Comparative Analysis of Gene Regulatory Networks Concepts in Normal and Cancer Groups

Rosa Aghdam[a]*, Pegah Khosravi[a], Elnaz Saberi Ansari[a]

[a] School of Biological Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Abstract. Cancer is a main health problem in many countries, so the study of cancer is a challenging task in Bioinformatics. Recently, various methods have been proposed to detect the significant genes by analyzing gene expression data. Almost every approach has its own advantages and limitations and there is still a great space for current methods to be improved. By using various approaches, the precision of finding significant and non-significant genes can be considerably improved. In this paper, we are interested in using different methods to detect the significant and non-significant genes with focus on Liver and Gastric cancers. Then, the gene regulatory networks for significant and non-significant genes in both states (cancer and normal) are compared. In order to compare and analyze the resulted networks, known measures such as network diameter, characteristic path length and network heterogeneity are considered. Among network concepts, the network heterogeneity is a suitable parameter to compare networks. Results show that the network heterogeneity is higher in the cancer networks. Also, based on this parameter, resulted networks for non-significant genes in normal and cancer groups are more similar in comparison with networks for significant genes in normal and cancer groups.

Keywords: Gene expression data; Gene regulatory network; Significant genes; Non-significant genes; Network concepts; Network heterogeneity.

1. Introduction

Cancer has been described as a heterogeneous disease including various subtypes. The early finding and forecast of a cancer type have ended up essential in disease research, as it can help the subsequent clinical treatment of patients [1]. Measuring the expression of all genes in a sample is presently conceivable by starting new strategies in microarray technology. These strategies can be applied to measure the expression of all genes through various types of tumors. With respect to these gene expression data, it is conceivable to stratification of disease patients into high or generally safe gatherings utilizing Bioinformatics approaches [2]. One suitable way to study cancer more significantly is analyzing differential concepts between normal and cancer interaction networks.structures in the data [1].

Analysis of genome-scale gene networks utilizing large-scale gene expression data gives remarkable chances to reveal gene connections and regulatory networks required in different various of biological processes and formative projects, prompting quickened revelation of novel learning of different biological processes, pathways and systems. The Context Likelihood of Relatedness (CLR) algorithm is one of the most well-known method to reconstruct networks from data [3]. The CLR algorithm is based on the Mutual Information (MI) for scoring the similarity of gene pairs. In this research, we are focused on two types of cancers including Liver and Gastric cancers. The gene expression data set for normal and cancer states are available via the World Wide Web (http://www.ncbi.nlm.nih.gov/). Three well-known methods such as Differential Expression via Distance Summary (DEDS), MULTiple TESTing (Multtest), and Significance Analysis of Microarrays (SAM) methods which are available as part of the Bioconductor project, are used to decipher the significant and non-significant genes in the Liver and Gastric cancers. Furthermore, the CLR algorithm is used to reconstruct co-expression network and Gene Regulatory Network (GRN) from gene expression data through various states.

2. Materials and methods

2.1. Data sets

Whole genome-based microarray data are downloaded from the Gene Expression Omnibus (GEO) database [4],

* Corresponding author.
E-mail address: rosa.aghdam@ipm.ir

accession number GSE45436 [5] and GSE64951 for Liver and Gastric cancers, respectively. The Liver data set contains 144 samples, including 39 samples from normal Liver tissue and 95 samples from tumor Liver tissue. Besides, Gastric data set comprises 94 samples, containing 31 samples from normal tissue and 63 samples from tumor tissue.

Microarray data are preprocessed and analyzed using the LIMMA package in R [6] which is originally developed for di_erential expression analysis of microarray data. Quantile normalization [7] is used to normalize data. More detailed descriptions of the method can be found in the original publications.

2.2. Significant genes

To detect significant genes, three well-known approaches are applied to data sets. In this section, these methods are introduced briey. The Differential Expression via Distance Summary (DEDS) selects differentially expressed genes by integrating and summarizing a set of statistics using a weighted distance approach [8]. Moreover, MULTiple TESTing (Multtest) approach is a Non-parametric bootstrap and permutation resampling-based multiple testing procedures for controlling the False Discovery Rate (FDR) [9, 10]. Furthermore, the Significance Analysis of Microarrays (SAM) is a tool to identify crucial genes using either a modified t-statistic or a Wilcoxon rank statistic [11].

Although usually significant genes are detected based on one method, significant genes are detected using all three methods in our paper. Here, one gene is called significant if it is considered important by all mentioned methods. Similarly, one gene is called non-significant if all methods detect this gene as a non-significant gene. The precision of finding significant and non-significant genes can be considerably improved by considering three well-known approaches rather than one method. In this paper, we are interested in comparing the gene regulatory networks for significant and non-significant genes in both states (cancer and normal). So, we just study genes which are considered significant or non-significant by three well-known approaches.

2.3. Reconstruct networks and analysis

In order to compare the interaction networks between normal and cancer group, we use a high performance web-based platform (DeGNServer), for genome-scale gene networks construction and subnetwork extraction. This platform is equipped for analyzing gene expression information with high dimensionality of gene space and extensive number of gene expression profiles. To reconstruct interaction networks, DeGNServer is capable of using several of microarray profiles of human (36,000 genes) within 30 minutes. The DeGNServer is as exact and delicate as the original CLR algorithm and runs hundreds to thousands times quicker [12].

Then, we use NetworkAnalyzer [13] plugin in Cytoscape [14] which is the most useful structural analysis software to analyze and compare networks. NetworkAnalyzer efficiently computes some topological parameters, including number of nodes, number of edges, clustering coefficient, connected components, network diameter, network radius, network centralization, characteristics path length, average number of neighbors, network density and network heterogeneity.

2.4. Networks concepts

Network concepts or network statistics or network indices include connectivity, mean connectivity, density, variance of the connectivity (related to the heterogeneity) etc. For descriptive statistics related to the networks some concepts can be utilized. Some network ideas concepts, found vital, use in biology and a few network ideas concepts are less fascinating to biologists. For example, network heterogeneity is an important concept in biology, but network centralization is not very useful to biologists [15]. Many measures of network heterogeneity are based on the variance of the connectivity, and author›s approaches differ on how to scale the variance. One definition of the network heterogeneity is the coefficient of variation of the connectivity distribution, i.e. [15].

$$Heterogeneity = \frac{variance}{mean}. \qquad (1)$$

This heterogeneity measure is scale invariant with respect to multiplying the connectivity by a scalar. Biological networks tend to be very heterogeneous: while some ‹hub› nodes are highly connected, the majority of nodes tend to have very few connections. Describing the heterogeneity (inhomogeneity) of the connectivity (degree) distribution has been the focus of considerable research in recent years.

For example, in this work network heterogeneity is an important concept to distinguish between cancer and normal groups in cancer issue, but network centralization is not very useful for comparison of these groups. So, our results concur with earlier findings in recognizing of suitability of this measure to compare cancer and normal networks. In order to compare networks known measures such as network diameter, characteristic path length and network heterogeneity are considered. In the following, we review mentioned network concepts. The length of a path is the number of edges forming it. There is also multiple paths connecting two given nodes. The shortest path length, also called distance, between two nodes n and m is denoted by L (n;m). The largest distance between two nodes is called network diameter. When a network is disconnected, its network diameter is the maximum of all diameters of its connected elements. Besides, eccentricity is the maximum non-infinite length of a shortest path between node n and another node in the network. So, the diameter is also the maximum node eccentricity. Additionally, the characteristic path length denotes the average shortest path length. For two connected nodes the characteristic path length gives the expected distance between them. Furthermore, the network heterogeneity mirrors the propensity of a network to include highly connected “hub” nodes [15]. Hub genes are played an imperative part in arranging the conduct of biological networks [16, 17]. To detect biologically significant genes in cancer, connectivity has been observed to be a vital reciprocal gene

screening variable [18, 19] and primate brain development [20]. So, considering network heterogeneity for comparing networks in cancer and normal groups is exceptionally helpful. It is expected that this parameter received higher value for cancer group in comparison with those of normal group.

3. Results

In this work, the significant and non-significant genes for Liver cancer are detected by combining three well known methods. Then, the GRNs for normal and cancer groups based on resulted significant and non-significant genes are reconstructed. Finally, the network concepts are utilized to analyze and compare the networks. With a speci_c end goal to assess our research we apply these steps on Gastric cancer. The networks concepts related to gene regulatory network for Liver cancer are given in Table 1. Table 2 is related to the GRNs concepts for Gastric cancer.

As shown by Tables 1 and 2 network statistics for cancer group contain higher values in comparison with those of normal group. Regarding to the previous studies, network heterogeneity is a suitable measure to compare two networks that leads to many unique properties of complex networks [21]. In Liver and Gastric cancers this measure receives higher value for GRN for significant genes in cancer group. In addition, the difference between network heterogeneity for cancer and normal groups related to significant genes is much higher than the difference for non-significant genes. Fig. 1 shows the GRNs for cancer state. Fig. 1 (a) illustrates the GRN for significant genes, the significantly important genes are shown by blue and green colors. The green nodes display the common genes between normal and cancer states. The GRN for non-significant genes is shown in fig. 1 (b), the significantly not important genes are indicated by red and yellow colors. The yellow nodes show the common genes between normal and cancer states.

Similarly, the GRNs for normal state is shown in Fig. 2. Fig. 2 (a) shows the GRN for significant genes, the significantly important genes are illustrated by blue and green colors. The common genes between normal and cancer states are shown by green color. The GRN for non-significant genes is shown in fig. 2 (b), the significantly not important genes are denoted by red and yellow colors. The common genes between normal and cancer states are shown by yellow color.

4. Discussion and some future works

In the past decades, using high-throughput technologies, cancer researchers have collected a huge amount of data on the differences between cancer cells and their healthy counterparts, with the ultimate purpose of identifying therapeutic targets [22]. This has led to the identification of genes casually involved in human cancer [23, 24].

In this work, three methods are applied to detect significant and non-significant genes for Liver gene expression data set. The genes are called significant if there are selected as important by all three methods and genes are called non-significant if these genes are considered not be important by all three methods. After significant genes are detected the GRNs for significant genes in cancer and normal groups are reconstructed by CLR algorithm. Also, these networks for two groups are considered for non-significant genes. The network heterogeneity reflects the tendency of a network to include hub nodes. Studying the hub nodes is very important in biological networks, so the network heterogeneity measure is considered to compare cancer and normal groups to assess the complexity of the network. By analysing network concepts, it can be concluded that the network heterogeneity for cancer group is higher than that of normal group. In addition, the differences between this parameter for networks related to the significant genes is higher than networks for non-significant genes. It means that the trend progresses toward higher complexity over GRNs related to important genes. The Gastric cancer is considered to evaluate the results of this research and similar results are obtained for this cancer. Finally, there remains uncertainly information to evaluate network heterogeneity parameter for different kinds of cancers. The future research may minimize this problem.

Table 1. The networks concepts for liver cancer.

GRN	network diameter	characteristic path length	network heterogeneity
NSGC	**8**	**3.656**	**0.596**
NSGN	6	3.407	0.445
NNSGC	**11**	**4.754**	**0.589**
NNSGN	7	3.548	0.573

Note: In this table NSGC, NSGN, NNSGC and NNSGN denotes Network for Significant Genes in Cancer group, Network for Significant Genes in Normal group, Network for non-Significant Genes in Cancer group and Network for non-Significant Genes in Normal group, respectively.

Table 2. The networks concepts for gastric cancer.

GRN	network diameter	characteristic path length	network heterogeneity
NSGC	**8**	**3.818**	**0.866**
NSGN	7	3.561	0.528
NNSGC	**9**	**4.106**	**0.498**
NNSGN	7	3.76	0.445

Note: In this table NSGC, NSGN, NNSGC and NNSGN denotes Network for Significant Genes in Cancer group, Network for Significant Genes in Normal group, Network for non-Significant Genes in Cancer group and Network for non-Significant Genes in Normal group, respectively.

Acknowledgments

Rosa Aghdam, Pegah Khosravi and Elnaz Saberi Ansari have been supported by the School of Biological Sciences of Institute for Research in Fundamental Sciences (IPM).

References

1. Kourou, Konstantina, et al. “Machine learning applications in cancer prognosis and prediction.” Computational and structural biotechnology journal 13 (2015): 8-17.
2. Nascimento, André, et al. “Mining rules for the automatic selection process of clustering methods applied to cancer gene expression data.” Artificial Neural Networks–ICANN 2009 (2009): 20-29.

3. Faith, Jeremiah J., et al. "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." PLoS biol 5.1 (2007): e8.
4. Barrett, Tanya, et al. "NCBI GEO: archive for functional genomics data sets—update." Nucleic acids research 41.D1 (2012): D991-D995.
5. Landi, Maria Teresa, et al. "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival." PloS one 3.2 (2008): e1651.
6. Smyth, Gordon K., et al. "LIMMA: linear models for microarray data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Statistics for Biology and Health." (2005).
7. Bolstad, Benjamin M., et al. "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics 19.2 (2003): 185-193.
8. Pepe, Margaret Sullivan, et al. "Selecting differentially expressed genes from microarray experiments." Biometrics 59.1 (2003): 133-142.
9. Dudoit, Sandrine, Juliet Popper Shaffer, and Jennifer C. Boldrick. "Multiple hypothesis testing in microarray experiments." Statistical Science (2003): 71-103.
10. Ge, Youngchao, Sandrine Dudoit, and Terence P. Speed. "Resampling-based multiple testing for microarray data analysis." Test 12.1 (2003): 1-77.
11. Schwender, Holger. "Modifying Microarray Analysis Methods for Categorical Data—SAM and PAM for SNPs." Classification—the Ubiquitous Challenge. Springer Berlin Heidelberg, 2005. 370-377.
12. Li, Jun, Hairong Wei, and Patrick Xuechun Zhao. "DeGNServer: deciphering genome-scale gene networks through high performance reverse engineering analysis." BioMed research international 2013 (2013).
13. Assenov, Yassen, et al. "Computing topological parameters of biological networks." Bioinformatics 24.2 (2007): 282-284.
14. Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." Genome research 13.11 (2003): 2498-2504.
15. Dong, Jun, and Steve Horvath. "Understanding network concepts in modules." BMC systems biology 1.1 (2007): 24.
16. Albert, Réka, Hawoong Jeong, and Albert-László Barabási. "Error and attack tolerance of complex networks." nature 406.6794 (2000): 378-382.
17. Carlson, Marc RJ, et al. "Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks." BMC genomics 7.1 (2006): 40.
18. Horvath, S., et al. "Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target." Proceedings of the National Academy of Sciences 103.46 (2006): 17402-17407.
19. Carter, Scott L., et al. "Gene co-expression network topology provides a framework for molecular characterization of cellular state." Bioinformatics 20.14 (2004): 2242-2250.
20. Oldham, Michael C., Steve Horvath, and Daniel H. Geschwind. "Conservation and evolution of gene coexpression networks in human and chimpanzee brains." Proceedings of the National Academy of Sciences 103.47 (2006): 17973-17978.
21. Wu, Jun, et al. "A new measure of heterogeneity of complex networks based on degree sequence." Unifying Themes in Complex Systems. Springer Berlin Heidelberg, 2010. 66-73.
22. Khosravi, Pegah, et al. "Network-based approach reveals Y chromosome influences prostate cancer susceptibility." Computers in biology and medicine 54 (2014): 24-31.
23. Hornberg, Jorrit J., et al. "Cancer: a systems biology disease." Biosystems 83.2 (2006): 81-90.
24. Hecker, Michael, et al. "Gene regulatory network inference: data integration in dynamic models—a review." Biosystems 96.1 (2009): 86-103.