ASEC 2022

MDPI

Type of the Paper (Proceedings, Abstract, Extended Abstract, Editorial, etc.)

Gene re-ranking and controllability analysis of protein – protein network for discovery potential drug target of breast cancer at different stages

Ha Nguyen [a], Nhung Nguyen [a], Quang Le [a], Nghia Tran [a] and Uyen Bui [b]

[a] General and Inorganic Department, Hanoi University of Pharmacy, 13 – 15 Le Thanh Tong Street, Hoan Kiem District, Hanoi, Vietnam (hantn@hup.edu.vn, nhungnp@hup.edu.vn, quangld@hup.edu.vn, nghiatd@hup.edu.vn)
[b] Hanoi Amsterdam High School for the Gifted (buituuyen.ams@gmail.com)

Abstract: The protein-protein interaction network (PPIN) is essential for functional processing and mechanism of multiple complex diseases. Recently, control theory has applied to protein interaction with the aims of identify the minimum set of nodes that can drive the whole network to the desired state. Here, we use different statistic network inference methods to generate the highest-scored re-ranking gene list as the source for constructing protein-protein interaction network. Then we characterize structural controllability of directed and weighted PPINs for breast cancer stages. The maximum matching (MM) approach for controllability analysis allows classifying nodes into three categories: critical, intermittent and redundant. This leads to identifying the most important proteins as critical nodes for each stage of breast cancer. In total, 70 critical nodes as drug targets have been-revealed across stages in this study.

Keywords: control theory; breast cancer; drug targets; maximum matching; protein-protein interaction network; critical; intermittent and redundant nodes

Citation: Nguyen a, H.; Nguyen a, N.; Le a, Q.; Tran a, N.; Bui b, U. Gene re-ranking and controllability analysis of protein – protein network for discovery potential drug target of breast cancer at different stages. **2022**, *4*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s):

Received: date
Accepted: date
Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Breast cancer is the most common type of cancer and it remains the primary cause of death in the women population. 1.67 million new breast cancer cases were diagnosed in 2012 that accounts for 25% cancer cases in women (Akram et al. 2017). Breast cancer is a complicated disease because of its heterogeneity with the contribution of many risk factors such as sex, ageing, estrogen, family history, gene mutations and unhealthy lifestyle (Y. S. Sun et al. 2017).

Among the many therapeutic options, the use of small molecules that modulates cancer-related protein/gene targets remains as one of the leading anticancer approaches (Hoelder, Clarke, and Workman 2012). Specifically, the telomerase enzyme has become an attractive target for new and more effective anticancer agents (Mengual Gomez et al. 2016). This is a reverse transcriptase responsible for the addition of a repetitive DNA sequence to the ends of linear eukaryotic chromosomes, and so, involved in the phenomenon of cellular immortalization (Feng et al. n.d.; Riou et al. 2002). This enzyme is active in more than 85% of tumor cells, but is not active in the majority of normal cells, with some

notable exceptions such as stem cells and germ line cells. Telomerase is therefore a promising target for the treatment of malignancies (Sprouse, Steding, and Herbert 2012).

Assessing breast cancer mechanisms by means of cross-disciplinary bioinformatics would become rapidly developing. Hence, to identify breast cancer-related genes and significant genes for prognosis and treatment, many bioinformatics pipelines for data analysis have been investigated. However, ineffectiveness of current drug therapy illustrates the compensation of cancer cells to drug inhibition by multiple pathways. For this, cancer genes need to be represented as a part of a network where each interaction could affect the disease (Emmert-Streib et al. 2014).

On the other hand, the extremely large size of cancer networks makes it nearly impossible to access all the interactomes. For example, a cancer human network representing the cancer genes and their connection to 1st neighbor contains 3068 nodes and 6572 edges (Ibrahim et al. 2011). Thus, the reduction of network size by prioritizing genes is necessary. Differentially expressed genes (DEG) analysis is one of the most common applications to prioritizes genes using RNA-sequencing (RNA-seq) data. The concept of DEG is to identify the genes with expression levels determined to be significantly differentially expressed across two experimental conditions (e.g. cancer versus normal). The significance of difference needs to be evaluated by a statistical test and mostly ANOVA, t-test. Nevertheless, data of microarray experiments is not always satisfactory to statistical test. The sample size of normal cases is frequently much less than tumor cases leading to violation of assumptions of statistical tests. Therefore, Bourdakou et al have proposed a protocol for revealing essential genes by multiple network inference strategies coupled with PageRank algorithm (Bourdakou, Athanasiadis, and Spyrou 2016). PageRank reconciles co-expression network which is interfered by a variety of mathematical algorithms and biological methods to re-rank all genes (Poirel et al. 2013). As a result, the top genes obtained from the re-ranked list were used as seed proteins to build up the PPIN. PPIN describes the physical interactions between proteins in genome-wide scale. It provides a directed and weighted network which helps us understand human signaling circuitry. Recently, control theory has emerged as a mathematical framework for understanding how best to control a dynamical system (Uhart, Flores, and Bustos 2016). A system considers controllable when the minimum number of input, termed as "driver" nodes, can drive the system from any initial state to any desired final state in a finite time (Y. Y. Liu, Slotine, and Barabási 2011). Furthermore, each node in network has been classified into three categories: (i) critical if in its absence more driver nodes are controlled to drive the network; (ii) intermittent if in its absence there is no change in driver nodes; (iii) redundant if in its absence fewer driver node can offer full control the network (Y. Y. Liu, Slotine, and Barabási 2011). In this study, we exploited microarray data from Gene Expression Omnibus (GEO) database (Edgar, Domrachev, and Lash 2002) and The Cancer Genome Atlas (TCGA) database (http://gdac.broadinstitute.org/) for co-expression network inference of telomere related genes per each stage of breast cancer, prioritizing significant genes by Page rank algorithm. To elucidate the impact of top 100 genes from each gene list over sample discrimination, we carried out a validation scheme using the data from GEO as test sets and TCGA data as train sets. Afterwards, the gene list from this method was employed to construct protein-protein interaction networks. Here we explore the role of each individual node by classifying each node into three categories: critical, intermittent and redundant in the established model by maximum matching method. Finally, we obtained a list of potential drug targets for each stage of breast cancer.

2. Methods

With the input of telomere genes list, different algorithms and methods were applied to finally reconstruct an either mathematical or real gene interaction network. In total, there are 42 network inference algorithms. All genes in the network were re-ranked by PageRank algorithm to select top 100 genes of each method. Finally, the validation scheme to classify normal and cancer sample using the data from GEO as test sets and TCGA data

as train sets applied to score the quality of each prioritized gene list. Genes of the highest score ranked list were used as seed proteins to build up the PPIN of each stage.

3. Results and Discussion

Results

Evaluation of gene re-ranking

We obtained 43 ranked gene lists from re-ranking methods along with traditional DEG ranking. To validate each approach, we applied a holdout validation scheme and calculated the mean classification accuracy for each method per stage. The median accuracy values of all methods are greater than 60% in stage I, 55% in stage II, 50% in stage III and 80% in stage IV (Fig 1). Aracnea.bsN3, clr.bsN3 and WGCNA have the least standard deviations throughout four stages.

Based on the maximum achieved mean classification accuracy across datasets, we calculated a score for each method. Table V indicates the score of each re-ranking method for all stages. It is observed that mean scores of ranking methods varied mostly from 0.6 to 0.9. In addition, these scores across four stages are comparable, low standard deviations are shown in most methods except aracnem.bsN3 algorithm.

The highest average score for breast cancer stage was achieved by Adlasso network inference algorithm with an average score of 0.9025 in which 0.81, 0.9, 0.9 and 1 are respectively scores of stage I, II, III and IV. After Adlasso, Initial method is the second-highest average score with the score 0.9 for all stages. The lowest average score is of aracnem.bsN3 method. Therefore, we chose the rank list from Adlasso network inference method to construct the protein-protein interaction network.

The mean accuracy rates of Adlasso throughout the top 100 genes. The mean accuracy of Adlasso remains mostly stable for about the first 14 genes at the highest score value, after that, the accuracy varies from 0.6 to 0.9 except stage IV. The accuracy of stage IV mostly fluctuated around 1. It reconfirmed that the Adlasso method is suitable for re-ranking genes.

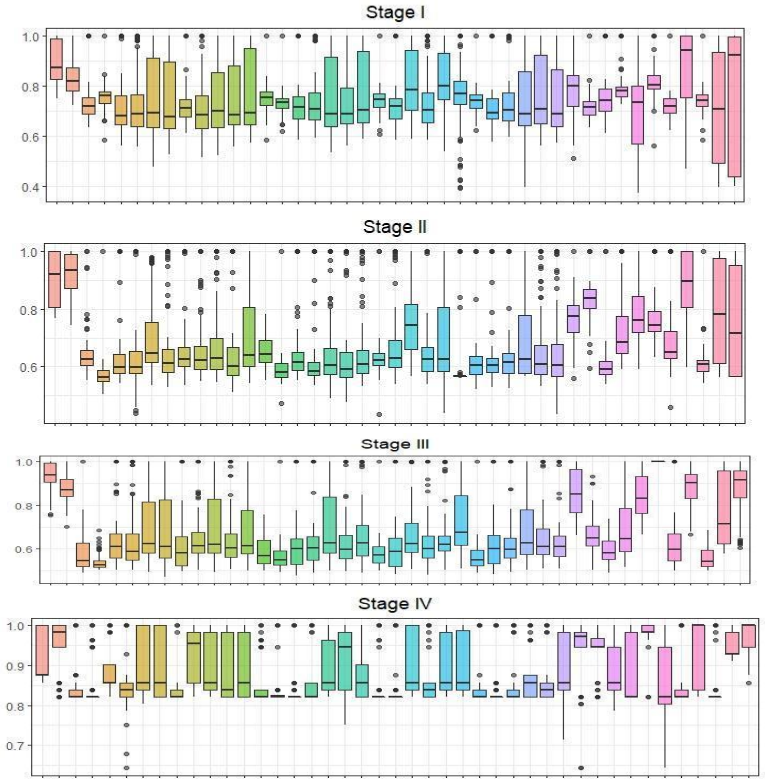

Figure 1. Boxplots of mean accuracy rates of the top 100 sequential genes of all re-ranked gene lists for breast cancer stages. The accuracy value is ranged from 0 to 1, shown on y axis. The x axis indicates different methods corresponding with different colors. Dots represent outlying value in each method.

Controllability analysis of weighted and directed biological network

First, we constructed four PPINs consisting of 76 nodes and 177 edges in stage I, 81 nodes and 181 edges in stage II, 78 nodes and 196 edges in stage III, 71 nodes and 183 edges in stage IV. We applied algorithms: MM to classify the nodes into three types of nodes (critical, intermittent, redundant). The PPIN network of stage I illustrates in Figure 2. In general, most of the nodes identified was redundant node and 70 critical nodes were classified in total. The number of critical nodes of each stage, respectively are 16, 20,12 and 21. Interestingly, it is observed that the highest-degree nodes of stage I, stage II, stage III and stage IV with 18, 21, 20, 19 degrees respectively are redundant nodes. Furthermore, 40% in total critical nodes have degrees of 1 and only 6% have the degree that is greater than 10.

Overall, our aim was to identify which nodes are of importance in controlling the network. Clearly critical nodes are the most important in controlling the network and become potential drug targets. The list of all critical nodes of breast cancer stage achieved by MM method is shown in Table I.

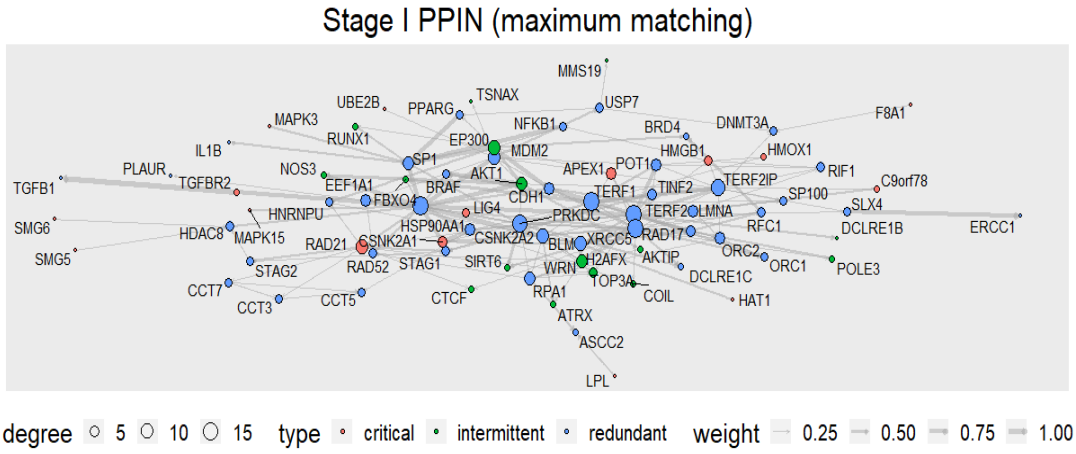

Figure 2. Protein-protein interaction network of stage I breast cancer by MM. The degree of a node which is the number of connections that it has to other nodes in the network is illustrated by the different size of the circle. The critical, intermittent and redundant nodes are displayed in red, green and blue, respectively. The different widths of the arrow represent different weights.

Table I: Critical nodes breast cancer stages

Stage 1	Stage 2	Stage 3	Stage 4
CSNK2A1	EXOSC10	C9orf78	BICD1
TGFBR2	ATRX	TFIP11	INHBA
SMG5	PTGES3	MAP2K6	BARD1
C9orf78	C9orf78	GADD45B	CIB1
HMGB1	POLD4	TSN	EP300
LIG4	HMGB1	APEX1	CCT4

UBE2B	MAP2K6	CTSB	RPA1
HAT1	CKAP4	LPL	LIG4
APEX1	RPA2	PAPSS1	CKAP4
HMOX1	GADD45B	TNKS	TELO2
LPL	UBE2B	RAD9A	UPF1
RAD21	XRN1	PPP1R10	UBE2B
MAPK3	CTSB	MAPK3	CTSB
SMG6	LPL		SF3B1
F8A1	RAD21		RAD21
MAPK15	XRCC1		RAP1A
	TNKS		LEP
	MAPK3		ALDH1B1
	BRCA2		DCLRE1B
	SMG1		PTGS2
			F8A1

Discussion

In this work, we used 43 network inference methods to reconstruct the co-expression network and prioritize genes to identify the significant gene in different breast cancer stages (I-IV). The result demonstrated that the gene list of Adaptive lasso, a correlation-based method, achieved the highest score. However, the simple statistical method (Initial) also got a high score, after Adaptive Lasso. Actually, it has been observed that the simple statistical gave slightly better score compared to other methods in the case of breast cancer stage genes. However, in case of molecular subtypes, it showed that re-ranking methods improved both score and accuracy (Bourdakou, Athanasiadis, and Spyrou 2016). Raeisi Shahraki et al. (2016) applied Adaptive Lasso to identify the most efficient genes from the data of 25 patients with bladder cancers, too (Shahraki et al. 2016).

The controllability analysis to assess critical nodes in complex biology network has become a trend in computational biology. Recently, MM has been proposed as a method to identify driver nodes in a directed network (Vinayagam et al. 2016). Liu and Pan studied controllability of human signaling network, they also applied MM methods to classify driver nodes into three categories critical, intermittent and redundant with the fraction 30.32, 30.24 and 39.44 percent respectively (X. Liu and Pan 2015).

Gol et al found that the essential human gene tends to encode hub protein (high-degree node in network) (Goh et al. 2007) and through investigation of topology feature, they concluded that cancer protein likely to have more degrees (J. Sun and Zhao 2010). However, in our study, possessing a high degree might not guarantee a critical node. In the study of controllability of human immunodeficiency virus type 1 (HIV) network, the result has shown that the critical nodes identified by MM tend to be peripheral in network due to their low in-degree value (Vinayagam et al. 2016). The study conducted by Liu and Pan about the controllability of human signaling network also concluded that critical driver nodes tend to have low in-degrees (X. Liu and Pan 2015).

4. Conclusions

In conclusion, using control theory to analyze complex networks in the context of breast cancer has proved to be a useful tool for biology system research. Besides that, re-ranking methods have successfully reduced the amount of our workload when using the genes in the highest score list as seed proteins for protein-protein interaction network. We

hope that our result can suggest fundamental insight on various genes related breast cancer players hidden inside. In the near future, virtual screening to develop potential drug candidates on the list of drug targets is perfectly possible.

References

Akram, Muhammad, Mehwish Iqbal, Muhammad Daniyal, and Asmat Ullah Khan. 2017. "Awareness and Current Knowledge of Breast Cancer." *Biological Research*. BioMed Central Ltd.

Bourdakou, Marilena M., Emmanouil I. Athanasiadis, and George M. Spyrou. 2016. "Discovering Gene Re-Ranking Efficiency and Conserved Gene-Gene Relationships Derived from Gene Co-Expression Network Analysis on Breast Cancer Data." *Scientific Reports* 6 (October 2015): 1–29.

Edgar, Ron, Michael Domrachev, and Alex E. Lash. 2002. "Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository." *Nucleic Acids Research* 30 (1): 207–10.

Emmert-Streib, Frank, Ricardo de Matos Simoes, Paul Mullan, Benjamin Haibe-Kains, and Matthias Dehmer. 2014. "The Gene Regulatory Network for Breast Cancer: Integrated Regulatory Landscape of Cancer Hallmarks." *Frontiers in Genetics* 5 (FEB): 15.

Feng, Junli, Walter D. Funk, Sy-Shi Wang, Scott L. Weinrich, Ariel A. Avilion, Choy-Pik Chiu, Robert R. Adams, et al. n.d. "The RNA Component of Human Telomerase." *Science*. American Association for the Advancement of Science. Accessed August 12, 2020.

Goh, Kwang Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert László Barabási. 2007. "The Human Disease Network." *Proceedings of the National Academy of Sciences of the United States of America* 104 (21): 8685–90.

Hoelder, Swen, Paul A. Clarke, and Paul Workman. 2012. "Discovery of Small Molecule Cancer Drugs: Successes, Challenges and Opportunities." *Molecular Oncology*. John Wiley and Sons Ltd.

Ibrahim, Shady S, Maha Ar Eldeeb, Mona Ah Rady, Karim M Abdel Hady, Mohamed S Lotfy, Noha S Farag, Stephan Verleysdonk, and Christoph P Bagowski. 2011. "The Role of Protein Interaction Domains in the Human Cancer Network." *Network Biology*. Vol. 1. www.iaees.orgArticle.

Liu, Xueming, and Linqiang Pan. 2015. "Identifying Driver Nodes in the Human Signaling Network Using Structural Controllability Analysis." *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12 (2): 467–72.

Liu, Yang Yu, Jean Jacques Slotine, and Albert László Barabási. 2011. "Controllability of Complex Networks." *Nature* 473 (7346): 167–73.

Mengual Gomez, D.L., R.G. Armando, C.S. Cerrudo, P.D. Ghiringhelli, and D.E. Gomez. 2016. "Telomerase as a Cancer Target. Development of New Molecules." *Current Topics in Medicinal Chemistry* 16 (22): 2432–40.

Poirel, Christopher L., Ahsanur Rahman, Richard R. Rodrigues, Arjun Krishnan, Jacqueline R. Addesa, and T. M. Murali. 2013. "Reconciling Differential Gene Expression Data with Molecular Interaction Networks." *Bioinformatics* 29 (5): 622–29.

Shahraki, Hadi Raeisi, Mansooreh Jaberipoor, Najaf Zare, and Ahmad Hosseini. 2016. "The Role of 22 Genes Expression in Bladder Cancer by Adaptive LASSO." *International Journal of Cancer Management* 9 (6).

Sprouse, Alyssa A., Catherine E. Steding, and Brittney Shea Herbert. 2012. "Pharmaceutical Regulation of Telomerase and Its Clinical Potential." *Journal of Cellular and Molecular Medicine* 16 (1): 1–7.

Sun, Jingchun, and Zhongming Zhao. 2010. "A Comparative Study of Cancer Proteins in the Human Protein-Protein Interaction Network." *BMC Genomics* 11 (SUPPL. 3).

Sun, Yi Sheng, Zhao Zhao, Zhang Nv Yang, Fang Xu, Hang Jing Lu, Zhi Yong Zhu, Wen Shi, Jianmin Jiang, Ping Ping Yao, and Han Ping Zhu. 2017. "Risk Factors and Preventions of Breast Cancer." *International Journal of Biological Sciences*. Ivyspring International Publisher.

Uhart, Marina, Gabriel Flores, and Diego M. Bustos. 2016. "Controllability of Protein-Protein Interaction Phosphorylation-Based Networks: Participation of the Hub 14-3-3 Protein Family." *Scientific Reports* 6 (December 2015): 1–11.

Vinayagam, Arunachalam, Travis E. Gibson, Ho Joon Lee, Bahar Yilmazel, Charles Roesel, Yanhui Hu, Young Kwon, et al. 2016. "Controllability Analysis of the Directed Human Protein Interaction Network Identifies Disease Genes and Drug Targets." *Proceedings of the National Academy of Sciences of the United States of America* 113 (18): 4976–81.