Methods 110 (2016) 81–89

Contents lists available at ScienceDirect

Methods

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

Protein interaction network (PIN)-based breast cancer subsystem identification and activation measurement for prognostic modeling

S. Lim [a], Y. Park [a], B. Hur [a], M. Kim [a], W. Han [b,c], S. Kim [a,d,e,*]

[a] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea
[b] Department of Surgery, Seoul National University College of Medicine and Hospital, Seoul, Republic of Korea
[c] Cancer Research Institute, Seoul National University College of Medicine, Seoul, Republic of Korea
[d] Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea
[e] Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 23 March 2016
Received in revised form 31 May 2016
Accepted 17 June 2016
Available online 18 June 2016

Keywords:
Breast cancer
Subsystem
Subsystem activation score
Cohort stratification
Protein interaction network
Prognostic modeling

ABSTRACT

Genome-wide gene expression information has been very useful for understanding cancer at the molecular level. In particular, breast cancer has been widely studied by utilizing a large amount of transcriptome data. Although statistical selection of differentially expressed genes, e.g., PAM50, has been successful to classify breast cancer subtypes, understanding breast cancer in terms of biological functions or pathways is still limited. Thus, it is essential to develop a tailored model that unravels breast cancer mechanisms by identifying disease-specific functional units of biological pathways and apply the model for breast cancer prognosis.

In this paper, a systematic characterization of breast cancer functional units or 'subsystems' is presented. We propose a novel concept of decomposing biological pathways into subsystems by utilizing protein interaction network, pathway information, and RNA-seq data. Subsystem activation score (*SAS*) was developed to measure the degree of activation for each subsystem and each patient. This method revealed distinctive genome-wide activation patterns or landscape of subsystems that are differentially activated among samples and among breast cancer subtypes. Then, we used *SAS* information for prognostic modeling by performing the classification and regression tree (CART) analysis. Eleven subgroups of patients, defined by 10 most significant subsystems, were identified with the maximal discrepancy in survival outcome. Our model not only defined patient subgroups with similar survival outcomes, but also provided patient-specific decision paths determined by subsystem activation status, suggesting functionally informative gene sets of breast cancer.

1. Introduction

Prognostication and prediction of patients' survival are one of the major goals of breast cancer research. Practical decision making of the breast cancer treatment plan is based on clinicopathological features such as tumor size, lymph-node metastasis, histological grade and three receptor (ER, PR, and HER2) responses to endocrine therapy [1]. Although these methods have been widely and successfully used since 1970s, they are not effective for diagnosis of the cancer at earlier stages and precise clinical decisions requires more than the clinicopathological features [2]. Thus investigation on the genome-wide landscape of molecular level features in breast cancer has been extensively performed [3–6]. These efforts initiated a new paradigm of clustering patients followed by annotating characteristic labels on the clusters of patient groups in terms of survival outcome [7].

In an effort to develop a model for grouping patients, there have been several array-based gene expression studies grouping patents based on a set of genes that are differentially expressed among the cohort, which results in molecular subtypes based on patient clusters [3,8–15].

Surprising discovery from these studies was that only a small number of genes were sufficient to characterize patient groups at the molecular level. In addition, genes selected by different studies show similarities in terms of gene expression levels. These gene expression signatures proved themselves as determinants to survival outcome without resorting to anatomical prognostic variables such as tumor size or nodal status [16]. Most of the

* Corresponding author at: Department of Computer Science and Engineering, Seoul National University, Seoul, Republic of Korea.
E-mail address: sunkim.bioinfo@snu.ac.kr (S. Kim).

http://dx.doi.org/10.1016/j.ymeth.2016.06.015
1046-2023/

methods showed equipotent performances in terms of prognostic modeling with a high concordance rate [17]. Among them, PAM50 method became standardized as the fundamental requirement for molecular diagnosis of breast cancer, of which assigns subtypes by incorporating microarray expression values to the centroids of 50 genes [18].

However, even PAM50 subtypes remained heterogeneous in receptor status; for an example, in basal-like subtype patients, 17 % of the samples were not in both ER-negative and HER2-negative statuses, despite that being accepted as typical clinical-pathological features of basal-like subtype [19]. In another study by [18], it was shown that luminal B subtype can be divided into at least five subgroups. One reason for this would be from the fact that the selection of genes in PAM50 was not guided by accurate gene expression profiles that are measured by microarray technologies. This can be resolved by leveraging RNA sequencing technologies as demonstrated in a study by [20]. In comparison with microarray data, RNA-seq produced more accurate gene expression measurements at the whole transcriptome level by showing that RNA-seq data had much higher concordance rate with expression profiles measured by qRT-PCR and also that RNA-seq achieved much better sensitivity for low-abundant genes.

Another technical issue for characterizing biological mechanisms underlying breast cancer is that we should consider relational nature of dysregulated genes with context. [21] used random forests for prioritizing important pathways in several diseases such as breast and lung cancers, rather than simply listing important genes for the diseases. Another popular technique is to use network. Recently a consortium of network biology was formed to analyze multi-dimensional genomic data [22]. Protein–protein interaction network (PIN) is one of the most widely used network based analysis techniques to cover true relational characteristics [23]. For example, [24] used PIN as a template to diffuse the significance of somatic mutation profiles and discovered biological modules crucial for identifying patient clusters of several cancers. This was consistent with previous studies that mutational events are localized to certain area (modular structure) of a network, hardly perturbing the whole biological structures [25,26].

1.1. Importance of pathway and network utilization

[27] classified module identification methods into three categories: expression-based, pathway-based, and network-based approaches and this categorization was recently revisited and well summarized by [28].

As biological knowledge discovery moving toward deciphering the functions of cooperative machinery rather than individual differentially expressed genes (DEGs), identifying the cluster or gene set modules became one of the popular research topics [29–31]. These methods mostly used machine learning or statistical techniques to identify systems of coordinated genes by utilizing gene expression profiles. In addition, several studies focused on the measurement of activity or level of perturbation using pathway information and expression profiles [32]. It is desirable to use multi-dimensional omics data to precisely measure the activity of a pathway as performed in [33]. However, the integrated analysis of multi-omics data needs to be further developed.

Fortunately, there are many studies that used only gene expression data to measure the degree of distorting the original (trained) distribution of gene set or metagene scores [34–36]. For example, a Bayesian regression model introduced by [34] used a set of 100 genes that maximally discriminates the ER status of breast cancer. This approach was extended to examine the status of several oncogenic pathways by using metagene concept [35]. A further analysis of 18 representative pathways was successful to classify human breast cancer subtypes [37].

In addition, there are a number of studies that utilized well curated networks other than biological pathways [38]. Among biological networks, PIN is widely used. As PIN covers a lot more number of genes than biological pathways, there have been several studies that initiated the identification of prognostic signatures using PIN [39,40]. [40] did a seminal work that incorporated gene expression information to PIN. In their analyses, edge weights in PIN were defined by using microarray based gene expression data and then network modules were identified by using the MCL clustering method. This study produced many false positives because using the microarray data do not have accurate gene expression information and co-expression information was not explicitly used. Furthermore, activation status of a module was simply calculated by averaging the gene expression values in the module without incorporating the relationships among the genes. This drawback can be remedied by utilizing RNA-seq expression data to use more accurate gene expression information and also by defining network modules in a stringent way [41]. Another work [42] was focusing on dealing with both nodes and edges and defined network biomarkers for differentiating cancer stages. Their method yielded edge-based biomarkers of improved accuracy on survival categorization with significant enrichment of biological information.

1.2. Necessity of subsystems

As we discussed in the previous subsection, gene expression or transcriptome data can be better analyzed in terms of biological pathways. Commonly used pathway databases are KEGG [43], REACTOME [44] and NCI cancer pathway [45]. A pathway is defined to model a series of actions among molecules in a cell that leads to a certain product or a change in a cell. As a result, a pathway consists of multiple complex biochemical functions, rather than a single biological function. This led to several research efforts to define multiple coherent units of a pathway. [46] pioneered the use of a subsystems approach to annotate genomes by categorizing genes into single functional groups. [47] proposed a strategy of decomposing pathway information into smaller modular structures. All these studies assure that defining functional units of a pathway is desirable and useful. However, there is no systematic study on defining subsystems of a specific disease using transcriptome data measured from many samples.

The goal of our study is to reveal biological mechanisms underlying breast cancer in terms of pathways. To achieve this goal, we need to develop a computational method to define functional units or subsystems of a pathway using transcriptome data. We illustrate our approach using PI3K-Akt Signaling pathway that consist of 293 genes in Fig. 1. The widely used DEG analysis results in too many statistically significant genes that can be mapped to many pathways, so the DEG approach does not distinguish core pathways from many activated pathways when expression values of all DEGs are mapped to pathways. To measure the activation status of a pathway, when expression values of all genes in the pathway were aggregated to a single value, the difference in the activation status of the pathway was not clear among cancer subtypes (Fig. 1A–C). However, our approach of decomposing the pathway into a set of distinct subsystems was effective to explain the differential activation status of the pathway among cancer subtypes (Fig. 1D).

1.3. Outline of this work

In this work, a novel method of generating a set of subsystems of breast cancer is proposed. Our method utilizes both PIN and pathways with accurate gene expression information measured