Gene 590 (2016) 250–262

Contents lists available at ScienceDirect

Gene

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

Research paper

Construction of synergy networks from gene expression data related to disease

Prantik Chatterjee, Nikhil Ranjan Pal *

Electronics and Communication Sciences Unit, Indian Statistical Institute, Calcutta, India

ARTICLE INFO

Article history:
Received 15 October 2015
Received in revised form 11 March 2016
Accepted 17 May 2016
Available online 22 May 2016

Keywords:
Measures of synergy
Gene interaction network
Collaborative genes
Mutual information

ABSTRACT

A few methods have been developed to determine whether genes collaborate with each other in relation to a particular disease using an information theoretic measure of synergy. Here, we propose an alternative definition of synergy and justify that our definition improves upon the existing measures of synergy in the context of gene interactions. We use this definition on a prostate cancer data set consisting of gene expression levels in both cancerous and non-cancerous samples and identify pairs of genes which are unable to discriminate between cancerous and non-cancerous samples individually but can do so jointly when we take their synergistic property into account. We also propose a very simple yet effective technique for computation of conditional entropy at a very low cost. The worst case complexity of our method is $O(n)$ while the best case complexity of a state-of-the-art method is $O(n^2)$. Furthermore, our method can also be extended to find synergistic relation among triplets or even among a larger number of genes. Finally, we validate our results by demonstrating that these findings cannot be due to pure chance and provide the relevance of the synergistic pairs in cancer biology.

1. Introduction

Genes are the most fundamental unit of life. The origin of a disease may be found by analysing the expression pattern of genes. The relation of a gene with a disease may often be straightforward. For example, a gene may be under expressed in diseased subjects but over expressed in normal ones or vice versa. In this case, the gene is a biomarker and has a linear relation to the disease. There exist several methods to identify linear or non-linear disease specific biomarkers and gene regulatory networks from microarray or sequence data (Lin et al., 2013a; Alisoltani et al., 2015; Vineetha et al., 2012; Lin et al., 2013b) However, not all genetic interactions are so obvious. For example, there are pairs of genes, where the individual member of a pair does not exhibit any relation to a particular disease, but the pair, taken together, does reveal a strong relation to the disease. Such interactions are called synergistic interactions. Here we have proposed a new measure of synergy, which is more consistent with the concept of *synergy* compared to other measures. We have used this new measure to develop an algorithm for identifying synergistic gene-pairs and the synergistic network associated with prostate cancer. We have also demonstrated the relevance of the synergistic genes found by our method in cancer biology. Moreover, our algorithm is computationally more efficient and can identify better synergistic pairs compared to a state-of-the-art method.

Several approaches have been explored for identification of gene interactions and gene regulatory mechanisms in the past (Wu et al., 2012; Rizvi and Jauhari, 2014; Muraro et al., 2013). Boolean network models (Shmulevich et al., 2002), probabilistic boolean networks (Shmulevich et al., 2003; Zhou et al., 2003; Zhou et al., 2004), relevance networks (Butte and Kohane, 2000) and Bayesian networks (Friedman et al., 2000; Yu et al., 2002; Pearl, 1988) are good approaches to model and infer causal relationships from microarray data. However, they are limited by their inherent requirement of usage of greedy or heuristic algorithms on the regulatory network topology (Watkinson et al., 2009). Other methods such as pairwise mutual information (Butte and Kohane, 2000; Margolin et al., 2006; Zhao et al., 2008), regression techniques (Gardner et al., 2003) and graphical Gaussian models (Kishino and Waddell, 2000; Schäfer and Strimmer, 2005) have also been used to identify gene-gene interactions. However, our objective in this work is to discover interactions between pairs of genes which collaborate between themselves with respect to a particular disease (such as cancer) given a set of gene expression data on several samples both in presence and absence of the disease. This problem is fundamentally different from identifying linear biomarkers for a disease (Anastassiou, 2007; Leung and Hung, 2010; Zhu et al., 2010) But, what do we really mean by two genes cooperating with each other with respect to cancer and how do we properly quantify it?

* Corresponding author.
E-mail address: nikhil@isical.ac.in (N.R. Pal).

http://dx.doi.org/10.1016/j.gene.2016.05.029
0378-1119/

1.1. Understanding the concept of synergy

First, let us demonstrate the cooperative behaviour, found between two genes with the help of some examples from a publicly available prostate cancer data set (Singh et al., 2002) containing the gene expression values of 12,558 genes in 102 samples of which the first 50 are non-cancerous and the rest are cancerous. Looking at Fig. 1(a), we can say that the expression value of HPN in a sample is sufficient by itself to determine whether that particular sample is cancerous or non-cancerous. By thresholding expression values of HPN in Fig. 1(a), we can more or less separate the cancerous and non-cancerous samples which is shown by the vertical line on the X-axis. However, looking at the plots of PTGDS in Fig. 1(b) and XBP1 in Fig. 1(c), we can see that no such clear separation exists between the expression values of cancerous and non-cancerous samples, i.e., neither PTGDS nor XBP1 alone is sufficient to determine whether a sample is cancerous or non-cancerous. But when the expression values of PTGDS and XBP1 are plotted together, the samples can be easily separated into their respective classes as is evident from Fig. 1(d). So, while neither PTGDS nor XBP1 alone is sufficient to determine whether a sample is cancerous or non-cancerous, when taken together they can separate the two classes of samples. This behaviour of PTGDS and XBP1 suggests that their joint association to cancer is much stronger than their respective individual association, which must be due to some cooperation between them in the cancer pathway.

Recent studies show that prostate cancer is linked with cellular damage from oxidative stress and the inhibition of the anti-apoptotic mechanisms (Varadan and Anastassiou, 2006; Ouyang et al., 2005). This is an example of synergistic behaviour among genes related to oxidative stress and genes related to anti-apoptotic mechanisms. Another example of synergistic behaviour is the interaction between CTLA-4 and HLA-DRB4 which affects thyroid function in Japanese population (Terauchi et al., 2003). CTLA-4 also interacts synergistically with HLA-DR15, another gene of the HLA family, in multiple sclerosis (Alizadeh et al., 2003). In these examples, a pair of genes collaborate with each other to regulate the manner in which a certain disease or phenotype is expressed.

In this context it is worth mentioning that there exists another kind of genetic interaction which is known as epistasis. In an epistatic interaction, the effects of an allele at a gene hide or suppress the effects of an allele in another gene. Synergistic interactions can be treated as a special case of epistatic interactions where genes or mutations are regulated(either enhanced or suppressed) by actions of other genes (Mackay, 2013; Cordell, 2002). Suppose, mutation of gene A produces phenotype X and mutation of gene B produces phenotype Y, but mutations of both genes A and B produce phenotype Z. Here gene A is epistatically related to gene B.

From the above discussion, it is clear that synergy refers to a situation where the entirety of a system can produce an effect beyond that

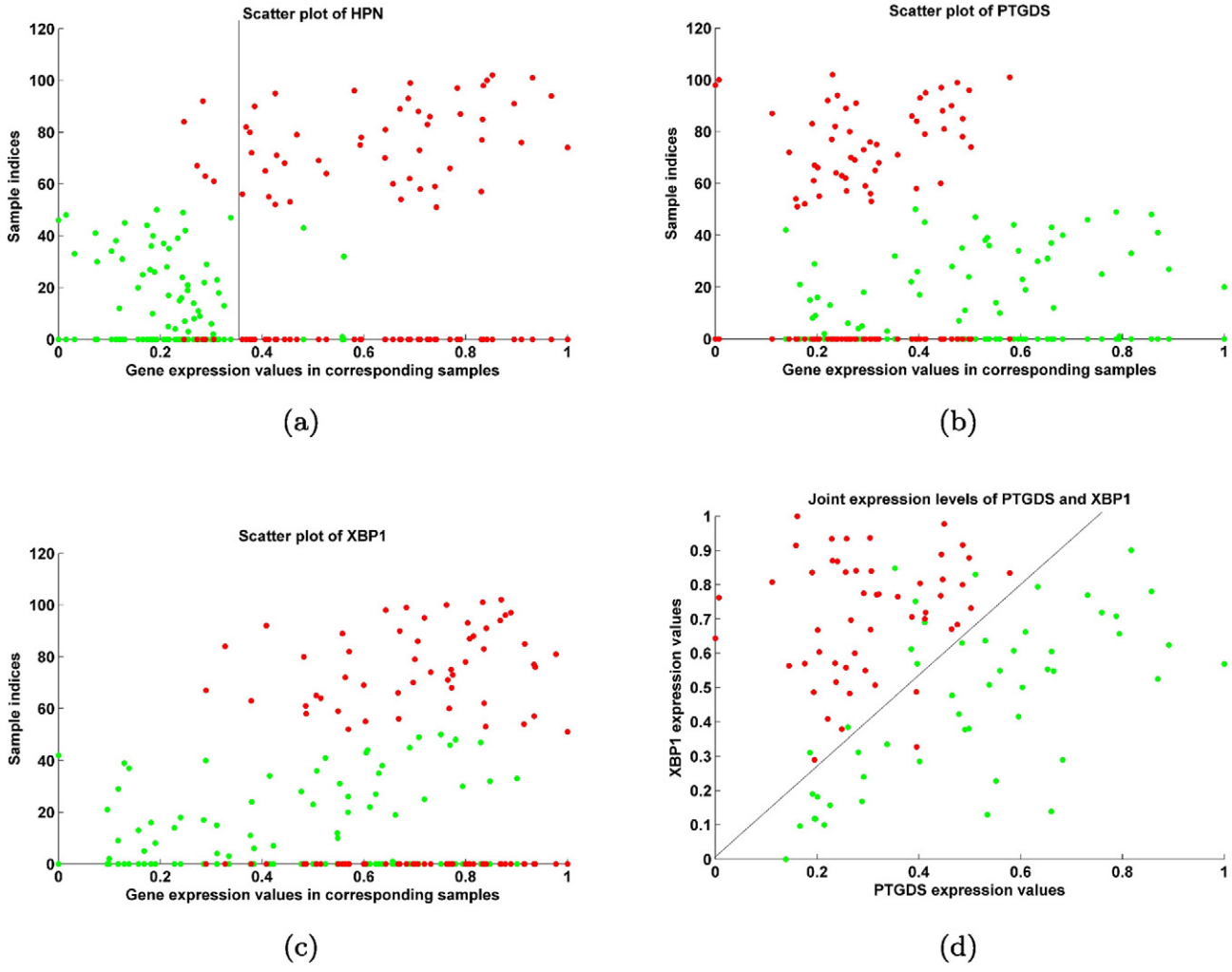

Fig. 1. The gene expression values of genes HPN, PTGDS and XBP1 are plotted along the X-axis and then projected on the X-axis for better understanding in panels (a), (b) and (c) respectively. The 50 non-cancerous samples are represented by green dots while the remaining 52 cancerous samples are represented by red dots. The joint expression levels of genes PTGDS and XBP1 exhibit a synergistic behaviour in panel (d).