Gene 551 (2014) 138–151

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Analysis and identification of essential genes in humans using topological properties and biological information

Lei Yang [a], Jizhe Wang [a], Huiping Wang [a], Yingli Lv [a], Yongchun Zuo [b], Xiang Li [c,*], Wei Jiang [a,*]

[a] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China
[b] The National Research Center for Animal Transgenic Biotechnology, Inner Mongolia University, Hohhot 010021, PR China
[c] Department of Neurology, The Fourth Affiliated Hospital of Harbin Medical University, Harbin 150081, PR China

ARTICLE INFO

*Article history:*
Received 12 May 2014
Accepted 25 August 2014
Available online 27 August 2014

*Keywords:*
Human essential genes
Protein–protein interaction
Statistical test
Support vector machine

ABSTRACT

Genes that are indispensable for survival are termed essential genes. The analysis and identification of essential genes are very important for understanding the minimal requirements of cellular survival and for practical purposes. Proteins do not exert their function in isolation of one another but rather interact together in PPI networks. A global analysis of protein interaction networks provides an effective way to elucidate the relationships between proteins. With the recent large-scale identifications of essential genes and the production of large amounts of PPIs in humans, we are able to investigate the topological properties and biological properties of essential genes. However, until recently, no one has ever investigated human essential genes using topological and biological properties. In this study, for the first time, 28 topological properties and 22 biological properties were used to investigate the characteristics of essential and non-essential genes in humans. Most of the properties were statistically discriminative between essential and non-essential genes. The F-score was used to estimate the essentiality of each property. The GO-enrichment analysis was performed to investigate the functions of the essential and non-essential genes. Finally, based on the topological features and the biological characteristics, a machine-learning classifier was constructed to predict the essential genes. The results of the jackknife test and 10-fold cross validation test are encouraging, indicating that our classifier is an effective human essential gene discovery method.

## 1. Introduction

Essential genes are indispensable for the survival of an organism (Seringhaus et al., 2006; Wang et al., 2013). These genes constitute the minimal gene set that is required for a living cell (Glass et al., 2009; Juhas et al., 2011). Therefore, the functions of these genes are essential, and the deletion of only one of the essential genes is sufficient to result in infertility or lethality (Zhang and Lin, 2009; Zhang and Zhang, 2008). Moreover, essential genes are important for defense against pathogens in humans, and essential genes have been found to be correlated with human disease genes (Furney et al., 2006). Therefore, the identification and analysis of essential genes have been a major focus in genomic research and in drug design. The experimental identification of essential genes is usually applied by single gene deletions (Giaever et al., 2002; Kobayashi et al., 2003), RNA interference (Cullen and Arndt, 2005; Kamath et al., 2003), antisense RNA (Ji et al., 2001) and transposon mutagenesis (Gallagher et al., 2007). However, there are several limitations to these experimental methods: a large investment of time and resources is required by these experimental methods, the speed of identifying essential genes falls far behind the speed of genome sequencing, these experimental techniques are limited to a few species, and the experimental conditions impact the identification of essential genes. Consequently, developing a fast and effective way to computationally identify essential genes would be of great value.

Because the majority of proteins interact with each other for proper function in a cell, knowledge about the interactions between proteins is essential to understand the molecular and cellular functions (Chaurasia et al., 2007; Rual et al., 2005; Stelzl et al., 2005; Yang et al., 2014). Therefore, the study of PPI networks provides many new insights into protein function and a global view of the biological systems in the context of a network. However, most of the PPI networks are too complex to be easily understood. This problem can be overcome using theoretical graph concepts to investigate the topological properties of the PPI networks. Topological properties have been used to study social networks in social sciences (Wasserman and Faust, 1994). Furthermore, topological properties have been used to evaluate the properties of PPI networks. Xu and Li used five topological properties to describe disease genes in the PPI network; the topological properties were statistically discriminative between the disease genes and the non-disease genes (Xu and Li, 2006).

*Abbreviations:* PPI, protein–protein interaction; LC, literature-curated; BN, bottleneck; EPC, edge percolation component; MNC, maximum neighborhood component; DMNC, density of maximum neighborhood component; ASP, average shortest path; Ka, the number of non-synonymous substitutions per non-synonymous site; Ks, the number of synonymous substitutions per synonymous site; Ka/Ks, evolutionary rate.

* Corresponding authors.
*E-mail addresses:* lixianghyd@163.com (X. Li), jiangwei@hrbmu.edu.cn (W. Jiang).

http://dx.doi.org/10.1016/j.gene.2014.08.046
0378-1119/© 2014 Elsevier B.V. All rights reserved.

Based on the work of Xu and Li, a new topological property was proposed by Zhang et al. to calculate the statistical significance (Zhang et al., 2010). The work of Zhu et al. showed that the topological properties of drug targets were significantly different from those of non-drug-targets in the human PPI networks (Zhu et al., 2009), and Wang et al. found that these differences were mainly caused by mir-drug-targets and that there was no difference in the topological properties between non-mir-drug-targets and non-drug-targets (Wang et al., 2011). In 2009, 10 topological properties and four sequence-based properties were used by Hwang et al. to describe the essential genes in the *Saccharomyces cerevisiae* and *Escherichia coli* PPI networks (Hwang et al., 2009). There were significant differences in these properties between the essential and non-essential genes. Network-based topological and sequence-based analyses were also used by other researchers in different networks (Coulomb et al., 2005; Feldman et al., 2008; Florez et al., 2010; Goh et al., 2007; Han et al., 2013a, 2013b; He and Zhang, 2006; Hwang et al., 2008; Joy et al., 2005; Kotlyar et al., 2012; Sualp and Can, 2011; Wachi et al., 2005; Xu et al., 2011; Yıldırım et al., 2007; Zotenko et al., 2008). However, until recently, neither network-based topological analysis nor sequence-based analysis has been used in the dataset of human essential genes.

Recently, many computational approaches have been proposed for predicting essential genes or proteins by integrating sequence-based and topology-based features. Chen and Xu integrated the protein evolutionary rate, paralogy, protein size and the degree of centrality to predict the essential proteins in five species including the budding yeast *S. cerevisiae*, the fission yeast *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, *Drosophila melanogaster* and *Caenorhabditis elegans* (Chen and Xu, 2005). The yeast essential proteins were also predicted by the k-nearest neighbor and SVM methods (Saha and Heber, 2006). Gustafson et al. built a naive Bayes classifier to predict the essential proteins in the yeast *S. cerevisiae* and *E. coli* (Gustafson et al., 2006). Network-based topological and sequence-based properties have also been used by other researchers to predict essential genes (Acencio and Lemke, 2009; Deng et al., 2011; Hwang et al., 2009; Lin and Zhang, 2011; Plaimas et al., 2010; Seringhaus et al., 2006).

In this study, the LC human PPI network was obtained from BIND (Bader et al., 2003), HPRD (Peri et al., 2003) and MINT (Ceol et al., 2010). A total of 28 topological properties, including five new common indices, were calculated for each node in the PPI network. In addition, 22 biological properties, such as GO scores, motif number, subcellular compartments, expression stage and phyletic age, were calculated for each gene. Significant differences were found between the topological properties and the biological properties of human essential genes and those properties of human non-essential genes. In addition, the F-score was also used to estimate the essentiality of each property (Chen and Lin, 2006). The functional uniqueness of the essential and non-essential genes was investigated by DAVID (Huang et al., 2008). Finally, based on the parameters of the 28 topological properties and 22 biological properties, a machine-learning approach was proposed to predict the human essential genes. Good performances were obtained using the jackknife test and 10-fold cross validation test. These performances indicate that our model could be a powerful tool for predicting essential genes. The workflow of our study is shown in Fig. 1.

## 2. Materials and methods

### 2.1. Dataset

Human PPI datasets were downloaded from the Online Predicted Human Interaction Database (OPHID) (version 1.95) (Brown and Jurisica, 2005). To obtain high-quality human PPIs, only the literature-curated human PPIs were used. The entire LC network comprises 12,265 nodes and 83,818 interactions. After removing the self-loops and duplicate edges, the final network comprises 12,265 nodes and 61,170 interactions. This network contains 228 connected components, and the main component comprises 11,952 nodes and 61,081 interactions. Because the topological properties are incalculable for proteins that do not belong to the main component, only the main component was considered in this study.

The human essential genes were downloaded from the OGEE database (build: 304) (Chen et al., 2012), the conditional essential and conditional non-essential genes were not used in this study. The protein product of an essential gene was regarded as an essential protein. There were 1528 human essential genes and 17,934 human non-essential genes in the OGEE database, and 1292 protein products of human essential genes and 7970 protein products of human non-essential proteins were mapped into the PPI network. The detailed information is shown in Table 1.

Random sampling was also performed to characterize the properties of the essential and non-essential genes. First, the dataset was split into two subsets: (1) essential gene set containing 1292 essential gene entries; (2) non-essential gene set containing 7970 non-essential gene entries. Second, 1292 entries were selected from the essential gene set and 1292 entries were randomly selected from the non-essential gene set.

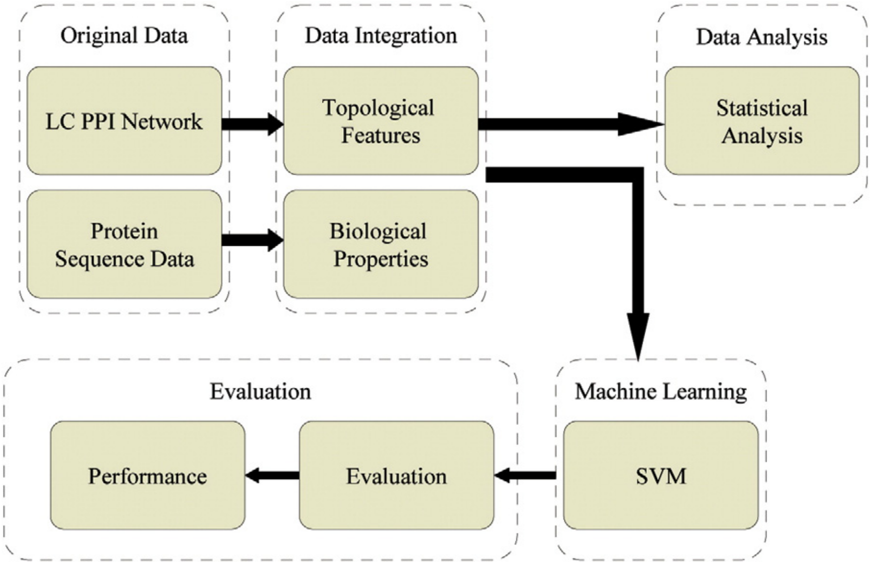

**Fig. 1.** Flow chart of our work.