International Journal of Data Warehousing and Mining
Volume 19 • Issue 2

Combining BPSO and ELM Models for Inferring Novel lncRNA-Disease Associations

Wenqing Yang, The Academy of Digital China, Fuzhou University, China

Xianghan Zheng, The Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, China

QiongXia Huang, The Academy of Digital China, Fuzhou University, China*

Yu Liu, The Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, China

Yimi Chen, The Fujian Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou University, China

ZhiGang Song, The Academy of Digital China, Fuzhou University, China

ABSTRACT

It has been widely known that long non-coding RNA (lncRNA) plays an important role in gene expression and regulation. However, due to a few characteristics of lncRNA (e.g., huge amounts of data, high dimension, lack of noted samples, etc.), identifying key lncRNA closely related to specific disease is nearly impossible. In this paper, the authors propose a computational method to predict key lncRNA closely related to its corresponding disease. The proposed solution implements a BPSO based intelligent algorithm to select possible optimal lncRNA subset, and then uses ML-ELM based deep learning model to evaluate each lncRNA subset. After that, wrapper feature extraction method is used to select lncRNAs, which are closely related to the pathophysiology of disease from massive data. Experimentation on three typical open datasets proves the feasibility and efficiency of our proposed solution. This proposed solution achieves above 93% accuracy, the best ever.

KEYWORDS

Binary Particle Swarm Optimization, Expression Profile, Extreme Learning Machine, Long Non-Coding RNA

INTRODUCTION

Bioinformatics research of Long non-coding RNAs (lncRNAs) has attracted much attention in academia and industry because of the important role of gene expression in the genome. lncRNAs are defined as transcripts larger than 200nt in length with limited protein-coding potential. LncRNAs cover a large part of the non-coding information of the human DNA, representing over 90% of the whole genome. Furthermore, recent studies showed that lncRNAs are involved in the pathophysiology in various ways, e.g., gene expression, transcription, and post-translational processing.

DOI: 10.4018/IJDWM.317092

*Corresponding Author

The initial lncRNA bioinformatics research mainly focuses on sequence acquisition and data collection, e.g., the functionalities to collect and annotate lncRNAs. However, with the deepening understanding of the datasets, more and more research has been transferred to data analysis and application. For instance, via the hypothesis of "Expression-related genes have a relevant function," "Interacting molecules have a relevant function," it is possible to evaluate the similarity between different lncRNAs and thus predict the relationship between lncRNA and corresponding disease. However, there are several pending technical challenges:

1. **Feature Extraction Challenge:** High-throughput genomics data has specific features, e.g., high dimension features and lack of noted samples. Therefore, the key technical challenge is the exploration of data distribution, characteristic patent, and potential relationships based on prior knowledge from a few labeled samples.
2. **Computation Challenge:** Under the circumstances of the huge amount of genomics data, the design of computation models, especially lightweight, intelligent, and efficient computation algorithms, is waiting for an urgent answer.
3. **Transfer Learning Challenge:** In case of data distribution changes (for instance, gene expression data change from one species to another), the seamless transfer from the previous training model to another field is another technical challenge.

This paper investigates lncRNA-related issues and proposes a generic, lightweight, intelligent, and efficient computing model to predict key lncRNA related to disease pathophysiology. There are three contributions in our work:

1. We proposed a Binary PSO-based algorithm for selecting possible lncRNA subsets based on extracted features and logical connections. As a result, it is possible to acquire optimal lncRNA subset via multiple and iterative optimization.
2. ELM-based classification model is imported and implemented to evaluate each lncRNA's influence on disease. The evaluation result is used to guide future selection preferences.
3. We selected three datasets for experiment and evaluation: breast invasive carcinoma, carcinoma of the colon, and lung adenocarcinoma data. The result shows that our proposed solution achieves 93.6% classification accurate, which is the best.

The rest of the paper is organized as follows. We first present the relationship between lncRNA and disease (especially in the field of cancer), and then describes existing machine learning-based lncRNA research. Next, we introduce the lncRNA data collection, pretreatment, and noise filtering. The next section introduces the proposed BPSO-ML-ELM solution for lncRNA function prediction. We then illustrate the corresponding experiment, evaluation, and discuss the proposed solution in three datasets. Finally, the conclusion and future works are suggested.

RELATED WORKS

LncRNA and Disease

LncRNAs are a group of RNA transcripts ranging in length from 200 nt to 100 kilobases (kb), yet lack significant open reading frames (ORFs) and have no protein-coding capacity. However, recent research has found their aberrant regulation in various diseases, especially in different cancers. Biological experiments have shown that mis-regulated lncRNAs expression across many cancer types shows that aberrant lncRNAs expression is a major contributor to tumorigenesis. Conversely, many lncRNA that are up-regulated in cancers play an important role in promoting cell proliferation, invasion, and metastasis, such as H19, HOTAIR, MALAT1, and HULC.

However, with the rapid development of high throughput sequencing and lncRNA chip technology, more and more lncRNAs have been found. In the meantime, traditional biological experiments show more and more technical limitations, especially in time consumption and laborious work. Therefore, the import of machine learning concepts in the research of genomics data for predicting key lncRNA, is urgent.

LncRNA Bioinformatics Research

The initial bioinformatics research in lncRNA is to obtain sequence data and construct a database from mainly two aspects: the experimental data collected from published literature (for example, lncRNA sequence and relational database LncRNAdb, LncRNA2Target, and LncRNADisease), and high-throughput gene databases, such as lncRNA gene expression information database ChIPBase, lncRNA-protein interaction database NPInter, and lncRNA-cancer control database starBase.

After that, lncRNA bioinformatics research shifted from data acquisition to data application. For example, the literature imports network inference algorithm into collected lncRNA-gene regulation data to predict potential regulation relationships, and establish a new lncRNA-gene regulation database LncReg. Literature establishes lncRNA and gene co-expression network based on genechip data and then uses network propagation algorithm to carry out the lncRNA function annotation to improve prediction accuracy. Upon the priori data of LncRNADisease database, the literature implements a semi-supervised learning framework to predict the relationship between lncRNA and disease and achieves 81.3% -84.7% accuracy.

DATASET COLLECTION

DataSet Selection

We crawled the experimental datasets from Cancer RNA-Seq Nexus (CRN), the first public database providing phenotypes-specific coding transcript/lncRNA expression profiles and mRNA-lncRNA coexpression networks in cancer cells. CRN contains 54 human cancer RNA-Seq data sets, including 326 phenotype-specific subsets and 11030 samples. Each subset is a group of RNA-Seq samples associated with a specific phenotype or genotype, e.g., breast cancer stage II, ER+ breast cancer, and Her2+ breast cancer. CRN also provides a user-friendly interface to efficiently organize and visualize coding-transcript/lncRNA expression profiles.

Inside the CRN database, we choose breast invasive carcinoma, colon adenocarcinoma, and lung adenocarcinoma as the experimental datasets for three reasons. First, the three datasets have representativeness. For instance, breast cancer is the leading cause of cancer deaths among women worldwide; colon adenocarcinoma is thought to have the third highest incidence of cancer worldwide; lung cancer has been one of the most malignant diseases with the highest morbidity and mortality and the greatest threat to people’s health and life. Second, the three data sets have relatively mature preliminary research and a reliable labeled sample. Third, the computational models available for three different gene expression profiles shows good generalization capability.

DateSet Crawler

We developed a data crawler to achieve three datasets from CRN via three steps:

1. We selected breast invasive carcinoma (BRCA), colon adenocarcinoma (COAD), and lung adenocarcinoma (LUAD), and from these three datasets, we selected two control subsets with a high-quality and suitable scale from each dataset.
2. We set appropriate selection parameters, then crawled the chosen six subsets from the expression profiles page on CRN.

3. We divided crawled data into several parts through lncRNA's name, transcript ID, and FPKM (Fragments Per Kilobase per Million). Then, classified storage facilitates the next preprocessing.

Finally, three types of datasets were achieved, and the description is illustrated in Table 1. Specifically, we show the details of the three types of data sets in Table 2.

An early cancer diagnosis is crucial to effective treatment of the patients and higher survival rates. So that we crawled cancer subsets at the early stage (Stage I) and normal stage to make our experiment more meaningful in early cancer prediction.

Table 1. Three Typical Open Data Sets

Disease	LncRNAs number	Subset Number	Total Samples	Selected Subset	Samples
Breast invasive carcinoma	1286	13	1191	Stage I	90
				Normal	111
Colon adenocarcinoma	1221	10	479	Stage I	74
				Normal	40
Lung adenocarcinoma	1044	9	572	Stage IA	133
				Normal	59

Table 2. The Detail of Three Typical Open Data Sets

Selection	Breast invasive carcinoma		Colon adenocarcinoma		Lung adenocarcinoma	
	Subset Name	Samples	Subset Name	Samples	Subset Name	Subset Name
1	Stage I	90	Stage I	Stage I	Stage IA	133
2	Normal (adjacent normal)	111	Normal (adjacent normal)	Normal (adjacent normal)	Normal (adjacent normal)	59
	Metastatic Stage IIB	3	Stage II	Stage II	Stage I	5
	Stage IA	84	Stage IIA	Stage IIA	Stage IB	140
	Stage IB	9	Stage IIB	Stage IIB	Stage IIA	51
	Stage II	3	Stage III	Stage III	Stage IIB	73
	Stage IIA	360	Stage IIIA	Stage IIIA	Stage IIIA	73
	Stage IIB	246	Stage IIIB	Stage IIIB	Stage IIIB	11
	Stage IIIA	152	Stage IIIC	Stage IIIC	Stage IV	27
	Stage IIIB	29	Stage IV	Stage IV		
	Stage IIIC	65	Stage IV	Stage IV		
	Stage IV	22				
	Stage X	14				

COMPUTATION MODEL

System Framework

Figure 1 illustrates the proposed computation model, with a description in the following three steps:

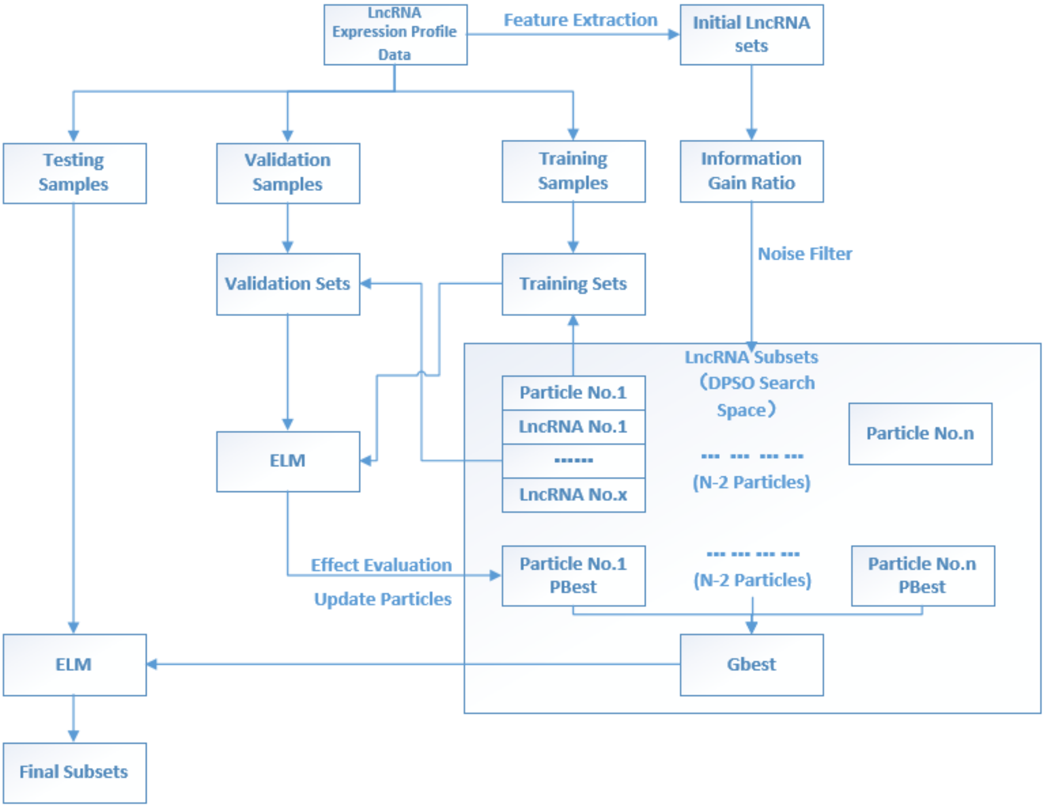
Step 1: Data Pre-Processing. The information gain ratio is proposed to reduce high dimensional features in lncRNA expression data and generate lncRNA candidate subsets.
Step 2: Optimum lncRNA Subset Selection. A binary PSO approach (with particle position and velocity parameters randomly initialized) is imported to select the possible optimum lncRNA subsets in the search space.
Step 3: Key lncRNA prediction. ML-ELM based approach is implemented to evaluate and predict possible optimum lncRNA subset selected.

Note that steps (3) and (2) could be iterated until the optimal classification accuracy reached a predefined threshold value.

Feature Extraction

To extract features from raw data, we import the approach of information gain ratio for dealing with high-dimensional data sets. First, the information gain ratio method can quickly filter many non-critical noise characteristics and narrow the optimization feature subset search range. Second, this

Figure 1. Computational Model



approach is more reliable than the original information gain, which may contain too much deviation value. We illustrate the computational formula as follows.

Inside, Gain(A) could be calculated:

$$Gain - ratio = Gain\left(A\right) / I \tag{1}$$

$$Gain\left(S,A\right) = E\left(S\right) - E\left(S,A\right) \tag{2}$$

$$E\left(S\right) = -\sum_{i=1}^{c} p_i \log_2\left(p_i\right) \tag{3}$$

$$E\left(S,A\right) = \sum_{v \in Value\left(A\right)} \frac{\left|S_v\right|}{\left|S\right|} E\left(S_v\right) \tag{4}$$

where A refers to a specific attribute, S refers to a sample set, S_v refers to the sample subset whose attribute A is equal to v in S.

Therefore, the information gain rate for each lncRNA could be computed to obtain an expression ratio value representing lncRNA-Disease interaction. We set a predefined threshold value as 0.6, and finally, 312 lncRNA are selected.

Binary Particle Swarm Optimization

Particle Swarm Optimization (PSO) algorithm is primarily used to solve numerical calculation issues. In PSO, particles flying through the multidimensional space to find the optimum typically model the swarm. We represent a potential solution to a problem as a particle with coordinates X_i and rate of change V_i in a multidimensional space. If the search space is D-dimensional and there are s particles in the swarm, the position of particle i is represented by the D-dimensional vector $X_i = \left(x_{i1}, \ldots, x_{id}, \ldots, x_{iD}\right)$. The velocity of this particle can be denoted as another vector $V_i = \left(v_{i1}, \ldots, v_{id}, \ldots, v_{iD}\right)$, which determines the flying direction and distance of the particle. The velocity of each particle and its new position can be updated according to formulas (5) and (6):

$$v_{id}\left(t+1\right) = wv_{id}\left(t\right) + c_1 r_1 \left[p_{id}\left(t\right) - x_{id}\left(t\right)\right] + c_2 r_2 \left[p_{gd} - x_{id}\left(t\right)\right] \tag{5}$$

$$x_{id}\left(t+1\right) = x_{id}\left(t\right) + v_{id}\left(t+1\right) \tag{6}$$

where t = 1, 2, ...; d = 1, 2, ..., D; i = 1, 2, ..., s; w is the inertia weight; c_1 and c_2 are the acceleration constants; r_1 and r_2 are random numbers uniformly distributed in [0, 1]; $v_{id}\left(t\right)$ and x_{id} are the speed and the position of particle i in the dth dimensional component at the t-th iteration, respectively

Furthermore, BPSO extends the application of PSO algorithm to optimize the discrete combinatorial problem [14]. In BPSO, particles represent the positions in a binary space. Each particle's position vector component receives a binary value, 0 or 1. Formally, $x_i \in b_n$, or $x_{id} \in \left\{0,1\right\}$. In a binary space, we may see a particle to move within a hypercube by flipping various numbers of bits; their velocity represents the probability that a bit will be in one state or the other. Since it is a probability, it should be constrained to the interval [0,1]. A sigmoid function $sig\left(v_{id}\right)$ can be used to accomplish this modification as:

$$sig\left(v_{id}\right) = \frac{1}{1 + e^{-v_{id}}} \tag{7}$$

The resulting position change is redefined as:

$$x_{id}\left(t+1\right) = \begin{cases} 1, & rand() < sig\left(v_{id}\left(t+1\right)\right) \\ 0, & else \end{cases} \tag{8}$$

where rand () is a random number selected from a uniform distribution interval [0, 1].

Considering that selecting optimal subsets in the massive lncRNA portfolio is a discrete combinatorial optimization problem, BPSO will be a typically suitable solution. Specifically, we consider all lncRNA as the input in the search space of BPSO. An optimal binary particle (an optimum lncRNA subset) could eventually be acquired via iterative particle swarm learning and self-improvement.

Multi-Layer Extreme Learning Machine

Furthermore, we propose a Multi-Layer Extreme Learning Machine based approach to evaluate the prediction accuracy of selected lncRNA subsets, and to give feedback on adjusting fitness function for BPSO. The Extreme Learning Machine (ELM) is a simple and effective learning algorithm based on single hidden layer feed-forward neural networks (SLFNs). Associate professor Huang Guangbin of Nanyang Technology University proposed the method in 2004. Unlike conventional neural network algorithms, ELM uses SLFNs to achieve fast training and overcome the over-fitting problem (Figure 2).

Suppose there are N random samples $\left(x_i, t_i\right)$, in which:

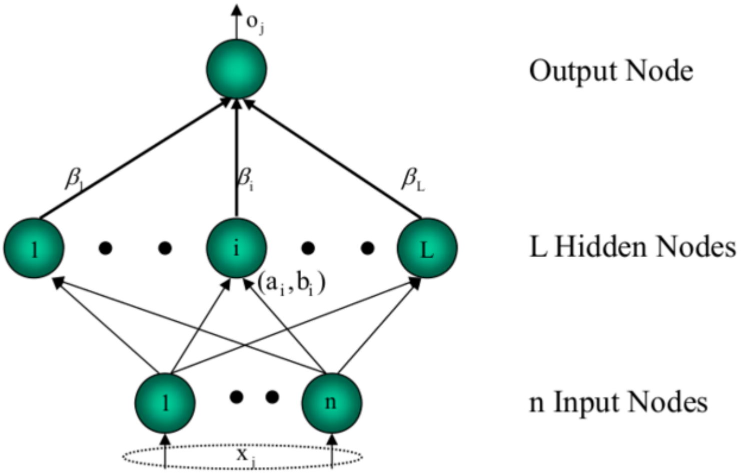
$$[X_i = x_{i1}, x_{i2}, \ldots, x_{in}]^T \in R^n \tag{9}$$

$$[t_i = t_{i1}, t_{i2}, \ldots, t_{im}]^T \in R^m \tag{10}$$

One SLFN contains L hidden layer nodes can be shown as:

$$\sum\nolimits_{i=1}^{L} \beta_i g\left(w_i \cdot x_j + b_i\right) = o_j, j = 1, 2, \ldots, N \tag{11}$$

Figure 2. The Structure of SLFN

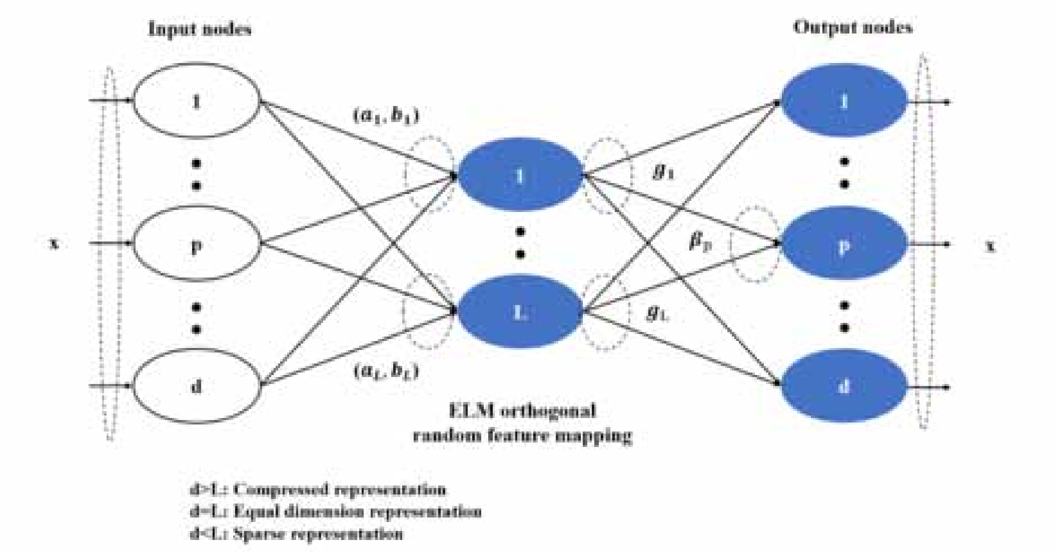


In formula (4.10), $g(x)$ is inspirit function, $[W_i = w_{i1}, w_{i2}, ..., w_{in}]^T$ is input weight, $[\beta_i = \beta_{i1}, \beta_{i2}, ..., \beta_{im}]^T$ is output weight, b_j is the bias of, $w_i \cdot x_j$ is the inner product of w_i and x_j.

The ML-ELM is an efficient algorithm combined with deep learning concepts for self-encoding. It can realize the feature expression of high latitude, equal dimension, and low latitude for original data. Widrow et al. (2013) proposed an ELM Automatic Encoder (ELM-AE) based on the least mean square method. Orthogonalization of randomly generated hidden parameters can enhance the generalization capability of ELM-AE.

The ELM-AE features a slight improvement from the ELM model. The network structure of the ELM-AE is shown in Figure 3. The hidden layer's input weights and bias parameters can be randomly

Figure 3. The Structure of ML-ELM-AE



set and normalized, which enables the mapping of input data into different dimensions to realize the expression of distinct features. Compared to the traditional deep learning network, the ML-ELM is based on initializing the hidden-layer weights according to the ELM-AE. However, the ML-ELM does not require an iterative adjustment. The network architecture of the ML-ELM is shown in Figure 4.

The ML-ELM hidden-layer activation function can be linear or nonlinear. If the number of nodes L_k in the k_{th} hidden layer is equal to the number of nodes $L_{k\text{-}1}$ in the $k-1_{th}$ hidden layer, then the activation function can be linear; otherwise, it is nonlinear, such as the sigmoidal function:

$$H^k = g\left(\left(\beta^k\right)^T H^{k-1}\right) \tag{12}$$

where H^k is the i_{th} hidden-layer output matrix. We selected some key lncRNAs as classification features in our experiments, then used ML-ELM to build the classifier. All health and cancer samples were divided into two parts: training samples and testing samples. ML-ELM was used on the training samples to generate the classification model. Meanwhile, ML-ELM was used on the testing sample to generate prediction results. According to the prediction results, we evaluated the performance of the classification model built by key lncRNAs subsets.

We show the flow of the whole experiment in Figure 5.

EXPERIMENT AND ANALYSIS

Experimental Setup

To verify the validity of the algorithm, we carried out a simulation experiment on the breast cancer data sets. We implemented all simulation experiments in MATLAB r2014a. Our computer used an Intel(R) Core (TM) i7-7700HQ CPU and had 8.00GB RAM. We had 201 breast cancer data sets samples, including 101 training sets, 50 validation sets, and 50 testing sets. We set the number of

Figure 4. The Structure of ML-ELM

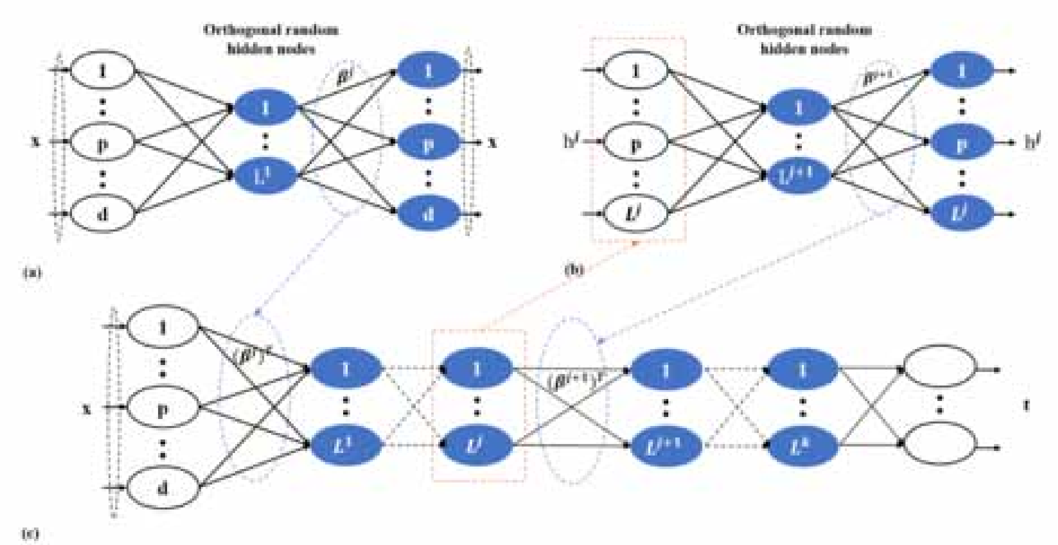
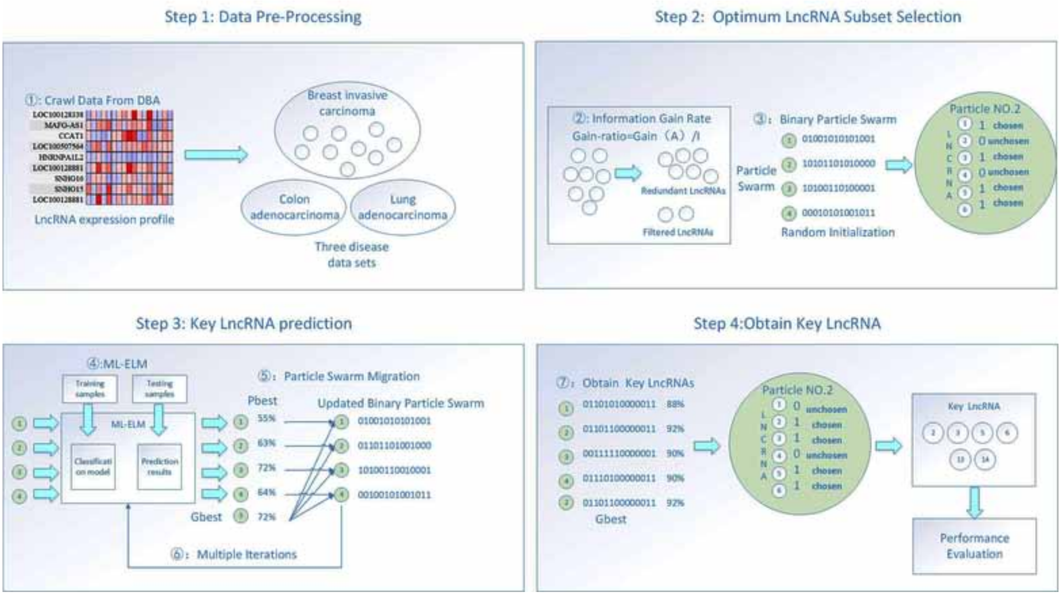


Figure 5. The Flow of the Whole Experiment



particles to 10, the maximum number of iterations as 30, and the number of key lncRNA to 10. The activation function used by the ML-ELM classifier is sigmoid, and we increase the number of nodes in the hidden layer from one to equal to the number of the training sample. We divided the lncRNA expression profiles into training, testing, and validation sets at 50%, 25%, and 25%, respectively (Table 3).

Performance

Figures 6 and 7 show that with the increasing number of hidden layer nodes, the prediction accuracy and precision of training are gradually rising. When hidden layer nodes are 40, our algorithm achieves the best possible accuracy, and then the accuracy decreases gradually. Therefore, we set the number of hidden layer nodes as 40 in the following experiments. In the figures below, each colored point represents a set of experimental accuracy, and the broken line represents the overall trend of experimental accuracy with the change in the number of hidden nodes.

Figure 8 shows that the classification accuracy of the randomly generated lncRNA sets is very low, but with the increase in the number of iterations, the iterative precision increases and finally becomes stable.

We conducted a series of experiments and showed the influence of a different number of key lncRNA on the experiment accuracy (shown in Table 4). Finally, we selected a group of 10 key lncRNAs and achieved a 99.47% training accuracy and a prediction accuracy of 93.81% on the breast cancer data set.

Table 3. The Number of Samples in Each Subsets

Subset name	Total subset	Training sets	Testing sets	Validation sets
Breast invasive carcinoma	201	100	51	50
Colon adenocarcinoma	114	57	29	28
Lung adenocarcinoma	192	96	48	48

Figure 6. Comparison of Prediction Accuracy of Different Number of Hidden Nodes

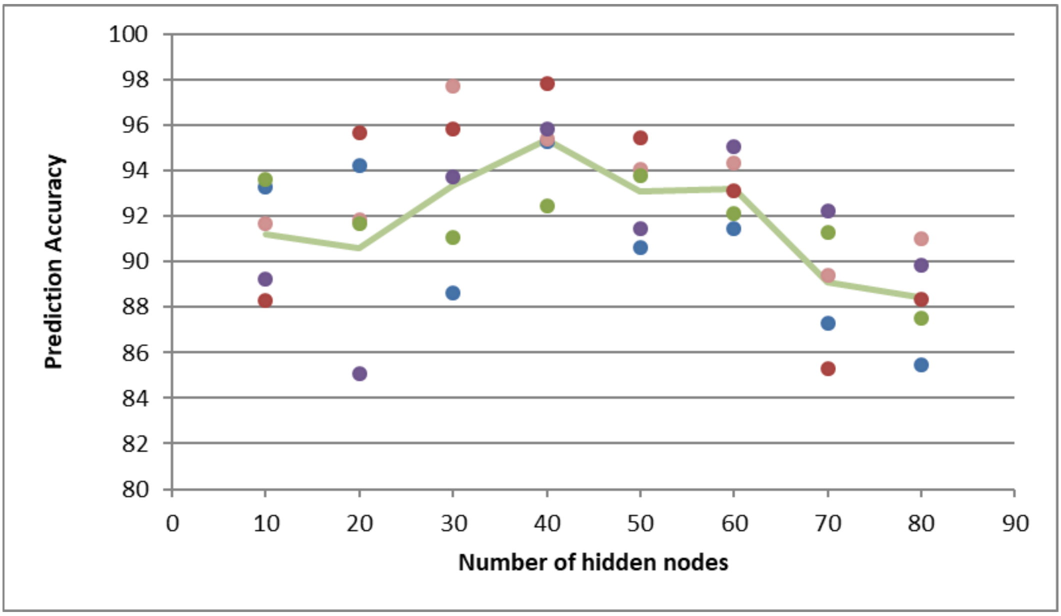
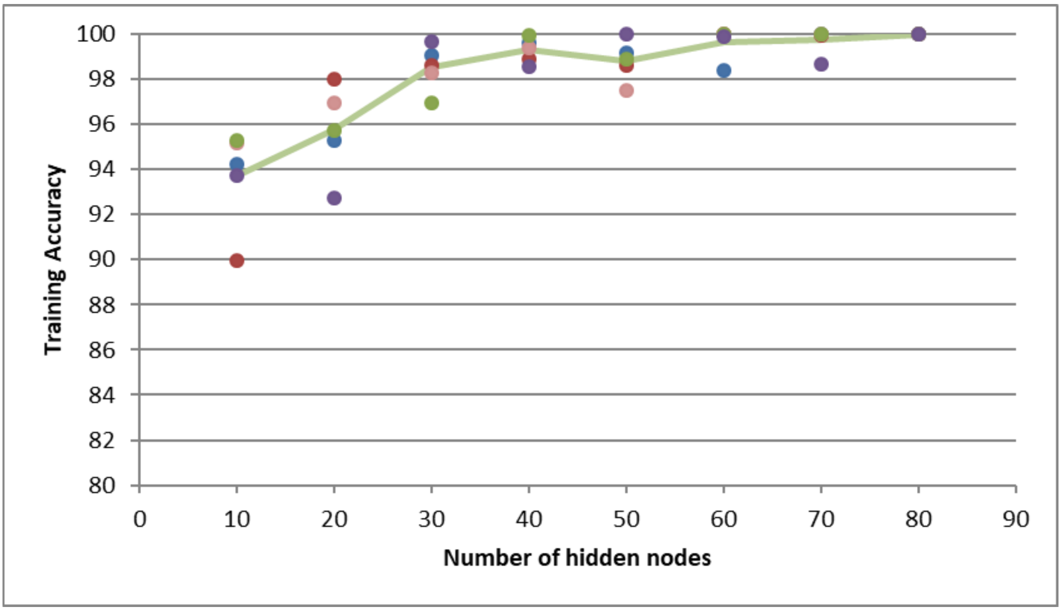


Figure 7. Comparison of Training Accuracy of Different Number of Hidden Nodes



Algorithm Comparison

Finally, we evaluated the effect of ML-ELM by replacing the ML-ELM model with a conventional machine learning method (KNN algorithm).

We compared our prediction model with other classical methods. To ensure the reliability of algorithm comparison, we used other classical methods to predict the same data set and compared the

Figure 8. Comparison of Prediction Accuracy of Different Number of Iterations

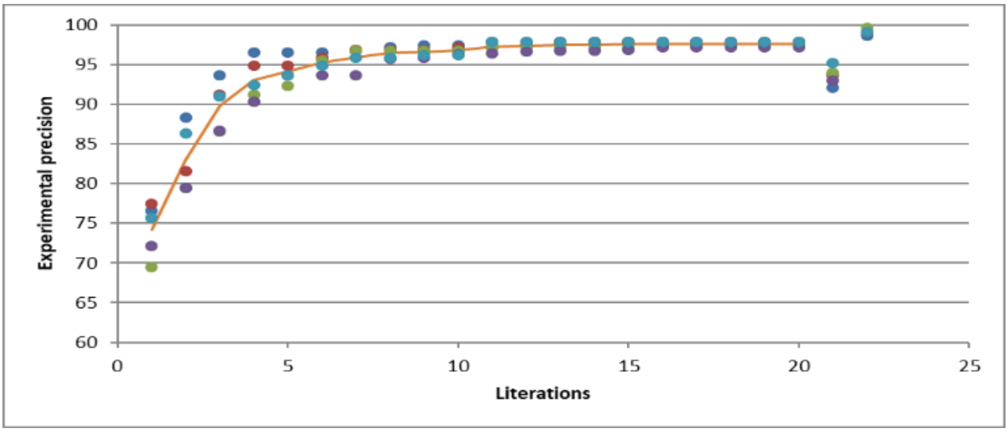


Table 4. The Influence of a Different Number of Key lncRNA on the Experiment Accuracy

The number of key lncRNA	Prediction accuracy	Training accuracy
5	91.49	99.12
10	93.81	99.47
15	93.69	98.57
20	90.45	98.73

prediction accuracy and computing speed. We chose PSO-KNN to compare with our core predictive model (Table 5).

K-nearest neighbor classification algorithm (KNN) is one of the most classical methods in data mining classification. PSO-KNN is the classical algorithm in feature selection and has been applied in gene selection research. Figure 9 and Figure 10 show that the PSO-ELM algorithm can achieve better accuracy. Besides, the computation time is significantly reduced in our proposed approach.

Table 5. PSO-ELM VS PSO KNN

Classification Method	Iterations	Breast invasive carcinoma		Colon adenocarcinoma		Lung adenocarcinoma	
		Time	Accuracy	Time	Accuracy	Time	Accuracy
PSO-ML-ELM	10	1.8267	88.03	2.3723	94.20	3.5187	88.67
	20	7.1833	92.21	5.3427	96.09	6.9730	92.87
	30	7.94	93.26	8.1543	96.10	10.5090	94.13
PSO-KNN	10	12.3967	88.67	8.7833	84.94	8.7773	91.11
	20	19.6593	89.33	16.7633	90.29	17.0587	92.11
	30	34.1763	90.00	25.3207	92.48	25.8733	91.67

Figure 9. Comparison of the Time of ML-ELM and KNN

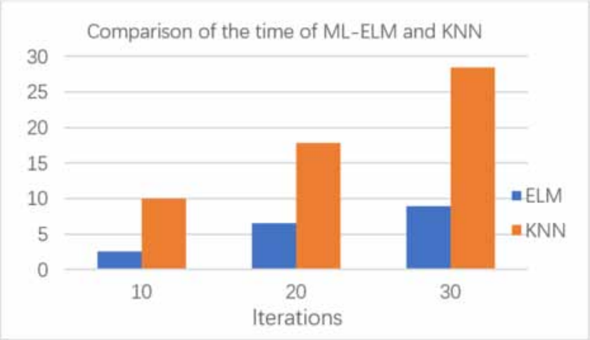
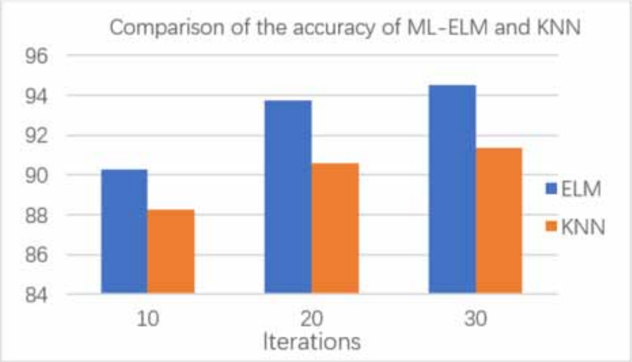


Figure 10. Comparison of the Accuracy of ML-ELM and KNN



Performance Evaluation

To quantify the classification performance, we used seven kinds of standard metrics: Sensitivity (TPR), False negative rate (FNR), False positive rate (FPR), Specificity (TNR), Positive predictive value (PPV), Accuracy, and F-score, which are defined as follows:

$$Sensitivity = \frac{TP}{\left(TP + FN\right)} \tag{13}$$

$$FNR = \frac{FN}{\left(TP + FN\right)} \tag{14}$$

$$FPR = \frac{FP}{\left(FP + TN\right)} \tag{15}$$

$$Specificity = \frac{TN}{\left(TN + FP\right)} \quad (16)$$

$$PPV = \frac{TP}{\left(TP + FP\right)} \quad (17)$$

$$Accuracy = \frac{TP + TN}{P + N} \quad (18)$$

$$F - score = \frac{2 \cdot Sensitivity \cdot PPV}{Sensitivity + PPV} \quad (19)$$

In these formulas, T is the number of cancer samples, and P is the number of healthy samples. TP is the number of correctly classified cancer samples, TN is the number of correctly classified healthy samples, FP is the number of falsely classified cancer samples, and FN is the number of falsely classified healthy samples. Sensitivity is the proportion of correctly classified cancer samples in the set of all cancer samples. FNR is the proportion of falsely classified cancer samples in the set of all cancer samples. FPR is the proportion of falsely classified healthy samples in the set of all healthy samples. Specificity is the proportion of correctly classified healthy samples in the set of all healthy samples. PPV is a ratio of true cancer samples to combined true and false cancer samples. Accuracy is the ratio of correctly classified samples in all samples. F-score is the harmonic mean of sensitivity, and we can use PPV as a single measure for the overall classification performance.

The performance evaluation result in Table 6 illustrates that our model is accurate and sensitive.

Through the performance evaluation of the experimental results, we find that our prediction model can get desired results in the three data sets, especially in the colon adenocarcinoma data set, we get an accuracy of 0.956 and the misdiagnosis rate of only 0.041, which proves that the prediction model is very reliable.

DISCUSSION

From the computation model description and the experiment work, the following crucial points should be noted. Given a specific disease and corresponding lncRNA, selecting key lncRNA that is closely related to the specific disease is feasible. Reversely, according to the change of some specific

Table 6. Performance Evaluation

Cancer	Sensitivity	FNR	FPR	Specificity	PPV	Accuracy	F-score
Breast invasive carcinoma	0.944	0.056	0.072	0.928	0.914	0.935	0.929
Colon adenocarcinoma	0.959	0.041	0.050	0.950	0.973	0.956	0.966
Lung adenocarcinoma	0.940	0.060	0.051	0.949	0.977	0.943	0.958

lncRNAs, it may be also possible to predict possible disease. This outcome will be typically important for researching the pathogenesis of the disease and therapy.

Besides, through querying LncRNA-disease, we found that some of the key lncRNA we acquired (such as DLEU2, SNHG3, and TINCR) were labeled as cancer-related lncRNA (Table 7). The description of three key lncRNA we gained in LncRNA-disease further proved that our approach can provide guidance for biological discovery of lncRNAs closely related to cancer in the future.

The performance of the proposed computation model depends on a few features. First, the reliable noted dataset in lncRNA-disease relationship is the foundation for the accuracy of model training and later testing; Second, the computation time mainly depends on the complexity of the proposed machine learning algorithm.

Theoretically, the proposed computation model is a generic and lightweight solution applicable in any "lncRNA-disease" annotation use case. Furthermore, we have also tested and proved the experiment's feasibility and correctness in three specific and representative cancer datasets. Therefore, it is convincible that the model contains good transfer capability.

CONCLUSION AND FUTURE WORKS

This paper proposes a computation model for the annotation between key lncRNA and corresponding disease. The proposed solution is based on the combination of BPSO and ML-ELM algorithms. The theoretical analysis and experimental results show that the proposed solution can predict an optimal lncRNA subset closely related to specific diseases and may also predict specific diseases according to

Table 7. The Description of Three Key lncRNA we Acquired in LncRNA-Disease

LncRNA name	Disease name	Dysfunction type	Description	Chr	Start	End	Strand	Species
DLEU1	chronic lymphocytic leukemia	Locus	In 13q14.3, where several tumor suppressor genes, including the miRNA genes miR-16-1 and miR-15a, are co-regulated by the two long non-coding RNA genes DLEU1 and DLEU2 that span the critical region.	chr13	50082169	50528643	+	Human
SNHG3	hepatocellular carcinoma	Expression	SNHG3 correlates with malignant status and poor prognosis in hepatocellular carcinoma.	chr1	28505943	28510892	+	Human
TINCR	squamous cell carcinoma	Expression	Interestingly, the lncRNA TINCR, which is highly induced during keratinocyte differentiation, is repressed in squamous cell carcinoma specimens compared to the normal adjacent epidermis.	chr19	5558167	5568034	-	Human

the change in lncRNA expression profile. The work is constructive with the development of disease precision in medical therapy.

ACKNOWLEDGMENT

This research was supported by the R&D Plan of Guangdong Province in key areas [2020B0101090005]; the specific research fund of The Innovation Platform for Academician of Hainan Province [YSPTZX202145]; Fujian Provincial Science and Technology Department Guided Project [2022H0012]; and the local special projects supported by The Central Government [2022L3003].

REFERENCES

Adams, B. D., Parsons, C., Walker, L., Zhang, W. C., & Slack, F. J. (2017). Targeting noncoding RNAs in disease. *The Journal of Clinical Investigation*, *127*(3), 761–771. doi:10.1172/JCI84424 PMID:28248199

Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Xhang, Q., Yan, G., & Cui, Q. (2012). LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Research*, *41*(D1), D983–D986. doi:10.1093/nar/gks1099 PMID:23175614

Chen, J. S., Wang, Y. F., Zhang, X. Q., Lv, J. M., Li, Y., Liu, X. X., & Xu, T. P. (2016). H19 serves as a diagnostic biomarker and up-regulation of H19 expression contributes to poor prognosis in patients with gastric cancer. *Neoplasma*, *63*(2), 223–230. doi:10.4149/207_150821N454 PMID:26774144

Chen, Y., & Zhou, J. (2017). LncRNAs: Macromolecules with big roles in neurobiology and neurological diseases. *Metabolic Brain Disease*, *32*(2), 281–291. doi:10.1007/s11011-017-9965-8 PMID:28161776

Colameco, S., & Elliot, M. A. (2017). Non-coding RNAs as antibiotic targets. *Biochemical Pharmacology*, *133*, 29–42. doi:10.1016/j.bcp.2016.12.015 PMID:28012959

Guo, X., Gao, L., Liao, Q., Xiao, H., Ma, X., Yang, X., Luo, H., Zhao, G., Bu, D., Jiao, F., Shao, Q., Chen, R., & Zhao, Y. (2013). Long non-coding RNAs function annotation: A global prediction method based on bi-colored networks. *Nucleic Acids Research*, *41*(2), e35–e35. doi:10.1093/nar/gks967 PMID:23132350

Hao, Y., Wu, W., Li, H., Yuan, J., Luo, J., Zhao, Y., & Chen, R. (2016). NPInter v3.0: An upgraded database of noncoding RNA-associated interactions. *Database (Oxford)*, *2016*. doi:10.1093/database/baw057 PMID:27087310

Jiang, Q., Wang, J., Wu, X., Ma, R., Zhang, T., Jin, S., Han, Z., Tan, R., Peng, J., Liu, G., Li, Y., & Wang, Y. (2015). LncRNA2Target: A database for differentially expressed genes after lncRNA knockdown or overexpression. *Nucleic Acids Research*, *43*(D1), D193–D196. doi:10.1093/nar/gku1173 PMID:25399422

Kar, S., Sharma, K. D., & Maitra, M. (2015). Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. *Expert Systems with Applications*, *42*(1), 612–627. doi:10.1016/j.eswa.2014.08.014

Kazimierczyk, M., Kasprowicz, M. K., Kasprzyk, M. E., & Wrzesinski, J. (2020). Human long noncoding RNA interactome: Detection, characterization and function. *International Journal of Molecular Sciences, 21*(3), . doi:1027

Koch, L. (2017). Screening for lncRNA function. *Nature Reviews. Genetics*, *18*(2), 70. doi:10.1038/nrg.2016.168 PMID:28045101

Lee, N. K., Lee, J. H., Ivan, C., Ling, H., Zhang, X., Park, C. H., Calin, G. A., & Lee, S. K. (2017). MALAT1 promoted invasiveness of gastric adenocarcinoma. *BMC Cancer*, *17*(1), 1–12. doi:10.1186/s12885-016-2988-4 PMID:28077118

Li, J., Chen, Z., Tian, L., Zhou, C., He, M. Y., Gao, Y., Wang, S., Zhou, F., Shi, S., Feng, X., Sun, N., Liu, Z., Skogerboe, G., Dong, J., Yao, R., Zhao, Y., Sun, J., Zhang, B., Yu, Y., & He, J. (2014). LncRNA profile study reveals a three-lncRNA signature associated with the survival of patients with oesophageal squamous cell carcinoma. *Gut*, *63*(11), 1700–1710. doi:10.1136/gutjnl-2013-305806 PMID:24522499

Li, J. R., Sun, C. H., Li, W., Chao, R. F., Huang, C. C., Zhou, X. J., & Liu, C. C. (2016). Cancer RNA-Seq Nexus: A database of phenotype-specific transcriptome profiling in cancer cells. *Nucleic Acids Research*, *44*(D1), D944–D951. doi:10.1093/nar/gkv1282 PMID:26602695

Liao, Q., Liu, C., Yuan, X., Kang, S., Miao, R., Xiao, H., Zhaoi, G., Luo, H., Bu, D., Zhao, H., Skogerbø, G., Wu, Z., & Zhao, Y. (2011). Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Research*, *39*(9), 3864–3878. doi:10.1093/nar/gkq1348 PMID:21247874

Ma, X., Huang, C., Luo, D., Wang, Y., Tang, R., Huan, X., Zhu, Y., Xu, Z., Liu, P., & Yang, L. (2016). Tag SNPs of long non-coding RNA TINCR affect the genetic susceptibility to gastric cancer in a Chinese population. *Oncotarget*, *7*(52), 87114–87123. doi:10.18632/oncotarget.13513 PMID:27893425

Quek, X. C., Thomson, D. W., Maag, J. L., Bartonicek, N., Signal, B., Clark, M. B., Glass, B. S., & Dinger, M. E. (2015). lncRNAdb v2. 0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Research*, *43*(D1), D168–D173. doi:10.1093/nar/gku988 PMID:25332394

Ramanathan, L., Dhanda, S., & Kumar, D. (2013). Predicting students' performance using modified ID3 algorithm. *IACSIT International Journal of Engineering and Technology*, *5*(3), 2491–2497.

Sanchez Calle, A., Kawamura, Y., Yamamoto, Y., Takeshita, F., & Ochiya, T. (2018). Emerging roles of long non-coding RNA in cancer. *Cancer Science*, *109*(7), 2093–2100. doi:10.1111/cas.13642 PMID:29774630

Sasa, G. B., Xuan, C., Lyu, G., Ding, X., & Meiyu, F. (2022). Long non-coding RNA ZFPM2-AS1: A novel biomarker in the pathogenesis of human cancers. *Molecular Biotechnology*, *64*(7), 725–742. doi:10.1007/s12033-021-00443-3 PMID:35098483

Sousa-Ferreira, I., & Sousa, D. (2017). A review of velocity-type PSO variants. *Journal of Algorithms & Computational Technology*, *11*(1), 23–30. doi:10.1177/1748301816665021

Tan, C., Zuo, F., Lu, M., Chen, S., Tian, Z., & Hu, Y. (2022). Identification of potential genes correlated with breast cancer metastasis and prognosis. *All Life*, *15*(1), 126–133. doi:10.1080/26895293.2021.2021302

Torre, L. A., Siegel, R. L., & Jemal, A. (2016). Lung cancer statistics. Lung cancer and Personalized Medicine, 1-19. Springer. doi:10.1007/978-3-319-24223-1_1

Wang, J., Lu, S., Wang, S. H., & Zhang, Y. D. (2021). A review on extreme learning machine. *Multimedia Tools and Applications*, *81*, 41611–41660. doi:10.1007/s11042-021-11007-7

Widrow, B., Greenblatt, A., Kim, Y., & Park, D. (2013). The no-prop algorithm: A new learning algorithm for multilayer neural networks. *Neural Networks*, *37*, 182–188. doi:10.1016/j.neunet.2012.09.020 PMID:23140797

Wu, X., Li, J., Wang, Z., Hansen, H., & Zhao, Y. (2020). Cancer-associated fibroblasts derived extracellular vesicles promote pancreatic cancer tumor plasticity and metastatic capacity. *Pancreatology*, *20*(1), S142. doi:10.1016/j.pan.2020.07.269

Xu, B., Mei, J., Ji, W., Bian, Z., Jiao, J., Sun, J., & Shao, J. (2020). LncRNA SNHG3, a potential oncogene in human cancers. *Cancer Cell International*, *20*, 536. doi:10.1186/s12935-020-01608-x PMID:33292213

Yang, J. H., Li, J. H., Shao, P., Zhou, H., Chen, Y. Q., & Qu, L. H. (2011). starBase: A database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Research*, *39*(suppl_1), D202–D209. doi:10.1093/nar/gkq1056 PMID:21037263

Yu, X., Zheng, H., Chan, M. T., & Wu, W. K. K. (2017). HULC: An oncogenic long non-coding RNA in human cancer. *Journal of Cellular and Molecular Medicine*, *21*(2), 410–417. doi:10.1111/jcmm.12956 PMID:27781386

Zhang, H., Wang, X., Zhang, Q., Ma, Y., & Wang, Z. (2020). DLEU2 participates in lymphovascular invasion and inhibits cervical cancer cell proliferation, migration, and invasion. *International Journal of Clinical and Experimental Pathology*, *13*(8), 2018–2026. PMID:32922596

Zhang, Y. Y., Huang, S. H., Zhou, H. R., Chen, C. J., Tian, L. H., & Shen, J. Z. (2016). Role of HOTAIR in the diagnosis and prognosis of acute leukemia. *Oncology Reports*, *36*(6), 3113–3122. doi:10.3892/or.2016.5147 PMID:27748863

Zhou, K. R., Liu, S., Sun, W. J., Zheng, L. L., Zhou, H., Yang, J. H., & Qu, L. H. (2017). ChIPBase v2. 0: Decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Research*, *45*(D1), D43–D50. doi:10.1093/nar/gkw965 PMID:27924033

Zhou, Z., Shen, Y., Khan, M. R., & Li, A. (2015). LncReg: A reference resource for lncRNA-associated regulatory networks. *Database (Oxford)*, *2015*. doi doi:10.1093/database/bav083 PMID:26363021