabolites and diseases, our method utilizes a heterogeneous network constructed by integrated similarity of diseases and Gaussian kernel similarity of metabolites and puts it into GCN to extract features, and the GCN is combined with GAT to improve predictive accuracy. We compare GCNAT with the above five algorithms by conducting 5-fold CV, and the results are shown in Figure 2.

From Figure 2, the AUC value of GCNAT is 0.950, while the values of RWR, PageRank, KATZ, DNN and EKRRMDA are 0.646, 0.786, 0.871, 0.906 and 0.938. It is 30.4, 16.4, 7.9, 4.4 and 1.2% higher than other comparison methods, respectively. According to the results, we can see that the predictive accuracy of RWR and PageRank are much lower than that of GCNAT. The reason for the low predictive accuracy of these two models may be that they did not use enough similarity and heterogeneous information in the prediction process. The values of KATZ, DNN and EKRRMDA are also lower than that of GCNAT. Because KATZ and EKRRMDA are both network measurement methods, they use the number and length of walks between nodes as an effective measure of similarity to achieve the purpose of prediction. In the process of feature extraction, GCNAT can better obtain

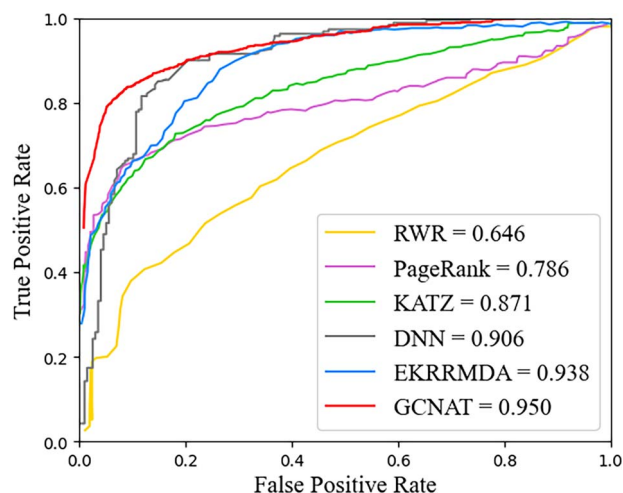


Figure 2. ROC curves of GCNAT and comparison methods on 5-fold CV under the same data set.

high-level statistics of the data through linear and non-linear operations. DNN has an advantage in the depth of the network, but it does not introduce an attention mechanism to make the model pay more attention to the information of adjacent nodes. In order to reflect the result of GCNAT more intuitively, we also give PR curves of GCNAT and comparison methods on 5-fold CV under the same data set in Figure S1. In Figure S1, the AUPR value of DNN is higher than that of GCNAT. This is because we take into account the huge amount of data and the construction of the model when conducting comparative experiments. We partition a sample-balanced validation set to examine the performance of the model. In GCNAT, such a validation set does not exist. From ROC and PR curves, we can see that the performance of our model is better than other five methods.

Case studies

From the above studies, it can be concluded that GCNAT can more accurately predict the correlations between metabolites and diseases. To further prove the accuracy of GCNAT in identifying novel metabolite–disease associations, our model is implemented on case studies of three complex human diseases, namely, Alzheimer’s disease, colorectal cancer and obesity. In the course of the case study, we hide the associations of these three diseases separately, using the remaining known associations to put into our model for feature learning. Then, we predict these three diseases by GCNAT and list the top 10 disease-related metabolites inferred by our model. Alzheimer’s disease is a progressive neurodegenerative disease with insidious onset, which may be a heterogeneous group of diseases. The disease occurs under the action of many factors [38, 39], and its main manifestations are cognitive decline, behavioral disorders and the gradual decline in the ability to daily living. The top 10 metabolites related to Alzheimer’s disease are confirmed by HMDB and recent biological experiments (Table 3). For example, L-asparagine is an amino acid

Table 3. Top 10 potential metabolites associate with Alzheimer’s disease

Alzheimer’s disease			
Rank	Metabolite name	Evidences	PubMed Unique Identifier (PMID)
1	1-Methylhistidine	HMDB0000001	17 031 479
2	2-Hydroxybutyric acid	HMDB0000008	23 857 558
3	Cis-Aconitic acid	HMDB0000072	23 857 558
4	Glycerol 3-phosphate	HMDB0000126	23 857 558
5	Fumaric acid	HMDB0000134	9 693 263
6	L-Asparagine	HMDB0000168	17 031 479
7	L-Histine	HMDB0000177	17 031 479
8	Uridine	HMDB0000288	23 857 558
9	Creatinine	HMDB0000562	23 857 558
10	24-Hydroxycholesterol	HMDB0001419	15 061 359

Table 4. Top 10 potential metabolites associate with colorectal cancer

Colorectal cancer			
Rank	Metabolite name	Evidences	PMID
1	2-Hydroxybutyric acid	HMDB0000008	25 105 552
2	Alpha-ketoisovaleric acid	HMDB0000019	27 275 383
3	Adenosine	HMDB0000050	27 015 276
4	Fumaric acid	HMDB0000134	27 015 276
5	L-Asparagine	HMDB0000168	25 037 050
6	D-mannose	HMDB0000169	27 015 276
7	L-isoleucine	HMDB0000172	20 156 336
8	L-histidine	HMDB0000177	20 156 336
9	Ornithine	HMDB0000214	25 105 552
10	Norepinephrine	—	None

that can be synthesized from central metabolic pathway intermediates in humans. Past experiments have shown that asparagine is detectable in both the serum and saliva of Alzheimer’s patients [40, 41]. In addition, 24-hydroxycholesterol is formed almost exclusively in the brain. In humans, its levels decrease with age. A growing body of evidence points to a potentially important link between it and Alzheimer’s disease [42].

Colorectal cancer is a common malignant tumor of the digestive tract that occurs in the colon or rectum. Its incidence is mainly related to high-fat and low-fiber diets, and chronic colorectal diseases are also an incentive [43, 44]. HMDB records that people with the chronic colorectal disease are more likely to develop cancer than normal people. As shown in Table 4, 9 of the top 10 potential colorectal cancer-associated metabolites validated by GCNAT could be certified. For example, fumaric acid, a dicarboxylic acid, is recently identified as a carcinogenic metabolite or an endogenous carcinogenic metabolite. This organic acid is present in high concentrations in tumors or biological fluids surrounding tumors. Experiments have shown that it can be detected in the paracancerous mucosa and feces of colorectal cancer patients [45].

Obesity is a metabolic disorder. When the human body eats more calories than it consumes, the excess calories

Table 5. Top 10 potential metabolites associate with obesity

Obesity			
Rank	Metabolite Name	Evidences	PMID
1	1-Methylhistidine	HMDB0000001	15 899 597
2	Glycine	HMDB0000123	OMIM IDS 601665
3	2-Ketobutyric acid	—	None
4	L-threonine	HMDB0000167	Doi:10.1007/s11306-013-0550-9
5	L-asparagine	—	None
6	L-glutamine	HMDB0000641	24 740 590
7	Putrescine	—	None
8	Stearoylcarnitine	HMDB0000848	601 665
9	Estrone sulfate	HMDB0001425	2 401 584
10	3,7-Dimethyluric acid	HMDB0001982	26 505 825

will be stored in the body in the form of fat. When the amount exceeds the normal physiological needs and reaches a certain value, it will evolve into obesity. Table 5 shows that 7 of the top 10 obesity infarction-related metabolites prediction results have been verified. For example, there is substantial evidence that a dietrich in L-glutamine is associated with positive gut effects [46, 47]. L-glutamine maintains intestinal barrier function, aids intestinal cell proliferation and differentiation and generally reduces the incidence of sepsis and symptoms of irritable bowel syndrome [48, 49]. The reason for this property is believed to be due to the higher rate of extraction of glutamine by the gut than other amino acids and is therefore considered the most viable option when trying to alleviate gastrointestinal-related disorders. Therefore, this metabolite has a certain relationship with bodyweight loss [45].

Ablation experiments

To further verify the generalization ability and robustness of GCNAT, we perform the ablation study using 5-fold CV on the main data set to evaluate the impact of each component on the model's predictive performance.

- GCNAT-OS: it uses a heterogeneous network constructed of SD, GIPM and known metabolite–disease associations as the input of GCNAT.
- GCNAT-OG: the input heterogeneous network consisting of GIPD, SD and known metabolite–disease correlations.
- GCNAT-No GAT (NG): it is a simplified version that has no GAT layer.
- GCNAT-No loss function (NL): it applies the de-weight cross-entropy loss function in GCNAT to test it.

Table 6 shows the performance comparison of GCNAT with its four variants in terms of AUC and AUPR. We observe that AUC and AUPR of GCNAT-OS and GCNAT-OG are not as high as those of GCNAT in when using a single similarity. GCNAT obtains the best AUC score of 0.950, and it is 2.03 and 11.05% higher than that of GCNAT-OS and GCNAT-OG, respectively. GCNAT obtains

the highest AUPR score of 0.405, outperforming 5.29 and 10.29% compared with the above two models. Although F1, ACC and specificity of GCNAT-OS are slightly higher than those of GCNAT, probably specificity represents the proportion of predicted negative samples to the actual number of samples. Using a single similarity predicts a higher number of negative samples, which also affects ACC and the recall. Although recall and precision are mutually exclusive, the value of recall decreases and the value of F1 increases accordingly. These results show that the disease semantic similarity and the disease GIP kernel similarity in the heterogeneous network contain useful information and give rise to the improved performance of GCNAT.

Compared with GCNAT-NG, GCNAT improves AUC by 3.3% and AUPR by 9.53%. This shows that the GAT layer has the effect of improving the predictive accuracy GCNAT. The index values of GCNAT are higher than those of GCNAT-NL. The AUC is 23.49% better, and the AUPR is 10.38% higher. The results show that the weighted cross-entropy loss function is important components of GCNAT. Based on the above discussion, all three components of GCNAT are critical for metabolite–disease associations' prediction.

Discussion and conclusion

In recent years, more and more studies related to metabolites have shown that they are closely related to human diseases [50], and the identification of disease-related metabolites can be used to diagnose diseases. However, biological experiments can test some hypotheses; they are often time-consuming and labor-intensive. Several algorithms exist to predict metabolite–disease associations, but the accuracy and reliability still need to be improved. Therefore, it is of great significance to develop a new predictive algorithm to overcome these drawbacks. In our study, based on the known association information between metabolites and diseases, Gaussian kernel similarity of metabolites and integrated similarity combining disease semantic similarity and Gaussian kernel similarity, we establish a deep learning algorithmic model named GCNAT to predict potential associations between metabolites and diseases. Furthermore, the inherent regularities and representation levels of sample data are learned through graph convolutional neural networks. Then, a GAT is used to combine the embeddings of multiple convolutional layers and assign them different weights. Finally, the prediction results are obtained by scoring the final synthetic embeddings. A 5-fold CV shows outstanding AUC (0.950) and AUPR (0.405) of GCNAT compared with previous methods and similar approaches. In addition, we also conduct case studies on Alzheimer's disease, colorectal cancer and obesity, and the results can show that GCNAT can be an effective tool to predict the potential correlations between metabolites and diseases.

The ideal predictive ability of GCNAT mainly depends on the following factors. First, through the known

Table 6. Comparison analysis between GCNAT and its ablation experiments on the same data set

Results						
Model	AUPR	AUC	F1	ACC	RE	SP
GCNAT	0.4049	0.9501	0.0950	0.9712	0.6712	0.9718
GCNAT-NG	0.3096	0.9171	0.0641	0.9513	0.6520	0.9601
GCNAT-NL	0.3011	0.7152	0.0602	0.8917	0.5617	0.9022
GCNAT-OS	0.3520	0.9298	0.0965	0.9774	0.6406	0.9780
GCNAT-OG	0.3020	0.8396	0.0706	0.9610	0.5604	0.9618

associations of metabolites and diseases, we obtain the Gaussian kernel similarity of the corresponding metabolites, and the integrated similarity combining the semantic similarity and Gaussian kernel similarity of diseases. By combining the two similarities of the disease, the sparsity of the data is reduced and the predictive results are more accurate. Second, from the perspective of algorithms, deep learning is an important research direction of machine learning, which makes machine learning closer to artificial intelligence. By building a deep learning model, the graph convolutional neural network is used to learn the inherent laws and representation levels of sample data. Third, using the attention layer of the graph, the embeddings of multiple convolutional layers are combined. Through this method, each node will obtain a weight about the neighbor nodes and obtain the comprehensive representation of the node through summation. In this way, the graph neural network can pay more attention to important nodes, thus reduce the influence of edge noise and improve the predictive accuracy of GCNAT.

However, our model still suffers from several limitations. First, our data set is not perfect at the moment. It contains far fewer known metabolite–disease samples than unknown ones, and the ratio of positive samples versus negative samples can reach 1:100. Second, the similarities of metabolites and diseases have a critical impact on model performance. By integrating more biological information to conduct experiments, a more reliable similarity measure can be obtained. Third, the choice of parameter values in our model can be further investigated. With more and more experimental verified metabolite–disease correlations being collected and new parameter optimization algorithms are applied in the future, we believe that the predictive property of GCNAT can be further improved.

Key Points

- We present a new deep learning algorithm based on graph convolutional network (GCN) with graph attention network (GAT) (GCNAT) for predicting metabolite–disease associations.
- GCNAT utilizes a GCN to capture structural information from the heterogeneous network composed of associations and similarities, and the convolutional layers are

fed into the GAT to obtain more information representations of metabolites and diseases.

- Compared with existing state-of-the-art methods, our method achieves higher predictive accuracy.

Abbreviations

GCN, graph convolutional network; GAT, graph attention network; 5-fold CV, 5-fold cross validation; ROC, receiver operating characteristic; TPR, true positive rate; FPR, false positive rate; AUC, area under the ROC curve; AUPR, precision-recall curve; DAG, directed acyclic graph; HMDB, Human Metabolome Database; MeSH, Medical Subject Headings

Data availability statement

The codes and data sets are available online at <https://github.com/zhaqi106/GCNAT>.

Funding

This study was supported by National Natural Science Foundation of China (Grant No. 11805091) and Foundation of Education Department of Liaoning Province (Grant No. LJKZ0280).

References

1. Dhanya M, Hegde S. Salivary glucose as a diagnostic tool in type II diabetes mellitus: a case-control study. *Niger J Clin Pract* 2016;**19**(4):486–90.
2. Ajouz H, Mukherji D, Shamseddine A. Secondary bile acids: an underrecognized cause of colon cancer. *World J Surg Oncol* 2014;**12**:164.
3. Chiang JY. Bile acid regulation of hepatic physiology: III. Bile acids and nuclear receptors. *Am J Physiol Gastrointest Liver Physiol* 2003;**284**(3):G349–56.
4. Stadler J, Sing Yeung K, Furrer R, et al. Proliferative activity of rectal mucosa and soluble fecal bile acids in patients with normal colons and in patients with colonic polyps or cancer. *Cancer Lett* 1988;**38**(3):315–20.
5. Costarelli V, Sanders TA. Plasma deoxycholic acid concentration is elevated in postmenopausal women with newly diagnosed breast cancer. *Eur J Clin Nutr* 2002;**56**(9):925–7.
6. Nobuoka A, Takayama T, Miyanishi K, et al. Glutathione-S-transferase P1-1 protects aberrant crypt foci from apoptosis induced by deoxycholic acid. *Gastroenterology* 2004;**127**(2):428–43.

7. Bonita JS, Mandarano M, Shuta D, et al. Coffee and cardiovascular disease: in vitro, cellular, animal, and human studies. *Pharmacol Res* 2007;**55**(3):187–98.
8. Marquez-Martin A, Puerta RDL, Fernandez-Arche A, et al. Modulation of cytokine secretion by pentacyclic triterpenes from olive pomace oil in human mononuclear cells. *Cytokine* 2006;**36**(5–6): 211–7.
9. Bruni F, Puccetti L, Pasqui AL, et al. Different effect induced by treatment with several statins on monocyte tissue factor expression in hypercholesterolemic subjects. *Clin Exp Med* 2003;**3**(1): 45–53.
10. Tonelli M, Isles C, Craven T, et al. Effect of pravastatin on rate of kidney function loss in people with or at risk for coronary disease. *Circulation* 2005;**112**(2):171–8.
11. Hu H, Zhang L, Ai H, et al. HLPi-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol* 2018;**15**(6):797–806.
12. Liu H, Ren G, Chen H, et al. Predicting lncRNA-miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowledge-Based Syst* 2019;**191**:105261.
13. Zhang L, Yang P, Feng H, et al. Using network distance analysis to predict lncRNA-miRNA interactions. *Interdiscip Sci: Comput Life Sci* 2021;**13**(3):535–45.
14. Chen X, Sun LG, Zhao Y. NCMCMDA: miRNA-disease association prediction through neighborhood constraint matrix completion. *Brief Bioinform* 2021;**22**(1):485–96.
15. Chen X, Li TH, Zhao Y, et al. Deep-belief network for predicting potential miRNA-disease associations. *Brief Bioinform* 2021;**22**(3):bbaa186.
16. Chen X, Zhu CC, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput Biol* 2019;**15**(7):e1007209.
17. Wang CC, Han CD, Zhao Q, et al. Circular RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2021;**22**(6):bbab286.
18. Zhao Q, Yang Y, Ren G, et al. Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans Nanobioscience* 2019;**18**(4):578–84.
19. Hu Y, Zhao T, Zhang N, et al. Identifying diseases-related metabolites using random walk. *BMC Bioinform* 2018;**19**(Suppl 5):116.
20. Lei X, Zhang C. Predicting metabolite-disease associations based on KATZ model. *BioData Min* 2019;**12**:19.
21. Lei X, Zhang C, Wang Y. Predicting metabolite-disease associations based on spy strategy and ABC algorithm. *Front Mol Biosci* 2020;**7**:603121.
22. Zhang C, Lei X, Liu L. Predicting metabolite-disease associations based on LightGBM model. *Front Genet* 2021;**12**:660275.
23. Ma Y, Ma Y. Hypergraph-based logistic matrix factorization for metabolite-disease interaction prediction. *Bioinformatics* 2021;btab652. <https://doi.org/10.1093/bioinformatics/btab652>.
24. Zhao T, Hu Y, Cheng L. Deep-DRM: a computational method for identifying disease-related metabolites based on graph deep learning approaches. *Brief Bioinform* 2021;**22**(4):bbaa212.
25. Wishart DS, Guo AC, Oler E, et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;**50**(D1):D622–d631.
26. Wang D, Wang J, Lu M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics (Oxford, England)* 2010;**26**(13): 1644–50.
27. Vanunu O, Magger O, Ruppin E, et al. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol* 2010;**6**(1):e1000641.
28. Yu Z, Huang F, Zhao X, et al. Predicting drug-disease associations through layer attention graph convolutional network. *Brief Bioinform* 2021;**22**(4):bbaa243.
29. Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks. *International Conference on Learning Representations (ICLR)*, Canada, 2018.
30. Huang YA, Hu P, Chan KCC, et al. Graph convolution for predicting associations between miRNA and drug resistance. *Bioinformatics* 2020;**36**(3):851–8.
31. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Sardinia, Italy, 2010, pp. 249–56.
32. Kingma DP, Ba J. Adam: a method for stochastic optimization. In: *International Conference on Learning Representations (ICLR)*, San Diego, 2015.
33. Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting. *J of Mach Learn Res* 2014;**15**(1):1929–58.
34. Takaya S, Marc R, Guy BJPO. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015;**10**(3): e0118432.
35. Yates EJ, Dixon LC. PageRank as a method to rank biomedical literature by importance. *Source Code Biol Med* 2015;**10**:16.
36. Hinton GE, Salakhutdinov RRJS. Reducing the dimensionality of data with neural networks. *SCIENCE* **313**(5786): 504–7.
37. Peng LH, Zhou LQ, Chen X, et al. A computational study of potential miRNA-disease association inference based on ensemble learning and kernel ridge regression. *Front Bioeng Biotechnol* 2020;**8**:40.
38. Abe T, Tohgi H, Isobe C, et al. Remarkable increase in the concentration of 8-hydroxyguanosine in cerebrospinal fluid from patients with Alzheimer's disease. *J Neurosci Res* 2002;**70**(3): 447–50.
39. Redjems-Bennani N, Jeandel C, Lefebvre E, et al. Abnormal substrate levels that depend upon mitochondrial function in cerebrospinal fluid from Alzheimer patients. *Gerontology* 1998;**44**(5): 300–4.
40. Fonteh AN, Harrington RJ, Tsai A, et al. Free amino acid and dipeptide changes in the body fluids from Alzheimer's disease subjects. *Amino Acids* 2007;**32**(2):213–24.
41. Tsuruoka M, Hara J, Hirayama A, et al. Capillary electrophoresis-mass spectrometry-based metabolome analysis of serum and saliva from neurodegenerative dementia patients. *Electrophoresis* 2013;**34**(19):2865–72.
42. Leoni V, Masterman T, Mousavi FS, et al. Diagnostic use of cerebral and extracerebral oxysterols. *Clin Chem Lab Med* 2004;**42**(2): 186–91.
43. Ni Y, Xie G, Jia W. Metabonomics of human colorectal cancer: new approaches for early diagnosis and biomarker discovery. *J Proteome Res* 2014;**13**(9):3857–70.
44. Goedert JJ, Sampson JN, Moore SC, et al. Fecal metabolomics: assay performance and association with colorectal cancer. *Carcinogenesis* 2014;**35**(9):2089–96.
45. Brown DG, Rao S, Weir TL, et al. Metabolomics and metabolic pathway networks from human colorectal cancers, adjacent mucosa, and stool. *Cancer & Metabolism* 2016;**4**:11.
46. Reinehr T, Wolters B, Knop C, et al. Changes in the serum metabolite profile in obese children with weight loss. *Eur J Nutr* 2015;**54**(2):173–81.

47. Wahl S, Yu Z, Kleber M, et al. Childhood obesity is associated with changes in the serum metabolite profile. *Obes Facts* 2012;**5**(5): 660–70.
48. Gronwald W, Klein MS, Kaspar H, et al. Urinary metabolite quantification employing 2D NMR spectroscopy. *Anal Chem* 2008;**80**(23):9288–97.
49. Hong YS, Hong KS, Park MH, et al. Metabonomic understanding of probiotic effects in humans with irritable bowel syndrome. *J Clin Gastroenterol* 2011;**45**(5):415–25.
50. Ugorski M, Laskowska A. Sialyl Lewis(a): a tumor-associated carbohydrate antigen involved in adhesion and metastatic potential of cancer cells. *Acta Biochim Pol* 2002;**49**(2):303–11.