Predicting Cancer Regulatory Networks

Kevin Tee, Michael Liang

Department of Computer Science
University of California, Berkeley
CS294-108 Final Project

{kevintee,michaelliang}@berkeley.edu

For many human diseases, such as cancer, gene regulatory networks can provide insight into disease etiology and pathogenesis. Recent technological advances in the measurement of genome-wide gene expression allow for many computational inferences of such networks. The high dimensionality of gene expression data and insufficient sample measures often resulted in the predicted regulatory relations that were not supported by biological evidence. The lack of consensus among these inference methods is another concern. Thus, validation of these inferred networks became crucial. To that end, we compared the inferred gene regulatory networks (GRN) from MERLIN and ???. Both methods were applied to publicly available breast tumor microarray data from the TCGA database. We found [summarize similarities and differences: overall gene regulatory network visualization is similar; % consensus and % disagreement in terms edges and clustering?]. Based on Gene Oncology analysis, approximately % of regulatory relations inferred by MERLIN and % by ??? showed evidence of biological relevance. The minor (or remarkable)

differences in the GRN inferred by these two methods pointed to the possible validation approach through comparisons of multiple inference methods and Gene Oncology analysis.

Introduction

Cancer is a disease of the genome that is associated with accumulation of mutational events that lead to disregulation of the cellular system. Genes involved in cancer development and progression are classified as oncogenes, tumor suppressor genes and genomic stability genes [1]. These genes have a key role in the regulation of the cell-cycle, proliferation and cell differentiation, and in the regulation of apoptosis [2]. Mutations in over 100 genes are known to drive tumorgenesis, affecting a broad classes of proteins such as transcription factors, chromatin remodelers, growth factors(e.g.,EGFR), growth factor receptors (e.g.,HER2),signal transducers, regulators of apoptosis and DNA repair genes [2]. Within any given tumor there are between 2-8 mutated driver genes modulating the activity of critical molecular pathways [3]. Pilot studies by TCGA and others demonstrated that patients harbor genomic alterations or aberrant expression in different genes and these genes often participate in a common pathway [4,5], indicating that pathway-level genomic perturbations are key features in the underlying cancer biology. Gene regulatory network (GRN) in these cancer pathways has been recognized as potentially important prognosis markers in risk of metastasis or ultimately the target for personalized treatments [6].

High-throughput technologies have produced enormous amount of genome-wide gene expression data. Many data driven mathematical and computational approaches have been developed for probing the molecular interactions of cancer pathways. [Should we review some of these methods, in the context of why we choose two of these methods to compare?] The molecular interactions between the genes and their products are complex and dynamic. Due to the high dimensionality of gene expression data and insufficient sample measures, the GRN with high prediction accuracy may not reveal the true biological relations. Further, different inference methods may yield different GRNs that are difficult to validate. Thus, the translational potential of gene expression profiling in cancer diagnosis, prognosis and in the development of personalized medicine may not be fulfilled due to the lack of consensus among the inferred GRNs and poor understanding of the underlying mechanisms.

In this paper, we aim to compare two methods using ChIP-seq data as gold standard and validate the clustering of genes into modules using Gene Ontology. The first method is modular regulatory network learning with per gene information (MERLIN), which assumes a conditional Gaussian distribution for the conditional relationships between regulators and genes and is based on a probabilistic graphical model representation of a regulatory network. The second method is [???]. [Why we choose these two specific methods for comparison?] Both methods will be applied to the publicly available breast tumor microarray data from the TCGA database to infer GRN. We will investigate the differences in the regulatory networks inferred from these two methods based on the established cancer pathway databases and Gene-Ontology analysis. [In what way, understanding these differences can help us?]

Methodology

Example

Discussion

References

[1] Vogelstein, B. and Kinzler, K.W. Cancer genes and the pathways they control. Nat Med 10, 789-99 (2004).

[2] Croce, C.M. Oncogenes and cancer. N Engl J Med 358, 502-11 (2008).

[3] Vogelstein, B. et al. Cancer genome landscapes. Science 339, 1546-58 (2013).

[4] Parsons, D.W. et al. An integrated genomic analysis of human glioblastoma multiforme. Science 321, 1807-12 (2008).

[5] Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 1061-8 (2008).

[6] Drier, Y., Sheffer, M. and Domany, E. Pathway-based personalized analysis of cancer. Proc Natl

Acad Sci U S A 110, 6388-93 (2013).