Computers in Biology and Medicine 67 (2015) 29–40

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A probabilistic approach for automated discovery of perturbed genes using expression data from microarray or RNA-Seq

Gopinath Sundaramurthy *,[1], Hamid R. Eghbalnia **,[1]

Department of Molecular and Cellular Physiology, University of Cincinnati, Cincinnati, OH, USA

ARTICLE INFO

Article history:
Received 3 April 2015
Accepted 30 July 2015

Keywords:
Biomarker discovery
Probabilistic modeling
Network analysis
Microarrays
RNA-Seq
Genomics
Breast cancer
Complex disease
Robust
Signaling and regulatory network

ABSTRACT

Background: In complex diseases, alterations of multiple molecular and cellular components in response to perturbations are indicative of disease physiology. While expression level of genes from high-throughput analysis can vary among patients, the common path among disease progression suggests that the underlying cellular sub-processes involving associated genes follow similar fates. Motivated by the interconnected nature of sub-processes, we have developed an automated methodology that combines ideas from biological networks, statistical models, and game theory, to probe connected cellular processes. The core concept in our approach uses probability of change (POC) to indicate the probability that a gene's expression level has changed between two conditions. POC facilitates the definition of change at the neighborhood, pathway, and network levels and enables evaluation of the influence of diseases on the expression. The 'connected' disease-related genes (DRG) identified display coherent and concomitant differential expression levels along paths.

Results: RNA-Seq and microarray breast cancer subtyping expression data sets were used to identify DRG between subtypes. A machine-learning algorithm was trained for subtype discrimination using the DRG, and the training yielded a set of biomarkers. The discriminative power of the biomarkers was tested using an unseen data set. Biomarkers identified overlaps with disease-specific identified genes, and we were able to classify disease subtypes with 100% and 80% agreement with PAM50, for microarray and RNA-Seq data set respectively.

Conclusions: We present an automated probabilistic approach that offers unbiased and reproducible results, thus complementing existing methods in DRG and biomarker discovery for complex diseases.

1. Background

Cancer is a prototype of a complex disease in which emergent phenotypes arise from the breakdown of complex cellular systems that are induced by multiple molecular and environmental perturbations [1–5]. While complex diseases may be the result of a range of molecular perturbations that can strongly vary between patients, they often dysregulate similar components of the cellular system [1,6–8]. Gene expression array platforms, such as microarray and RNA-Seq, have been routinely used in the study of complex diseases [9–12]. These platforms have been useful in assessing the collective behavior of cellular systems in response to disease perturbations by identifying a subset of significantly changing genes, which are likely to play a role in the disease [11,13]. Expression platforms produce high-dimensional data by simultaneously screening thousands of genes [13–15]. To reduce the dimensionality of gene expression data and identify the differentially expressed genes, 'frequentist' approaches employ fold change cutoffs and statistical significance levels [46]. The basis of selecting threshold values for differential genes selection is empirical, and variations in cutoff levels have been shown to significantly alter the interpretation of the results [17]. Recognition that use of thresholds and assumptions of independence may lead to loss of potentially important players in impacted pathways [14], and alternative approaches that rely on networks and pathways have emerged. Pathway analysis is a term broadly used to describe approaches, in which prior knowledge about biological networks is used in order to reduce the complexity of the data set and to extract features relevant to the biological changes [12,15,16]. Among common features in pathway-based approaches, one may include the use of pathway knowledge bases, creation of mathematical or statistical models for input data, and methods for scoring or evaluating the level of changes in pathways [17]. The majority of these approaches obtain their signaling, metabolic and

* Corresponding author. Tel.: +1 513 497 5149.
** Corresponding author.
E-mail addresses: sundargh@mail.uc.edu (G. Sundaramurthy), eghbalhd@ucmail.uc.edu (H.R. Eghbalnia).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.compbiomed.2015.07.029
0010-4825/

protein–protein interaction network data from publically available knowledge bases such as, KEGG [18], Pathway Commons [19], PANTHER [20] and Reactome [21].

List-based network analysis is the most common methodology, which uses a subset of expression array genes called 'seed' to create the network [22,23]. These methods are useful because using the full set of genes for classification (disease vs. healthy, for example), or multi-way classification (cancer sub-typing, for example) is ill advised; due to the high dimensionality of the data sets and low sample numbers [18]. The basis of picking threshold values in list-based analysis for seed genes selection in these approaches is empirical, and variations in cutoff levels have been shown to significantly alter the interpretation of the results [22]. Use of thresholds may also result in the elimination of relevant but sub-threshold genes, thereby increasing the potential for incomplete or incorrect results [22]. Although their performance does not depend on statistical thresholds, their efficacy deteriorates due to increased noise sensitivity that comes from the (very many) genes that do not change much between the phenotypes.

In addition, though seed gene identities can be combined as binary values (present or absent) in the network context, there is no canonical mathematical basis for combining p-values and fold changes as graded values that could represent changes at a sub-network level [19,17]. In order to strike a balance between sensitivity and selectivity of genes in pathway-based methods, it is necessary to delicately evaluate pathway level changes. The diversity of assumptions that underpin these approaches limit the options for meaningful comparison—for example, different biological investigations may use different levels for significant fold change, or threshold for *p*-value, utilize specific clustering methodology, and employ unique pathway selection approaches.

We propose a global network-based probabilistic approach to enhance the detection of differentiating properties in an automated manner. This approach does not make primary use of thresholds. However, we note that our computational steps make secondary use of threshold in order to speed up processing, and to streamline interpretation of final path analysis results. A comparative overview in Fig. 1 identifies the key differences of our probabilistic approach with the commonly used list-based, or direct, methodology. In order to assign the probability values, the idea is to assign a probabilistic response to the following question: given a gene and two sets of gene expression values for the gene, corresponding to change between two conditions (set A and set B), what is the probability that a given (unlabeled) expression values can be correctly assigned to the correct set (set A or set B)? This probability is called 'Probability of Change' (POC), and the name is suggestive of the rationale. If expression values in sets A and B are sufficiently different (i.e. a change has occurred because of the disease), a robust discrimination model can be built. The graded response represented by POC can be adapted readily to the network setting. Moreover, we show that POC succinctly captures the essential features of the frequentist gene-expression measures (*p*-value and fold change) into a single score.

The architecture of biological networks has been the subject of intense study. These networks have been shown to be scale-free, and wherein the empirical degree distribution has heavy tails. For example, it has been heuristically noted that most nodes in the network have low degree (most genes interact with at most 5 other genes) while a few nodes have a very high degree. Several methods for identification of hubs have been proposed and compared by Vallabhajosyula et al. [24]. A useful heuristic for capturing all hubs in networks is to capture the top 10% of the overall nodes formed the heavy tail of the distribution—in most cases studies, hubs are within the top 5–8%. To identify dysregulated genes, we traverse the biological networks using 'hub' genes and proteins, which are well-conserved proteins that have been shown to play an important role in the robustness of biological networks by acting as 'problem distributors' [3,17]. Dysregulation of hubs and the paths between hubs have been shown to be responsible for the fragility of biological networks [21]. As a result of these vital network properties, disease-related genes have been shown to be in close proximity to hubs [1,9,25].

We tested our platform using breast cancer subtyping data sets—a well-studied complex disease [26]. Our results identify disease-related

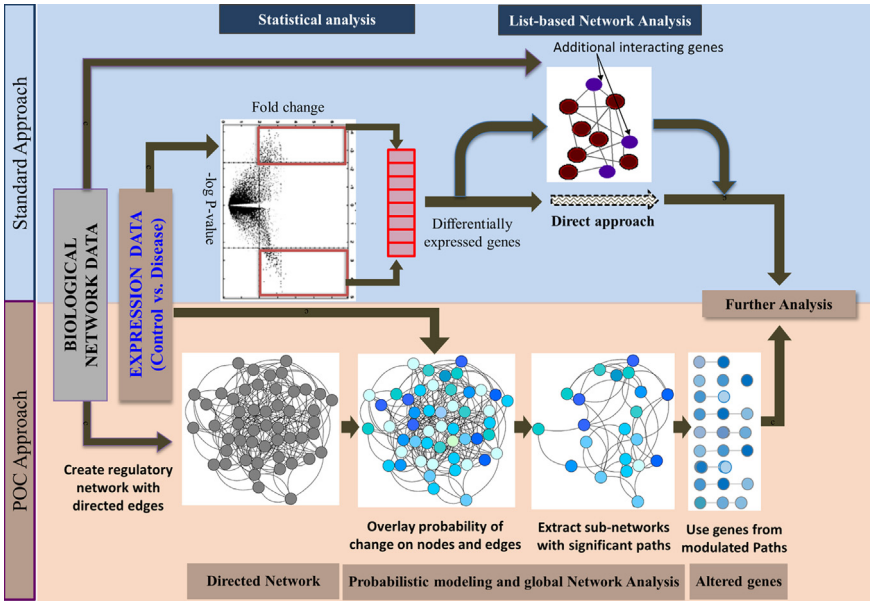

Fig. 1. In the standard approach (top portion of figure) genes are analyzed for fold-change and *p*-value (statistical analysis) in order to identify differentially expressed genes. These feature genes are used, in some approaches, to seed a network and include additional genes before the list of genes is submitted for further analysis. In some cases, the list is directly used for further analysis, which may include discriminative machine learning and functional exploration using pathway databases such as DAVID. The bottom portion of the figure illustrates the POC approach. The POC approach begins by building a putative regulatory network using a combination of database. This network is overlaid with node and edge probabilities obtained from the POC analysis. A series of network analysis algorithms leads to selection of maximally altered paths (sequence of directly connected genes). The genes in the set of identified paths is then submitted for further analysis. In our analysis, the list of genes is further analyzed using a standard machine learning algorithm to evaluate the discriminatory power of the genes obtained by POC.