Mining influential genes based on deep learning

Lingpeng Kong[1], Yuanyuan Chen[2], Cong Pian[2], Mingmin Xu[1], Zutan Li[1], Jingya Fang[1], Liangyun Zhang[2*]

1 College of Agriculture, Nanjing Agricultural University, Jiangsu, Nanjing 210095, China

2 Department of Mathematics, College of Science, Nanjing Agricultural University, Nanjing 210095, China

* To whom correspondence should be addressed: zlyun@njau.edu.cn

Abstract

Background: Currently, large-scale gene expression profiling has been successfully applied to the discovery of functional connections among diseases, genetic perturbation, and drug action. To address the cost of an ever-expanding gene expression profile, a new, low-cost, high-throughput reduced representation expression profiling method called L1000 was proposed, with which one million profiles were produced. Although a set of ~1,000 carefully chosen landmark genes that can capture ~80% of information from the whole genome has been identified for use in L1000, the robustness of using these landmark genes to infer target genes is not satisfactory. Therefore, more efficient computational methods are still needed to deep mine the influential genes in the genome.

Results: Here, we propose a computational framework based on deep learning to mine a subset of genes that can cover more genomic information. Specifically, an AutoEncoder framework is first constructed to learn the non-linear relationship between genes, and then DeepLIFT is applied to calculate gene importance scores. Using this data-driven approach, we have re-obtained a landmark gene set. The result shows that our landmark genes can predict target genes more accurately and robustly than that of L1000 based on two metrics (mean absolute error (MAE) and Pearson correlation coefficient (PCC)). This reveals that the landmark genes detected by our method contain more genomic information.

Conclusions: We believe that our proposed framework is very suitable for the analysis of biological big data to reveal the mysteries of life. Furthermore, the landmark genes inferred from this study can be used

for the explosive amplification of gene expression profiles to facilitate research into functional connections.

Keywords: landmark genes, deep learning, AutoEncoder, DeepLIFT

Background

One of the fundamental challenges that has emerged throughout biomedicine is the need to establish relationships between disease, physiological processes and the role of small molecule therapies. To address this problem, a genomic signature is required that should have sufficiently high complexity to provide a rich description for all biological states, including those that are physiological, related to disease, or induced with a chemical, and that should be generated in a low-cost and high-throughput way. Gene expression profiling has been widely applied in medicine and biology to elucidate the response mechanism of cells to diseases, genetic interference and drug therapy[1, 2]; using this technique, the Connectivity Map (CMap) project has been proposed a systematic approach to discover functional connections among diseases, genetic perturbation, and drug action. Meanwhile, this study also suggested the value of a large-scale community CMap project[3].

Higher requirements have been put forward for the scale of the CMap project, and a diversity of chemical perturbations, genetic perturbations, and cell types await to be characterized. Unfortunately, although the price of commercial gene expression microarrays has been decreasing steadily, the high cost of profiling thousands of samples makes this prospect difficult. Therefore, how to reduce the cost of acquiring gene expression profiles is the first problem to be solved.

Previous studies have shown that although there are a large number of genes in the genome, most of their expression patterns are highly correlated[4, 5]. Cluster analysis of single-cell RNA-Seq indicated that genes from the same cluster showed similar expression patterns under different conditions[6]. Given such high similarity, researchers from the Library of Integrated Network-Based Cellular Signatures (LINCS) program hypothesized that it is possible to capture any cellular state at a low cost by measuring a reduced

representation of the transcriptome[7]. Using Affymetrix HG-U133A microarray data from the Gene Expression Omnibus (GEO)[8], these researchers applied an iterative peel-off procedure based cluster analysis to identify the subset of universally informative transcripts termed ‘landmark genes’. According to the LINCS analysis, a set of ~1,000 genes was finally identified as landmark genes, which was sufficient to recover 82% of the information in the full transcriptome. Then, the expression profile of the target genes was inferred by a linear regression algorithm, which was subsequently improved several times to improve the reliability of prediction[9, 10]. Finally, based on the ~1,000 landmark genes, a new, low-cost, high-throughput reduced representation expression profiling method called L1000 was proposed, with which one million profiles were reported for the first time.

Cluster analysis mostly measures the similarity between variables by linear distance, such as Euclidean distance. As nonlinear regulatory relationships between genes are very common in biology[11], it is difficult for the ~1,000 landmark genes inferred by cluster analysis to fully represent genomic information. Therefore, a new computational method with the capacity to capture the non-linear relationships of genes is needed to re-mine the influential genes that cover more information about the genome.

Deep learning, a non-linear network structure using multi-layer non-linear functions, has recently emerged based on big data, and academic interest has increased rapidly since the early 2000s[12]. Furthermore, the recent success of deep learning in diverse fields such as image and speech recognition[13, 14], natural language processing[15, 16], and bioinformatics[17, 18] suggests its ability to learn hierarchical nonlinear patterns on large data sets. Deep learning can be divided into supervised learning and unsupervised learning. The former mainly includes deep neural network (DNN), convolutional neural network (CNN) and recurrent neural network (RNN) and is mainly used for classification tasks such as transcription factor binding site prediction[19], promoter prediction[20] and predicting the effects of noncoding variants[21]. The most representative of the latter is AutoEncoder, which is commonly used for dimension reduction[22] to analyse high-dimensional gene expression data[23, 24] and to integrate heterogeneous data[25, 26]. As a non-linear feature extraction method, AutoEncoder is capable of learning more useful features than linear feature extraction methods, such as principal component analysis (PCA).

Despite deep neural networks become increasingly popular, there is still a "black box" nature that hinders their application when interpretability is paramount. Understanding how an input feature affects a particular input can lead to new scientific discoveries. Therefore, multiple studies have been conducted to explain this “black box”[27-29]. Similarly, DeepLIFT is an efficient and effective method for computing importance scores in a neural network by comparing the activation of each neuron to a reference activation[30]. This method has been successfully applied to visualize splice site-related motifs from a trained CNN model[31].

Here, we present a deep learning framework to mine a gene set that can cover more genomic information. Specifically, we first constructed an AutoEncoder framework using ~130,000 gene expression profiles from the GEO Affymetrix microarray platform for training to learn the complex regulatory relationships across genes. Using this model, ~22,000 dimensional expression data were reduced to only 100. Clustering analysis of lung cancer showed that these 100 dimensional features well represent the biological information of gene expression data. Then, DeepLIFT was applied to measure the impact of each input layer neuron on the bottleneck layer neurons by providing an importance score. Using this data-driven approach, we obtained a list of genes that were sorted based on the importance score. By extracting genes from top to bottom, a new landmark gene set with the same number of genes as the original set from L1000 was finally identified. To compare the two landmark gene sets, we next used D-GEX[10] as a prediction model to infer the expression profiles of the target genes (besides the landmark genes) based on the landmark genes. The result shows that our landmark gene set can predict target genes more accurately and reliably than that of L1000 by comparing two performance metrics, MAE and PCC. Therefore, the landmark genes inferred by our method truly contain more information about the genome and are more suitable for expanding the scale of the CMap project.

Results and discussion

Performance evaluation of the AutoEncoder model

After training the AutoEncoder model with GEO-based training samples (99909), we use reserved test samples (11100) to evaluate its predictive power in both gene and sample dimensions. In terms of genes, we use MAE and PCC to measure the prediction error and similarity of each gene. As shown in Figure

1A, the average MAE and PCC of all genes are 0.2222 and 0.7627, respectively, and the permutation test shows that there is a significant high similarity between the predicted value and the real value of almost all genes (21696/22268). In terms of samples, we collect 237 lung cancer samples from the GEO database as new test samples, including 49 normal samples, 58 lung adenocarcinoma (ADC) samples and 130 lung squamous cell carcinoma (SCC) samples. Then, we take the expression profiles of these samples as the input of the trained AutoEncoder and use the output of the bottleneck layer to cluster the samples. Figure 1B shows that the low dimensional space mapped by the trained AutoEncoder well retains the biological information of the samples. All of these results show that our trained AutoEncoder can learn the non-linear relationships between genes well.

Figure 1. Performance evaluation of the AutoEncoder model in both gene (A) and sample dimensions (B). (A) The density plots of the predictive error (MAE) and the similarity (PCC) of all genes. (B) The circular diagram of clustering for three types of samples, including normal (Normal), lung adenocarcinoma (ADC) and lung squamous cell carcinoma (SCC).

Comparison of the landmark genes

First, we analyse the degree of overlap between our landmark genes (called D1000) and the landmark genes from L1000 (called L1000) and find that only 129 genes are shared. In addition, to evaluate the performance of the landmark genes inferred by our method, we use them as input to infer the expression profile of the target genes using a deep learn-based method, D-GEX. Then, we also use the MAE and PCC of each common target gene (9163) to compare D1000 with L1000. We define MAE and PCC of the target genes inferred from L1000 and D1000 as MAE_{L1000}, MAE_{D1000}, PCC_{L1000}, and PCC_{D1000}, respectively. As shown in Figure 2A and 2B, compared with MAE_{L1000} with a value of 0.1129-1.0524, the MAE_{D1000} value range is 0.0994-0.6681, and the paired t-test shows that MAE_{D1000} is significantly

lower than MAE_{L1000} ($p<0.01$). Similarly, as shown in Figure 2C and 2D, compared with PCC_{L1000} with a value of 0.0006-0.9875, the PCC_{D1000} value range is 0.4764-0.9905, and the paired t-test shows that PCC_{D1000} is significantly higher than PCC_{L1000} ($p<0.01$). Furthermore, all PCC_{D1000} pass the permutation test, but 44 target genes fail in PCC_{L1000}. These results show that the new landmark genes inferred from our method can predict target genes more accurately and robustly than the old landmark genes.

Figure 2. The density plot (A, C) and scatter plot (B, D) are used for comparison of the landmark genes inferred from our method (labelled as "D1000") and that of L1000 (labelled as "L1000") in terms of MAE (A, B) and PCC (C, D). In B and D, each dot represents a predicted target gene, and the red dot indicates that D1000 is better than L1000.

Cross-platform generalization analysis of the landmark genes

RNA-Seq is another high-throughput sequencing platform that has gradually become the standard for gene expression profiling. Next, to explore the ability to use landmark genes inferred from the microarray-based GEO dataset to infer target genes from the RNA-Seq-based expression profiling, we download a RNA-Seq-based gene expression profiling containing 2,921 samples from GTEx database, and the predicted target genes are analysed. The results indicate that the average MAE and PCC of all target genes are 0.4590 and 0.7790 (Figure 3), respectively, and that 92.51% of the target genes pass the permutation test, which shows that the landmark genes have excellent cross-platform generalization.

Figure 4. Cross-platform generalization analysis of the landmark genes inferred from our method.

Functional analysis of the landmark genes

Finally, to analyse whether the landmark genes suggested by our data-driven approach based on the analysis of 129,158 samples are enriched in particular known biological categories, we study their molecular functions from the perspective of Gene Ontology (GO). Given that the landmark genes cover most information about the genome, we infer that the landmark genes, when considered as a set, are dominated by either very few functions or many functions.

To test this inference, we use the R Bioconductor package clusterProfiler (v3.10.1) to apply hypergeometric statistics between the 943 landmark genes and a database of 1,645 gene sets that come from molecular function terms compiled in Gene Ontology. As shown in Figure 4, we observe only 34 functional categories, most of which tend to be basic and generic, such as “DNA binding transcription factor binding”, "GDP binding", "enzyme inhibitor activity" and "protease binding", and contain only a small fraction of the landmark genes (e.g., "cell adhesion molecule binding" contains 61 of 943 landmarks). The results show that no particular functional category dominates the landmark genes.

Figure 4. Enriched GO molecular functions term by using the landmark genes as a set.

Conclusion

The central dogma of molecular biology states that the flow of genetic information is "DNA to RNA to protein". Current biological studies, such as genomic studies including variable splicing and single nucleotide polymorphisms, and epigenomic studies including methylation and histone modification, are all ultimately concerned with the regulation of gene expression. Therefore, gene expression patterns can reflect almost every aspect of life activities and can be used as genomic signatures to discover the functional connections among diseases, genetic perturbation, and drug action.

In this study, we proposed a deep learning-based method to detect influential genes in the genome to obtain large-scale expression profiles at lower costs. In a nutshell, this is a question of feature selection. The computing framework we designed combines AutoEncoder and DeepLIFT to assess the impact of each gene in the genome, which was carried out using a data-driven approach in an unbiased manner rather than selecting transcripts based on prior biological knowledge. AutoEncoder is a nonlinear feature extraction method, which is transformed into a feature selection method by the application of DeepLIFT. The results show that using our landmark gene set can predict target genes more accurately and robustly than the gene set inferred from cluster analysis and reflects the advantages of deep learning in nonlinear computation.

We believe that our proposed framework is very suitable for the analysis of biological big data to reveal the mysteries of life. Furthermore, the landmark genes inferred from this study can be used for the explosive amplification of gene expression profiles to facilitate research into functional connections.

Methods

In this study, our goal is to extract ~1,000 influential genes from ~22,000 genes, which is a feature selection problem. Although many feature selection methods such as subset selection[32] and random forest[33], which are usually used in classification tasks, can effectively filter out redundant features, they cannot effectively capture the nonlinear relationship between features. In view of the above problems, we designed a computational framework as follows.

Data sources

In Table 1, three publicly available datasets are used for our analysis: the microarray-based GEO dataset, the RNA-Seq-based GTEx dataset and the lung cancer subtype dataset. The first two were downloaded from https://cbcl.ics.uci.edu/public_data/D-GEX/; the latter, from the GEO database.

First, the microarray-based GEO dataset is used to train AutoEncoder. This dataset contains 129,158 gene expression profiles, each of which contains 22,268 probes corresponding to 978 landmark genes and 21,290 target genes. The original expression data are quantile normalized to a range of values between 4 and 15 to remove technical variation[34]. Considering that a dataset containing a large number of redundant samples with high similarity corresponds to low statistical representativeness[35], the k-means

clustering program is used to remove duplicated profiles. Finally, the remaining 111,009 samples are randomly divided into ~90% (99909) for training and ~10% (11100) for testing.

Next, cross-platform performance can be evaluated based on the RNA-Seq dataset from GTEx, which contains 2,921 gene expression profiles of various tissue samples produced on RNA-Seq platform in the format of reads per kilobase per million (RPKM). We refer to the pre-processing protocol used in D-GEX for cross-platform data matching and joint quantile normalization. The 22,268 probes are finally matched to 10463 genes based on Gencode V12 annotations, including 943 landmark genes and 9520 target genes.

Finally, the lung cancer subtype dataset is used to verify whether AutoEncoder can effectively learn biological information. This dataset contains 237 gene expression profiles from the GSE4573 and GSE10072 microarray datasets, including 49 normal samples, 58 lung adenocarcinoma (ADC) samples and 130 lung squamous cell carcinoma (SCC) samples.

Table 1. Three expression datasets from the GEO and GTEx databases.

Dataset	Sample size	Platform	Database
1	129,158	Microarray	GEO
2	2,921	RNA-Seq	GTEx
3	237	Microarray	GEO

A brief summary of the computational framework

Our computational framework mainly consists of two parts, AutoEncoder-based and DeepLIFT-based (Figure 5). In the AutoEncoder-based part, we use ~130,000 gene expression profiles to train an AutoEncoder that is composed of two steps, encoder and decoder. However, AutoEncoder is a feature extraction method that transforms data from the original, high-dimensional space to a relatively low-dimensional space. In other words, new features are generally different from original features. Here, the encoder compresses the 22268 dimensional samples to 100 dimensions. In the DeepLIFT-based part, we use DeepLIFT to compute the importance scores of each input layer neuron on the bottleneck layer

neurons. Then, we rank the genes based on the average importance scores, and the new landmark genes (see Additional file 1) can be identified by selecting the top 943 genes (the same number as the L1000).

Figure 5. The workflow for mining influential genes based deep learning. (A) The architecture and parameter settings of AutoEncoder. (B) Application of DeepLIFT to compute the importance scores in the Encoder network and use of D-GEX as a baseline method to predict target genes for performance evaluation.

AutoEncoder framework

AutoEncoder is a multi-task unsupervised feed-forward neural network with multiple stacked hidden layers, which is composed of two parts, an encoder and a decoder (Figure 1A). Considering a dataset X with m samples and n features, the encoder $E|_{X\to Y}$ aims to map the original data X to the reduced representation Y through the bottleneck layer, and the purpose of the decoder $D|_{Y\to X}$ is tuned to reconstruct the original data X from the low-dimensional representation Y by minimizing the difference between X and $\hat{X}$.

Specifically, we use the Python Keras library to implement an AutoEncoder with three hidden layers of 500, 100, and 500 nodes. For a given layer l, we use sigmoid as the activation function.

$$o = f_l(x) = sigmoid(W_l x + b_l)$$

Where x is an input vector of size d, W_l is the weight matrix of size $p \times d$, and b_l is an intercept vector of size p. Given a set of gene expression profiles $S = (s_1, ..., s_m)$ containing m samples, where $s_m = (g_{m1}, ..., g_{mn})$ denotes each gene expression profile containing n genes, the input vector s_m is reconstructed to $\hat{s}_m$ through a series of matrix transformations of multiple network layers. Training an AutoEncoder involves finding parameters $\theta = (W, b)$ minimizing a specific loss function. Here, we use Mean Absolute Error (MAE) as the loss function.

$$MAE(g, \hat{g}) = \frac{1}{n}\sum_{i=1}^{n}|g_i - \hat{g}_i|$$

To control overfitting, we add an L2 regularization penalty α=1e-6 on the weight vector. Thus, the loss function above becomes:

$$loss(g,\hat{g}) = \frac{1}{n}\sum_{i=1}^{n}\left|g_i - \hat{g}_i\right| + \sum_{j=1}^{k}\alpha\left\|W_j\right\|_2^2$$

Finally, AutoEncoder is trained using the Adam[36] optimization algorithm with 100 epochs and 10% dropout.

Evaluation metrics

Given a test set $T = (s_1,...,s_m)$ containing m samples, we use two different metrics for the evaluation of predicted expression. For each gene g_j, the definition of MAE is:

$$MAE_j = \frac{1}{m}\sum_{i=1}^{m}\left|g_{ij} - \hat{g}_{ij}\right|$$

The following equation shows the definition of PCC:

$$PCC_j = \rho(g_j,\hat{g}_j) = \frac{\sum_{i=1}^{m}(g_{ij}-\mu_j)(\hat{g}_{ij}-\hat{\mu}_j)}{\sqrt{\sum_{i=1}^{m}(g_{ij}-\mu_j)^2}\sqrt{\sum_{i=1}^{m}(\hat{g}_{ij}-\hat{\mu}_j)^2}}$$

where PCC_j indicates the Pearson correlation coefficient for the *j*-th predicted gene and μ_j, $\hat{\mu}_j$ are the mean of g_j, $\hat{g}_j$ respectively.

The Pearson correlation coefficient, an absolute measure of similarity between genes, does not in itself reflect how uncommon that similarity is. Hence, we apply a permutation test to aid in the interpretation of similarity. Briefly, in addition to computing the PCC_j between g_j and $\hat{g}_j$, we also compute the PCC_{null} between the $\hat{g}_j$ and any gene other than g_j as a reference distribution of similarity values. After

that, we compare PCC_j to PCC_{null}, and if the fraction of PCC_{null} that is higher than PCC_j is lower than 0.01, g_j and $\hat{g}_j$ are considered to be significantly correlated.

Additional files

Additional file 1: A file in TXT format including a list of genes sorted based on contribution scores.

Abbreviations

CMap: Connectivity Map; MAE: mean absolute error; PCC: Pearson correlation coefficient; GEO: Gene Expression Omnibus; ADC: lung adenocarcinoma; SCC: lung squamous cell carcinoma.

Declaration

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Three publicly available datasets are used for our analysis: the microarray-based GEO dataset, the RNA-Seq-based GTEx dataset and the lung cancer subtype dataset. The first two were downloaded from https://cbcl.ics.uci.edu/public_data/D-GEX/; the latter, from the GEO database with GSE4573 and GSE10072.

Competing interests

The authors declare that they have no competing interests.

Funding

This work is supported by the Natural Science Foundation of China (11571173).

Authors' contributions

LPK: Conceived and designed the experiments; LPK and ZTL: Performed the experiments; YYC and JYF: Analyzed the data; PC and MMX: Contributed reagents/materials/analysis tools; LPK and LYZ: Wrote the paper. All authors read and approved the final manuscript.

Acknowledgments

This work was supported by the high-performance computing platform of Bioinformatics Center, Nanjing Agricultural University.

References

1. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, Hayden Gephart MG, Barres BA, Quake SR: **A survey of human brain transcriptome diversity at the single cell level**. *Proc Natl Acad Sci U S A* 2015, **112**(23):7285-7290.

2. Calon A, Lonardo E, Berenguer-Llergo A, Espinet E, Hernando-Momblona X, Iglesias M, Sevillano M, Palomo-Ponce S, Tauriello DV, Byrom D *et al*: **Stromal gene expression defines poor-prognosis subtypes in colorectal cancer**. *Nat Genet* 2015, **47**(4):320-329.

3. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN *et al*: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease**. *Science (New York, NY)* 2006, **313**(5795):1929-1935.

4. Ntranos V, Kamath GM, Zhang JM, Pachter L, Tse DN: **Fast and accurate single-cell RNA-seq analysis by clustering of transcript-compatibility counts**. *Genome Biol* 2016, **17**(1):112.

5. Heimberg G, Bhatnagar R, El-Samad H, Thomson M: **Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing**. *Cell Syst* 2016, **2**(4):239-250.

6. Shah S, Lubeck E, Zhou W, Cai L: **In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus**. *Neuron* 2016, **92**(2):342-357.

7. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, Gould J, Davis JF, Tubelli AA, Asiedu JK *et al*: **A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles**. *Cell* 2017, **171**(6):1437-1452 e1417.

8. Edgar R, Domrachev M, Lash AE: **Gene Expression Omnibus: NCBI gene expression and hybridization array data repository**. *Nucleic Acids Res* 2002, **30**(1):207-210.

9. Wang X, Ghasedi Dizaji K, Huang H: **Conditional generative adversarial network for gene expression inference**. *Bioinformatics* 2018, **34**(17):i603-i611.

10. **Gene expression inference with deep learning**. *Bioinformatics* 2016.

11. Brunel H, Gallardo-Chacon JJ, Buil A, Vallverdu M, Soria JM, Caminal P, Perera A: **MISS: a non-linear methodology based on mutual information for genetic association studies in both population and sib-pairs analysis**. *Bioinformatics* 2010, **26**(15):1811-1818.

12. Min S, Lee B, Yoon S: **Deep learning in bioinformatics**. *Brief Bioinform* 2017, **18**(5):851-869.

13. Krizhevsky A, Sutskever I, Hinton GE: **ImageNet classification with deep convolutional**

neural networks. *Communications of the ACM* 2017, **60**(6):84-90.

14. Chorowski J, Bahdanau D, Serdyuk D, Cho K, Bengio Y: **Attention-Based Models for Speech Recognition**. *Adv Neur In* 2015, **28**.

15. Li JW, Luong MT, Jurafsky D: **A Hierarchical Neural Autoencoder for Paragraphs and Documents**. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, Vol 1* 2015, **1**:1106-1115.

16. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P: **Natural Language Processing (Almost) from Scratch**. *J Mach Learn Res* 2011, **12**:2493-2537.

17. Kelley DR, Snoek J, Rinn JL: **Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks**. *Genome Res* 2016, **26**(7):990-999.

18. Kalkatawi M, Magana-Mora A, Jankovic B, Bajic VB: **DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions**. *Bioinformatics* 2019, **35**(7):1125-1132.

19. Zhou J, Lu Q, Gui L, Xu R, Long Y, Wang H: **MTTFsite: cross-cell type TF binding site prediction by using multi-task learning**. *Bioinformatics* 2019, **35**(24):5067-5077.

20. Umarov R, Kuwahara H, Li Y, Gao X, Solovyev V: **Promoter analysis and prediction in the human genome using sequence-based deep learning models**. *Bioinformatics* 2019, **35**(16):2730-2737.

21. Zhou J, Troyanskaya OG: **Predicting effects of noncoding variants with deep learning-based sequence model**. *Nat Methods* 2015, **12**(10):931-934.

22. Gligorijevic V, Barot M, Bonneau R: **deepNF: deep network fusion for protein function prediction**. *Bioinformatics* 2018, **34**(22):3873-3881.

23. Chen L, Cai C, Chen V, Lu X: **Learning a hierarchical representation of the yeast transcriptomic machinery using an autoencoder model**. *BMC Bioinformatics* 2016, **17 Suppl 1**:9.

24. Khalili M, Alavi Majd H, Khodakarim S, Ahadi B, Hamidpour M: **PREDICTION OF THE THROMBOEMBOLIC SYNDROME: AN APPLICATION OF ARTIFICIAL NEURAL NETWORKS IN GENE EXPRESSION DATA ANALYSIS**. *ARCHIVES OF ADVANCES IN BIOSCIENCES (JOURNAL OF PARAMEDICAL SCIENCES)* 2016, **7**(2):15-22.

25. Miotto R, Li L, Kidd BA, Dudley JT: **Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records**. *Sci Rep* 2016, **6**:26094.

26. Chen Q, Song X, Yamada H, Shibasaki R: **Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference**; 2016.

27. Zeiler M, Fergus R: **Visualizing and Understanding Convolutional Neural Networks**, vol. 8689; 2013.

28. Springenberg J, Dosovitskiy A, Brox T, Riedmiller M: **Striving for Simplicity: The All Convolutional Net**. 2014.

29. Simonyan K, Vedaldi A, Zisserman A: **Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps**. *preprint* 2013.

30. Shrikumar A, Greenside P, Kundaje A: **Learning Important Features Through Propagating Activation Differences**. 2017.

31. Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W: **SpliceRover: interpretable convolutional neural networks for improved splice site prediction**. *Bioinformatics* 2018,

34(24):4180-4188.

32. John GH, Kohavi R, Pfleger K: **Irrelevant Features and the Subset Selection Problem**. In: *Machine Learning Proceedings 1994*. Edited by Cohen WW, Hirsh H. San Francisco (CA): Morgan Kaufmann; 1994: 121-129.

33. Liaw A, Wiener M: **Classification and Regression by RandomForest**. *Forest* 2001, **23**.

34. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias**. *Bioinformatics* 2003, **19**(2):185-193.

35. Chen W, Lv H, Nie F, Lin H: **i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome**. *Bioinformatics* 2019, **35**(16):2796-2800.

36. Kingma D, Ba J: **Adam: A Method for Stochastic Optimization**. *International Conference on Learning Representations* 2014.