Editorial: Machine learning-based methods for RNA data analysis—Volume II

Lihong Peng1,2*, Jialiang Yang3, Minxian Wang4,5 and Liqian Zhou1**

1 College of Life Sciences and Chemistry, Hunan University of Technology, Zhuzhou, China, 
2 School of Computer, Hunan University of Technology, Zhuzhou, China, 
3 Geneis (Beijing) Co Ltd., Beijing, China, 
4 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China, 
5 University of Chinese Academy of Sciences, Beijing, China

KEYWORDS
machine learning, lncRNA, microRNA, circRNA, mRNA, gene expression

RNAs regulate multiple biological processes including RNA transcription, splicing, stability, and translation. They play significant roles in cell biology (Connelly et al. (2016); Licatalosi and Darnell (2010); Mukherjee et al. (2022); Chen et al. (2018b)). The Encyclopedia of DNA elements project reported that only 1.5% of human genome is translated into proteins, while approximately 70%–90% is transcribed to RNAs (Falese et al. (2021)). RNAs greatly expand the range of targets from proteins to RNAs by re-targeting mutated targets (Yu et al. (2019); Chen et al. (2020); Li et al. (2022); Yang et al. (2022)). Particularly, noncoding RNAs have dense linkages with human diseases including cancers. Now, RNAs have been diagnostic or prognostic markers of complex diseases (Hui et al. (2011); Xu et al. (2022); Peng et al. (2022a); Shen et al. (2022); Zhang T. et al. (2022); Chai et al. (2022)). In this topic, we aim to analyze diverse RNA data to provide clues for the diagnosis and therapy of various diseases (Dal Molin et al. (2022); Wang S. et al. (2022); Li J. et al. (2019); Liu et al. (2020)). Long noncoding RNAs (lncRNAs) regulate many significant biological processes (such as immune response and embryonic stem cell pluripotency) by linking to RNA-binding proteins (Wapinski and Chang (2011); Chen and Huang (2017); Ping et al. (2018); Wang et al. (2020)). Wang et al. (2021 W.); Peng et al. (2020)). They have been important biomarkers for cancers (Wu et al. (2022a); Banerjee et al. (2020); Zhang S. et al. (2021); Zhou G. et al. (2021); Peng et al. (2022a); Liang et al. (2022b); Peng et al. (2021); Zhou L. et al. (2021)). For example, IncRNAs AFAP1-AS1, CCAT1, CYTOR, GAS5, HOTAIR, and PVT1 are molecular regulators of lung cancer (Aftahi et al. (2021)). KCNQ1OT1 may be a prognostic biomarker in colorectal cancer (Lin et al. (2021)). lncRNAs are also oncogenes (such as MKL1N-AS, GHT1, LASP1-AS, MALAT1, HULC, HOTAIR, and PAPAS) and tumor suppressors (such as CASC2, DGCR5, MEG3, GAS5, and NRON) in hepatocellular carcinoma (Guo et al. (2021)). Many machine learning methods have been proposed to...
infer new LncRNA-Disease Associations (LDAs). For example, graph convolutional completion with conditional random (Fan et al. (2022)), heterogeneous graph attention network with meta-paths (Zhao et al. (2022)), graph convolutional auto-encoders (Silva and Spinosa (2021)), multi-view attention graph convolutional network and stacking ensemble (Liang et al. (2022b)), and learning to rank-based model (Wu et al. (2022a)) are widely used methods for LDA prediction.

In this research topic, Sun et al. developed a weighted graph-regularized matrix factorization approach (LPI-WGRMF) to identify possible LncRNA-protein interactions (LPIs) based on known biological information and LPI matrix. LPI-WGRMF obtained an AUC of 0.9012 and AUPR of 0.7324 on LPI dataset provided by Zhang et al. (Zhang et al. (2018)) based on 5-fold cross validation. They predicted that LncRNAs SNHG3, SFPQ, and PRPF31 may interact with proteins Q9NU5L5, Q9NU5L5, and Q9UKV8, respectively. Yao et al. designed a random walk with restart algorithm (MHRWRRLDA) to infer LDAs on multiplex and heterogeneous networks. MHRWRRLDA computed an AUC of 0.6874 under leave-one-out cross validation, and inferred that LncRNA BCYRN1 may associate with colon cancer and hepatocellular carcinoma. Cheng et al. considered that the recurrence rate of nonfunctioning pituitary adenoma is relatively high after surgical resection and built LncRNA signatures for its prognosis. They obtained microarray sequencing profiles of LncRNA expressions from 66 patients who suffered from nonfunctioning pituitary adenoma. Univariable Cox regression analysis and random survival forests-variable hunting were applied to filter LncRNAs. They found that three LncRNAs, LOC101927765, RP11-23N2.4, and RP4-533D7.4, have dense associations with tumor recurrence and inferred that the three LncRNAs may be potential therapeutic targets of nonfunctioning pituitary adenoma.

MicroRNAs (miRNAs) are a class of endogenous noncoding RNAs with a length of approximately 22 nucleotides (Sun et al. (2022); Chen et al. (2019b, 2018b); Zhang L. et al. (2021)). miRNAs regulate many biological activities and influence almost all genetic pathways (Chen et al. (2018c); Peng et al. (2017); Chen et al. (2018a)). Thus, miRNAs have been a class of tumor suppressor genes in clinical medicine (Chen et al. (2019a); Peng et al. (2018)). For example, miR-940 is a potential biomarker of prostate cancer (Rajendiran et al. (2021)). Urinary exosome microRNA signatures are noninvasive prognostic markers for prostate cancer (Shin et al. (2021)). Recently, machine learning methods have been widely used to identify possible MicroRNA-Disease Associations (MDAs). For example, tensor decomposition with relational constraints (Huang et al. (2021)), similarity constrained matrix factorization (Li L. et al. (2021)), tensor factorization and label propagation (Yu et al. (2022)), deep attributed network embedding model (Ji et al. (2021)), and multi-view multichannel attention graph convolutional network (Tang et al. (2021)) are popular methods in MDA prediction.

In this topic, Qu et al. explored a computational model (BRWRMHMDA) for MDA inference combining enforcing degree-based biased random walk with restart. BRWRMHMDA computed an AUC of 0.8310 under leave-one-out cross validation. They predicted that hsa-let-7i and hsa-mir-30e may associate with esophageal neoplasms and breast neoplasms, respectively. Zhou et al. proposed a pseudogene-miRNA association identification method (PMGAE) by integrating feature fusion, graph autoencoder, and eXtreme gradient boosting. First, they computed three types of similarities for pseudogenes and miRNAs, that is, Pearson similarity, cosine similarity, and Jaccard similarity. Second, the above similarities were fused to build a similarity profile for each node. Third, the similarity profiles and pseudogene-miRNA associations are further aggregated to depict each node as a low-dimensional vector through a graph autoencoder. Finally, the feature vector was fed into eXtreme gradient boosting for pseudogene-miRNA association prediction. PMGAE computed better AUC of 0.8634 and AUPR of 0.8966. The results from PMGAE showed that miRNAs hsa-miR-34c-5p, hsa-miR-199b-5p, and hsa-miR-103a-3p may associate with pseudogenes RPLP0P2, HLA-H, and HLA-J, respectively.

Circle RNAs (circRNAs) is a class of novel endogenous noncoding RNAs with a covalently closed loop structure (Wang C.-C. et al. (2021); Li G. et al. (2019); Wang et al. (2021b)). circRNAs have more stable expressions due to their resistances to RNA exonuclease degradation (Li et al. (2020); Wang et al. (2021c,b)). They can regulate protein binding, miRNA sponges, alternative splicing and transcription, and generate pseudogenes (Wang C.-C. et al. (2021); Chen (2020)). In addition, they demonstrate close associations with cancers, cardiovascular and nervous system diseases (Wang C.-C. et al. (2021); Li G. et al. (2019, 2020); Wang et al. (2021c,C,b)). Therefore, various computational models have been developed to detect possible CircRNA-Disease Associations (CDAs). For example, network embedding and subspace learning method (Xiao et al. (2021)), knowledge attention network (Lan et al. (2022)), multi-source feature fusion-based machine learning framework (Wang L. et al. (2022)), and robust nonnegative matrix factorization model (Peng et al. (2022c)) are widely used in CDA prediction.

Furthermore, Li et al. developed a computational CDA identification method (GATGCGN) based on graph attention network and graph convolutional network. First, they fused several biomedical data from different sources through the centered kernel alignment model. Second, graph attention network was deployed to obtain latent representation of circRNAs and diseases. Finally, graph convolutional network was explored to infer CDAs. GATGCGN computed better an AUC of 0.951 under leave-one-out cross validation and an AUC of 0.932 under 5-fold cross-validation. They found that circRNAs hsa_circRNA_404833, hsa_circ_0013509, hsa_circRNA_2149,
circR_284, and circR_284 have the highest association scores with lung cancer, diabetes retinopathy, prostate cancer, cholangiocarcinoma, and clear cell renal cell carcinoma, respectively.

A large quantity of transcriptomic data enable us to investigate complex biological processes at single-cell resolution levels (Peng et al. (2022b); Liang et al. (2022a); Zhang et al. (2022b); Xu et al. (2020). Therefore, Miao et al. (2021) considered specific noises and computing efficiency, and then designed biologically interpretable integration strategies to integrate multi-omics single-cell data. Zhou P. et al. (2021) used multiscale stochastic dynamics to dissect transition cells from transcriptome data. Ye et al. (2022) used combinatorial hybrid sequencing to construct the axolotl cell landscape at single-cell resolution. McKellar et al. (2021) detected transitional progenitor states in mouse skeletal muscle regeneration based on single-cell transcriptomic data. Wu et al. (2022b) exploited a stacking ensemble learning-based model to implement single-cell Hi-C classification.

In particular, Panchy et al. analyzed large-scale transcriptome datasets using non-negative principal component analysis and non-negative matrix factorization. The results showed that the above two methods provided low-dimensional features for the progression of biological processes. They found that gene expression signatures from conserved epithelial-mesenchymal transition can be applied to depict the stages in multiple cell lines. Lang et al. evaluated the performance of two sequencing platforms (Nextseq500 and MGISEQ-2000) using the same capture DNA libraries built by the Illumina protocol. The results demonstrated that a significant loss of fragment occurred in the range of 101–133 bp sizes on MGISEQ-2000 for Illumina libraries while not for the capture DNA libraries. Bao et al. considered that it is crucial to differentiate the transcriptomic and proteomic profiles between unstable and stable atherosclerotic plaques. They obtained 5 unstable and 5 stable human carotid atherosclerotic plaques by carotid endarterectomy to identify IncRNA-targeted genes and circRNA-originated genes. The results indicated that 293 proteins, 488 lncRNAs, 91 circRNAs, and 202 mRNAs are differentially expressed between unstable and stable atherosclerotic plaques. Furthermore, CD5L, S100A12, CKB, CEMIP, and SH3GL1 may be key genes in regulating the stability of atherosclerotic plaques. In addition, Zheng et al. used a series matrix file search method and obtained data related to breast cancer from the ArrayExpress and Gene Expression Omnibus databases. They found that RSK2 is a possible biomarker in breast cancer.

RNA sequencing data have been broadly applied to screen therapeutic strategies for various diseases (Przybyla and Gilbert (2022); Zhang Y. et al. (2021); Li C.-x. et al. (2021)). Chen et al. (2022) used RNA sequencing to explore the mechanism of oxygen-boosted sonodynamic therapy for the treatment of hepatocellular carcinoma. Zhang et al. (2022c) integrated single-cell and bulk RNA sequencing data to probe a pan-cancer stemness signature. Sammut et al. (2022) combined multi-omics data including DNA and RNA sequencing and machine learning technique to predict breast cancer therapy response. Based on RAN sequencing data, Ma et al. first downloaded RNA sequencing data related to gliomas from the TCGA database. Then they used DESeq2, key driver and weighted gene correlation network to identify differentially expressed genes. They observed that Paclitaxel, Cidofovir, 6-benzyladenine, Erlotinib, Bilirubin, Oxaliplatin, Nutlins, Valproic acid, and Fenofibrate may be potential drugs in inhibiting the recurrence of gliomas. Similarly, Xiang et al. detected gene expression and network differences between limited and advanced stages for the diffuse large B-cell lymphoma (DLBCL) patients to predict potential agents against DLBCL. First, they collected RNA sequencing data from the DLBCL patients at different clinical stages from the TCGA database. Second, they used DESeq2 to identify differentially expressed genes and weighted gene correlation network and differential modules to analyze variations between different stages. Finally, they extracted important genes using key drivers and identified potential agents for DLBCL patients using gene-expression perturbations and the CREEDS database. The results indicated that the thistle1 module had high association with the clinical stage of DLBCL. In addition, MOCOS, RAB6C, ACCSL, MMP1, and RGS21 were highly linked to the occurrence and development of DLBCL.

RNAs are a carrier of genetic information and have broad roles in regulating gene expression and other biological processes. Furthermore, the majority of noncoding RNAs are highly associated with diseases including cancers and nontumorigenic diseases. Thus, RNA data analysis contributes to prioritizing previously unrecognized therapeutic targets. We anticipate that this topic can provide clues for the diagnose and prognosis of complex diseases especially cancers.

**Author contributions**

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

**Conflict of interest**

Author JY was employed by Geneis (Beijing) Co Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.
