Computational Biology and Chemistry 53 (2014) 191–197

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Research Article

Hierarchical closeness efficiently predicts disease genes in a directed signaling network

Tien-Dzung Tran [a,b], Yung-Keun Kwon [a,*]

[a] School of Computer Engineering and Information Technology, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan 680-749, South Korea
[b] Department of Information Technology, Center for Research and Development, Hanoi University of Industry, Tu Liem, Hanoi, Viet Nam

ARTICLE INFO

Article history:
Received 25 February 2014
Received in revised form 13 August 2014
Accepted 25 August 2014
Available online 19 September 2014

Keywords:
Hierarchical closeness
Disease gene prediction
Signaling network
Boolean network

ABSTRACT

Background: Many structural centrality measures were proposed to predict putative disease genes on biological networks. Closeness is one of the best-known structural centrality measures, and its effectiveness for disease gene prediction on undirected biological networks has been frequently reported. However, it is not clear whether closeness is effective for disease gene prediction on directed biological networks such as signaling networks.

Results: In this paper, we first show that closeness does not significantly outperform other well-known centrality measures such as Degree, Betweenness, and PageRank for disease gene prediction on a human signaling network. In addition, we observed that prediction accuracy by the closeness measure was worse than that by a reachability measure, but closeness could efficiently predict disease genes among a set of genes with the same reachability value. Based on this observation, we devised a novel structural measure, hierarchical closeness, by combining reachability and closeness such that all genes are first ranked by the degree of reachability and then the tied genes are further ranked by closeness. We discovered that hierarchical closeness outperforms other structural centrality measures in disease gene prediction. We also found that the set of highly ranked genes in terms of hierarchical closeness is clearly different from that of hub genes with high connectivity. More interestingly, these findings were consistently reproduced in a random Boolean network model. Finally, we found that genes with relatively high hierarchical closeness are significantly likely to encode proteins in the extracellular matrix and receptor proteins in a human signaling network, supporting the fact that half of all modern medicinal drugs target receptor-encoding genes.

Conclusion: Taken together, hierarchical closeness proposed in this study is a novel structural measure to efficiently predict putative disease genes in a directed signaling network.

1. Introduction

Genes and their regulatory interactions form a large-scale cellular interaction network, and a multitude of studies have examined the structural characteristics of these networks for insight into the association between genes and diseases (Wu et al., 2008; Zhao and Li, 2010, 2012). For example, it was suggested that disease genes are often centrally distributed as hub nodes (i.e., nodes with high connectivity) on the network. Indeed, genes related to neurodegenerative disease (Panda et al., 2012), breast cancer (Chand and Alam, 2012), and hereditary disease (Xu and Li, 2006) were shown to have higher regulatory interactions than non-disease genes. In contrast, other studies reported that disease genes tend to be non-hubs (Barabasi et al., 2011; Goh et al., 2007). These conflicting results emphasize the necessity of investigating various other structural centrality measures. Closeness (Sabidussi, 1966), a structural centrality measure in which a node is defined as the inverse of the total sum of the shortest distance to all the other nodes in an undirected network, has been frequently used to predict the disease risk of genes on undirected biological networks with satisfactory performance (Erten et al., 2011; Gottlieb et al., 2011; Hsu et al., 2011; Wu et al., 2008). The closeness definition can be also slightly modified to be properly used in a directed network (Opsahl et al., 2010). However, it may not be useful for disease gene prediction on a directed biological network because it does not fully employ direction-related information on the network. In particular, we note that the functional importance of a node can be

* Corresponding author at: Complex Systems Computing Laboratory, School of Electrical Engineering, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan 680-749, South Korea. Tel.: +82 52 259 2728; fax: +82 52 259 1687.
E-mail addresses: trantd.vn@gmail.com (T.-D. Tran), kwonyk@ulsan.ac.kr (Y.-K. Kwon).

http://dx.doi.org/10.1016/j.compbiolchem.2014.08.023
1476-9271/

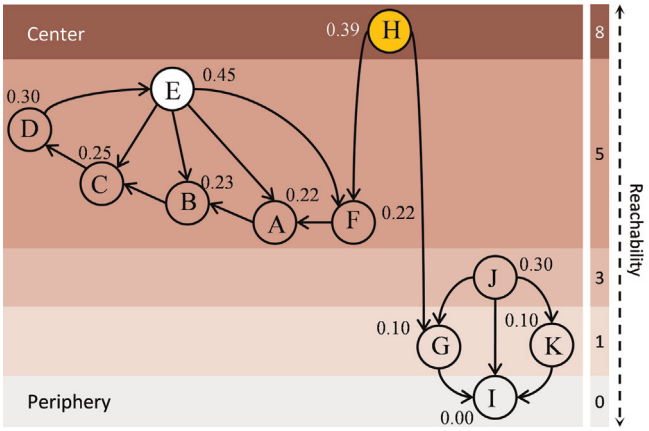

Fig. 1. An illustrative example to explain the concept of hierarchical closeness in a directed network. Reachability values represent the hierarchical level of nodes, ranging from 0 to 8. Closeness values denoted beside circle nodes range from 0 to 1. A subset of nodes with the same reachability is further ranked by the closeness value denoted beside a circle node. Six nodes, *A* through *F*, form a reachability-tied group and node *E* with the highest closeness is locally most central in that group. Node *H* is globally most central whereas *I* is most peripheral in terms of HC measure.

proportional to the reachability of the node, i.e., the subset of connected nodes from it, on a directed network. This concern led us to investigate the effectiveness of closeness on a directed biological network.

In this study, we first observed that reachability is better than closeness in predicting putative disease genes on a signaling network, particularly for top-ranked genes. In addition, it was observed that a gene with higher closeness is more likely to a disease gene within a set of tied genes with the same reachability. Inspired by these observations, we proposed a novel structural measure, hierarchical closeness (HC), by combining reachability and closeness in such a way that the reachability first ranks all genes and then the closeness plays a role as a tie-breaking measure. To demonstrate the effectiveness of HC, we compared HC and four other well-known structural centrality measures, including Degree, Closeness, Betweenness, and PageRank, with respect to disease gene prediction on a human signaling network and discovered that HC outperforms all the other measures, particularly for cancer, hereditary, immune, and neurodegenerative disease-related genes. Interestingly, we also found that the set of highly ranked genes in terms of HC is clearly different from the set of hub genes. It was also interesting that all of these findings are general properties conserved in random networks. Finally, we found that genes with high HC values are significantly likely to encode proteins in the extracellular matrix and receptor proteins in a human signaling network, explaining why half of all modern medicinal drugs target receptor-encoding genes.

2. Materials and methods

2.1. Datasets of disease genes and biological networks

In this work, we examine the topological distribution of genes in a human signaling network, which is a directed network, and a protein–protein interaction network, which is an undirected network. To this end, we selected 4350 disease genes extracted from OMIM database (Online Mendelian Inheritance in Man) in NCBI (Amberger et al., 2009, 2011) (see Table S1 in Supplementary Information) and mapped them into a human cellular signaling network composed of 1953 nodes and 8579 links obtained from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kim et al., 2011) and a human protein–protein interaction network (HPPI) composed of 7535 nodes and 22,052 interactions (Goh et al., 2007). In particular, the KEGG signaling network published in (Kim et al., 2011) was constructed by integrating all the pathways of *Homo sapiens* (human) which can be represented by a directed graph: for example, pathways about metabolism, environmental information processing, cellular process, human disease, and so on. All the same identifiers of different pathways were merged into one node and redundant or neutral links were removed. In addition, an interaction from a gene/protein G to a group of genes/proteins $\{G_1,G_2,\ldots,G_k\}$ in the original KEGG pathways was transformed into k different interactions $G\rightarrow G_1$, $G\rightarrow G_2,\ldots,$ and $G\rightarrow G_k$ in the signaling network.

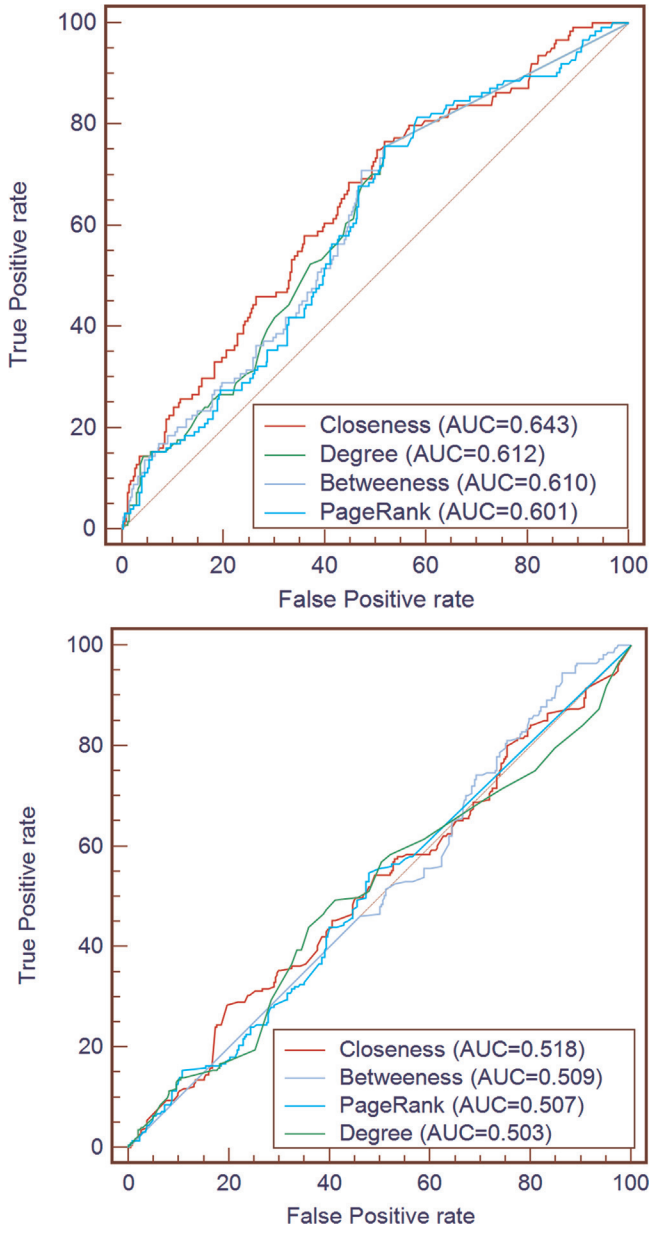

Fig. 2. Comparison of closeness and other centrality measures in terms of the prediction performance of disease genes on an undirected network (HPPI) and a directed network (KEGG). (A) Result on the HPPI network. The AUC value of closeness is significantly higher than that of Degree, Betweenness, and PageRank (all p-values $\leq$0.05). (B) Result of the KEGG network. The AUC value of closeness is not significantly higher than those of all the other centrality measures (all p-values >0.60).