Revealing the Cancer Epigenome: Computational Analysis of Biological Networks

Abstract

The last decade has witnessed extensive developments in the area of biological instrumentation and sequencing technologies. The advent of Next Generation Sequencing (NGS) techniques resulted in the piling up of specialized data in various biological databases. Today, an unimaginably huge volume of information relevant for genomic studies is available in scientific literature and molecular databases, thanks to the emergence of Bioinformatics as an interdisciplinary field of science. While this information provides a goldmine for elucidating the mysteries of life, it also poses a daunting challenge of analyzing and interpreting these data.

Epigenetics has recently emerged as a new wave of research, which attempts to study the mechanism of inheritance not involving changes in the DNA sequence. Epigenetic events are characterized by methyl, acetyl, or histone modifications to the chromosome. All these chemical tags taken together, constitute the *epigenome*. The epigenome information is currently viewed as a means to improve clinical diagnosis and precise molecular classification of diseases such as cancer in humans. Existing epigenetic studies on pathways fail to consider the interactions between pathways, and treat them simply as a functionally cohesive set of genes. Identifying epigenetic patterns that are conserved across cancer pathways is worth investigating, as it helps in molecular classification of cancer. However, existing computational models try to align protein-protein interaction networks based on sequence similarity information, which is of little use in epigenetic profiling, as the changes in epigenome are not reflected in the DNA sequence. Computational methods to identify gene clusters, integrate multiple heterogeneous omic data types into a single scaffold network. However, the exact relationship between gene expression and methylation has not been elucidated clearly. Keeping in mind the above aspects, this thesis proposes three significant contributions in order to reveal the complex epigenetic mechanism underlying one of the most intricate and dreadful diseases affecting humans, namely cancer.

The first contribution is a novel machine learning framework to identify biological

pathways that are dysregulated by epigenetic mechanism. The proposed computational model abstracts gene-level details and considers interaction between pathways. The dysregulated pathways are identified by formulating the problem as a feature selection task in machine learning. The objective is to find a set of pathways that can best discriminate cancer samples from the normal ones. Experiments were conducted on benchmark cancer datasets from the National Center for Biotechnology Information (NCBI). Comparison with state-of-the-art pathway identification methods reveal the effectiveness of the proposed approach.

The second contribution is a deep embedding framework to extract cancer-specific and across-cancer epigenetic signatures from pathways. Here, a Deep Neural Network is trained to learn an encoded representation of a set of pathways. The encoded pathways are then aligned for topological correctness and functional consistency. Experiments on benchmark cancer datasets from the National Center for Biotechnology Information (NCBI) produced promising results. Comparison with recent network alignment methods clearly suggests that the proposed method obtains highly coherent signatures. A web-based tool called the Deep Encoded Epigenetic Pathway Aligner (DEEPAligner) has been developed based on the proposed computational method.

As a third contribution, a consensus-based clustering framework is proposed to identify communities of differentially methylated patterns of genes. The proposed method takes multiple gene networks as input and identifies regions of differential epigenetic activity within these networks. Initially, individual networks are constructed to model heterogeneous omic data types. Clustering is then applied on this network. Consensus clustering is applied to obtain a single community structure from the individual community structures. Experiments on benchmark cancer datasets from The Cancer Genome Atlas (TCGA) reveal that multi-network approaches produce more discriminative epigenetic communities than integrated approaches.

The proposed computational methods will definitely advance recent research efforts in the area of epigenetic epidemiology, which deals with investigating the influence of epigenetic changes in the etiology of complex diseases. A thorough study of these changes, juxtaposed with other epigenetic modifications such as chromatin and non-coding RNAs, are sure to lead us way beyond.