A Deep Learning Framework for Predicting Human Essential Genes by Integrating Sequence and Functional data

Wangxin Xiao[1#], Xue Zhang[2,3#*], Weijia Xiao[4]

[1]Faculty of Transportation Engineering, Huaiyin Institute of Technology, Huai'an, Jiangsu 223003, China
[2]Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an, Jiangsu 223003, China
[3]School of Medicine, Tufts University, Boston, MA 02111, USA
[4]Boston Latin School, Boston, MA 02115, USA

Abstract

Motivation: Essential genes are necessary to the survival or reproduction of a living organism. The prediction and analysis of gene essentiality can advance our understanding to basic life and human diseases, and further boost the development of new drugs. Wet lab methods for identifying essential genes are often costly, time consuming, and laborious. As a complement, computational methods have been proposed to predict essential genes by integrating multiple biological data sources. Most of these methods are evaluated on model organisms. However, prediction methods for human essential genes are still limited and the relationship between human gene essentiality and different biological information still needs to be explored. In addition, exploring suitable deep learning techniques to overcome the limitations of traditional machine learning methods and improve the prediction accuracy is also important and interesting.

Results: We propose a deep learning based method, DeepSF, to predict human essential genes. DeepSF integrates four types of features, that is, sequence features, features from gene ontology, features from protein complex, and network features. Sequence features are derived from DNA and protein sequence for each gene. 100 GO terms from cellular component are used to form a feature vector for each gene, in which each component captures the relationship between a gene and a GO term. Network features are learned from protein-protein interaction (PPI) network using a deep learning based network embedding method. The features derived from protein complexes capture the relationships between a gene or a gene's direct neighbors from PPI network and protein complexes. The four types of features are integrated together to train a multilayer neural network. The experimental results of 10-fold cross validation show that DeepSF can accurately predict human gene essentiality with an average performance of AUC about 94.35%, the area under precision-recall curve (auPRC) about 91.28%, the accuracy about 91.35%, and the F1 measure about 77.79%. In addition, the comparison results show that DeepSF significantly outperforms several widely used traditional machine learning models (SVM, Random Forest, and Adaboost), and performs slightly better than a recent deep learning model (DeepHE).

Conclusions: We have demonstrated that the proposed method, DeepSF, is effective for predicting human essential genes. Deep learning techniques are promising at both feature learning and classification levels for the task of essential gene prediction.

Keywords: Essential gene prediction, deep learning, gene ontology, network embedding, sequence feature, protein complex

Availability and Implementation: The python code will be freely available at https://github.com/xzhang2016/DeepSF.

#These authors have contributed equally to this work

*Corresponding author: Xue Zhang (lindajia03@gmail.com)

Introduction

There are more than 20,000 genes of human being, which forms a redundant and high fault-tolerant system. Among these genes, some are vital for the survival and reproduction of us, but others are not. These two groups of genes are named as essential genes and nonessential genes. Essential genes are a group of fundamental genes necessary for a specific organism to survive in a specific environment. For human, essential genes refer to a subset of genes that are indispensable for the viability of individual human cell types [1, 2]. These genes encode conservative functional elements which mainly contribute to DNA replication, gene translation, gene transcription, and substance transportation. The identification and analysis of essential genes is very important for understanding the minimal requirements of basic life, finding disease genes, and developing new drugs.

There are two ways to identify essential genes, wet lab experimental methods and computational methods. For example, gene direct deletion and transposon-based randomized mutagenesis have been used to identify essential genes for bacteria and yeast in the genome scale [3]; microinject KO and nuclear transfer techniques have been used to identify essential genes in mice [4]; the CRISPR/Cas9 genome editing system has been used to identify essential genes from human cell lines [1,2,5]. Experimental methods are often costly, time-consuming, and laborious. The accumulation of essential gene datasets and the sequence data as well as multiple functional data enables researchers to explore the relationships between gene essentiality and different genomics data, and to develop effective models to predict gene essentiality. These computational methods can greatly reduce the cost and time involved in for finding essential genes which further boosts our understanding to basic life and to human diseases, and helps quickly finding new drug targets and developing drugs.

Computational methods can be classified into two groups. One focus is to design a centrality measure to rank proteins/genes, while the other focus is to integrate multiple features using machine learning to predict gene essentiality. The most widely used centrality measures include degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, to name a few. These centrality measures have been found having relationship with gene essentiality in multiple model organisms and human [6, 7], however they can only differentiate a subset of essential genes from nonessential genes. One reason might be the incompleteness and false positive/false negative interactions in the protein-protein interaction (PPI) networks, the other reason might be the fact that gene essentiality relates to multiple biological factors rather than only the topological characteristics. Due to these reasons, researchers have proposed some more complicated centrality measures by integrating network topological properties with other biological information. For example, Zhang et al. proposed a method, CoEWC, to capture the common properties of essential genes in both date hubs and party hubs by integrating network topological property with gene expression profiles, which showed significant improvement of prediction ability compared to those only based on PPI networks [6]. An ensemble framework was also proposed based on gene expression data and PPI network, which can greatly enhance the

prediction power of common used centrality measures [8]. Luo et al. proposed a method LIDC to predict essential proteins by combining local interaction density with protein complexes [9]. Zhang et al. proposed OGN by integrating network topological properties, gene expression profiles, and orthologous information [10]. GOS was proposed by Li et al. by integrating gene expression, orthologous information, subcellular localization and PPI networks [11]. These proposed integrated centrality measures have improved prediction power over the ones only based on PPI network topological properties. However, they still have limited prediction accuracy since gene essentiality relates to many complicated factors which make it impossible to represent them by a scalar score. At this point, machine learning is a good choice to fully utilize multiple features for predicting essential genes.

Many machine learning models and deep learning frameworks have been proposed and successfully applied in different fields. For example, convolutional neural network (CNN) was used for biomedical image segmentation [12]; density-based neural network was used for pavement distress image classification [13]; CNN was also used for drug-target prediction [14]; active learning and transductive k-nearest neighbors were used for text categorization [15, 16]; Support Vector Machines (SVM) was used for essential gene prediction [17, 18]. In the research field of essential gene prediction, many machine learning based prediction methods have been proposed to integrate features from multiple omics data [19]. As shown in the review article [19], traditional machine learning methods were used to predict gene essentiality and most of them were evaluated on data from model organisms. In these methods, topological features together with features from sequence and other functional genomics data were extracted manually and then used to train the models. How to extract informative features is very important and challenging, which requires ample domain knowledge as well as the understanding of what relationship exist between gene essentiality and each omics data. Usually we only know that an omics data would contribute to gene essentiality, but we don't know what attributes of it have such effect and how to represent it. This puts a limitation on traditional machine learning methods to obtain good prediction accuracy.

In recent years, deep learning techniques have been used to automatically extract features and to train a more powerful classification model for predicting essential genes. Grover et al. proposed a network embedding method based on deep learning, node2vec, to learn a low-dimension representation for each node [20]. This method has been used to extract topological features from PPI networks for predicting essential genes, and these features are more informative than those obtained via some popular centrality measures [7, 21, 22]. CNN was used to extract local patterns from time-serials gene expression profiles from *S. cerevisiae* [21] and Zeng et al. also used bidirectional long short term memory (LSTM) cells to extract features from the same gene expression profiles [22]. The automatically learned features are combined with other manually extracted ones to train a deep learning model for human essential gene prediction [7]. A six hidden-layers neural network was designed to predict essential genes in microbes by only using manually extracted features from sequence data [23].

Recently, human essential genes have been identified in several human cancer cell lines by

utilizing CRISPR-Cas9 and gene-trap technology [1,2,24]. These identified essential genes provide a clear definition of the requirements for sustaining the basic cell activities of individual human tumor cell types, and can be regarded as targets for cancer treatment. These essential gene datasets together with other available biological data sources enable us to test one important and interesting assumption that human gene essentiality might be accurately predicted using computational methods. In this paper, we propose a deep learning framework, DeepSF, to predict human essential genes. DeepSF integrates features derived from nucleotide sequence and protein sequence data, features learned from PPI network, features encoded using gene ontology (GO) enrichment scores, and features from protein complexes to train a multi-layers deep learning model. We show that DeepSF can accurately predict human essential genes.

The Proposed Deep Learning Method

DeepSF consists of two modules, feature extraction module and classification module. In its feature extraction module, it extracts/learns features from multiple omics data and integrate these features together as input to the classification module. In this paper, we mainly consider four types of features, although other types of features can be easily integrated into our model.

Features encoded with gene ontology (GO) enrichment scores are calculated as follows. We choose 100 GO terms from cellular component part to encode the genes. For each gene, we first obtain its direct neighbors from PPI network to form a gene set consisting the gene and its neighbors, then do gene ontology enrichment analysis for this gene set against the 100 GO terms using hypergeometric test. The enrichment score is calculated as –log10(p-value) for each GO term. In this way, we get a 100-dimension feature vector for each gene.

From protein complexes, we extract two features. The first feature is the number of protein complexes a gene involved in. The second feature is calculated as follows. For a gene, we first get the gene set N consisting of its direct neighbors in a PPI network. Suppose there are M neighbors of this gene involved in a protein complex. We calculate a score $s = |M|/|N|$ as the ratio of its neighbors involved in a protein complex. The second feature is the sum of the ratios across all the protein complexes.

Network features are learned based on a network embedding method, node2vec. Each gene is represented by a 64-dimension feature vector learned from the PPI network.

Sequence features consist of codon frequency, maximum relative synonymous codon usage (RSCU*max*), codon adaptation index (CAI), gene length, GC content, amino acid frequency, and protein sequence length. There are 89 sequence features in total. For more details about how these features are calculated, please refer to [7].

The classification module of DeepSF is based on the multilayer perceptron structure. It includes one input layer, three hidden layers, and one output layer. We use rectified linear unit (ReLU) as the activation function for all the hidden layers, while the output layer uses sigmoid activation function to perform discrete classification. The loss function in DeepSF is binary cross-entropy. A dropout layer is followed each hidden layer to make the network less sensitive to noise in the training data and increase its generalization ability. In order to address the imbalanced learning issue inherent in essential gene prediction problem, we utilize class weight to train a weighted neural network, which will allow the model to pay more attention to instances from minority class than those from majority class.

Results and discussion

Datasets

Human essential genes are downloaded from DEG database [25]. There are 16 human essential gene datasets among which 13 datasets are from [1,2,24]. We choose the genes contained at least in 5 datasets as our essential gene dataset. Excluding the genes annotated to essential genes in DEG, the other genes are considered as nonessential genes. The DNA sequence data and protein sequence data are downloaded from Ensembl [26] (release 97, July 2019). We download the PPI data from BioGRID [27] (release 3.5.181, February 2020). Only physical interactions between human genes are used. After filtering out self-interactions and several small separated subgraphs, we obtain a protein-protein interaction network with 17,762 nodes and 355,647 edges. This interaction network is used to learn embedding features for each gene. Gene Ontology data are downloaded from Gene Ontology website [28,29] and protein complex data are downloaded from CORUM [30]. The genes having sequence features, network embedding features as well as GO enrichment scores are used for the following classification performance evaluation. In total 2009 essential genes and 8416 nonessential genes are used for the analysis.

Evaluation metrics

We use multiple metrics to evaluate the performance of DeepSF. The first metric is the area under the receiver operating characteristic (ROC) curve (AUC). ROC plot represents the trade-off between sensitivity and specificity for all possible thresholds. The second metric is the area under the precision-recall curve (auPRC). Precision-Recall (PR) curves summarize the trade-off between the true positive rate and the positive predictive value using different probability thresholds. ROC curves are appropriate for balanced classification problems while PR curves are more appropriate for imbalanced datasets. The third metric is Matthews correlation coefficient (MCC) [31], and the fourth metric is F1 measure. We also give the following performance measures: sensitivity (Sn), specificity (Sp), positive predictive value (PPV), and accuracy (Ac), which are defined in (1) - (4), MCC and F1 in (5) and (6), where *TP*, *TN*, *FP*, and *FN* are the number of true positives, true negatives, false positives, and false negatives, respectively.

$$S_n = \frac{TP}{TP+FN} \tag{1}$$

$$S_p = \frac{TN}{FP+TN} \tag{2}$$

$$PPV = \frac{TP}{TP+FP} \tag{3}$$

$$Ac = \frac{TP+TN}{TP+FN+TN+FP} \tag{4}$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}} \tag{5}$$

$$F1 = 2 * PPV * Sn/(PPV + Sn) \tag{6}$$

Performance evaluation

In the following experiments, we use the parameters for DeepSF as follows. Adam with initial learning rate 0.001 is used as the optimizer. The sizes for hidden layers are 1024, 512, 256. The dropout rate is 0.2. The training runs for 200 epochs using early stopping criteria with patience =

15. In order to cope with the imbalanced learning issue, we set class weight to 4.5 for essential genes, and 1 for nonessential genes. The 10-fold cross validation is used to evaluate the performance of DeepSF, and 10% of the training data is used as the validation data in the training process.

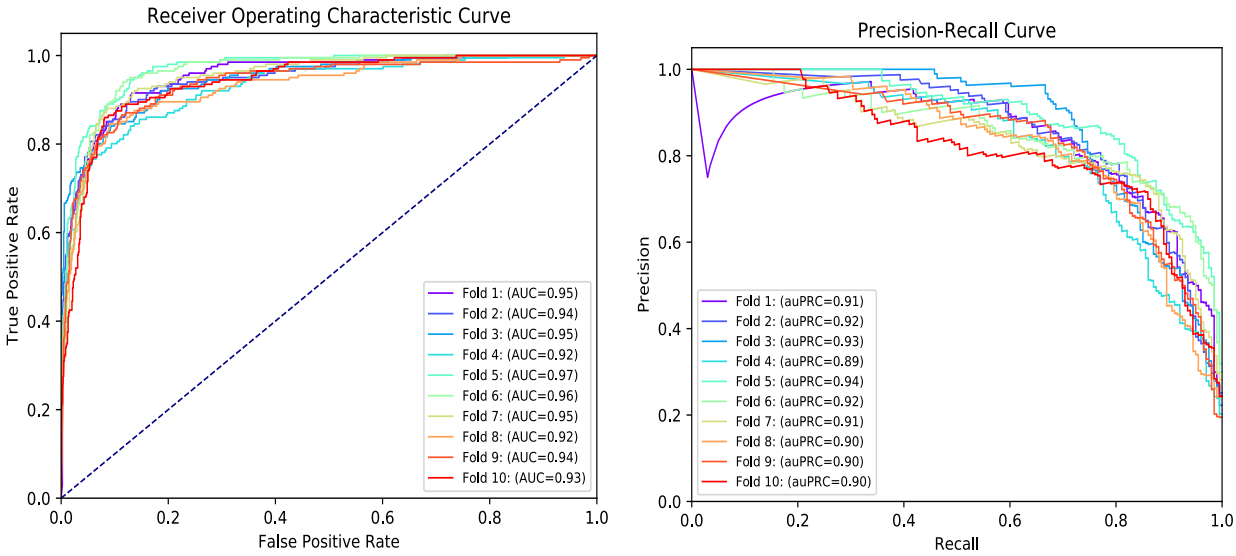

Figure 1 The ROC and PR curves of DeepSF.

Figure 1 presents the ROC and auPRC curves of DeepSF across the 10-fold cross validation. From figure 1 we can see that DeepSF reaches its best performance at fold 5 with AUC = 0.97 and auPRC = 0.94. The average AUC of DeepSF across the 10-fold cross validation is 94.35% with standard deviation STD = 0.014, and the average auPRC is 91.28% with STD = 0.015. In addition, the performance of DeepSF is quite stable since the difference is less than 5% between its best and worst AUC (auPRC) scores. The worst auPRC score is still above 89% which indicates that DeepSF is very effective for predicting human essential genes.

In order to demonstrate the superior of DeepSF, we also compare it with three popular traditional machine learning models (SVM, RF, and Adaboost) as well as a recent deep learning based model DeepHE. For fair comparison, DeepSF and the three traditional machine learning models use the same input features so that the only difference here is the classification module. DeepSF uses deep learning techniques while the three methods are all traditional machine learning models. For DeepHE, we use the features presented in [7] and set the hidden layers to 128, 256, 512 as used in [7] while keeping the training data and other settings same with DeepSF.

Figure 2 shows the box plot for the performance comparison between DeepSF and the four compared models across 10-fold cross validation. From figure 2 we can see that the AUC and auPRC scores of DeepSF are significantly higher than those of the three traditional machine learning models, and are slightly better than those of DeepHE. The average AUC of DeepSF increases 0.85%, 7.52%, 22.89%, and 13.86% compared with that of DeepHE, SVM, RF, and Adaboost respectively. The average auPRC of DeepSF increases 1.95%, 43.26%, 63.15%, and 51.32% compared with that of DeepHE, SVM, RF, and Adaboost respectively. In terms of the

other two comprehensive metrics MCC and F1, DeepSF achieves similar performance with SVM, but obvious better than DeepHE, RF and Adaboost. For example, the average MCC of DeepSF increases 5.67%, 0.68%, 14.34%, and 6.57% compared with that of DeepHE, SVM, RF, and Adaboost respectively. The average F1 of DeepSF increases 4.18%, 0.35%, 15.49%, and 5.41% compared with that of DeepHE, SVM, RF, and Adaboost respectively. Overall, our deep learning based model is superior than the traditional machine learning models.

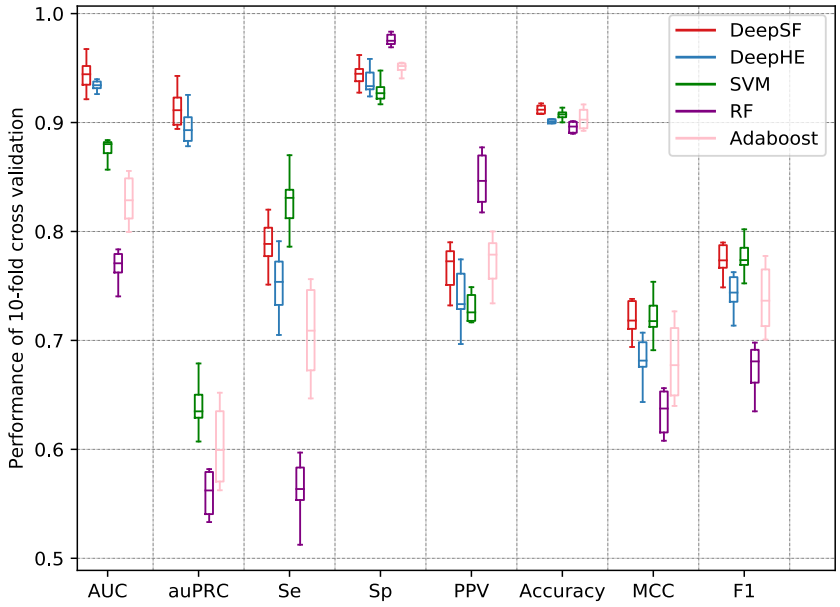

Figure 2 Performance comparison of DeepSF, DeepHE, SVM, RF, and Adaboost

In order to see the contribution of each type of features used in DeepSF, we also evaluate it by removing one type of features each time. The experiments use same settings except the input features. Figure 3 gives the box plot for DeepSF with different input features, which tells us that DeepSF with the integration of all four types of features performs best which further confirms the contribution and complement of these four types of features. By removing one type of features, we can evaluate the contribution of each type of features to the overall performance. From figure 3 we can see that the combination of N + G + C performs worst which might be due to the fact that these three types of features all utilize the PPI network information so that they have less complement effect. The other three combinations with sequence features perform only slightly worse than that of using all the four types of features, which tells us that sequence features have big complement with the other types of features.

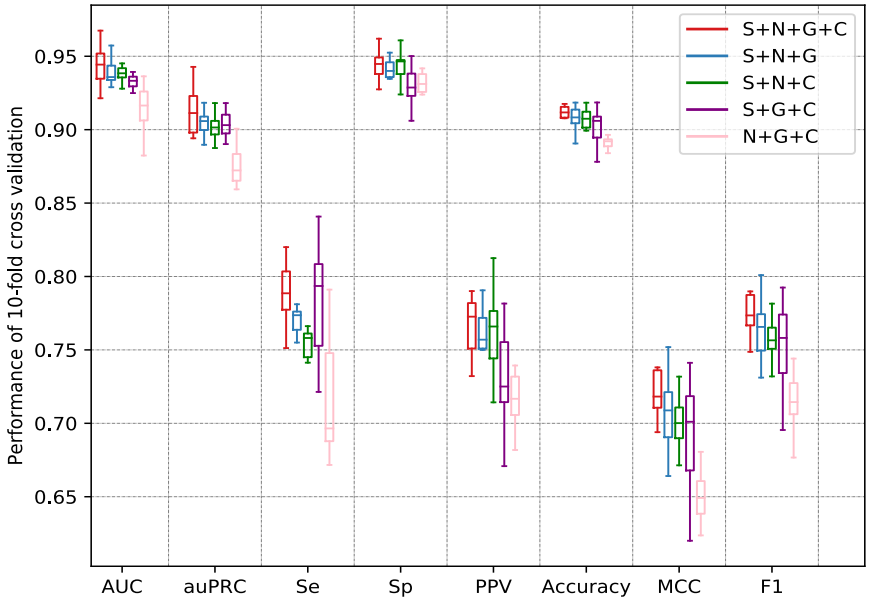

Figure 3 Performance comparison of DeepSF with different features. S: sequence features; N: network embedding features; G: gene ontology features; C: protein complex features.

Conclusions

In this paper, we propose a deep learning based method, DeepSF, to predict human essential genes. DeepSF integrates four types of features extracted/learned from sequence and functional genomics data, and utilizes deep learning techniques to train a cost-sensitive classifier to enhance prediction accuracy of gene essentiality. Our 10-fold cross validation experiments demonstrate that deep learning is more superior than the traditional machine learning models, and that the integration of features from different omics data can enhance the prediction accuracy. Our experiments demonstrate that DeepSF is effective for predicting human essential genes.

In the future, we are interested in how to use deep learning to automatically learn features from different types of biological data. For example, how to learn a low-dimension representation if we use all GO terms to encode genes instead of the subset of selected GO terms. Exploring and integrating more biological data into the learning and classification model is another interesting direction, such as protein domain and epigenomics data. Predicting cancer cell line specific essential genes by designing effective deep learning models and integrate cell line specific information would be very interesting. In addition, we are also interested in testing whether data editing [32] and clustering aided techniques [33] are useful for imbalanced learning problem.

Authors' contributions

XZ and WXX conceived and designed the experiments. XZ and WJX performed the experiments and analysis. XZ and WXX drafted and revised the manuscript. All authors approved the final manuscript.

Funding Statement

This work was supported by the National Natural Science Foundation of China, No. 61402423, XZ; National Natural Science Foundation of China, No.51678282, WXX; National Natural Science Foundation of China, No.51378243, WXX; Guizhou Provincial Science and Technology Fund with grant No. [2015]2135, XZ. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability

All data used in this study are third party and freely accessible from public databases. Protein-protein interaction data are available from BioGRID database at http://thebiogrid.org/download.php. Essential genes data and the corresponding sequence data from DEG database are available at http://tubic.tju.edu.cn/deg/. DNA sequence and protein sequence data are available at https://useast.ensembl.org/Homo_sapiens/Info/Annotation. Gene Ontology data are available at http://current.geneontology.org/products/pages/downloads.html. Protein complex data are available at http://mips.helmholtz-muenchen.de/corum/.

References

1. Wang T, Birsoy K, Hughes NW, Krupczak KM, Post Y, et al. Identification and characterization of essential genes in the human genome. Science, 2015, 350(6264): 1096-101. doi: 10.1126/science.aac7041.
2. Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. Cell, 2015, 163(6): 1515-26. doi:10.1016 /j.cell.2015.11.015.
3. Grazziotin AL, Vidal NM, Venancio TM. Uncovering major genomic features of essential genes in Bac- teria and a methanogenic Archaea. FEBS J. 2015; 282(17):3395–411. https://doi.org/10.1111/febs. 13350
4. Liao BY, Zhang JZ. Mouse duplicate genes are as essential as singletons. Trends in Genetics. 2007; 23(8):378–81. https://doi.org/10.1016/j.tig.2007.05.006
5. Morgens DW, Deans RM, Li A, Bassik MC. Systematic comparison of CRISPR/Cas9 and RNAi screens for essential genes. Nature Biotechnology. 2016; 34(6):634–6. https://doi.org/10.1038/nbt.3567
6. Zhang X, Xu J, Xiao W-x. A New Method for the Discovery of Essential Proteins. PLoS ONE. 2013; 8 (3): e58763. https://doi.org/10.1371/journal.pone.0058763.
7. Zhang X, Xiao WX, Xiao WJ. DeepHE: Accurately predicting human essential genes based on deep learning. bioRxiv, Feb 2020. doi: https://doi.org/10.1101/2020.02.14.950048
8. Zhang X, Xiao W, Acencio ML, Lemke N, Wang X. An ensemble framework for identifying essential proteins. BMC Bioinformatics. 2016; 17:322. https://doi.org/10.1186/s12859-016-1166-7.
9. Luo J, Qi Y (2015) Identification of Essential Proteins Based on a New Combination of Local Interaction Density and Protein Complexes. PLoS ONE 10(6): e0131418. doi:10.1371/journal. pone.0131418
10. Zhang X, Xiao W, Hu X. Predicting essential proteins by integrating orthology, gene expressions, and PPI networks. PLoS ONE, 2018, 13(4): e0195410. https://doi.org/10.1371/journal. pone.0195410.

11. Li G, Li M, Wang J, Wu J, Wu F, Pan Y. Predicting essential proteins based on subcellular localization, orthology and PPI networks. BMC Bioinformatics. 2016; 17(Suppl 8):279. https://doi.org/10.1186/ s12859-016-1115-5.

12. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, LNCS, Vol.9351: 234--241, 2015

13. Xiao W, Yan X, Zhang X. Pavement distress image automatic classification based on density-based neural network. International Conference on Rough Sets and Knowledge Technology (RSKT), July 2006: 686-692. Lecture Notes in Computer Science (LNCS volume 4062).

14. Ozturk H, Ozgur A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatcs, 34, 2018, i821-i829. Doi: 10.1093/bioinformatics/bty593.

15. Zhang X, Zhao D, Chen L, Min W. Batch mode active learning based multi-view text classification. The sixth international conference on fuzzy systems and knowledge discovery, August 2009, 7: 472-476.

16. Xiao W, Zhang X. Active transductive KNN for sparsely labeled text classification. The 6th international conference on soft computing and intelligent systems, jointly the 13th international symposium on advanced intelligence systems. November 2012, 2178-2182.

17. Guo F, Dong C, Hua H, Liu S, Luo H, Zhang H, Jin Y, Zhang K. Accurate prediction of human essential genes using only nucleotide composition and association information. Bioinformatics, 33(12), 2017, 1758 - 1764. Doi: 10.1093/bioinformatics/btx055.

18. Chen L, Zhang YH, Wang SP, Zhang YH, Huang T, Cai YD. (2017) Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways. PLoS ONE 12(9): e0184129. https://doi.org/10.1371/journal. pone.0184129

19. Zhang X, Acencio ML and Lemke N. Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review. Front. Physiol. 2016; 7:75. https:// doi.org/10.3389/fphys.2016.00075 PMID: 27014079.

20. Grover A, Leskovec J (2016). node2vec: Scalable Feature learning from networks. KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. August 2016, pp 855-864. https://doi.org/10.1145/2939672.2939754

21. Zeng M, Li M, Wu FX, Li YH, Pan Y. DeepEP: a deep learning framework for identifying essential proteins. BMC Bioinformatics 2019, 20(Suppl 16):506. https://doi.org/10.1186/s12859-019-3076-y.

22. Zeng M, Li M, Fei Z, Wu F, Li Y, Pan Y, Wang J. A deep learning framework for identifying essential proteins by integrating multiple types of biological information. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2019 Feb 5. doi: 10.1109/TCBB.2019.2897679.

23. Hasan MA, Lonardi S. DEEPLYESSENTIAL: A deep neural network for predicting essential genes in microbes. BioRxiv, 2019. http://dx.doi.org/10.1101/607085.

24. Blomen VA, Májek P, Jae LT, Bigenzahn JW, Nieuwenhuis J, et al. Gene essentiality and synthetic lethality in haploid human cells. Science, 2015, 350(6264):1092-6. doi: 10.1126/science.aac7557.

25. Luo H, Lin Y, Gao F, Zhang CT, and Zhang R. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res., 42 (Database issue): D574–80, Jan. 2014.

26. Ruffier M, Kähäri A, Komorowska M, Keenan S, Laird M, et al. Ensembl core software resources: storage and programmatic access for DNA sequence and genome annotation. Database, 2017, doi: 10.1093/

database/bax020.

27. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. Biogrid: A General Repository for Interaction Datasets. Nucleic Acids Res. Jan 1, 2006; 34: D535-9.

28. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.

29. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. Jan 2019;47(D1):D330-D338.

30. Giurgiu M, Reinhard J, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Ruepp A. CORUM: the comprehensive resource of mammalian protein complexes-2019. Nucleic Acids Res. 2018 Oct 24. doi: 10.1093/nar/gky973.

31. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975; 405(2):442–51. Epub 1975/10/20.

32. Zhang X, Xiao W. Active semi-supervised framework with data editing. Computer Science and Information Systems. 2012; 9(4): 1513–1532. doi: 10.2298/CSIS120202045Z

33. Zhang X, Xiao W. Clustering based two-stage text classification requiring minimal training data. Computer Science and Information Systems. 2012; 9(4):1627–1643. doi: 10.2298/CSIS120130044Z