Journal of Theoretical Biology 383 (2015) 20–27

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Network-based identification of reliable bio-markers for cancers

Shiguo Deng [a], Jingchao Qi [a], Mutua Stephen [a,b], Lu Qiu [a], Huijie Yang [a,*]

[a] Business School, University of Shanghai for Science and Technology,Shanghai 200093, China
[b] Computer Science Department, Masinde Muliro University of Science and Technology, P.O. Box 190-50100, Kakamega, Kenya

HIGHLIGHTS

- A spanning-tree based threshold is proposed to reconstruct gene networks from microarray data.
- Structural changes of gene networks are jointly used to seek high-confident bio-markers from candidates.
- From a total of 34 candidates collected from the literature we identified 16 high-confident informative genes, which perform best in distinguishing cancer samples from normal ones.

ARTICLE INFO

Article history:
Received 26 February 2015
Received in revised form
9 May 2015
Accepted 28 July 2015
Available online 4 August 2015

Keywords:
Gene expressions
Bio-marker
Complex network

ABSTRACT

Finding bio-markers for complex disease from gene expression profiles attracts extensive attentions for its potential use in diagnosis, therapy, and drug design. In this paper we propose a network-based method to seek high-confident bio-markers from candidate genes collected in the literature. The algorithm includes three consequent steps. First, one can collect the proposed bio-markers in literature as being the preliminary candidate; Second, a spanning-tree based threshold can be used to reconstruct gene networks for normal and cancer samples; Third, by jointly using of degree changes and distribution of the candidates in communities, one can filter out the low-confident genes. The survival candidates are high-confident genes. Specially, we consider expression profiles for carcinoma of colon. A total of 34 preliminary bio-markers collected from literature are evaluated and a set of 16 genes are proposed as high confident bio-markers, which behave high performance in distinguishing normal and cancer samples.

1. Introduction

Cancer originates from genome mutations induced by physical or chemical carcinogens (Croce, 2008; Lopez-Lazaro, 2010). Change in composition and/or sequence of base-pair in genes is displayed phenomenally in abnormities of gene distribution and expression profiles of genes. People have collected a large amount of gene expression data for samples of normal and cancer organisms as benchmark (to cite an example, ftp://smd-ftp.stanford.edu/pub/smd/). Consequently, how to category and identify (sub-)type of cancer from gene expression profiles becomes one of the essential problems at present time. The key step is to identify bio-markers representing (sub-)type of cancer, which will be great helpful in diagnosis, therapy, and drug design.

More than ten years of efforts have made great achievements (Diamandis et al., 2010; Mikula et al., 2011), but there exist still several problems to be dealt with. Mathematically, identifying bio-markers is a under-determined problem, because micro-array data may present simultaneously expression profiles for several thousands genes while only several tens of experiments are available. What is more, expression records are perturbed by unacceptable noises. Correlation coefficients from expression profiles show that generally many genes are closely correlated with each other. One must separate bio-markers from the genes having strong correlations with them (Alipanahi and Frey, 2013; Barzel and Barabasi, 2013; Feizi et al., 2013).

Different methods extract information from different viewpoints, and propose consequently sets of bio-markers which are different significantly (van de Kooy et al., 2002; Wang et al., 2005). Sometimes, there exist few overlapping genes in sets of bio-markers proposed by different methods. Hence, how to evaluate and integrate proposed gene markers is an essential problem.

Extensive researches (Chuang et al., 2007; Taylor et al., 2009) show that cancer is not a consequent result of simple superposition of mutations in cancer-related genes. One kind of cancer is usually related with many genes, while one gene may participate in occurrences of several kinds of cancers. The genes are networked into a complicated system and they function cooperatively to realize

[*] Corresponding author.
E-mail address: hjyang@ustc.edu.cn (H. Yang).

http://dx.doi.org/10.1016/j.jtbi.2015.07.026
0022-5193/© 2015 Elsevier Ltd. All rights reserved.

biological processes (Chen et al., 2012; Liu et al., 2014). Accordingly, present attentions are being focusing on changes of cellular networks in healthy and different cancer stages (Gu et al., 2014; Qin and Zhao, 2014; Wong et al., 2014; Zhang et al., 2014; Wu et al., 2014, 2014; Hu et al., 2014; Lu et al., 2014).

Identifying bio-markers based on networks has been well studied (e.g., the works reported in references Liu et al., 2012a, 2012b). In the present paper, we proposed a new network-based method to seek high-confident bio-markers from the candidates proposed in the literature. First, from the literature we collect bio-markers suggested by using different methods, as being the candidates of bio-markers; Second, from gene expressions we calculate cross-correlations between each pair of the genes and construct consequently the spanning-tree of the genes; Third, let us mark the candidates on the spanning-tree. If a pair of the candidates are linked on the spanning-tree, we record the linking strength. The minimal value among the recorded strengths is proposed herein as the threshold to filter out weak links in the correlation matrix, by using of which the cross-correlation matrix is mapped to a network. The survival links is believed to be high-confident (Alipanahi and Frey, 2013; Wang and Huang, 2014); Finally, by comparing structural behaviors of the healthy and cancer networks we filter out low-confident candidates and suggest a set of high-confident bio-markers.

As a typical example, we collect a total of 34 proposed bio-markers for carcinoma of colon in the literature. By comparison of structural patterns for normal and cancer networks (including degree changes and distribution in communities), we evaluate the collected bio-markers, and propose accordingly a set of 16 high confident bio-markers. Performance test show that the high-confident bio-markers can separate the healthy and cancer samples with highest precisions.

2. Materials and methods

2.1. Subset of genes related with carcinoma of colon

We consider micro-array data for carcinoma of colon (Alon et al., 1999), which contains expression profiles for 22 normal and 40 carcinoma of colon organisms (see Table A1 in Appendix). Each sample includes totally 2000 genes. Herein, we filter out the genes whose expressions are independent with sample types, i.e., there is not distinguished differences between the expression levels in cancer and normal samples. The concept of information index classification (RUAN, 2005) is used to measure the difference of expression levels, which is defined as,

$$IIC(i)=\frac{1}{2}\frac{|\mu_1(i)-\mu_2(i)|}{\sigma_1(i)+\sigma_2(i)}+\frac{1}{2}\ln\left(\frac{\sigma_1(i)^2+\sigma_2(i)^2}{2\sigma_1(i)\sigma_2(i)}\right), \qquad (1)$$

where $\mu_1(i), \mu_2(i)$ and $\sigma_1(i), \sigma_2(i)$ are average expression levels and standard square deviations of gene i for normal and carcinoma of colon samples, respectively. Generally speaking, the difference between two expression levels can be measured by two contributions. The primary contribution is the average of expression levels (as described by the average over all samples in the first term), while the secondary contribution is the details in distributions of expression levels (as measured by the standard deviation of distribution of expression levels).

Table 1
A total of 34 preliminary candidates of bio-markers collected from the literature.

Gene ID	Node ID	GenBank Acc. No	Mapped region
14	101	H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN).
245	7	M76378	Human cysteine-rich protein (CRP) gene, exons 5 and 6.
249	5	M63391	Human desmin gene, complete cds.
415	26	T60155	ACTIN, AORTIC SMOOTH MUSCLE (HUMAN).
493	1	R87126	MYOSIN HEAVY CHAIN, NONMUSCLE (Gallus gallus)
513	13	M22382	MITOCHONDRIAL MATRIX PROTEIN P1 PRECURSOR (HUMAN).
581	79	T51571	P24480 CALGIZZARIN.
625	10	X12671	Human gene for heterogeneous nuclear ribonucleoprotein (hnRNP) core protein A1.
792	153	R88740	ATP SYNTHASE COUPLING FACTOR 6, MITOCHONDRIAL PRECURSOR (HUMAN).
822	17	T92451	TROPOMYOSIN, FIBROBLAST AND EPITHELIAL MUSCLE-TYPE (HUMAN).
897	12	H43887	COMPLEMENT FACTOR D PRECURSOR (Homo sapiens)
1060	18	U09564	Human serine kinase mRNA, complete cds.
1115	55	R97912	SERINE/THREONINE-PROTEIN KINASE IPL1 (Saccharomyces cerevisiae)
1227	102	T96873	HYPOTHETICAL PROTEIN IN TRPE 3'REGION (Spirochaeta aurantia)
1346	114	T62947	60S RIBOSOMAL PROTEIN L24 (Arabidopsis thaliana)
1387	92	L05144	PHOSPHOENOLPYRUVATE CARBOXYKINASE, CYTOSOLIC (HUMAN);contains Alu repetitive element;contains element PTR5 repetitive element.
1400	216	M59040	Human cell adhesion molecule (CD44) mRNA, complete cds.
1423	11	J02854	MYOSIN REGULATORY LIGHT CHAIN 2, SMOOTH MUSCLE ISOFORM (HUMAN);contains element TAR1 repetitive element .
1472	67	L41559	Homo sapiens pterin-4a-carbinolamine dehydratase (PCBD) mRNA, complete cds.
1473	91	R54097	TRANSLATIONAL INITIATION FACTOR 2 BETA SUBUNIT (HUMAN).
1494	19	X86693	H.sapiens mRNA for hevin like protein.
1570	301	H81558	PROCYCLIC FORM SPECIFIC POLYPEPTIDE B1-ALPHA PRECURSOR (Trypanosoma brucei brucei)
1635	4	M36634	Human vasoactive intestinal peptide (VIP) mRNA, complete cds.
1668	15	M82919	Human gamma amino butyric acid (GABAA) receptor beta-3 subunit mRNA, complete cds.
1671	70	M26383	Human monocyte-derived neutrophil-activating protein (MONAP) mRNA, complete cds.
1771	16	J05032	Human aspartyl-tRNA synthetase alpha-2 subunit mRNA, complete cds.
1772	14	H08393	COLLAGEN ALPHA 2(XI) CHAIN (Homo sapiens)
1843	6	H06524	GELSOLIN PRECURSOR, PLASMA (HUMAN).
1892	85	U25138	Human MaxiK potassium channel beta subunit mRNA, completecds.
1897	74	U19969	Human two-handed zinc finger protein ZEB mRNA, partial cds.
1917	259	M91463	Human glucose transporter (GLUT4) gene, complete cds.
1924	86	H64807	PLACENTAL FOLATE TRANSPORTER (Homo sapiens)
1935	188	X62048	H.sapiens Wee1 hu gene.
1967	21	T60778	MATRIX GLA-PROTEIN PRECURSOR (Rattus norvegicus)