Austin
Publishing Group

Editorial

Challenges in Gene Network Problems for Cancer Treatment

Sung Won Han*

Department of Population Health, New York University, USA

*Corresponding author: Sung Won Han, Division of Biostatistics, Department of Population Health, School of Medicine, New York University, USA

Editorial

Cancer is still an important disease since it is one of the leading causes of death, and there is no treatment for curing it completely. Thus, monitoring cancer, and discovering the best treatments for increasing survival and quality of life for patients, is important research subjects. Conventionally, for cancer treatment, people use surgery, chemotherapy, and radiation therapy. Recently, other technologies such as immunotherapy, photodynamic therapy, and target cancer therapies such as gene therapy have been developed or are being developed. Especially, gene therapy becomes more important nowadays since this technique follows the mechanism of cancer development.

The growth of cancer is related to production of a protein, which is a functioning unit organism. Such a protein is produced through several steps. The DNA sequence inside of the human cell is a map of the human body. A part of the DNA is duplicated and becomes RNA sequences; this is called transcription process. Such RNA is called messenger RNA (mRNA). Several small organisms such as ribosome and t RNA attach to the mRNA, and produce amino acid; this is called translation process. During the process, micro RNA (mi RNA) also sticks to the mRNA sequence and regulates the process. After the translation process, the amino acids are folded into a certain type of protein. If some part of this entire process malfunctions, the odds of cancer development increase.

Cancer, which is also called a malignant neoplasm, occurs for various reasons. From exogenous agents such as radiation or chemicals, the DNA can be damaged. In addition, due to diet or mental stress, more oncogene and fewer tumors suppressor genes might be switched on. At sometime, DNA repair genes are epigenetically altered. In addition, mi RNA is known to affect the DNA repair genes. Thus, to find the mechanism of cancer development, understanding the interaction between genes or between gene and protein is important. The cancer treatment based on the blocking or switching-on of certain genes is gene therapy. Finding the gene pathway and interaction for gene therapy is called the gene network problem for cancer.

The gene network problem is very challenging from a mathematical point of view. It covers statistics/probability, optimization, and graph theory. The Directed Acyclic Graph (DAG) is a useful tool to find the gene pathway or estimate gene networks. A directed acyclic graph with probability distribution is called a Bayesian network. There are many approaches to Bayesian network problems. There are three categories for estimating the Bayesian Networks: a score-and-search approach to find a solution in the space of the Bayesian network, a constraint-based approach that tests conditional independencies which can be identified through the data, and a hybrid approach which combines the two approaches. The purpose of a score-and-search approach is to find a structure which gives a good score function value, and apply a heuristic algorithm to find the optimal value of the score function. The purpose of a constraint-based approach is to use a statistical test of conditional independence to find a skeleton as well as directionality of the network. One of the well-known methods for the constraint-based approach is the PC-algorithm [1]. In high-dimensional data, [2] showed that the PC-algorithm can estimate sparse DAGs within a reasonable computational time. Hybrid search approaches incorporating the above two approaches have been proposed. One known method is a max-min hill-climbing algorithm [3]. Those methods have been used to estimate DAGs with even a moderate number of nodes.

However, for the estimation of a DAG in gene network problems, various challenges are confronted. The existing approaches did not fully consider all challenging issues in the estimation of a DAG in gene network problems, such as unknown variable ordering, unknown and unequal variance of latent variables, convexity, acyclicity, equivalence class, high dimensionality, and solution quality. For example, if the score function is non-convex, search algorithms that have been proposed find only a local optimum. Furthermore, the acyclic condition makes it hard for the algorithm to find a global optimal solution. In addition, observational equivalence does not allow us to determine a unique solution since multiple solutions give the same value of the score function. Most existing solution search algorithms are ad-hoc or rudimentary, and their solution quality is not sufficiently justified in the viewpoint of optimization. Finally, depending on the nature of the problems, we need to utilize the prior information of the structure such as partial ordering or target response, so that the model and algorithm should be easy to customize. Thus, one issue is how to estimate a DAG in high dimensions with a reasonable computational time and a good guaranteed solution quality under such challenging circumstances. Especially, for high dimensional data, the conventional Bayesian approach requires excessive computational time.

To deal with such high dimensionality in gene network problems, the approach based on the Lasso framework has been developed [4] applied a Lasso approach to find the neighborhood of a node for high dimensional DAGs. To estimate directed acyclic graphs under known variable order, [5] applied the L1-penalized likelihood to the data, and showed that the penalized likelihood can be transformed

into separable Lasso problems [6] applied L1-penalized likelihood to estimate DAGs under unknown variable order. However, their score function is non-convex, so it might cause multiple local optimal solutions [7] proposed the lasso-based score function, and showed that their approach is competitive with other methods especially under hub network structure. In addition, for the next generation sequencing technology, most of the gene expression data such as RNA seq are discrete [8] developed a Lasso-based approach to estimate DAGs under the count data assumption. Despite the improvement in methodology evident in those works, we have still a challenging issue in estimating DAGs to show complete information for gene networks. We welcome the submission of papers dealing with this interesting problem.

References

1. Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search, MIT Press, Cambridge, MA. 2000.

2. Kalisch M, Buhlmann P. "Estimating high-dimensional directed acyclic graphs with the PC-algorithm." The Journal of Machine Learning Research. 2007; 8: 613-636.

3. Tsamardinos I, Brown L, Aliferis C. "The max-min hill-climbing Bayesian network structure learning algorithm." Machine Learning. 2006; 65: 31-78.

4. Meinshausen N, Buhlmann P. "High-dimensional graphs and variable selection with the Lasso." The Annals of Statistics. 2006; 34: 1436-1462.

5. Shojaie A, Michailidis G. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. Biometrika. 2010; 97: 519-538.

6. Fu F, Zhou Q. "Learning Sparse Causal Gaussian Networks With Experimental Intervention: Regularization and Coordinate Descent." Journal of the American Statistical Association. 2013; 108: 288-300.

7. Han SW, Chen G, Belousov A, Essioux L, Zhong H. "Estimation of directed acyclic graphs through a lasso framework for gene network inference," Submitted. 2014.

8. Han SW, Zhong H. "Estimation of sparse directed gene network for count data with the lasso framework," Submitted. 2014.