EasyChair Preprint

№ 8913

Graph Random Forest: a Graph Embedded Algorithm for Identifying Highly Connected Important Features

Leqi Tian and Tianwei Yu

October 3, 2022

Graph Random Forest: A graph embedded algorithm for identifying highly connected important features

Leqi Tian[1,2] and Tianwei Yu[1,2,3]

[1] School of Data Science, The Chinese University of Hong Kong – Shenzhen, Guangdong, China
[2] Shenzhen Research Institute of Big Data, Guangdong, China
[3] Warshel Institute, Shenzhen, Guangdong, China
leqitian@link.cuhk.edu.cn, yutianwei@cuhk.edu.cn

Abstract. Random Forest (RF) is a widely used machine learning method with good performance on classification and regression tasks. It can train on over parameterized datasets which benefits the applications in the field of biology. For example, gene expression data always has a considerable number of features (p) compared to the size of samples (n). Though the predictive accuracy using RF is high, there are some problems when selecting important genes from a large number of features. The important genes selected by RF are usually scattered on the gene network, which conflicts with the biological assumption of connectivity between effective features. To apply random forest better in the biological field with external topological information between features, we propose the Graph Random Forest (GRF) for identifying highly connected important features by involving an interactive network when constructing the forest. The algorithm can identify effective features that form a highly connected sub-graph and achieve equivalent classification accuracy to RF. To evaluate the capability of our proposed method, we conducted simulation experiments and applied the method to two real datasets – non-small cell lung cancer RNA-seq data from The Cancer Genome Atlas and human embryonic stem cell RNA-seq dataset (GSE93593). The resulting high classification accuracy, connectivity of selected sub-graph, and interpretable feature selection results suggest the method is a helpful addition to graph-based classification models and feature selection procedures.

Keywords: Feature selection · Random forest · Gene network.

1 Introduction

With the widespread use of high-throughput technologies, more gene expression datasets are available to study clinical outcomes such as cancer and other diseases. An accurate classification model to distinguish two cancer subtypes through gene expression data can be used to design specific treatments and medical plans. Identifying genes that help distinguish different subtypes is also important, which benefits understanding different biological mechanisms in disease development. Moreover, this kind of research is beneficial for discovering new drug targets[17]. Therefore, it is of great interest to develop a new method to predict disease outcomes and reveal the biological mechanisms using gene expression simultaneously. The challenge of analyses yields from the structure of the gene expression datasets, which often contain tens of thousands of features while only processing hundreds of observations. A natural solution is to use tree-based methods such as random forest and LightGBM, which perfectly solve the problem of over-parameterized and require little parameter tuning or data transformation ([5]). Not only that, random forest provides excellent convenience for feature selection with an inherent measure of feature importance, which leads to its wide use in the biological field[33]. In the task of distinguishing benign breast cancer from malignant one, an algorithm combining the backward elimination approach and random forest achieves 99% classification accuracy [25]. Also, random forest help detect biomarkers for prostate cancer progression on DNA methylation data [32] and can identify abnormal pap smear cervical cell images [31]. Besides classification tasks, random forest is commonly used in predicting gene regulatory network (GRN). GENIE3 [15] applies random forest to infer GRN by solving a regression model based on target genes and selecting the strongest predictor as the regulator.

Given the nature of gene expression data, functionally related genes that collectively contribute to disease outcomes tend to be dependent and close on the gene interaction network. Therefore, it

is natural to combine the network structure with gene expression. Networks with topological structures include gene interaction networks, metabolite networks, miRNA networks, protein-protein interaction (PPI) networks, etc. A general network consists of vertices and edges, in which an edge represents an interaction between two vertices. Previous research [22] shows that interactive network provides valuable information for disease prediction and can help improve predictive performance. Furthermore, it has been shown that a more accurate subnetwork with markers can be identified by integrating gene expression data with PPI networks [16]. Many existing studies utilize the network structure to provide ancillary information when constructing random forests. IRatNet [26] utilizes heterogeneous data, including PPI network, to derive preliminary information and integrate the information into a weighted sampling scheme under random forest to infer the final GRN. The method improves the performance of predicting TF-target gene regulations. In the task of identifying predictive disease-related long non-coding RNA, GAERF [36] first embeds graph information with observed expression data into low dimension using Graph Auto Encoder (GAE), and then performs random forest to predict the outcome. Many methods incorporate network structure with neural networks to achieve the goal of feature selection. GEDFN [19] embeds a gene interaction graph in a feed-forward neural network, and the accuracy of classifying breast cancer has been significantly improved. GLRP [8] uses a graph convolutional neural network to learn gene expression data with graph structure and then select important features measured through the layer-wise relevance propagation (LPR) method. Although neural networks have been commonly exploited, the over-parameterized problem in biological data is still a problem. More effort is needed during training the model due to the small sample size.

Most existing random forest methods simply integrate graph information in prior knowledge and do not change the procedure of building a tree. Moreover, the topological property of the sub-graph established using the selected features has not been emphasized. Barabási [4] designed a model for predicting disease using a network-based approach and illustrated that disease-related components tend to be near to the ones that had been identified. In a protein-protein interaction network, functionally related genes are closer or even connected to each other [11]. Therefore, we extracted the sub-graph of the selected features from the feature graph and examined its connectivity using different methods. A better selection method is expected to choose features that form cliques on the whole graph. However, our simulation studies and practical applications show that the important features selected by a random forest are scattered on the feature network, which is not consistent with our expectations. Therefore, we propose a Graph Random Forest (GRF) model, which involves the network information in the tree-building process and can find relatively clustered important features. Similar to our work, network-guided forest (NGF) [12] uses graph information to build up a tree. The first splitting node is determined randomly, and then for each splitting, possible genes are selected from the neighborhood of the ones already in the tree. Even though this approach has proven to work well in predicting cell type and disease state in breast cancer and glioma data, there are still some disadvantages. First, there is a loss of flexibility when limiting the splitting scope in the neighborhood since the neighborhood size for each gene is always not large. Second, a random selection of the splitting node may increase the randomness of the model performance. Third, the algorithm is time-consuming since it needs to determine the scope of available splitting nodes each time. Fourth, the graph's structure selected by important features is not studied. Based on these considerations, our approach focuses on improving the flexibility and robustness of random forest, and we examine the connectivity of feature selection sub-graphs.

The article is organized as follows: Section 2 illustrates our graph embedded random forest architecture. Section 3.1 compares the performance of our method with other models on synthetic datasets. Results of two real applications on RNA-seq datasets are presented in Section 3.2. Finally, Section 4 is the conclusion.

2 Method

2.1 Graph Random Forest

Our proposed method involves graph information in the process of generating each decision tree in a random forest. The intuition for this particular structure is based on the following assumptions. First, only a small fraction of features that form as a sub-graph effectively affects the outcome.

Second, features in a graph are dependent on their neighborhoods. These assumptions are concluded from real data and have been widely used in previous works reviewed in Section 1.

To flexibly merge graph information into the framework of random forest, we proposed a new method – Graph Random Forest (GRF). The key idea is to embed graph information in the process of building a decision tree. When establishing a decision tree, we considered features in the neighborhood of any number of hops to the head-splitting node. The head node for each decision tree was determined through a data-driven approach.

The overall architecture of Graph Random forest (GRF) is shown in Fig 1(a). Denoted the gene expression data as X with n observations on p genes and the corresponding output as Y. The gene network $G = (V, E)$ could be obtained from open source, where V was the collection of p vertices and E was the collection of edges between V. Instead of choosing the head node in a decision tree arbitrarily from V, we first trained a random forest containing 500 decision trees with depth one to determine the first splitting nodes in GRF. Based on the pre-train random forest, we counted the number of times each node appeared as the head node in the pre-trained forest and recorded the result as vector $h = [c_1, c_2, ..., c_p]^T$. The sum of the vector h was set as 500 in our experiments, and the value could be adjusted. For each node $v \in V$ with a non-zero value in h, we set up a decision tree with v being the head-splitting node and chose the next splitting nodes in the vicinity of v. Fig 1(b) illustrates the vicinity of one-step hop and two-step hops with the orange node representing v, and the blue dash circle and purple dash circle show the two vicinity areas. Denoted the vicinity of v within k hops as Neighbor(v, k) (including v), then c_v decision trees were built with each tree choosing part of the features from Neighbor(v, k). In this way, a random forest with $(c_1 + c_2 + ... + c_p)$ decision trees was set up, and the final result was based on the majority voting from all decision trees.

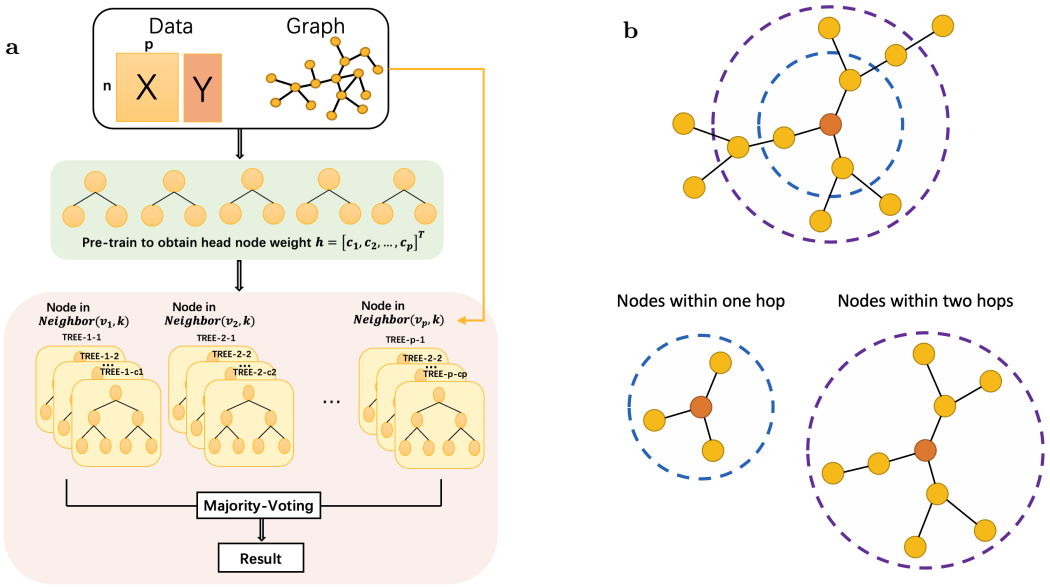

Fig. 1. (a) The overall architecture of Graph Random forest (GRF). (b)Visualization of one-step hop vicinity and two-step hops vicinity.

2.2 Evaluation of feature importance

Besides predicting the corresponding labels on testing data, it is important to find features that significantly contribute to the classification and help reveal biological mechanisms. Inheriting the Gini importance used in the random forest, we could similarly calculate feature importance in GRF.

In the implementation, we built up a random forest with c_i decision using features in the vicinity of i for each $i \in V$, so we could compute the importance easily by combining the importance obtained from all forests. Specifically, denoted the feature importance as $I = [\text{imp}_1, \text{imp}_2, ..., \text{imp}_p]^T$,

the initialization for each element in I was 0. Then for each node i with non-zero c_i, we built up a random forest F_i with c_i trees using nodes in Neighbor(v, k). From random forest F_i, we obtained the Gini importance $I_i = [\text{imp}_{ij}], \text{for} j \in \text{Neighbor}(v, k)$ and then updated I as $\text{imp}_j = \text{imp}_j + \text{imp}_{ij} \times c_i$ for $j \in \text{Neighbor}(v, k)$. In this way, we could access the feature importance for all variables, allowing us to select high-ranking features.

2.3 Details of model setting

The training of the GRF contained two parts. The first part was to train a simple random forest with each tree with a depth of one. Using a random forest fitted on the training dataset, we recorded the number of occurrences of each node as a head-splitting node and then embedded the information in GRF. The second part was training GRF on the training dataset. We set the number of trees in GRF as 500 and the other parameters as the default values. In this work, we did not fine-tune for the best model on specific data since our interest lay in evaluating the ability to find important variables and studying the clustering properties of the selected sub-graph.

3 Results and Discussion

3.1 Simulation experiments

To mimic disease classification, a series of simulation experiments were conducted using gene expression data and gene network. We compared the performance of GRF with the usual random forest. The simulation study aimed to examine whether GRF could identify an effective sub-graph more accurately.

Synthetic data generation To generate a scale-free graph for p features, we used the Barabási-Albert (BA) model [3]. The degree distribution of a biological graph can be captured by a power law which is characterized by m in a BA model. We set the power law as 0.5 in the simulation studies. Using the generated feature network, we denoted $D \in \mathcal{R}^{p \times p}$ as the shortest distance for all pairwise vertices. Then we derived the covariance matrix Σ for p features using

$$\Sigma_{i,j} = 0.8^{D_{i,j}}, i, j = 1, ..., p. \tag{1}$$

The diagonal elements of Σ were one since the distance from one node to itself was zero.

With the feature graph and covariance matrix Σ, we simulated expression data $X = [x_1, x_2, ..., x_n]^T$ with p features. A multivariate Gaussian distribution with mean zero was used to generate X, for each sample x_i,

$$x_i \sim N(0, \Sigma), i = 1, 2, ..., n. \tag{2}$$

In this way, two features tended to have more significant covariance if they were close to each other.

We selected a subset of features as true predictors to generate the outcome Y, which corresponded to different disease outcomes. Under the assumption that only a small part of genes that tended to form cliques in the big network was effective in disease outcomes, we first chose core features and then expanded the cliques by including part of their neighboring vertices in the graph. Considering that the scales and characteristics might differ between graphs, we used an average strategy to identify potential core features. Specifically, our goal was to find features with significantly higher node degrees. We first ranked features from high to low using their node degree and denoted the degree as $d_1 \geq d_2 \geq, ..., \geq n_p$, then we calculated the average degree in three steps as $\bar{d}_i = (d_i + d_{i-1} +$
$d_{i-2})/3, i = 3, 4, ..., p$. We determined the change point as the first feature whose degree d_t decreased more than 10% comparing to the averaged value, which was $(d_t - d_{t-1})/\bar{d}_t > -0.01$. Features with node degree larger than d_t were chosen as potential core features.

In the simulation study, we conducted experiments with core features randomly chosen from the pool of potential core features. To expand the clique starting from the core features, we iteratively chose part of the neighboring features. Specifically, we limited the selection of at most m features from the neighborhood each time. To avoid the selected sub-graph becoming too dense, we assigned an attenuation rate to parameter m. In this way, we obtained a collection of true predictors S,

which formed a sub-graph. Denoting the size of S as p_0, we sampled parameter $\beta = (\beta_1, \beta_2, ..., \beta_{p_0})^T$ from uniform distribution ranged in $(0.1, 0.8)$, and set part of them negative. The output Y was generated using a generalized linear model as

$$P(y_i = 1|x_i) = \sigma(x_i^T \beta + \beta_0), i = 1, 2, ..., n \tag{3}$$

$$y_i = I(P(y_i = 1|x_i) > 0.5), i = 1, 2, ..., n \tag{4}$$

where σ was logistic link function or absolute link function,

$$\sigma(x) = \frac{1}{e^x + 1} \text{ or } \sigma(x) = \text{abs}(x - \bar{x}) + 0.5. \tag{5}$$

Following the procedure ahead, we generated expression data X with 4000 features and 500 samples. We conducted our proposed method GRF and RF for comparison. In real gene data, the true predictors usually only accounted for a small part of genes. A widely observed rate was around 0.5%. Taking this into consideration, we used different numbers of true predictors to generate the corresponding Y, i.e., 30, 60, 90, 120, 150, 180, and 210 in our study. Besides the size of true predictors, we also considered different sub-graph shapes. Generally speaking, an effective sub-graph was clustered into one clique, but in some cases, the clustering shape looked like two cliques with little connection between the two clusters. So we considered two clustering shapes of the effective sub-graph, i.e., one core and two cores. In the case of two core features, we limited the distance between two core features should be no less than four so that we could obtain two cliques. Also, we tested different relationships between the X and Y using the logistic link function and absolute link function.

Simulation results In our simulation study, we first trained a simple random forest with one-depth trees to determine head nodes in GRF and then built up 500 graph-embedded decision trees. For vanilla random forest, we determined the number of trees and depth using grid search.

For each simulation setting, we generated ten datasets and then applied GRF and RF. Each dataset was divided into a training set and a testing set with a ratio of 7:3. The model was trained on the training set and made predictions on the testing set. For each experiment, the computation time for GRF was around 7s on a Linux workstation with a 5950x CPU, 128 Gb RAM, and a GTX 3060 GPU. In the simulation study, we were access to the specific list of true predictors contributing to the output Y. So we could evaluate the power of the estimated importance for each feature through the receiver operating characteristic (ROC) curve, which showed the performance of a classification model at all classification thresholds. The areas underneath the ROC curve (AUC) using different methods were recorded. We also compared the area under the precision-recall curve (PR AUC), which was a supplementary metric to AUC and helped for a complete picture when evaluating the performance.

Fig 2 shows the results with logistic link function. The first row demonstrates the performance with a true sub-graph extended from one core node, and the second row shows the results of a true sub-graph extended from two core nodes. The error bar represents the estimated standard deviation estimated from ten experiments. Fig 2(a) and Fig 2(d) show the test accuracy using RF and GRF with hopping steps 1, 2, and 3. In the case of one core node, GRF with hopping step 3 had the highest test accuracy when the number of true predictors was less than 100, and GRF 2 was higher when the size of true predictors grew larger. RF achieved the highest score in the setting with 210 true predictors. In the case of two core nodes, GRF with hopping steps 2 and 3 achieved the highest score in more than half of the cases. In general, there is little difference in test accuracy between different methods, and an upward trend in prediction accuracy can be observed as the number of true predictors increases. The second column and third column of Fig 2 are AUC and PR-AUC with different sizes of core nodes. The estimated feature importance and a list with true predictors as one and other features as 0 were used for calculating the two metrics. The results of AUC using two/three hopping step GRF in one core simulation were mostly higher than 0.9, which was a huge improvement compared to the value in RF, which is around 0.6-0.75. Simulation results using the absolute link function are shown in Fig 3, which shows similar patterns. Overall, the simulation results verify that GRF has the capability to identify more significant features without sacrificing prediction accuracy.

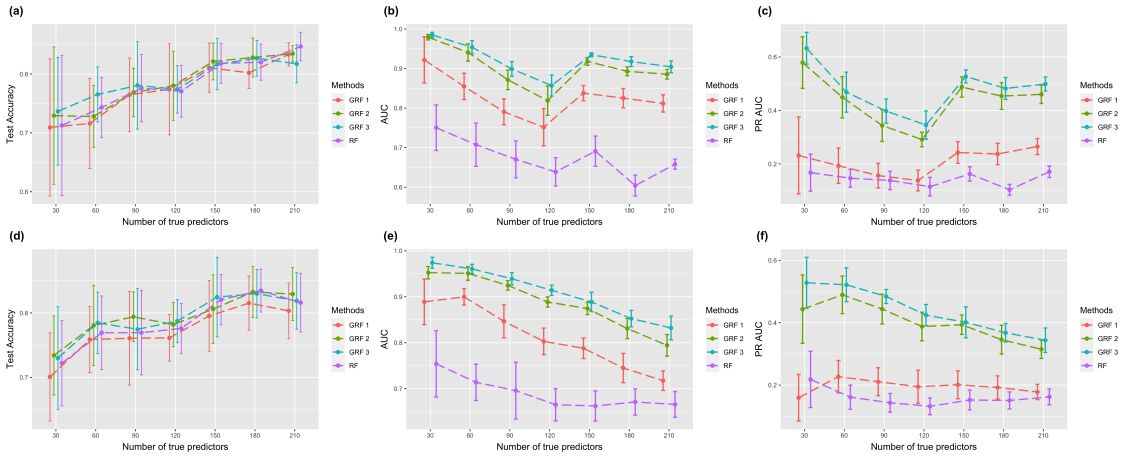

Fig. 2. Plots of classification and feature selection with the logistic link function. The first row corresponds to one core node, and the second corresponds to two core nodes. Error bars represent the mean value plus/minus the standard error. GRF with different numbers indicates various parameters of selective range.

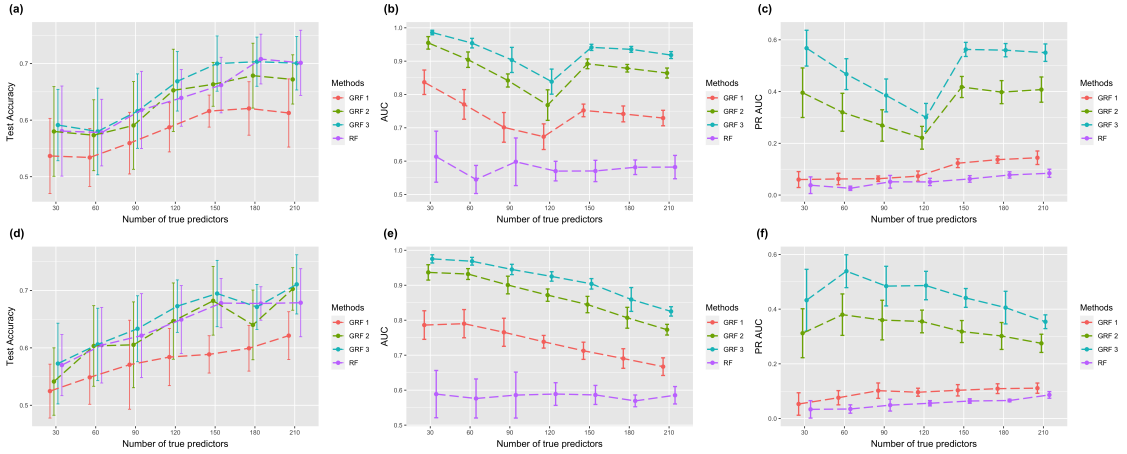

Fig. 3. Plots of classification and feature selection with the absolute link function. The first row corresponds to one core node, and the second corresponds to two core nodes. Error bars represent the mean value plus/minus the standard error. GRF with different numbers indicates various parameters of selective range.

To explore the property of the sub-graph, which consists of high-ranking features, we selected the top 100 most important features and extracted their connections from the simulated network. Fig 4 shows the sub-graph density, number of connected components, and size of the largest connected component in the setting with the logistic link function. The two rows correspond to settings of the true sub-graph extended from one or two core nodes. Fig 4 (a) and (d) show the graph density, which is defined as the ratio between the number of edges and the number of all possible edges. A larger density indicates a more connected graph. Results show that GRF has a larger density in all experiments. Fig 4 (b) and (e) present the number of connected component. A connected component means each pair of nodes in the component are connected through a certain pathway. A graph with more connected components indicates it is more scattered. Fig 4 (c) and (f) show the size of largest connected component in different experiments. It is clear from the results that sub-graphs selected by RF are more separate, and the largest connected components are smaller. On the contrary, sub-graphs generated using GRF are highly connected and have large connected components. When the hopping step equals two or three, the largest connected component contains more than 80 nodes in different experimental settings. Results of using the absolute link function are shown in Fig 5, which demonstrate similar patterns.

Concerning the robustness in reproducibility of feature selection using GRF, we were curious whether selected features were stable across different model training times. To explore this property,

we simulated a dataset with 500 samples, 4000 features, and 210 true features that determined the classes Y using a logistic link function. A GRF with hopping step two was used here. We repeated the training 20 times and recorded the top 100 features each time. In the 20 sets of most important features, 58 appeared more than 14 times, and 86 appeared more than ten times. Among the 58 features that appeared more than 14 times, 53 were on the list of true predictors, and the other five features were within the neighborhood of one step to true predictors. Among the 86 features that appeared more than ten times, 75 were true predictors, and the others were all around the true predictors in one step. For the union of the 20 lists of selected features, 46.4% of them were true predictors, and 89.3% of them were within one step neighborhood of true predictors.

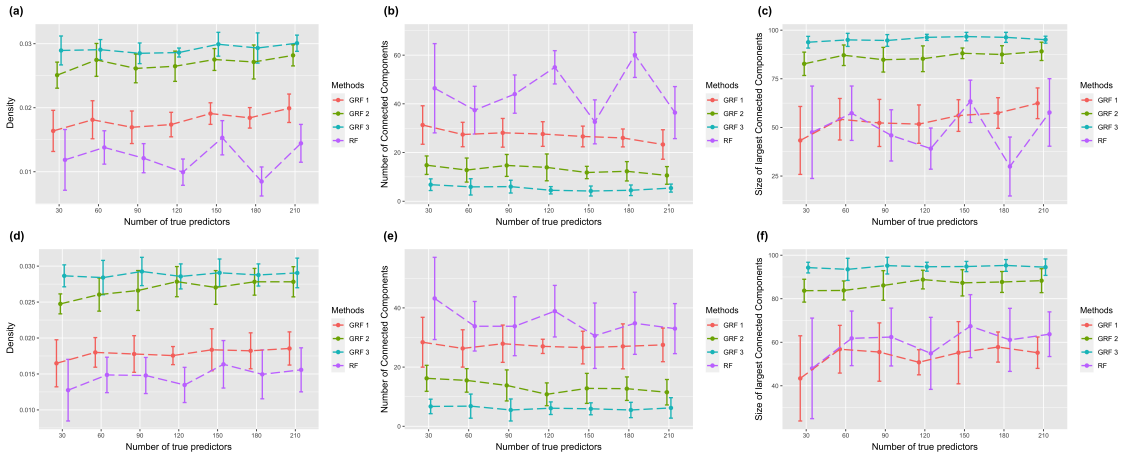

Fig. 4. Plots of sub-graph properties with the logistic link function. The first row corresponds to one core node, and the second corresponds to two core nodes. Error bars represent the mean value plus/minus the standard error. GRF with different numbers indicates various parameters of selective range.

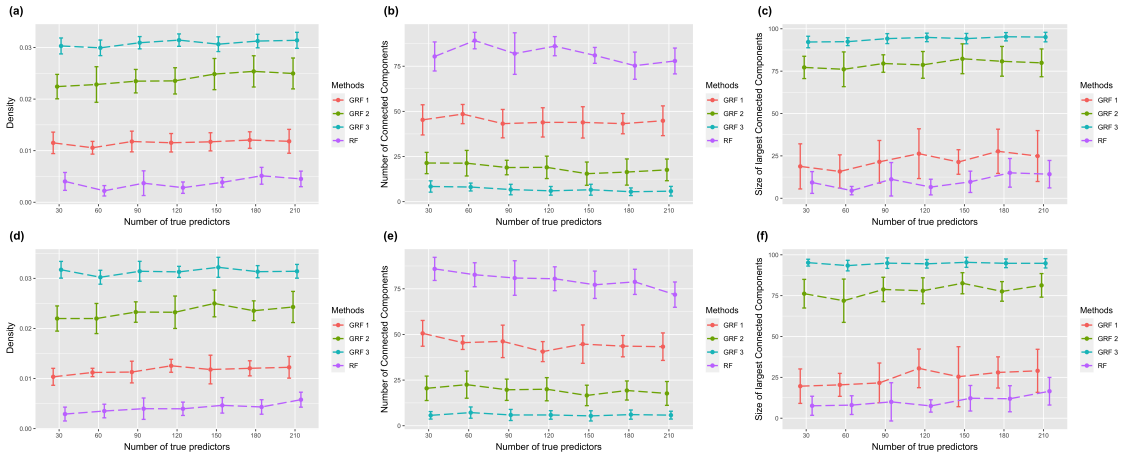

Fig. 5. Plots of sub-graph properties with the absolute link function. The first row corresponds to one core node, and the second corresponds to two core nodes. Error bars represent the mean value plus/minus the standard error. GRF with different numbers indicates various parameters of selective range.

3.2 Real data applications

Non-small cell lung cancer data We applied our GRF method to distinguish two types of most common subtypes of lung cancer – lung adenocarcinoma (LUAD) [24] and lung squamous cell carcinoma (LUSC) [23] from The Cancer Genome Atlas. Lung cancer is one of the deadliest cancer nowadays. However, it is still unclear how these two subtypes differ in biological mechanisms,

and they are still treated equally as non-small cell lung cancer (NSCLC). We tried to identify the differences between these two subtypes and analyze the biological mechanisms using GRF. LUAD dataset consisted of an expression matrix with 23032 miRNA in 524 patients, and LUSC dataset contained an expression matrix with 23652 miRNA expression in 496 observations. Combining the two datasets of different subtypes using overlapped features and selecting the largest connected component of miRNA network obtained from HINT ([11]), we eventually obtained an expression matrix with 9819 miRNA features of 1020 samples. The label for LUAD was marked as one and LUSC as zero, correspondingly. We used the processed data for downstream analysis.

Table 1. Classification result and properties of selected sub-graph for NSCLC data.

Methods	GRF[1]	RF[1]
Mean accuracy	0.9457(0.0116)	0.9483(0.0097)
Number of connected component	20.65 (3.63)	94.9 (1.92)
Size of the largest connected component	73.75 (6.63)	3.7 (1.22)
Average distance	4.29 (0.25)	1.38 (0.38)
Average distance in the largest component	4.31 (0.25)	1.53 (0.33)

[1] The values in brackets correspond to the standard deviations.

Using the data with its corresponding network, we conducted GRF and RF for comparison. When training the model, we split the dataset into the training set and testing set with the ratio of 7:3. We used GRF with hopping step two and evaluated the performance based on an averaged score. For each method, we conducted twenty experiments and summarized the performances. The first row in Table. 1 shows that the two methods had good performance on the classification task with high predicting accuracy. The accuracy of the testing dataset for GRF and RF was around 0.94, indicating a different expression pattern of miRNA. It was reasonable to distinguish the two subtypes of lung cancer using this dataset. From the result, we figured out that further accuracy improvement is difficult and meaningless since the performance almost reached an upper bound. Though our proposed GRF had a tiny gap in accuracy compared to RF, GRF had advantages in identifying potentially important features which had been proved in the simulation study. Moreover, features selected by GRF were more likely to fall into cliques on the original graph and could generate a sub-graph with fewer connected components. A highly connected graph was preferred because of its similarity to the ones found in experimental studies. By choosing 100 features with the highest importance scores, we generated the sub-graph shown in Fig 6(a). Table 1 also exhibits selected sub-graph properties through GRF and RF. The second row is the number of connected components, and the third row represents the size of the largest connected component. The fourth and fifth rows are the average shortest distance and the average distance in the largest components between each node. The results illustrate that the sub-graph selected by GRF had fewer connected components and possessed a larger connected component with a reasonable distance.

Functional analysis of genes selected by GRF was conducted by testing the enrichment of gene ontology (GO) biological processes using the clusterProfiler package ([38]). The biological processes with P-values less than 0.01 and adjusted P-values less than 0.05 were considered significant pathways in our study. The top 15 GO terms are shown in Table 2. Twelve of the 100 selected genes belonged to the regulation of DNA metabolic process, which was the most important GO term. In addition, 'DNA ligation', 'somatic DNA recombination', and 'regulation of DNA biosynthesis processes' were among the top terms. Changes in DNA function and damage affect cell proliferation and differentiation, which may influence cancer progression [28]. Meanwhile, overexpression of CDK2 and CDK16 in LUSC has been proved to cause abnormal regulation of cell cycle and promote cell proliferation. These effects may increase the malignant potential of the tumor and lead to a faster growth speed compared to the progression of LUAD [13] [20]. The second most important term was 'telomere maintenance', and there were many terms among the top ones related to telomere, such as 'telomere organization', 'negative regulation of telomere maintenance via telomerase', and 'telomere elongation control'. Limiting telomere from shortening is one of the significant mechanisms by which cancer cells gain resistance to inhibition. The ability to maintain telomere above critical length represents the degree of cell deterioration [21]. Cancer cells can resist

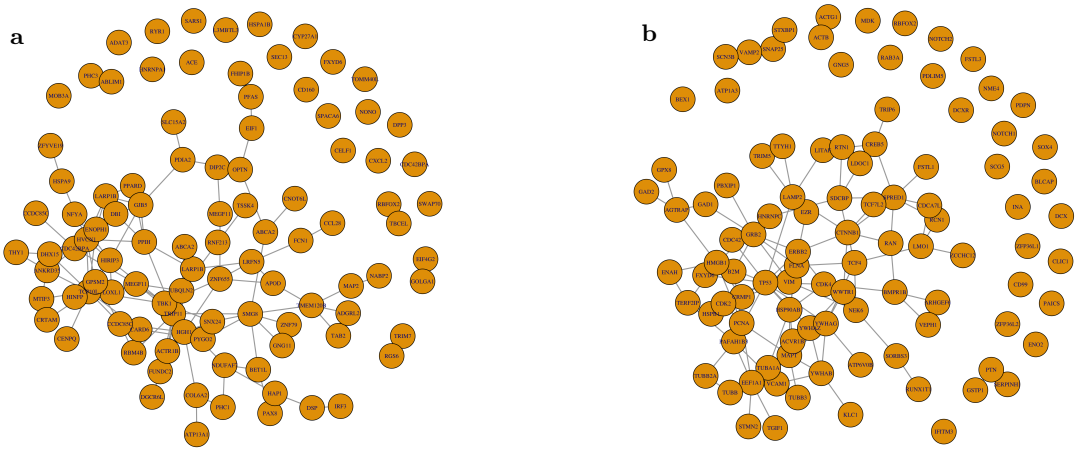

Fig. 6. Sub-graph selected by GRF on (a) NSCLC and (b) non-small cell lung cancer data.

Table 2. Top 10 GO biological process for the sub-graph selected by GRF on NSCLC data.

GOBPID[1]	Adj-P[2]	Term
GO:0051052	0.015	regulation of DNA metabolic process
GO:0000723	0.015	telomere maintenance
GO:0032069	0.015	regulation of nuclease activity
GO:0032200	0.015	telomere organization
GO:0032211	0.015	negative regulation of telomere maintenance via telomerase
GO:0051098	0.015	regulation of binding
GO:0048598	0.028	embryonic morphogenesis
GO:1904357	0.028	negative regulation of telomere maintenance via telomere lengthening
GO:0042098	0.028	T cell proliferation
GO:0006303	0.028	double-strand break repair via nonhomologous end joining

[1] Manual pruning of partially overlapping GO terms was performed.
[2] Adj-P represents the adjusted Pvalue.

death and realize the immortality of replication through activating the telomerase [14]. LUAD and LUSC have different expression levels of telomere-related genes, so telomere maintenance genes are considered to be potential biomarkers for two subtypes. At the same time, a vaccine against telomerase named GV1001 has been proved to be beneficial to immunotherapy for NSCLC patients [30].

The remaining significant GO terms include 'T cell proliferation and activation', 'immune response modulation', 'embryonic morphological development', 'endoderm development', and so on. Studies on immune-related genes (IRGs) have found that T-cell receptor signaling expresses differently in two subtypes [7]. MHC molecule, which is crucial for antigen processing in immune responses, and chemokine, which guides cell migration, are found to be inhibited more rapidly in LUSC. These observations confirm that LUSC grows faster by suppressing the immune system. HOX gene encoding is an important transcription factor in the embryonic development and differentiation of adult cells. Recent studies have shown that HOXA1 is significantly up-regulated and hypermethylated in LUAD [37]. The genes Hh and ErbB are found to be strongly correlated with two subtypes, which are related to lung development [2]. Hh maintains stem cells, responds to injury, and affects the formation of bronchial numbers. ErbB can cause defects of type II epithelial cells in the alveolar lining and reduce branching morphogenesis in embryos by affecting the expression level of anti-EGF antisense oligonucleotides. The whole table containing all GO terms for functional analysis using GRF and RF can be found in https://github.com/tianlq-prog/GRF/blob/main/Supplementary.pdf.

Human embryonic stem cell data We also applied GRF on the human embryonic stem cell RNA-seq dataset from GEO (GSE93593) [10]. The dataset contained 23045 genes from 1733 observations and their corresponding clinic information, including doublecortin (DCX) status and

days of culture. We obtained the gene network from the HINT database ([11]). After screening the genes in the HINT database and selecting the largest connected component of the network, 12215 genes were finally selected. A log transformation was conducted on the expression value of each gene. Our goal was to explore the relationship between gene expression and the status of DCX, whether positive or negative. Therefore, the task became a binary classification problem.

Doublecortin (DCX) is a microtubule-associated protein expressed explicitly by immature neurons in embryonic and adult cortical structures. It is necessary for neuron migration and differentiation and is closely related to the development of the central nervous system. Since the expression of DCX changes two weeks before the appearance of new neurons, the richness of neurogenesis in the brain cannot be directly quantified. So, DCX is a powerful tool for identifying early and immature neurons. Therefore, research on the transient expression of DCX to help understand the development of the nervous system has received extensive attention.

We tested GRF and RF on a human embryonic stem cell dataset with twenty repeated experiments. The computation time for each experiment using GRF on the workstation was around 100s. The classification accuracy results are shown in the first row of Table. 3. The mean accuracy was high for each method, and GRF had slightly lower accuracy than RF. However, as shown in Table 3, the sub-graphs generated by GRF with the top 100 most important genes had higher connected properties than the ones using RF. Specifically, when using GRF, the number of connected components was much smaller, and the sub-graphs had larger connected components with an average value achieving 67.

Table 3. Sub-graph property for GSE data. The values in brackets correspond to the standard deviations.

Methods	GRF[1]	RF[1]
Mean accuracy	0.9280(0.0089)	0.9301(0.008)
Number of connected component	31.15(4.83)	83.85 (3.73)
Size of the largest connected component	67.00(6.10)	7.95(3.32)
Average distance	3.67(0.30)	2.17(0.62)
Average distance in the largest component	3.68(0.30)	2.54(0.62)

[1] The values in brackets correspond to the standard deviations.

Fig 6(b) is the sub-graph of the top 100 genes with the highest averaged importance score using GRF. GO enrichment analysis was performed on the sub-graph, and the top 10 pathways are shown in Table 4. The top GO term was 'positive regulation of epithelial to mesenchymal transition'. In addition, 'homotypic cell-cell adhesion' and 'epithelial cell differentiation' were among the top terms. Epithelial cells have regular cell-cell contacts and adhesion to surrounding cellular structures, thus can avoid the separation of individual cells [6]. However, during embryonic development, cells need to migrate to adjacent tissues to form new organs, and tissues [35], so quiescent epithelial cells undergo epithelial-mesenchymal transition (EMT), thereby differentiate into motile mesenchymal cells [27] and possess the invasive ability. The response to EMT comes from stromal cells such as fibroblasts and mesenchymal stem cells. These stromal cells secrete a series of heterotypic signals, growth factors, platelet-derived growth factor (PDGF), and epidermal growth factor (EGF) [1]. This also explained the appearance of 'platelet aggregation', 'regulation of hematopoiesis', and 'platelet activation' in the top GO terms. Also, many of the genes involved in 'lung development,' 'respiratory system development,' and 'air duct development' are part of the response to growth factor stimulation, leading to the significance of these terms. The second most significant GO term was 'synaptic organization', and other important terms related to it included 'axon development' and 'axogenesis'. Synapses, responsible for transmitting information between neurons and target cells, play an essential role in nervous system development. The fetal brain begins to develop from the third week of gestation [29], neural precursor cells divide and form neurons and glia. Furthermore, the number of synapses keeps increasing in the first few years of life [34]. The fifth-ranked GO term was 'transmembrane receptor protein serine/threonine kinase signaling pathway'. The serine-threonine kinase Akt plays a central role in integrating cellular responses to growth factors [18], and it has been proven to maintain cellular integrity and protect 'tagged' from exposure. It is also involved in the phagocytic disposition of cells, in which it promotes neuronal

Table 4. Top 10 GO biological process for the sub-graph selected by GRF on DCX data.

GOBPID[1]	Adj-P[2]	Term
GO:0010718	0.0002	positive regulation of epithelial to mesenchymal transition
GO:0050808	0.0002	synapse organization
GO:0010717	0.0002	regulation of epithelial to mesenchymal transition
GO:0034109	0.0002	homotypic cell-cell adhesion
GO:0007178	0.0002	transmembrane receptor protein serine/threonine kinase signaling pathway
GO:0070527	0.0003	platelet aggregation
GO:0048667	0.0003	cell morphogenesis involved in neuron differentiation
GO:0048812	0.0003	neuron projection morphogenesis
GO:1903706	0.0003	regulation of hemopoiesis
GO:0001837	0.0003	epithelial to mesenchymal transition

[1] Manual pruning of partially overlapping GO terms was performed.
[2] Adj-P represents the adjusted Pvalue.

and vascular survival and prevents induction of programmed cell death [9]. Overall, in the DCX classification task, GRF could identify important and easily interpretable sub-graph.

4 Conclusion

We presented a new version of random forest which embeds graph information. It has the ability to identify important features with the property of high connectivity in the original information graph. A simulation study verified that our method could find features with higher AUC and PR-AUC without losing much classification accuracy. Two real applications showed the power of selection, which helps reveal meaningful biological mechanisms.

References

1. Abba, M.L., Patil, N., Leupold, J.H., Allgayer, H.: Microrna regulation of epithelial to mesenchymal transition. Journal of clinical medicine **5**(1), 8 (2016)
2. Anusewicz, D., Orzechowska, M., Bednarek, A.K.: Lung squamous cell carcinoma and lung adenocarcinoma differential gene expression regulation through pathways of notch, hedgehog, wnt, and erbb signalling. Scientific reports **10**(1), 1–15 (2020)
3. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. science **286**(5439), 509–512 (1999)
4. Barabási, A.L., Gulbahce, N., Loscalzo, J.: Network medicine: a network-based approach to human disease. Nature reviews genetics **12**(1), 56–68 (2011)
5. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
6. Cavey, M., Lecuit, T.: Molecular bases of cell–cell junctions stability and dynamics. Cold Spring Harbor perspectives in biology **1**(5), a002998 (2009)
7. Chen, M., Liu, X., Du, J., Wang, X.J., Xia, L.: Differentiated regulation of immune-response related genes between luad and lusc subtypes of lung cancers. Oncotarget **8**(1), 133 (2017)
8. Chereda, H., Bleckmann, A., Menck, K., Perera-Bel, J., Stegmaier, P., Auer, F., Kramer, F., Leha, A., Beißbarth, T.: Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. Genome medicine **13**(1), 1–16 (2021)
9. Chong, Z., Maiese, K.: Targeting wnt, protein kinase b, and mitochondrial membrane integrity to foster cellular survival in the nervous system. Histology and histopathology **19**(2), 495 (2004)
10. Close, J.L., Yao, Z., Levi, B.P., Miller, J.A., Bakken, T.E., Menon, V., Ting, J.T., Wall, A., Krostag, A.R., Thomsen, E.R., et al.: Single-cell profiling of an in vitro model of human interneuron development reveals temporal dynamics of cell type production and maturation. Neuron **93**(5), 1035–1048 (2017)
11. Das, J., Yu, H.: Hint: High-quality protein interactomes and their applications in understanding human disease. BMC systems biology **6**(1), 1–12 (2012)
12. Dutkowski, J., Ideker, T.: Protein networks as logic functions in development and cancer. PLoS computational biology **7**(9), e1002180 (2011)

13. Galimberti, F., Thompson, S.L., Liu, X., Li, H., Memoli, V., Green, S.R., DiRenzo, J., Greninger, P., Sharma, S.V., Settleman, J., et al.: Targeting the cyclin e-cdk-2 complex represses lung cancer growth by triggering anaphase catastrophe. Clinical Cancer Research **16**(1), 109–120 (2010)
14. Hanahan, D., Weinberg, R.A.: Hallmarks of cancer: the next generation. cell **144**(5), 646–674 (2011)
15. Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring regulatory networks from expression data using tree-based methods. PloS one **5**(9), e12776 (2010)
16. Ideker, T., Ozier, O., Schwikowski, B., Siegel, A.F.: Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics **18**(suppl_1), S233–S240 (2002)
17. Johannes, M., Brase, J.C., Fröhlich, H., Gade, S., Gehrmann, M., Fälth, M., Sültmann, H., Beißbarth, T.: Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. Bioinformatics **26**(17), 2136–2144 (2010)
18. Kandel, E.S., Hay, N.: The regulation and activities of the multifunctional serine/threonine kinase akt/pkb. Experimental cell research **253**(1), 210–229 (1999)
19. Kong, Y., Yu, T.: A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data. Bioinformatics **34**(21), 3727–3737 (2018)
20. Kumar, V., Abbas, A.K., Aster, J.C.: Robbins basic pathology e-book. Elsevier Health Sciences (2017)
21. Mason, P.J., Perdigones, N.: Telomere biology and translational research. Translational Research **162**(6), 333–342 (2013)
22. Navlakha, S., Kingsford, C.: The power of protein interaction networks for associating genes with diseases. Bioinformatics **26**(8), 1057–1063 (2010)
23. Network, C.G.A.R., et al.: Comprehensive genomic characterization of squamous cell lung cancers. Nature **489**(7417), 519 (2012)
24. Never-smoker, N.E.s.: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**, 543–50 (2014)
25. Nguyen, C., Wang, Y., Nguyen, H.N.: Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic (2013)
26. Petralia, F., Wang, P., Yang, J., Tu, Z.: Integrative random forest for gene regulatory network inference. Bioinformatics **31**(12), i197–i205 (2015)
27. Serrano-Gomez, S.J., Maziveyi, M., Alahari, S.K.: Regulation of epithelial-mesenchymal transition through epigenetic and post-translational modifications. Molecular cancer **15**(1), 1–14 (2016)
28. Situ, Y., Gao, R., Lei, L., Deng, L., Xu, Q., Shao, Z.: System analysis of fhit in luad and lusc: The expression, prognosis, gene regulation network, and regulation targets. The International Journal of Biological Markers p. 03936155221084056 (2022)
29. Stiles, J., Jernigan, T.L.: The basics of brain development. Neuropsychology review **20**(4), 327–348 (2010)
30. Storti, C.B., de Oliveira, R.A., de Carvalho, M., Hasimoto, E.N., Cataneo, D.C., Cataneo, A.J.M., De Faveri, J., Vasconcelos, E.J.R., Dos Reis, P.P., Cano, M.I.N.: Telomere-associated genes and telomeric lncrnas are biomarker candidates in lung squamous cell carcinoma (lusc). Experimental and molecular pathology **112**, 104354 (2020)
31. Sun, G., Li, S., Cao, Y., Lang, F.: Cervical cancer diagnosis based on random forest. International Journal of Performability Engineering **13**(4), 446 (2017)
32. Toth, R., Schiffmann, H., Hube-Magg, C., Büscheck, F., Höflmayer, D., Weidemann, S., Lebok, P., Fraune, C., Minner, S., Schlomm, T., et al.: Random forest-based modelling to detect biomarkers for prostate cancer progression. Clinical epigenetics **11**(1), 1–15 (2019)
33. Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., Boekhorst, J., Wels, M., van Hijum, S.A.: Data mining in the life sciences with random forest: a walk in the park or lost in the jungle? Briefings in bioinformatics **14**(3), 315–326 (2013)
34. Van Den Heuvel, M.P., Kersbergen, K.J., De Reus, M.A., Keunen, K., Kahn, R.S., Groenendaal, F., De Vries, L.S., Benders, M.J.: The neonatal connectome during preterm brain development. Cerebral cortex **25**(9), 3000–3013 (2015)
35. Varga, J., De Oliveira, T., Greten, F.R.: The architect who never sleeps: tumor-induced plasticity. FEBS letters **588**(15), 2422–2427 (2014)
36. Wu, Q.W., Xia, J.F., Ni, J.C., Zheng, C.H.: Gaerf: predicting lncrna-disease associations by graph auto-encoder and random forest. Briefings in bioinformatics **22**(5), bbaa391 (2021)
37. Yang, X., Deng, Y., He, R.Q., Li, X.J., Ma, J., Chen, G., Hu, X.H.: Upregulation of hoxa11 during the progression of lung adenocarcinoma detected via multiple approaches. International journal of molecular medicine **42**(5), 2650–2664 (2018)
38. Yu, G., Wang, L.G., Han, Y., He, Q.Y.: clusterprofiler: an r package for comparing biological themes among gene clusters. Omics: a journal of integrative biology **16**(5), 284–287 (2012)