DÜZCE ÜNİVERSİTESİ
BİLİM ve TEKNOLOJİ DERGİSİ
Düzce University Journal of Science & Technology
e-ISSN: 2148-2446
Cilt Volume 10 Sayı Issue 3 2022

Düzce University Journal of Science & Technology

Research Article

GEGE: Predicting Gene Essentiality with Graph Embeddings

Halil Ibrahim KURU[a], Yasin Ilkagan TEPELI[b], Oznur TASTAN [b,*]

[a] *Department of Computer Engineering, Faculty of Engineering, Bilkent University, Ankara, TURKEY*
[b] *Faculty of Natural Sciences and Engineering, Sabanci University, Istanbul, TURKEY*
** Corresponding author's e-mail address: otastan@sabanciuniv.edu*

DOI: 10.29130/dubited.1028387

ABSTRACT

A gene is considered essential if its function is indispensable for the viability or reproductive success of a cell or an organism. Distinguishing essential genes from non-essential ones is a fundamental question in genetics, and it is key to understanding the minimal set of functional requirements of an organism. Knowledge of the set of essential genes is also crucial in drug discovery. Several reports in the literature show that the gene location in a protein-protein interaction network is correlated with the target gene's essentiality. Here, we ask whether the node embeddings of a protein-protein interaction (PPI) network can help predict gene essentiality. Our results on predicting human gene essentiality show that node embeddings alone can achieve up to 88% AUC score, which is better than using topological features to characterize gene properties and other previous work's results. We also show that, when combined with homology information across species, this performance reaches 89% AUC. Our work shows that node embeddings of a protein in the PPI network capture the network connectivity patterns of the proteins and improve the gene essentiality predictions.

Keywords: *Graph representations, Node embeddings, Gene essentiality, Network topological features, Protein-protein interaction network*

GEGE: Çizge Gömülümleriyle Gen Esaslılığını Tahmin Etme

ÖZET

İşlevi, bir hücrenin veya organizmanın hayatta kalabilmesi veya üreme başarısı için vazgeçilmez olan genler, esaslı genler olarak kabul edilir. Esaslı genleri esaslı olmayanlardan ayırt etmek, bir organizmanın minimum fonksiyonel gereksinimlerinin anlaşılabilmesi için genetikte kilit bir sorudur. Esaslı genler küme bilgisi, ilaç tasarlanmasında da çok önemlidir. Literatürdeki, bir protein-protein etkileşim ağındaki gen konumunun, gen esaslılığı ile ilişkili olduğunu göstermiştir. Burada, bir protein-protein etkileşimi (PPI) ağının düğüm yerleştirmelerinin gen gerekliliğini tahmin etmeye yardımcı olup olamayacağını soruyoruz. İnsan geninin esaslığını tahmin etme konusundaki sonuçlarımız, düğüm gömülümlerinin tek başına %88'e kadar AUC skoruna ulaşabileceğini göstermektedir. Bu skor, gen özelliklerini karakterize etmek için topolojik özellikleri kullanılan modellerin başarımından ve önceki çalışma sonuçlarından daha iyidir. Ayrıca, türler arası homoloji bilgisi ile birleştiğinde, bu performansın %89 AUC skoruna ulaştığını gösteriyoruz. Çalışmamız, PPI ağındaki bir proteinin düğüm gömülümlerinin, proteinlerin ağ bağlantı modellerini yakaladığını ve gen esaslılık tahminlerini geliştirdiğini gösteriyor.

Received: 26/11/2021, Revised: 13/01/2013, Accepted: 01/03/2022

Anahtar Kelimeler: *Çizge gösterimleri, Düğüm gömülümleri, Gen esaslılığı, Ağ topolojik özellikleri, Protein-protein etkileşim ağı*

I. INTRODUCTION

A gene is considered essential if its function is indispensable for the viability or reproductive success of a cell or an organism [1]. Identifying essential genes is critical for understanding the minimal functional requirements of an organism or a cell [2], [4]. The knowledge of essential genes also has practical significance for drug target identification. The essential genes of a pathogen constitute potential drug targets for infectious diseases [5.], [6]. Similarly, a gene that is essential for a cancer cell but non-essential for a normal cell reveals a vulnerable point in cancer cells that can be targeted by drugs [7], [8].

Assessing the essentiality of a gene requires assessing the viability of the living system that entirely lacks that gene or in which the expression or function of that gene has been significantly compromised. There are small-scale experimental techniques for single gene knockouts [9]. To find all the essential genes in a cell requires disrupting the genes one at a time and assessing their individual effects on the target cell's viability. Single-gene knockout experiments [10], RNAi screens [11], and more recently CRISPR/Cas9 genome editing technologies [12] have been used for this purpose. While experimental methods provide powerful results, they are laborsome, time-consuming, and costly. Also, the results are highly dependent on the experimental conditions [13]. Computational tools enable predictions that are not otherwise attainable through experimental studies and shed light on the question of what makes a gene essential.

The earliest gene essentiality methods transferred gene essentiality annotations from bacterial species by using homology information [3]. Later, as the list of essential genes accumulated for model organisms, machine learning approaches were used for predicting the gene essentially. Diverse biological information compiled from experimental data or the properties of the genes can be used as features to predict gene essentiality. There exist various methods that make use of properties of the genes, such as gene sequence [14], [16], gene expressions [17], and functional annotation such as gene ontology [18].

With the availability of protein interaction data at a large scale, it is now possible to ask whether the positioning of a gene in the PPI has a relation to the essentiality of the gene coding that protein. Several studies in the literature report that topological network properties of a gene in the PPI network are related to gene essentiality [19], [27]. Towards this aim, centrality measures are explored. Among them, the local centrality quantifies the local connectivity patterns of a node, whereas betweenness centrality indicates whether a node is a key connector of different parts of the network. Early network analysis pointed out that there is a correlation between lethality and the degree of a node, where highly connected proteins in PPI networks tend to be essential [19]. Although others [28], [29] challenged this idea, several other studies that examined datasets in different organisms supported the original idea that proteins with high local centrality are correlated to gene essentiality [20], [22].

Hwang et al. [23] reported correlation with clustering coefficient, defined as the ratio of the number of edges connecting the neighbors of a node to the maximum number of possible edges among them. Other work reported that nodes with high betweenness centrality are likely to be essential [32], [33]. In the search to find the proper centrality measures related to gene essentiality, several other studies interrogated the relationship of gene essentiality with a series of centrality measures [24]. Additionally, studies in the literature have integrated the topological information with additional information to predict essential genes in a binary classification framework in a supervised learning setting [25], [27]. Another relevant work is ProtRank [28]. To arrive at a global topological measure for proteins in the PPI network, ProtRank uses Google's PageRank algorithm [29], which is based on random walks. ProtRank of a protein is computed by conducting random walks on the PPI network and measuring the amount of time the random walker spends on the protein in the PPI network. In this work, the performance of the ProtRank measure is evaluated on the essential yeast genes by computing the ratio of the essential genes found in the top-k ranked genes when the ProtRank measure ranks genes. The author showed differences compared to local topological features such as degree and betweenness centrality.

In this work, we ask whether node embeddings learned in a deep learning framework can represent the topological properties of a protein in the PPI better than the network node properties and help discriminate essential genes from non-essential genes. For this purpose, we learned the node embeddings of the genes in the human PPI network. We used this low-dimensional representation of the genes as features to train a binary classifier. Our results show that the deep graph embedding methods help find good features representations instead of pre-selected topological features. Additionally, we show that information on gene conservation across species improves adds value to these predictions.

II. METHODS

A. GEGE FRAMEWORK

In this study, we set out to predict the essentiality of genes by formulating this problem as a classification task over the nodes of a protein-protein interaction (PPI) network. We denote this dataset as D = $\{\mathbf{x}_i, y_i\}_{i=1:N}$, where N is the number of examples, $\mathbf{x}_i$ is a multi-dimensional numeric representation of the gene i and $y_i \in \{-1, 1\}$ is the class label for gene i. Here, 1 indicates the essential gene class and −1 indicates the non-essential gene class. The node classification task involves predicting the most probable label for the node.

The GEGE framework comprises two main steps: representing each gene with feature vectors based on node embeddings and building a classifier based on the learned representations of the genes. This feature vector can be augmented with additional information on the genes. Figure 1 summarizes this idea. In the following sections, we detail the methodology and data sources.

B. REPRESENTING GENES WITH NODE EMBEDDINGS

We represent the PPI network as a graph, $G = (V, E)$, where V is the set of vertices representing the genes coding for the proteins, and E denotes an edge between two such genes. The feature vector for a gene, $\mathbf{x}_i$, is created based on the node embeddings which we learn on G.

A graph embedding of node v returns a feature representation of this node in a d-dimensional space such that the local structures and the similarities between the nodes are conserved in this new feature space. This representation is learned based on the relationship of the nodes with each other; thus, the topology of the graph. In this representation, highly connected nodes belong to the same communities due to the homophily principle, and they are expected to be embedded closely. Additionally, the nodes that are not necessarily close in the network but have similar structural nodes (e.g., hubs) shall have similar embeddings. In short, these methods operate with homophily [34] and structural equivalence [35] principles.

We experiment with two different node embedding methods in the current study: DeepWalk [36] and node2vec [37]. Both methods derive an embedding based on the neighborhood of a node, wherein the neighborhood is based on random walks on a graph. They both aim to minimize the differences between the graph representation and the embedding representation. Random walks centered on a vertex v are used to derive a neighborhood of a given vertex v_i. In the literature, random walk approaches have been used for similarity measures and describing the local community information of a graph [38], [39]. On the other hand, using random walks to capture the local structures of a graph is a reasonable choice because it requires less computational power than the approaches that use the whole graph [36].

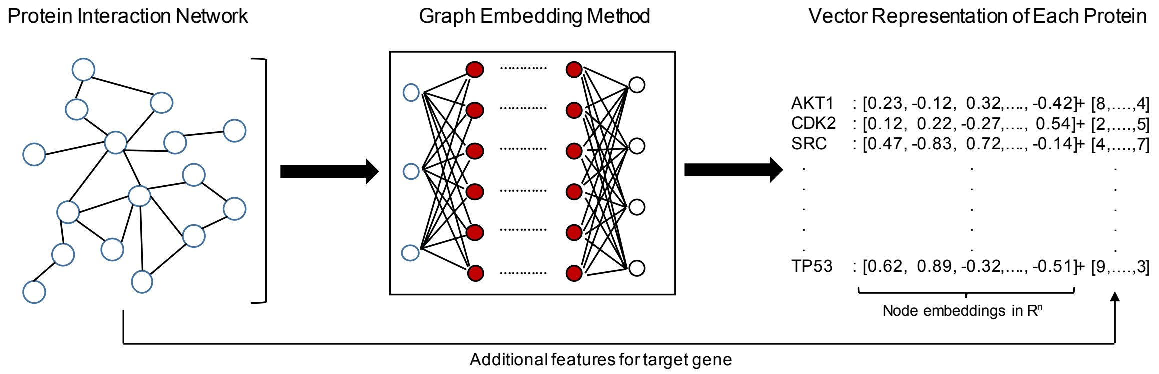
Protein Interaction Network
Graph Embedding Method
Vector Representation of Each Protein
AKT1 : [0.23, -0.12, 0.32,...., -0.42]+ [8,....,4]
CDK2 : [0.12, 0.22, -0.27,...., 0.54]+ [2,....,5]
SRC : [0.47, -0.83, 0.72,...., -0.14]+ [4,....,7]
TP53 : [0.62, 0.89, -0.32,...., -0.51]+ [9,....,3]
Node embeddings in Rⁿ
Additional features for target gene

Figure 1. *Schematic describing how the feature vectors are created based on node embeddings. Input is the PPI network, and the output is a latent low-dimensional representation of nodes in the network. Additional features about the genes can be concatenated to the node embedding vectors.*

C. NETWORK TOPOLOGY MEASURES

An alternative to the PPI network representation is to use a set of network topology measures to describe each node. Several centrality measures are known to be correlated with gene essentiality [17], [19], [20], [23], [32], [33], [40]. In the current study, we select four mostly used topological features, which are closeness centrality, degree centrality, betweenness centrality, and clustering coefficient. We use the SNAP library [41] to calculate each of these topology metrics.

D. SUPERVISED CLASSIFICATION

After representing each gene with a low-dimensional feature vector, a Support Vector Machine (SVM) classifier is used in the second step of our framework. We use SVM because of its effectiveness in a variety of tasks. The SVM model parameters, embedding size of the node embeddings, the number of walks, walk length, *p* and *q* parameters are tuned via grid search strategy in 10-fold cross-validation. We report the area under the curve (AUC), F1, and average precision (AP) scores.

E. DATASET

E. 1. Gene Essentiality Data

The information of whether a gene is essential or not is obtained from [16]. The origin of the data is from the DEG (http://tubic.tju.edu.cn/deg/) database, which compiled datasets from three different studies [12, 42, 43]. Guo et al. [16] obtain 11 different gene essentiality sets along with corresponding cell lines. They mark a gene as positive (essential) if it is reported as essential in more than half of the cell lines. The final dataset contains 12,015 genes. Among these 12,015 genes, 1,516 of them are essential. More details on the criteria of deciding which genes are essential can be found in [16].

E. 2. Protein Interaction Data

The protein-protein interaction network is obtained from InbioMap, which is publicly available at https://www.intomics.com/inbio/map.html. InbioMap specifies a confidence score for each edge, which represents the support of the interaction in the literature. The interactions that have lower than 0.1 confidence cut-off are eliminated from the network to remove noisy edges. The remaining network includes 17,653 genes and 625,641 interactions between these genes. Among the 12,015 genes that have information on their gene essentiality, 10,579 are in the PPI network. Of the 10,579 genes that are present in the PPI network genes, 1,514 genes are essential, which constitute the positive class in our data, while the remaining 9,065 genes are not essential, and they constitute the negative class.

E. 3. Homology Information

Homology information is obtained from the HGNC Comparison of Orthology Predictions (HCOP) database (https://www.genenames.org/cgi-bin/hcop). The data contain homologous gene information of human genes with other 19 species.

II. RESULTS AND DISCUSSION

This section presents the results of predicting gene essentiality using a different representation of gene topology in the PPI network.

A. PREDICTIVE PERFORMANCE

In these experiments, we apply the node2vec and the DeepWalk algorithms to generate gene node embeddings on the PPI network. We compare these results with the alternate representation of different topological features. Additionally, we counted the number of organisms in which a gene is conserved for each human gene and used this as an additional feature.

Table 1 shows the best performances for different feature settings and when SVM is run with linear or RBF kernels. When features are derived from conventional topological features that describe the gene's connectivity pattern in the PPI network, the best result obtained with SVM using the RBF kernel is 0.831 AUC score. Adding the homology feature improves the results slightly up to 0.846. These results are surpassed by the models that use node2vec and DeepWalk embeddings to represent network nodes. Node2vec alone reaches 0.850 AUC, and when homology information is added, it can achieve 0.868 AUC. DeepWalk representation is the one that yields the highest performance metrics. The node embedding features obtained from DeepWalk results with a 0.874 AUC score and addition of the homology feature raise this to 0.884. These results also hold when models are compared with accuracy, F1, and average precision (Table 1).

One interesting observation is that using a non-linear kernel instead of a linear kernel yields different predictive performance gains in models trained with network topological features and models trained with graph embedding features. Comparing linear vs. RBF kernel results, we observe that the use of non-linear kernels improves the network topology-based methods' performance drastically. F1 score increases from 0.395 to 0.62, and similar increases are observed in other performance metrics (Table 1). In contrast, the gain in performance with a non-linear kernel for graph-based feature representation is modest. This may be related to the fact that graph embedding methods can extract relevant non-linear features during model training; as a result, there is no additional benefit obtained using a non-linear kernel. On the other hand, the network topology-based features cannot capture this non-linearity.

To assess the robustness of our approach, we evaluate the configuration of our best performance with 100 random bootstrap samples. We randomly split our dataset into a test (20%) and trained (80%) 100 times. Our best performance in 10-fold cross-validation produces 0.884 mean AUC, 0.687 F1, and 0.514 average precision (AP), and we use the same parameters with the configuration of these results in 100 random bootstrapped samples. This experiment finds 0.881 mean AUC, 0.683 mean F1, and 0.508 AP.

***Table 1.** Gene essentiality prediction performances when different features are input to the SVM classifier, and the kernel of choice is varied.*

Embeddings	Kernel	ACC	AUC	F1	AP
DeepWalk	Linear	0.856	0.867	0.637	0.457
	RBF	0.871	0.874	0.661	0.483
DeepWalk + homology	Linear	0.875	0.883	0.672	0.497
	RBF	**0.885**	**0.884**	**0.687**	**0.514**
node2vec	Linear	0.856	0.840	0.588	0.405
	RBF	0.856	0.850	0.625	0.442
node2vec + homology	Linear	0.853	0.860	0.629	0.448
	RBF	0.880	0.868	0.669	0.491
Topological Features	Linear	0.584	0.736	0.395	0.244
	RBF	0.847	0.831	0.602	0.416
Topological Features + homology	Linear	0.802	0.824	0.553	0.371
	RBF	0.844	0.846	0.609	0.426
Guo et al. [16]	Linear	NA	0.845	NA	NA

These results are close to the 10-fold cross-validation results. Therefore, we claim that our framework is robust against test dataset selection.

Results with the additional homology feature reach 0.884 mean AUC, 0.687 mean F1, and 0.514 mean AP scores for DeepWalk embeddings with RBF kernel. These results are better than those reported in [16], who used the same gene essentiality dataset for their predictions and used nucleotide sequence features. The best performing model achieves 0.845 mean AUC in 5-fold cross-validation. Their 5-fold cross-validation results achieved a 0.885 mean AUC score with a feature selection. They applied a similar strategy to our bootstrap experiment. Therein, they randomly split the data into test and train with a 20% ratio and found 0.854 mean AUC across 100 samples. To sum up, our results indicate that node embeddings are highly predictive of gene essentiality.

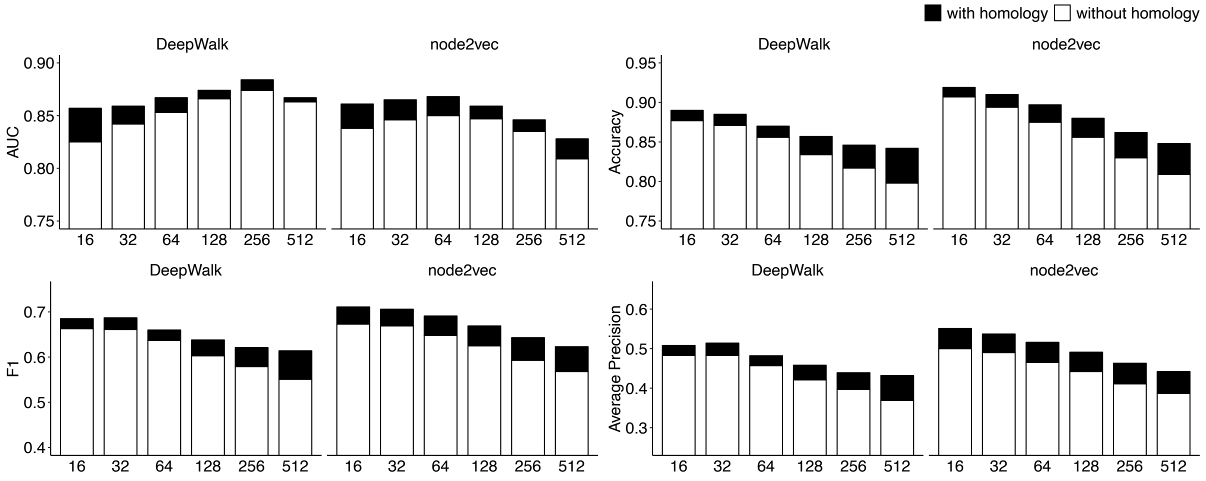

***Figure 2.** Performance change obtained when node2vec and DeepWalk methods are run with varying embedding sizes. The black portion of the bar indicates the performance gain due to the homology feature. F1, AUC, and Average Precision metrics are provided.*

B. THE EFFECT OF THE EMBEDDING SIZES

Node embedding methods have a number of parameters to control the trade-off between overfitting and overgeneralization. The dimension of the embedding space is the most important parameter. Node embedding methods return a feature representation in Rd where d is the dimension of embedding. In this experiment, we explored the effect of the embedding size on the DeepWalk and node2vec performances. We varied the embedding sizes while we fixed the other parameters to their best values. Therefore, we show the effect of embedding size on average performance under 10-fold cross-validation. Figure 2 shows how the performance changes when different embedding sizes are used. We find the best result as 0.884 mean AUC score among the 10-folds. The patterns from the figures reveal that DeepWalk embeddings perform better than node2vec embeddings in the adopted settings. For the kernel parameter of SVM, the RBF kernel gives about 1 % higher performance compared to the linear kernel. We find the best AUC score for the linear kernel with DeepWalk embeddings as 0.867 and 0.874 mean AUC score for the RBF kernel when the embedding size d is set to 256. The node2vec embeddings give their best performance when the embedding size d is set to 64 with RBF SVM, and it leads to 0.85 AUC score while the best performance with linear kernel achieves 0.84 AUC score. Nearly in all embedding sizes, DeepWalk embeddings consistently outperform node2vec embeddings.

C. PERFORMANCE WITH ADDITIONAL HOMOLOG GENES

We calculate the number of organisms that maintain genes homologous with the target gene, and we call this feature the homology feature. We add this feature to our graph embeddings for each node and apply the same procedure for assessing the essentiality. As shown in Table 1 and further evidenced in Figure 2, where we vary the embedding sizes and summarize the best overall results, homology brings complementary information and improves the results by about 2% in accuracy in all configurations. The DeepWalk algorithm's best performance with RBF SVM improves from 0.874 to 0.884, and the best performance with Linear SVM improves from 0.867 to 0.883. Similarly, node2vec's best performance with the RBF kernel improves from 0.850 to 0.868, while the best performance with the linear kernel improves from 0.84 to 0.86.

D. EXPLORING CONSISTENTLY MISCLASSIFIED AS ESSENTIAL

We examine the genes labeled as non-essential in the dataset but are consistently predicted as essential genes in our repeated bootstrap experiments. These constitute the false positive predictions of the classifier. As the experimental datasets are incomplete, these genes could indeed be essential genes in reality. We calculate the counts of false positive predictions in 100 bootstrap experiments for each gene. We refer to this fraction as the false positive rate. We examine the genes whose false positive rates are greater than 0.50. Among these genes, we find that some of the genes are actually reported as conditionally essential genes. In a given context, the gene is important for the organism's viability. For example, SERPINE1, AIMP1, FIGF, RPS6KA6, and PDK4 genes are listed as non-essential in our benchmark dataset. However, we find that [44] reports that these genes as essential. A study in [45] labels the DAZ2 gene as essential, while [46] labels KIAA0408 and ZCCHC13 genes as essential. These outcomes show that relations among the protein-protein interaction network may provide potential information about the essentiality of a gene. Current non-essential genes may be labeled as essential in future experiments, and interactions between proteins may lead to discoveries of new essential genes.

IV. CONCLUSION

In this study, we propose a framework called GEGE, which predicts gene essentiality based on node embeddings of the genes in the PPI. We learn a latent lower-dimensional representation of the nodes in the PPI network with two different graph embedding methods, DeepWalk [36] and node2vec [37]. By applying machine learning algorithms to this new representation of the genes, we show that the gene essentiality can be predicted with high success. We compare our predictions with a previously reported work that reports results on the same dataset, GEGE, which overperforms this method by 4%. We also compare our results to the alternative of representing each node with topological node features. Graph embeddings achieve significant improvements in all settings.

In our experiments, when compared to node2vec, DeepWalk embeddings achieve the best performance, but their results are very close. The framework also allows for the addition of other gene features. When we augment the node embeddings with homology information, we observe performance improvements in all settings. We perform a robustness analysis with 100 random bootstrap samples, which shows that the results are not affected by the selection of random test genes. We also investigate the genes whose true labels are in the benchmark dataset but are repeatedly predicted as essential genes in the 100 bootstrap samples. Some of these genes are reported to be conditionally essential genes; they can be conditional depending on the context. This work can be extended in different directions: (i) The gene essentiality predictions can be tested for other organisms, and (ii) other relevant features, in addition to homology information, can be incorporated into this framework.

ACKNOWLEDGEMENTS: Halil Ibrahim Kuru acknowledges Bilkent University for department fellowship. Yasin İlkağan Tepeli acknowledges TUBITAK-BIDEB for the 2210-A scholarship program.

V. REFERENCES

[1] G. Rancati, J. Moffat, A. Typas, N. Pavelka, “Emerging and evolving concepts in gene essentiality”, *Nature Reviews Genetics,* vol. 19, no.1, pp. 34, 2018.

[2] M. Itaya, “An estimation of minimal genome size required for life”, *FEBS Letters*, vol. 362, no.3, pp. 257–60, 1995.

[3] A. R. Mushegian, E.V. Koonin, “A minimal gene set for cellular life derived by comparison of complete bacterial genomes”, *Proceedings of the National Academy of Sciences*, vol. 93, no.19, pp. 10268–73, 1996.

[4] E.V. Koonin, “How many genes can make a cell: the minimal-gene-set concept”, *Annual Review of Genomics and Human Geneti*cs, vol. 1, no. 1, pp. 99–116, 2000.

[5] M.Y. Galperin, E.V. Koonin, “Searching for drug targets in microbial genomes”, *Current Opinion in Biotechnology*, vol. 10, no. 6, pp. 571–78, 1999.

[6] A.F. Chalker, R.D. Lunsford, “Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach”, *Pharmacology & Therapeutics*, vol. 95, no. 1, pp. 1–20, 2002.

[7] H. Farmer, N. McCabe, C.J. Lord, A.N. Tutt, D.A. Johnson, T.B. Richardson, et al. “Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy”, *Nature*, vol. 434, no. 7035, pp. 917, 2005.

[8] N.J. O’Neil, M.L. Bailey, P. Hieter, “Synthetic lethality and cancer”, *Nature Reviews Genetics,* vol. 18, pp. 10, pp. 613, 2017.

[9] A. Cho, N. Haruyama, A.B. Kulkarni, “Generation of transgenic mice”, *Current Protocols in Cell Biology*, vol. 42, no. 1, chapter. 19, unit. 11, 2009.

[10] G. Giaever, A.M. Chu, L. Ni, C. Connelly, L. Riles, S. V´eronneau, et al. “Functional profiling of the Saccharomyces cerevisiae genome”, *Nature*, vol. 418, no. 6896, pp. 387–91, 2002.

[11] J.M. Silva, K. Marran, J.S. Parker, J. Silva, M. Golding, M.R. Schlabach, et al. "Profiling essential genes in human mammary cells by multiplex RNAi screening", *Science*, vol. 319, no. 5863, pp. 617–20, 2008.

[12] T. Wang, K. Birsoy, N.W. Hughes, K.M. Krupczak, Y. Post, J.J. Wei, et al. "Identification and characterization of essential genes in the human genome", *Science*, vol. 350, no. 6264, pp. 1096–101, 2015.

[13] M.A. D'Elia, M.P. Pereira, E.D. Brown, "Are essential genes really essential?", *Trends in Microbiology,* vol. 17, no. 10, pp. 433–8, 2009.

[14] L.W. Ning, H. Lin, H. Ding, J. Huang, N.N.M. Rao, F.B. Guo, "Predicting bacterial essential genes using only sequence composition information", *Genetics and Molecular Research: GMR,* vol. 13, no. 2, pp. 4564–72, 2014.

[15] W.C. Wei, L.W. Ning, Y.N. Ye, F.B. Guo. "Geptop: A gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny", *PloS One*; 2013.

[16] F.B. Guo, C. Dong, H.L. Hua, S. Liu, H. Luo, H.W. Zhang, et al. "Accurate prediction of human essential genes using only nucleotide composition and association information", *Bioinformatics*, 33 12:1758–64, 2017.

[17] J. Deng, L. Deng, S. Su, M. Zhang, X. Lin, L. Wei, et al. "Investigating the predictability of essential genes across distantly related organisms using an integrative approach", *Nucleic Acids Research*, vol. 39. no. 3, pp. 795-807, 2011.

[18] L. Chen, Y.H. Zhang, S. Wang, Y. Zhang, T. Huang, Y.D. Cai, "Prediction and analysis of essential genes using the enrichments of gene ontology and KEGG pathways", *PloS One*, vol. 12, no. 9, e0184129, 2017.

[19] H. Jeong, S.P. Mason, A.L. Barabasi, Z.N. Oltvai, "Lethality and centrality in protein networks", Nature, vol. 411, no. 6833, pp. 41-2, 2001.

[20] M.W. Hahn, A.D. Kern, "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks", *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 803–6, 2004.

[21] N.N. Batada, L.D. Hurst, M. Tyers, "Evolutionary and physiological importance of hub proteins", *PLoS Computational Biology*, vol. 2, no. 7, e88, 2006.

[22] E. Zotenko, J. Mestre, D.P. O'Leary, T.M. Przytycka, "Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality", *PLoS Computational Biology*, vol. 4, no. 8, e1000140, 2008.

[23] Y.C. Hwang, C.C. Lin, J.Y. Chang, H. Mori, H. F. Juan, H.C. Huang, "Predicting essential genes based on network and sequence analysis", *Molecular BioSystems*, vol. 5, no.12, pp. 1672–78, 2009.

[24] J. Wang, M. Li, H. Wang, Y. Pan, "Identification of essential proteins based on edge clustering coefficient", *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 9, no. 4, pp. 1070–80, 2012.

[25] M.L. Acencio, N. Lemke, "Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information", *BMC Bioinformatics*, vol. 10, no. 1, pp. 290, 2009.

[26] J. Cheng, W. Wu, Y. Zhang, X. Li, X. Jiang, G. Wei, et al. "A new computational strategy for predicting essential genes", *BMC Genomics*, vol. 14, no. 910, 2013.

[27] M.C. Palumbo, A. Colosimo, A. Giuliani, L. Farina, "Functional essentiality from topology features in metabolic networks: a case study in yeast", *FEBS Letters,* vol. 579, no. 21, pp. 4642-6, 2005.

[28] T. Can, "ProtRank: A better measure for protein essentiality," in *Proceedings of the 3rd International Symposium on Health Informatics and Bioinformatics (HIBIT'08),* Istanbul, May 2008.
[29] L. Page, S. Brin, R. Motwani and T. Winograd, "The Pagerank Citation Ranking: Bringing Order to the Web," *Technical Report, Stanford University, Stanford*, 1998.

[30] S. Coulomb, M. Bauer, D. Bernard, M.C. Marsolier-Kergoat, "Gene essentiality and the topology of protein interaction networks", *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 272, no. 1573, pp. 1721–1725, 2005.

[31] X. He, J. Zhang, "Why do hubs tend to be essential in protein networks?"*, PLoS Genetics,* vol. no. 6, e88, 2006.

[32] H. Yu, P.M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, "The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics", *PLoS Computational Biolog*y, vol. 3, no. 4, e59, 2007.

[33] M.P. Joy, A. Brock, D.E. Ingber, S. Huang, "High-betweenness proteins in the yeast protein interaction network", *BioMed Research International*, vol. 2005, no. 2, pp. 96–103, 2005.

[34] M. McPherson, L. Smith-Lovin, J.M. Cook, "Birds of a feather: Homophily in social networks", *Annual Review of Sociology*, vol. 27, 1, 415–44, 2001.

[35] F. Lorrain, H.C. White, "Structural equivalence of individuals in social networks", *The Journal of Mathematical Sociology*, vol. 1, no. 1, pp. 49–80, 1971.

[36] B. Perozzi, R. Al-Rfou, S. Skiena, "DeepWalk: Online Learning of Social Representations", *KDD: Proceedings International Conference on Knowledge Discovery & Data Mining*, pp. 701–10, 2014.

[37] A. Grover, J. Leskovec, "node2vec: Scalable Feature Learning for Networks", *KDD: Proceedings International Conference on Knowledge Discovery & Data Mining*, pp.855–864, 2016.

[38] R. Andersen, F. Chung, K. Lang, "Local graph partitioning using PageRank vectors", IEEE, pp. 475–86, 2006.

[39] F. Fouss, A. Pirotte, J.M. Renders, M. Saerens, "Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation", *IEEE Transactions on Knowledge and Data Engineerin*g, vol. 19, no. 3, pp. 355–69, 2007.

[40] Y. Chen, D. Xu. "Understanding protein dispensability through machine-learning analysis of high-throughput data", *Bioinformatics*, vol. 21, no. 5, pp. 575–81, 2004.

[41] J. Leskovec, R. Sosic, "SNAP: A general-purpose network analysis and graph-mining library", *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 1, pp. 1, 2016.

[42] T. Hart, M. Chandrashekhar, M. Aregger, Z. Steinhart, K.R. Brown, G. MacLeod, et al. "High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities", *Cell*, vol. 163, no. 6, pp. 1515–26, 2015.

[43] V.A. Blomen, P. Majek, L.T. Jae, J.W. Bigenzahn, J. Nieuwenhuis, J. Staring, et al. "Gene essentiality and synthetic lethality in haploid human cells", *Science*, vol. 350, no. 6264, pp.1092–6. 2015.

[44] J.M. Silva, K. Marran, J.S. Parker, J. Silva, M. Golding, M.R. Schlabach, et al. "Profiling essential genes in human mammary cells by multiplex RNAi screening", *Science*, vol. 319, no. 5863, pp. 617–20, 2008.

[45] R. Marcotte, K.R. Brown, F. Suarez, A. Sayad, K. Karamboulas, P.M. Krzyzanowski et al. "Essential gene profiles in breast, pancreatic, and ovarian cancer cells", *Cancer Discovery*, vol. 2, no. 2, pp. 172–89, 2012.

[46] J. Luo, M.J. Emanuele, D. Li, C.J. Creighton, M.R. Schlabach, T. Westbrook, et al. "A Genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene", *Cell*, vol. 137, no. 5, pp. 835–48, 2009.