SCIENCE CHINA
Information Sciences

• LETTER •

July 2021, Vol. 64 179101:1–179101:3
https://doi.org/10.1007/s11432-018-1512-0

Logistic regression algorithm to identify candidate disease genes based on reliable protein-protein interaction network

Xiujuan LEI* & Wenxiang ZHANG

School of Computer Science, Shaanxi Normal University, Xi'an 710119*, China*

Received 15 August 2018/Revised 30 November 2018/Accepted 18 February 2019/Published online 17 May 2021

Citation Lei X J, Zhang W X. Logistic regression algorithm to identify candidate disease genes based on reliable protein-protein interaction network. Sci China Inf Sci, 2021, 64(7): 179101, https://doi.org/10.1007/s11432-018-1512-0

Dear editor,

Genetic diseases have seriously threatened human health; however, identifying candidate disease genes is expected to contribute to understanding complex genetic diseases. Several machine learning algorithms based on protein-protein interaction (PPI) network have been developed to prioritize candidate diseases [1,2]. However, most methods ignore the intrinsic properties of proteins, such as domain information and gene ontology (GO) annotations. Additionally, PPI networks generated by high-throughput technologies typically contain numerous false positives and false negatives. This study proposes a method that uses a logistic regression algorithm in a reliable PPI network (LR-RPN) to identify disease genes. First, a heterogeneous network is constructed based on the PPI network and keywords obtained from the universal protein resource (UniProt) database [3]. The keywords can be utilized to retrieve subsets of protein entries and create indexes of entries based on functions and structures. Further, keywords can be classified into 10 categories, including domain, biological process (BP), and cellular component (CC) [3]. A previous study [4] applied keywords to enrich the PPI network, and the results demonstrated that such keywords can improve prediction accuracy significantly. Subsequently, we use the random walk with restart algorithm on a heterogeneous network to calculate topological similarities. Furthermore, we extract a topological similarities matrix of the PPI network and normalize it by row. Finally, a reliable PPI network is constructed using the topological similarities matrix to connect pairs of nodes with a similarity greater than a given threshold. Next, a logistic regression algorithm is utilized to identify candidate disease genes based on multiple heterogeneous biological features on a reconstructed biological network. The multiple heterogeneous biological features are extracted from the direct neighbors of a disease-related gene in the reliable PPI network, human protein complexes, tissue expressions, and semantic similarity of genes based on GO terms.

Our innovations. We propose LR-RPN to identify candidate disease genes. The innovations of LR-RPN are summarized as follows.

(1) Reconstructing PPI network. PPI networks play a remarkable role in the identification of disease-related genes. Several studies [1,2,4] have demonstrated that proteins related to the same or similar diseases tend to exhibit common topological characteristics in a PPI network. However, PPI networks ignore cases where two proteins are not connected but have a certain biological relation. Additionally, PPI networks typically contain several false negative interactions and highly skewed degree distribution [4,5]. These two issues seriously impact the prediction of disease-related genes. Therefore, we first combine a PPI network and keywords to form a heterogeneous network to physically link some proteins with certain biological relations. Next, we apply the random walk algorithm to calculate the similarity between proteins based on this heterogeneous network. Forming such a reliable PPI network to reduce issues related to spurious interactions and highly skewed degree distributions is reliable.

(2) Integrating multiple heterogeneous data. Complex diseases are typically caused by various factors, such as genetic and environmental factors [6,7]. Considering only a PPI network is insufficient, thus, we require additional biological knowledge about proteins, such as their molecular functions and biological processes. Herein, keywords are utilized to construct a reliable PPI network. They are obtained from the UniProt database [3], which contains various biological aspects related to proteins. A previous study [4] demonstrated that including keywords in a PPI network can facilitate identification of disease-related genes. Additionally, heterogeneous features are extracted from the reliable PPI network, protein complexes, tissue expressions, and the semantic similarity of genes based on GO terms. Then, these features are utilized to identify disease-related genes using a logistic regression algorithm. Therefore, disease-related genes can be mined by integrating multiple heterogeneous biological data.

(3) Construction of 10 logistic regression model classifiers.

* Corresponding author (email: xjlei@snnu.edu.cn)

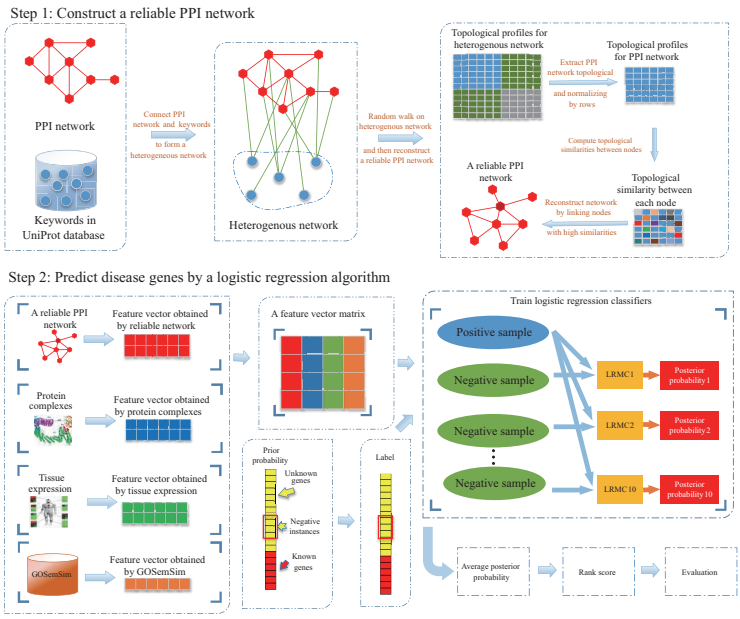

Figure 1 (Color online) Overall framework of LR-RPN for prioritizing disease-related genes. LR-RPN first constructs a reliable PPI network, and then predicts disease-related genes using a logistic regression algorithm.

Identifying disease-related genes is a class-imbalance problem, which means that the number of disease-related genes is much smaller than the number of unknown genes. This introduces bias to unknown genes in trained logistic regression model classifiers. Therefore, we employ the under-sampling method, in which the number of unknown genes is twice that of disease-related genes. Next, the selected unknown genes and disease-related genes in the under-sampling process are utilized to train the logistic regression model. However, under-sampling likely leads to loss of some important information by deleting unknown disease-related genes. Inspired by [8], we employ the under-sampling method to train 10 logistic regression classifiers. Then, the final posterior probability is equal to the average of the posterior probability for the 10 logistics regression classifiers.

Dataset. In this study, the PPI network contains 7311 human genes, and the protein complex data involves 2870 protein complexes that are associated with 3881 human genes. The gene-disease association data contain 12 disease classes associated with 815 human genes. The tissue expression data include 1110 proteins in 307 types of expression terms. The keyword data are collected from the UniProt database [3], which contains 632 types of terms with at least two proteins. We also select five multifactorial diseases from the OMIM database [9] (from December 2017), that are associated with greater than or equal to 10 known valid disease-related genes located in different genomic regions according to the OMIM file (morbidmap.xlsx).

LR-RPN algorithm. The overall framework to identify disease-related genes is shown in Figure 1. The proposed LR-RPN first constructs a heterogeneous network using the PPI network and keywords. Then, random walk with restart is utilized to obtain a topological profile for a heterogeneous network. Here, the goal is to construct a reliable PPI network; thus, we only extract the topological profiles of the PPI network. Inspired [5], a reliable PPI network is constructed based on the topological profiles of the PPI network. We then extract multiple heterogeneous biological features from the reliable PPI network, protein complexes, tissue expressions and the semantic similarity of genes based on GO terms. Finally, a logistic regression algorithm is employed to prioritize disease-related genes, in which we use the under-sampling method and train 10 logistic regression model classifiers (LRMC) to overcome the class-imbalance classification problem. Refer to Appendixes A–C for more details about the LR-RPN algorithm.

Experiment. The performance of the proposed LR-RPN was evaluated using leave-one-out cross validation (LOOCV). According to the LOOCV results, the receiver operating characteristic curve was plotted and the area under the curve value was computed. In addition, precision and recall were used to evaluate the performance of the proposed method. In the experiment, we first evaluated the effects of different forms of features in the original PPI network (Figure B1), which proved that integrating multiple heterogeneous data enhances the performance of identifying disease-related genes. Next, we analyzed the effect of the LR-RPN algorithm's parameters on identifying disease-related genes, as shown in Figures B2 and B3 and Appendix C, where the optimal parameters were extracted for the LR-RPN algorithm. We then compared the performance of prioritizing disease-related genes between the original PPI and reliable PPI networks, as shown Figure B4, to verify the reliability of the reconstructed PPI network. In addition, the LR-RPN algorithm was compared to other methods (Table B2, Figures B5–B7). The results indicate that the proposed LR-RPN algorithm performs better than existing algorithms. Finally, we conducted case studies to verify the effectiveness of the proposed LR-RPN relative to identification of disease-related genes (Table B3).

Conclusion. We have proposed a logistic regression algorithm to identify candidate disease-related genes based on multiple heterogeneous biological features on a reconstructed biological network. The proposed method integrates multiple heterogeneous data, such as tissue expressions, protein complexes, a PPI network, and GO annotations. In addition, a new heterogeneous network is constructed by connecting the PPI network and keywords from

the UniProt database, which helps reduce the number of spurious interactions and highly skewed degree distribution in PPI networks. To avoid bias to unknown genes for LRMC, we employed the under-sampling method, and then trained 10 LRMCs. The experimental results demonstrated that the proposed method is promising and can be applied to various disease classes to explore new disease-related genes.

Nevertheless, there are still some limitations that must be addressed. Although the proposed method can integrate various types of data, it has some disadvantages because we integrate some irrelevant or redundant biological information. Logistic regression is a machine learning algorithm, and as such, it suffers the typical shortcomings of machine learning algorithms. In addition, there are many categories of biological data, such as disease phenotype similarities, pathways, and gene expression profiles. In this study, we only integrated limited biological data; therefore, in future, we plan to conduct addition research to overcome these limitations.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61672334, 61972451, 61902230) and Fundamental Research Funds for the Central Universities (Grant No. GK201901010).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Li Y J, Patra J C. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics, 2010, 26: 1219–1224
2 Chen B, Li M, Wang J X, et al. A fast and high performance multiple data integration algorithm for identifying human disease genes. BMC Med Genomics, 2015, 8: 2
3 The UniProt Consortium. UniProt: a hub for protein information. Nucleic Acid Res, 2015, 43: 204–212
4 Kircali S A, Fang Y, Wu M, et al. Disease gene classification with metagraph representations. Methods, 2017, 131: 83–92
5 Lei C W, Ruan J H. A novel link prediction algorithm for reconstructing protein-protein interaction networks by topological similarity. Bioinformatics, 2013, 29: 355–364
6 Cardon L R, Bell J I. Association study designs for complex diseases. Nat Rev Genet, 2001, 2: 91–99
7 Maher B. The case of the missing heritability. Nature, 2008, 456: 18–21
8 Liu X Y, Wu J X, Zhou Z H. Exploratory undersampling for class-imbalance learning. IEEE Trans Syst Man Cybern B, 2009, 39: 539–550
9 McKusick V A. Mendelian inheritance in man and its online version, OMIM. Am J Human Genets, 2007, 80: 588–604