Computers in Biology and Medicine 76 (2016) 160–167

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Identifying epigenetically dysregulated pathways from pathway–pathway interaction networks

R. Visakh *, K.A. Abdul Nazeer

Department of Computer Science and Engineering, NIT Calicut, Kerala, India

ARTICLE INFO

Article history:
Received 23 May 2016
Received in revised form
29 June 2016
Accepted 30 June 2016

Keywords:
Epigenetics
Pathway–pathway interaction network
Molecular Pathological Epidemiology
Differential gene expression
Differential gene methylation
Feature selection

ABSTRACT

Background: Identification of pathways that show significant difference in activity between disease and control samples have been an interesting topic of research for over a decade. Pathways so identified serve as potential indicators of aberrations in phenotype or a disease condition. Recently, epigenetic mechanisms such as DNA methylation are known to play an important role in altering the regulatory mechanism of biological pathways. It is reasonable to think that a set of genes that show significant difference in expression and methylation interact together to form a network of pathways. Existing pathway identification methods fail to capture the complex interplay between interacting pathways.

Results: This paper proposes a novel framework to identify biological pathways that are dysregulated by epigenetic mechanisms. Experiments on four benchmark cancer datasets and comparison with state-of-the-art pathway identification methods reveal the effectiveness of the proposed approach.

Conclusion: The proposed framework incorporates both topology and biological relationships of pathways. Comparison with state-of-the-art techniques reveals promising results. Epigenetic signatures identified from pathway interaction networks can help to advance Molecular Pathological Epidemiology (MPE) research efforts by predicting tumor molecular changes.

1. Introduction

Pathway analysis is primarily meant to explore the impact of aberrant modifications in molecular pathways that can lead to a particular change in phenotype of interest such as a malignant tumor or abnormal cell growth [1]. A pathway encompasses a set of genes and gene products such as proteins to interact with probably, dozens of other pathways to form a complex network within a cell. This paper seeks to address the problem of identifying pathway dysregulations implicated by epigenetic changes, more specifically, methylation, to infer prognosis of diseases. Epigenetics studies the heritable changes in functions and behavior of genes that cannot be explained by changes in gene sequence [2]. DNA Methylation refers to addition of methyl (CH_3) group to the 5′ end of a string of Cytosine or Guanine nucleotides in human genome [3]. Patterns of epigenetic information are faithfully propagated over multiple cell divisions, which makes epigenetic regulation a key mechanism for cellular differentiation. Fig. 1 shows the mechanism of DNA methylation.

Epigenetic research has gained new momentum during the last five years especially due to the explosion of high throughput data available from various tools and databases. Epigenetic modifications such as DNA methylations, and Histone modifications, can intervene in the normal gene-protein *interactome*.[1] Epigenetic interaction networks are not well characterized in the literature. Recent efforts to picture the dynamics of epigenetic interaction network can be seen in [3–7].

In [3], Zhang et al. integrated whole genome methylation data across seven different cancers and performed hierarchical clustering to reveal cancer specific methylation patterns. MacNeil et al. [5] proposed a novel Gene Set Omic Analysis (GSOA) that can identify aberrant patterns of genes across multiple types of Omic data. Attempts to model the correlation between genes of a pathway and connecting nearby pathways based on differential correlation can be seen in [6,7]. The first successful step in this direction was done by Tarca et al. [6]. They proposed a novel methodology called Signaling Pathway Impact Analysis (SPIA) that combined the differential expression of genes in a single pathway with measured perturbation level of that pathway to analyze its significance. A powerful method for the analysis of dysregulated pathways is proposed by Han et al. [4]. Known as the Edge Set Enrichment Analysis (ESEA), this method captures the biological interdependence of pathways in terms of their component gene

* Corresponding author.
E-mail addresses: visakh_p140056cs@nitc.ac.in (R. Visakh), nazeer@nitc.ac.in (K.A. Abdul Nazeer).

[1] Whole set of molecular interactions in a cell.

http://dx.doi.org/10.1016/j.compbiomed.2016.06.030
0010-4825/

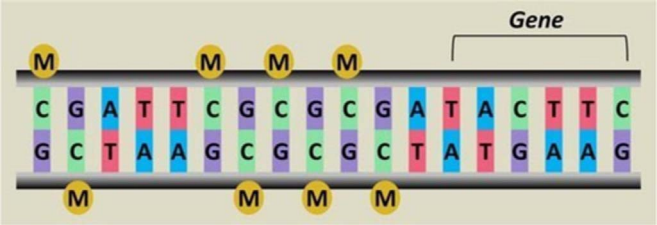

Fig. 1. Mechanism of DNA methylation (Source: Delaware Center for Neuroscience Research, US).

expression levels. Another method proposed by Liu et al. [7] address the issue of identifying dysregulated pathways as a feature selection problem.

It seems that epigenetic pathway dysregulation studies are still in its infancy. Except for a few notable works, none of them considers the impact of interactions among epigenetic subnetworks. A novel framework called Differential Methylation Pathway–Pathway Interaction Network (DMPIN) is proposed in this direction. A two-fold level of abstraction is used. First, a pathway is modeled as an integration of gene expression and methylation data. The next level of abstraction hides the gene-level details and considers a pathway interaction network as a set of pathways and their interactions. The dysregulated pathways are identified by modeling the problem as a feature selection task in machine learning. The task is to find a set of features (pathways) that can best discriminate between case and control samples.

The remainder of this paper is organized as follows: Section 2 gives the formal definition of the problem and a detailed description of the proposed methodology. Section 3 gives the experimental results. Section 4 gives the discussion of the results and comparison with existing methods. Section 5 concludes the paper with a remark on future prospects of epigenetic studies.

2. Materials and methods

2.1. DNA methylation, gene expression and pathway data

Human DNA methylation data were obtained from the publicly available NCBI Gene Expression Omnibus [8] which consisted of Illumina HumanMethylation27 and 450 BeadChip array data across 4 different cancers. These include lung cancer (Wilson et al., 2009 accession GSE19034), prostate cancer (Aryee M.J. et al., 2012, accession GSE38240), breast cancer (Di Cello F. et al., 2013, accession GSE44837) and colorectal adinoma (Vladimir Naumov et al., 2012, accession GSE42752). All values were normalized beta values. IlluminaHumanMethylation450k.db library was used to map probe IDs to Entrez IDs. Those probes which contained missing values and do not map to any gene ID were discarded.

Gene expression data were also obtained from the publicly available NCBI Gene Expression Omnibus [8]. The Affymetric array expression profiles of matched samples from 4 cancer types include lung cancer (GSE4115), prostate cancer (GSE6919), breast cancer (GSE 44836) and colorectal adenoma (GSE8671 and GSE35896). The platform details and sample information of methylation and expression datasets are given in Tables 1 and 2.

Pathway data were downloaded from the Molecular Signatures Database (MSigDB) [9]. 1000 cellular pathways were included in the study which includes signaling pathways from BioCarta [10], Reactome [11] and metabolic pathways from the Kyoto Encyclopedia of Genes and Genome (KEGG) [12].

Table 1
DNA methylation datasets.

Disease	GEO accession	Sample count (case/control)	Platform
Lung cancer	GSE19034	60 (30/30)	GPL8490 (Illumina HM27K)
Prostate cancer	GSE38240	12 (8/4)	GPL13534 (Illumina HM450K)
Breast cancer	GSE44837	26 (13/13)	GPL13534 (Illumina HM450K)
Colorectal cancer	GSE42752	63 (41/22)	GPL13534 (Illumina HM450K)

Table 2
Gene expression datasets.

Disease	GEO accession	Sample count (case/control)	Platform
Lung cancer	GSE4115	192 (102/90)	GPL96 (HG-U133A)
Prostate cancer	GSE6919	43 (25/18)	GPL8300 (Affy HG U95C v2.0)
Breast cancer	GSE44836	26 (13/13)	GPL6480 (Agilent WHG 4 × 44)
Colorectal adenoma	GSE8671, GSE35896	126 (94/32)	GPL570 (HG-U133 Plus 2)

2.2. Finding differential methylation and expression

Not all genes are methylated the same way. Some are over methylated, while some others lack sufficient methylation. This Differential Methylation (DM) can regulate gene expression or silencing. DM thus causes perturbations in biological pathways constituted by such genes. For each dataset, DM analysis of genes were carried out using R [13] limma [14] package. The *toptable* function was used to filter out genes showing differential methylation p value less than 0.01. False discovery rate adjustment method was used for p value correction. Differential expression (DE) of genes was also obtained using limma *toptable* function. For each pathway, the list of genes that were both differentially expressed and methylated was taken for further analysis.

2.3. Characterizing pathway activity

A pathway is modeled as an integration of gene expression and methylation data. To characterize the interactions between genes in a pathway, a connected, weighted and directed graph $G = \{V, E\}$ is constructed where V is the set of vertices (genes) of the graph and E is the set of edges that constitute the interactions between them. Weights on the edges denote the differential activity between a pair of genes. The differential activity is computed by integrating two scores: the differential score and the similarity score. A similar kind of scoring function is used in [15]. Differential score of a gene pair is a measure of discrepancy between expression or methylation behavior of those two genes across case and control samples. The differential score $D(x, y)$ between a pair of genes x and y is calculated

$$D(x, y) = \frac{1}{2n} \sum_{i=1}^{n} \delta(x^i) + \delta(y^i) \quad (1)$$

where n denotes the total number of Omic profile data available. Here, n is 2 since only methylation and expression profile data are considered. $\delta(x^i)$ and $\delta(y^i)$ are the mean log fold change in expression or methylation values normalized to lie in the interval