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/

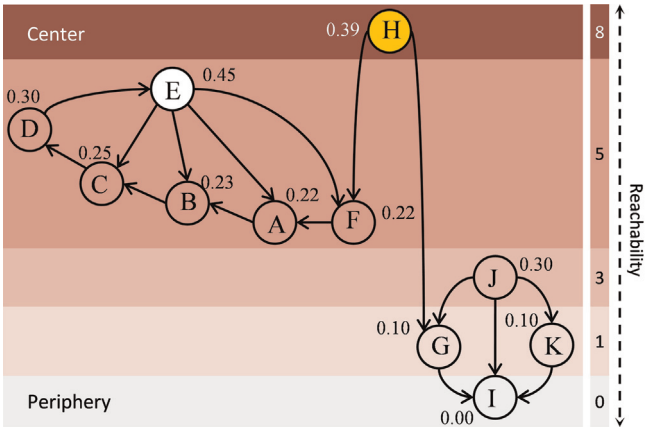

Fig. 1. An illustrative example to explain the concept of hierarchical closeness in a directed network. Reachability values represent the hierarchical level of nodes, ranging from 0 to 8. Closeness values denoted beside circle nodes range from 0 to 1. A subset of nodes with the same reachability is further 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.

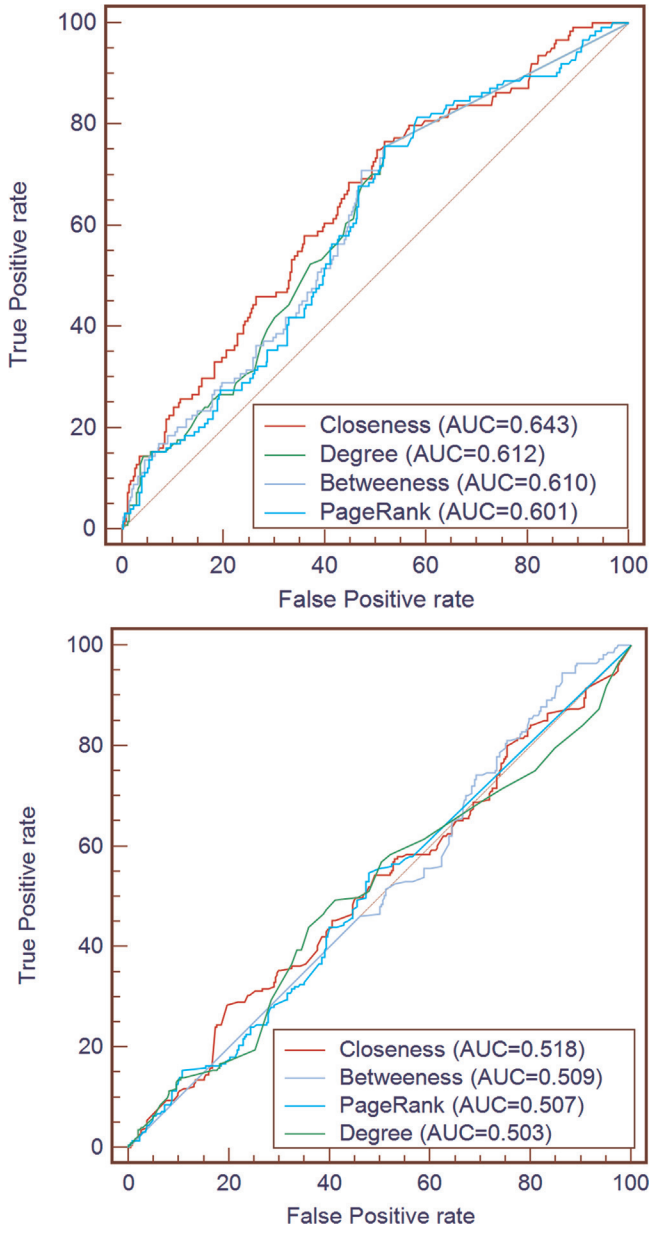

Fig. 2. Comparison of closeness and other centrality measures in terms of the prediction performance of disease genes on an undirected network (HPPI) and a directed network (KEGG). (A) Result on the HPPI network. The AUC value of closeness is significantly higher than that of Degree, Betweenness, and 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).

2.2. Structural centrality measures

Various studies have assessed the centrality of a node in a network, and here we briefly introduce four well-known structural centrality measures. It is assumed that a directed network $G(V, A)$ is given.

Degree: Degree has been applied in numerous previous studies to locate putative disease genes. The degree of a node $v \in V$ is defined as

$$C_{\text{deg}}(v) = |\{(v, w)|(v, w) \in A\}| + |\{(w, v)|(w, v) \in A\}|.$$

In other words, it denotes the number of in-coming or out-going interactions with respect to v.

Closeness: Closeness of a node v (Sabidussi, 1966) is defined as follows:

$$C_{\text{clo}}(v) = \frac{1}{\sum_{w \in V \setminus \{v\}} d(v, w)}$$

where $d(v, w)$ is the distance of the shortest path, if any, from v to w; otherwise, $d(v, w)$ is specified as an infinite value. This measure has been successfully used to prioritize disease candidate genes in a protein–protein interaction network (Gottlieb et al., 2011; Hsu et al., 2011). The definition of $C_{\text{clo}}(v)$ is not proper, though, in cases where there is a node j that is not reachable from v because $C_{\text{clo}}(v)$ eventually becomes zero. Thus, we used a variant definition of closeness (Opsahl et al., 2010) as follows:

$$C_{\text{clo}-v}(v) = \frac{1}{|V| - 1} \sum_{w \in V \setminus \{v\}} \frac{1}{d(v, w)}.$$

Betweenness: Betweenness of a node v (Freeman, 1977) is defined as

$$C_{\text{bet}}(v) = \sum_{s,t \in V \setminus \{v\}, s \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where σ_{st} is the total number of the shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of those paths that pass through v. The betweenness centrality was successfully used to investigate the relationship between structure and robustness in gene networks of glioma for renal cancer tissues (Sun et al., 2012). Proteins with high betweenness centrality in the pathway network were suggested as drug targets (Breitkreutz et al., 2012).

PageRank: PageRank (Page et al., 1999) is often used to predict disease genes (Chen et al., 2009; Winter et al., 2012). Assuming that there are n nodes, w_1, w_2, . . . , and w_n, which have an interaction going to v, PageRank of a node v is given as follows:

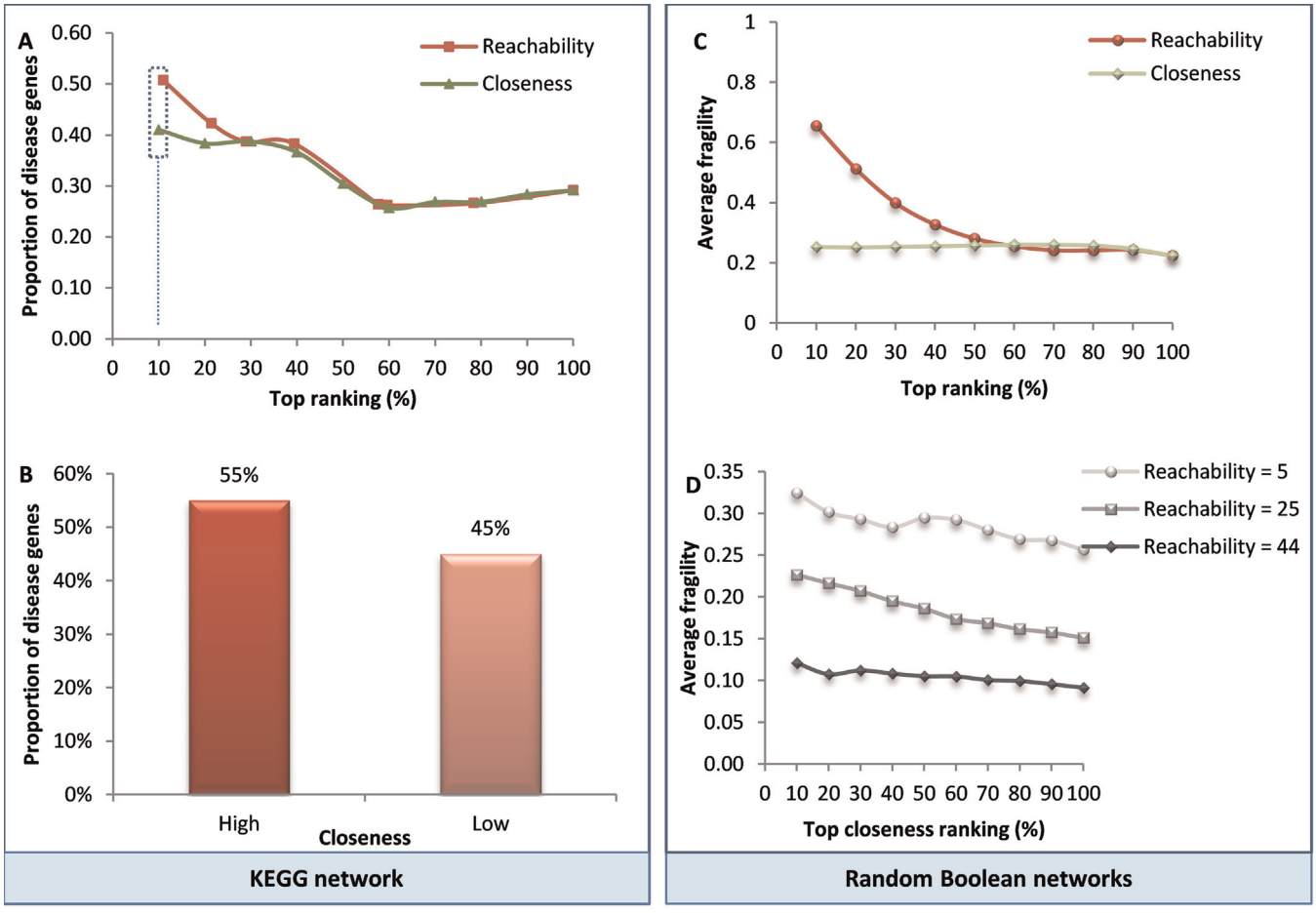

Fig. 3. Comparison between reachability and closeness in KEGG and random Boolean networks (A) change in the proportions of disease genes ranked by reachability and closeness in the KEGG network. The proportion of disease genes ranked by reachability is significantly higher than that ranked by closeness in approximately the top 10% ($p < 0.05$), whereas these proportions are similar to each other in larger top-rankings. (B) Comparison between high- and low-closeness groups with respect to the proportion of disease genes in the KEGG network. High- and low-closeness groups mean the sets of genes whose closeness is larger and lower, respectively, than the average closeness over the tied genes with the same reachability value. The proportion of disease genes of the high-closeness group is significantly higher than that of the low-closeness group ($p = 0.0185$). (C) Change in proportions of fragile nodes ranked by reachability and closeness in random Boolean networks. A total of 1000 random Boolean networks were generated with $|V| = 50$ and $49 \leq |A| \leq 100$. (D) Change in proportions of fragile nodes further ranked by closeness for three sets of tied nodes in random Boolean networks. Correlation coefficients for reachabilities equal to 5, 25 and 44 are 0.939 ($p = 0.005$), −1.0 ($p = 0.003$), and −0.964 ($p = 0.004$), respectively.

$$C_{PR}(v) = (1-d) + d\left(\frac{C_{PR}(w_1)}{C(w_1)} + \cdots + \frac{C_{PR}(w_n)}{C(w_n)}\right)$$

where d is a damping factor usually set to 0.85 and $C(w)$ is the number of interactions going out from w.

2.3. Hierarchical closeness

Although the original closeness measure partially denotes how centrally located a node is in a network, it does not explicitly include information about the range of other nodes that can be affected by the given node. In this regard, we propose hierarchical closeness of a node v, $C_{hc}(v)$, by combining reachability and closeness measures as follows:

$$C_{hc}(v) = N_R(v) + C_{\mathrm{clo}-v}(v)$$

where $N_R(v) \in [0, |V|-1]$ is the reachability of a node v defined by

$$N_R(v) = |\{w \in V | \exists \mathrm{apathfrom} v \mathrm{to} w.\}|.$$

In other words, $N_R(v)$ represents the number of nodes in V that can be reachable from v. It can also represent the hierarchical position of a node in a network (Jothi et al., 2009; Mones et al., 2012). We note that if $N_R(v)=0$, then $C_{hc}(v)=0$ because $C_{\mathrm{clo}-v}(v)$ is 0. In cases where $N_R(v)>0$, the reachability is a dominant factor because $N_R(v) \geq 1$ but $C_{\mathrm{clo}-v}(v)<1$. In other words, the first term indicates the level of the global hierarchy and the second term presents the level of the local centrality. Fig. 1 illustrates the hierarchical closeness notion. As shown in that example, all nodes are first ranked by the reachability value, and the set of nodes with the same reachability are further ranked by the closeness. Based on this definition, genes with the highest HC values eventually represent a set of geometrically central genes in a directed network that can reach to most other genes by relatively short paths. In this study, we hypothesize that the higher the hierarchical closeness of a node is, the more functionally important the node is on the directed network.

2.4. Boolean network dynamics

To evaluate network robustness we employed a Boolean network model, which has been intensively used to investigate the dynamics of various biological networks (Kauffman, 1969, 1993; Maki-Marttunen et al., 2013; Samaga and Klamt, 2013; Stern, 1999). A network is represented by a simple directed graph $G(V, A)$ where $V=\{v_1, v_2, \ldots, v_n\}$ is a set of Boolean variables and $A=\{(u, w)|u,\}$ is a set of directed interactions. Then, a network state $s(t)=(v_1(t), v_2(t), \ldots, v_n(t))$ at time t transits to the next state $s(t+1)$ according to a set of update rules $F=\{f_1, f_2, \ldots, f_n\}$; i.e., $s(t+1)=F(s(t))$ where a logical conjunction or disjunction is randomly selected for f_i with a uniform probability distribution. For example, if a Boolean variable v has activation relationships with v_1 and v_2, and an inhibition relationship with v_3, then the conjunction and disjunction update rules are $v(t+1) = v_1(t) \wedge v_2(t) \wedge \overline{v}_3(t)$ and $v(t+1) = v_1(t) \vee v_2(t) \vee \overline{v}_3(t)$, respectively. In the case of the conjunction, the value of v at time $t+1$ is 1 only if the values of v_1, v_2, and v_3 at time t are 1, 1, and 0, respectively. With these update rules, the generated Boolean network will operate in the ordered regime. The network eventually converges to a fixed state, or a limit-cycle attractor. We denote the converged attractor starting from state $s(t)$ as $\langle s(t)\rangle$. The network is termed robust against the mutation at v if $\langle s\rangle$ is equal to $\langle s_{\overline{v}}\rangle$ where $\overline{v}(=\neg v)$ indicates the state perturbation of s subject to v. This concept to measure robustness has been widely used (Ciliberti et al., 2007; Kitano, 2004; Kwon and Cho, 2008). Similarly, we employed the fragility of a node v, $\gamma(v)$, to represent the degree to which a node is not robust against the mutation subject to v as follows:

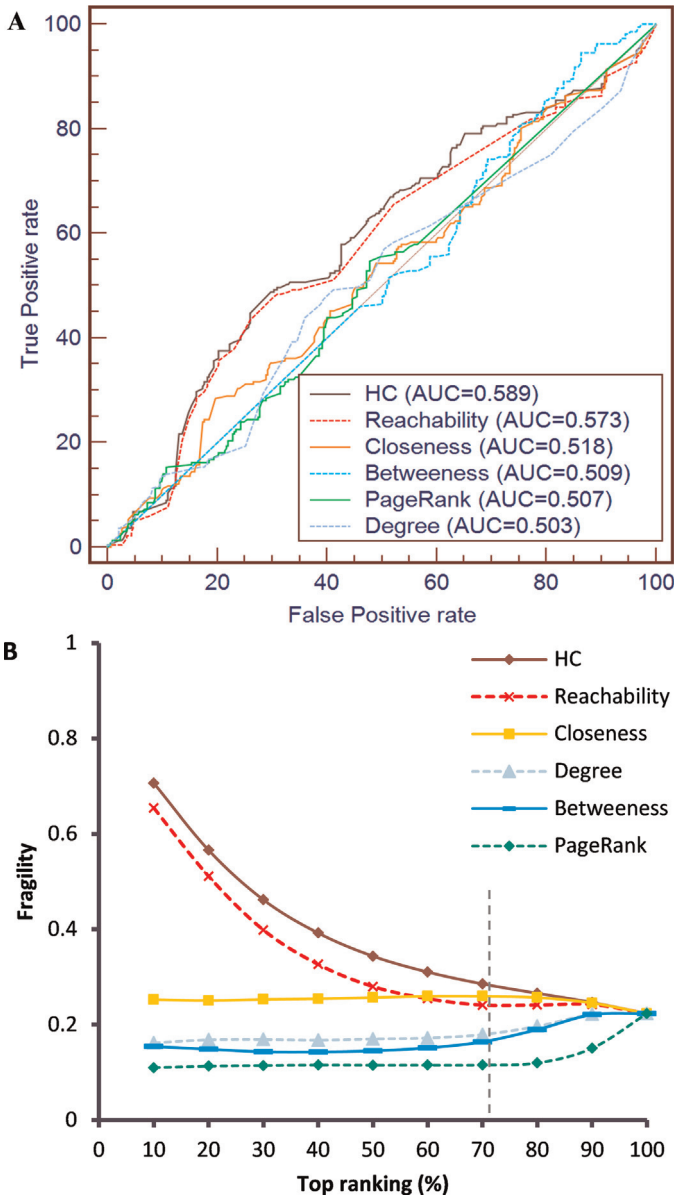

Fig. 4. Performance of HC for disease gene or fragile node prediction on KEGG and random Boolean networks. (A) Result of ACU values on the KEEG network. HC shows significantly better performance than other centrality measures (all p-values <0.05). (B) Change of proportions of fragile nodes ranked by HC in random Boolean networks. A total of 1000 random Boolean networks were generated with $|V|=50$ and $49 \leq |A| \leq 100$. HC shows significantly better performance than other centrality measures in the range of $K<70\%$ (p-value <0.0001).

$$\gamma(v) = \frac{1}{|S|}\sum_{s \in S} I(\langle s\rangle \neq \langle s_{\overline{v}}\rangle)$$

where S is a set of whole network states (here, $|S|=2^n$), and $I(\cdot)$ is an indicator function. A node is called a fragile node if the fragility is larger than zero. In this study, the dynamics of a human signaling network are compared with those of random networks. In this regard, we employed the Barabasi–Albert network-growth model[1] (Barabasi and Albert, 1999) to generate random directed networks.

[1] In this study, the Barabasi–Albert network-growth model was used to generate random networks with the scale-free property inducing a few hub nodes and many non-hub nodes, as observed in real signaling networks.

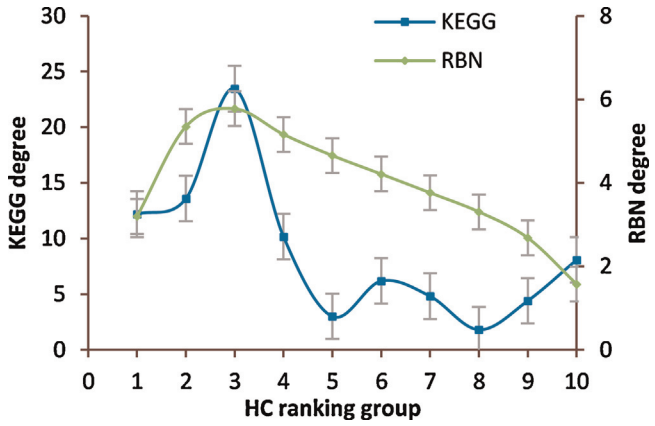

Fig. 5. Degree distribution of groups of nodes classified by HC on the KEGG and 1000 random Boolean networks (RBNs). All genes in a network are grouped into 10 subgroups of the same size. Points and error-bars of the *y*-axis represent the average and standard error values, respectively.

3. Results and discussion

3.1. Limitation of closeness in a directed signaling network

Closeness is known to be an effective centrality measure for prioritizing the disease candidate genes in protein–protein interaction networks, which are undirected networks (Gottlieb et al., 2011; Hsu et al., 2011). To verify this, we compared the predictive performance of closeness and other centrality measures including Degree, Betweenness, and PageRank, on the HPPI network (Fig. 2A and Table S2 for details; see Section 2 for dataset). For a reliable performance comparison, we drew the receiver operating characteristic and computed the area under the curve (AUC). As shown in Fig. 2A, closeness significantly outperforms Degree ($p = 0.042$), Betweenness ($p = 0.048$), and PageRank ($p = 0.026$). However, the usefulness of closeness has not been sufficiently proven on a directed network, therefore we examined the performance of those four centrality measures on the KEGG signaling network (Fig. 2B and Table S3 for details; see Section 2 for dataset). As depicted in Fig. 2B, closeness shows the best performance but does not significantly outperform Betweenness, Degree, or PageRank (all *p*-values >0.60). We note that closeness does not fully reflect the reachability property of a node in a directed network as shown in Fig. 1. Thus, we hypothesize that the hierarchical closeness can be an effective alternative to improve the performance of closeness in a directed network.

Table 1
Comparison between HC-high and HC-low gene groups with respect to GO term frequency. All of the genes in the KEGG network were classified according to HC value into HC-high and HC-low groups consisting of 586 and 1367 genes, respectively. The values in HC-high and HC-low columns indicate the percentage of genes involved in the corresponding term among each group of genes.

	Percentage of genes having GO term		
GO term	HC-high (%)	HC-low (%)	*p*-value
Postsynaptic membrane	3.07	0.73	2.07E-04
Membrane raft	8.36	2.34	7.28E-09
Integrin complex/laminin receptor protein	4.61	0	2.55E-17
Basement membrane	4.78	0.88	1.58E-07
Collagen	2.90	0.29	1.79E-06

3.2. Motivation and prediction performance of hierarchical closeness

As explained in Section 2.3, we expect that reachability is more useful than closeness overall in predicting disease genes and that closeness can be a tie-break measure for genes with the same reachability value. To explain the motivation for such a hypothesis, we investigated the prediction performance of reachability and closeness on the KEGG network. We computed the proportion of disease genes among the top-*K* genes ranked by the reachability or the closeness over the whole number of genes with *K* varying from 0% to 100% (Fig. 3A). As shown in Fig. 3A, the proportions of disease genes for both reachability and closeness show negative correlations against *K* (≤60%). This means that genes with a higher reachability or closeness ($K \leq 60\%$) tend to be disease genes. In addition, the reachability was significantly better than the closeness when *K* was approximately 10% ($p < 0.05$), whereas their accuracies were similar for larger top-ranking values. The reachability measure is not sufficient, however, because there can be many tied nodes with respect to the reachability value, as expected by the definition (see Fig. S1 for the frequency of reachability values in the KEGG network). Therefore, we considered the closeness as a tie-break measure. To validate this, we further compared the proportion of disease genes between high- and low-closeness groups defined by sets of genes whose closeness is larger and lower, respectively, than the average closeness over the tied genes with the same reachability (Fig. 3B). As shown in Fig. 3B, the proportion of disease genes of the high-closeness group is significantly higher than that of the low-closeness group ($p = 0.0185$). This implies that a gene with higher closeness is more likely to be a disease gene in each tied group, and thus closeness can be an efficient tie-break measure. To investigate whether this finding is an intrinsic principle in random networks, we generated 1000 random Boolean networks and examined the proportion of fragile nodes (see Section 2 for the definition) among the top-*K* nodes ranked by the reachability or the closeness with *K* varying from 0% to 100% (Fig. 3C). Similar to Fig. 3A, reachability outperformed closeness for some ranges of *K* (≤40%). We also examined the proportion of the fragile nodes, which are further ranked by the closeness for tied nodes (Fig. 3D). More specifically, three reachability values, 5, 25, and 44, with relatively high frequencies were chosen. In this Fig. 3D, all lines show a significantly negative relation against top closeness ranking (%) (all *p*-values <0.01). Taken together, the consistent results in random Boolean networks support the motivation to combine the reachability and the closeness to prioritize dynamically important nodes in a directed network.

As a result, the hierarchical closeness is proposed as a new centrality measure that combines reachability and closeness. To verify the efficiency of HC for disease gene prediction on a directed network, we evaluated the AUC value by HC prediction on the KEGG network (Fig. 4A). As shown, HC outperforms other centrality measures including Degree, Reachability, Closeness, Betweenness, and PageRank (all *p*-values <0.05). In addition, we examined the performance of HC for fragile node prediction on random Boolean networks (Fig. 4B). After generating 1000 random Boolean networks, we examined the proportion of fragile nodes among the top-*K* nodes ranked by HC with *K* varying from 0% to 100%. As shown in Fig. 4B, HC shows the best performance compared with the other centrality measures in ranges of $K < 70\%$ (all *p*-values <0.0001). This implies that HC can be a generally useful measure to predict dynamically important nodes in a directed network.

Furthermore, we investigated the prediction performance of HC on four subgroups of specific disease genes such as cancer, hereditary, immune, and neurodegenerative disease subgroups (Fig. S2). We found that HC outperforms the other centrality measures significantly for the top 10% or 20% ranked genes (Fig. S2A–D). On the other hand, HC was not efficient to predict all the other types of disease genes (Fig. S2E). This implies that HC cannot predict all types of disease genes.

3.3. Comparison between HC and hub centers

Many previous studies have shown that hub nodes are typically associated with disease genes (Panda et al., 2012). On the other hand, some studies indicate that disease genes tend to be non-hubs (Barabasi et al., 2011; Goh et al., 2007). In this regard, it is meaningful to compare the HC and hub centers. Therefore we investigated the degree distribution of node groups ranked by HC on the KEGG network (Fig. 5). More specifically, we classified all the genes into 10 subgroups of the same size according to HC ranking and examined the average degree and its standard deviation for each group. The lower the group number, the more central the node is with respect to the HC measure. It is interesting that the average degree of the third group, rather than the first group, is highest. This implies that the set of HC-center genes are different from the hub genes (with the largest interactions). We additionally examined the degree distribution on the random Boolean networks and observed a similar result. Taken together, the HC-centered but non-hub nodes can be considered dynamically important nodes on a directed network.

3.4. The biological functions of the HC-centered genes

To investigate the biological characteristics of the HC-center genes in the KEGG network, we classified all the genes into two subgroups by HC values – 'HC-high' (586 genes) and 'HC-low' (1367 genes) – and examined the percentage of genes involved in each gene ontology (GO) term with respect to biological process, cellular component, and molecular function using FatiGO software (Al-Shahrour et al., 2004) (Table 1). The result shows that the HC-high genes are related to the extracellular matrix and cell surface receptor terms (membrane receptors, transmembrane receptors) (Table 1). Cell surface receptors are specialized integral membrane proteins that participate in communication between the cell and the outside world. Alteration or deficiency of genes encoding membrane receptors can disrupt signal transduction and ultimately cause diseases such as cancer (Muller-Pillasch et al., 1998) and Alzheimer's (Scheuer et al., 1996). Proteins in the extracellular matrix are also considered as drug targets in pharmacotherapy (Huxley-Jones et al., 2008; Jarvelainen et al., 2009; Schaefer, 2010). Our finding suggests that cell surface receptor proteins are often associated with disease due to the topological characteristics of the genes encoding them, i.e., due to the central position of these genes in the signaling network. Although the HC-central genes are in optimal positions to reach most other genes, they are also the most fragile with respect to mutations.

Based on this observation, we also note that the membrane proteins can be an efficient indicator to identify disease genes because they act as signal sensors (Sanders and Myers, 2004; Sanders and Nagy, 2000). Through a further investigation, we found that 506 membrane genes (Group A) were included in the KEGG network by using Mouse Genome Informatics database (www.informatics.jax.org), and 207 genes among them were disease genes (i.e., the ratio is 0.409 (=207/506)) whereas there were 206 disease genes (i.e., the ratio is 0.407 (=206/506)) among the same number of genes having highest HC values (Group B). In addition, the number of genes in the intersection of Group A and B was 161 and there were 70 disease genes among them (i.e., the ratio is 0.435 (=70/161)). Taken together, membrane proteins are considerably efficient in disease gene prediction although all of them do not show the highest HC values.

4. Conclusions

Closeness is one of the best-known structural centrality measures, and its effectiveness for disease gene prediction has frequently been reported for undirected biological networks. In this study, we investigated whether closeness is equally effective on a directed network. We first showed 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, which is a directed network, compared with an undirected network. In addition, we observed that the prediction accuracy by closeness measure was worse than that by reachability, but closeness could efficiently predict disease genes among a set of genes with the same reachability degree. Therefore, we proposed a novel measure, hierarchical closeness, in which we combine 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 found that hierarchical closeness outperforms other structural centrality measures in disease gene prediction, particularly for cancer, hereditary, immune, and neurodegenerative disease-related genes. Moreover, the set of highly ranked genes in terms of hierarchical closeness is clearly different from the set of hub genes with relatively high connectivity. It is also interesting that these findings are consistently observed in random Boolean networks. 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, consistent with the fact that half of all modern medicinal drugs target receptor-encoding genes. These results suggest that HC-center genes should be seriously considered as putative dynamically important genes in a directed biological network.

Conflict of interest

None declared.

Acknowledgements

This work was supported by 2014 Research Funds of Hyundai Heavy Industries for University of Ulsan.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.compbiolchem.2014.08.023.

References

Al-Shahrour, F., Diaz-Uriarte, R., Dopazo, J., 2004. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics (Oxford, England) 20, 578–580.
Amberger, J., Bocchini, C., Hamosh, A., 2011. A new face and new challenges for online Mendelian inheritance in man (OMIM(R)). Hum. Mutat. 32, 564–567.
Amberger, J., et al., 2009. McKusick's online Mendelian inheritance in man (OMIM). Nucleic Acids Res. 37, D793–D796.
Barabasi, A.L., Albert, R., 1999. Emergence of scaling in random networks. Science 286, 509–512.
Barabasi, A.L., Gulbahce, N., Loscalzo, J., 2011. Network medicine: a network-based approach to human disease. Nat. Rev. Genet. 12, 56–68.
Breitkreutz, D., et al., 2012. Molecular signaling network complexity is correlated with cancer patient survivability. Proc. Natl. Acad. Sci. U. S. A. 109, 9209–9212.

Chand, Y., Alam, M.A., 2012. Network biology approach for identifying key regulatory genes by expression based study of breast cancer. Bioinformation 8, 1132–1138.
Chen, J., Aronow, B.J., Jegga, A.G., 2009. Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics 10, 73.
Ciliberti, S., Martin, O.C., Wagner, A., 2007. Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comput. Biol. 3, e15.
Erten, S., et al., 2011. DADA: degree-aware algorithms for network-based disease gene prioritization. BioData Min. 4, 19.
Freeman, L.C., 1977. A set of measures of centrality based on betweenness. Sociometry 40, 35–41.
Goh, K.I., et al., 2007. The human disease network. Proc. Natl. Acad. Sci. U. S. A. 104, 8685–8690.
Gottlieb, A., et al., 2011. PRINCIPLE: a tool for associating genes with diseases via network propagation. Bioinformatics (Oxford, England) 27, 3325–3326.
Hsu, C.L., et al., 2011. Prioritizing disease candidate genes by a gene interconnectedness-based approach. BMC Genomics 12 (Suppl. 3), S25.
Huxley-Jones, J., Foord, S.M., Barnes, M.R., 2008. Drug discovery in the extracellular matrix. Drug Discov. Today 13, 685–694.
Jarvelainen, H., et al., 2009. Extracellular matrix molecules: potential targets in pharmacotherapy. Pharmacol. Rev. 61, 198–223.
Jothi, R., et al., 2009. Genomic analysis reveals a tight link between transcription factor dynamics and regulatory network architecture. Mol. Syst. Biol. 5.
Kauffman, S.A., 1969. Metabolic stability and epigenesis in randomly constructed genetic nets. J. Theor. Biol. 22, 437–467.
Kauffman, S.A., 1993. The Origins of Order: Self-Organization and Selection in Evolution. Oxford University Press, New York.
Kim, J.R., et al., 2011. Reduction of complex signaling networks to a representative kernel. Science Signal. 4, ra35.
Kitano, H., 2004. Biological robustness. Nat. Rev. Genet. 5, 826–837.
Kwon, Y.K., Cho, K.H., 2008. Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. Bioinformatics (Oxford, England) 24, 987–994.
Maki-Marttunen, T., Kesseli, J., Nykter, M., 2013. Balance between noise and information flow maximizes set complexity of network dynamics. PLoS One 8, e56523.
Mones, E., Vicsek, L., Vicsek, T., 2012. Hierarchy measure for complex networks. PLoS One 7, e33799.
Muller-Pillasch, F., et al., 1998. Identification of a new tumour-associated antigen TM4SF5 and its expression in human cancer. Gene 208, 25–30.
Opsahl, T., Agneessens, F., Skvoretz, J., 2010. Node centrality in weighted networks: generalizing degree and shortest paths. Soc. Netw. 32, 245–251.
Page, L., et al., 1999. The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab.
Panda, A., Begum, T., Ghosh, T.C., 2012. Insights into the evolutionary features of human neurodegenerative diseases. PloS One 7, e48336.
Sabidussi, G., 1966. The centrality index of a graph. Psychometrika 581–603 %G English.
Samaga, R., Klamt, S., 2013. Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. Cell Commun. Signal. 11, 43.
Sanders, C.R., Myers, J.K., 2004. Disease-related misassembly of membrane proteins. Ann. Rev. Biophys. Biomol. Struct. 33, 25–51.
Sanders, C.R., Nagy, J.K., 2000. Misfolding of membrane proteins in health and disease: the lady or the tiger? Curr. Opin. Struct. Biol. 10, 438–442.
Schaefer, L., 2010. Extracellular matrix molecules: endogenous danger signals as new drug targets in kidney diseases. Curr. Opin. Pharmacol. 10, 185–190.
Scheuer, K., et al., 1996. Cortical NMDA receptor properties and membrane fluidity are altered in Alzheimer's disease. Dementia 7, 210–214.
Stern, M.D., 1999. Emergence of homeostasis and noise imprinting in an evolution model. Proc. Natl. Acad. Sci. U. S. A. 96, 10746–10751.
Sun, L., et al., 2012. Analysis of cascading failure in gene networks. Front. Genet. 3, 292.
Winter, C., et al., 2012. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. PLoS Comput. Biol. 8, e1002511.
Wu, X., et al., 2008. Network-based global inference of human disease genes. Mol. Syst. Biol. 4, 189.
Xu, J., Li, Y., 2006. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics (Oxford, England) 22, 2800–2805.
Zhao, S., Li, S., 2010. Network-based relating pharmacological and genomic spaces for drug target identification. PLoS One 5, e11764.
Zhao, S., Li, S., 2012. A co-module approach for elucidating drug-disease associations and revealing their molecular basis. Bioinformatics 28, 955–961.