Expert Systems with Applications 36 (2009) 11589–11594

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Reproducible gene selection algorithm with random effect model in cDNA microarray-based CGH data

Mijung Kim *

Institute for Mathematical Sciences, Yonsei University, 134 Shinchon-Dong, Seodaemun-Gu, Seoul 120-752, Republic of Korea

ARTICLE INFO

Keywords:
cDNA microarray
Comparative genomic hybridization (CGH)
Copy-number changes
Gastric cancer
Reproducibility
Random effect model

ABSTRACT

cDNA microarray-based CGH with 30 pairs of normal and tumor gastric tissues using cDNA microarrays containing 17,000 human genes was performed to delineate the individual genes that undergo copy-number changes. Frequency analysis is more efficient than mean analysis for detecting subtle differences in copy-number when most of the data are from low spot intensities, such as those seen when performing cDNA microarray-based CGH. This article studies on how to deal with variation of data in replicated measurements for application of frequency analysis. A reproducible gene selection algorithm was developed for minimizing variation across array measurements. This algorithm incorporates a measurement of reproducibility with a random effect model and collects individual genes with reproducible copy-number change as a filtering process. This algorithm controls both reproducibility and number of remaining genes by dropping genes with large variations and results in increased reproducibility. Application of this algorithm allows for obtaining a well-filtered set of genes, thus dealing with variation in frequency analysis of the replicated data.

1. Introduction

The Cancer Metastasis Research Center (CMRC) at Yonsei University conducted a cDNA microarray-based CGH study to investigate gastric cancer-related DNA copy-number changes; this type of study makes it possible to observe a diverse pattern of potential biomarkers at the DNA level. cDNA microarray-based CGH was performed on 30 pairs of normal and gastric tumor tissues, and direct comparisons were made to detect gastric cancer-related genes with copy-number changes. The primary purpose of this experiment was to identify copy-number changes in individual genes rather than in segments of genes.

For gene-by-gene identification of copy-number changes, Yang et al. performed simple frequency analysis and selected genes showing at least one alteration in gastric cancer (Yang, Seo, & Jeong, 2005). Seo et al. have also investigated individual genes for copy-number changes in bilateral breast cancer (Seo, Rha, & Yang, 2004); Cheng et al. analyzed array CGH based on a gene-by-gene search through array rank order to detect copy-number changes in human cancer (Cheng, Kimmel, Neiman, & Zhao, 2003).

cDNA microarray-based CGH data include special features such as more meaning for occurrence of copy-number changes rather than the quantity of means for intensity difference. In addition, many of the data have low signal-to-noise ratios. Since cDNA microarray-based CGH experiments produce low-intensity spots, several issues are raised with data analysis for detecting subtle differences in copy-number change. One such issue is that only a few genes are identified as altered due to their small mean values. Analysis using mean values, such as the *t*-test, does not always successfully identify 'altered gene' where the change in the mean copy-number should be defined higher than the minimal criterion for an alteration. To detect subtle differences in copy-number change, frequency analysis is more efficient than mean analysis.

A second issue is to deal with variations of gene over the arrays in frequency analysis because there are many genes with relatively large variations compared to the total variation of all the genes; frequency analysis, such as a 1.5- or 2-fold change cut-off, does not consider variations of the gene over the arrays. This article suggests that fold change cut-off system is incorporated after filtering genes with small variations by utilizing the developed algorithm which helps it possible to obtain a well-filtered set of genes.

As an application of the proposing algorithm, it is possible to select candidates for altered genes by choosing genes with a high frequency of alteration at a set filtered with the algorithm; the selected genes include genes identified by *t*-test, as well as altered genes with a high frequency of one-side alteration and small means that such mean-utilizing statistical tests rarely detect due to the fact that the minimal criterion for an alteration in mean copy-number change was not satisfied.

As another application of the proposing algorithm, it might be utilized for detecting genes that displayed subtle but 'consistent'

* Tel.: +82 2 2123 4093; fax: +82 2 363 4845.
E-mail address: mjkim@yonsei.ac.kr

0957-4174/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2009.03.034

differences in copy-number change, by increasing reproducibility of the data through lowering threshold of the algorithm; the term ‘consistent’ gene is used to denote either only gain or only loss of gene copy-number (that is, one side alteration) in all available arrays. This allows for ranking genes with their frequency of consistent alteration. Genes with relatively large copy-number variations compared to the total variation of all the genes often reveal mixtures of gain and loss, which are referred to ‘hampering genes’; when a matter of concern is not hampering but consistent genes the developed algorithm is applicable for selecting set of consistent genes by increasing reproducibility.

Several ways of minimizing variation in replicated measurements have been reported. Data with large standard deviation (sd.) within experiments have been filtered out (Alizadeh, Eisen, & Davis, 2000; Marton, Derisi, & Bennett, 1998; Ross, Scherf, & Eisen, 2000; White, Rifkin, Hurban, & Hogness, 1999), and Kodota et al. introduced PRIM (Preprocessing Implementation for Microarray) to filter out data with small mean correlations between any two replicates (Kadota et al., 2001).

In this study, a reproducible gene selection algorithm (RGSA) was developed as a solution to the issues discussed above. In RGSA, the variability of replicated—measurements was quantified via a random effect model and a measurement of reproducibility was incorporated using intra-class correlation coefficient. RGSA controls both reproducibility and number of remaining genes. The well-filtered set this article suggests has both reproducibility and number of remaining genes maximized. In addition to the recommending filtered set, more reproducible genes are collectable by lowering threshold in RGSA. In this case, genes showing both gain and loss with low reproducibility are dropped and it is possible to rank genes based on their frequency of consistent alteration when categorizing the data according to the criterion on alteration.

For dealing with variations of the replicated data in frequency analysis, this article suggests to perform the frequency analysis for the set filtered by RGSA.

cDNA microarray-based CGH data whose experiment was conducted at CMRC of Yonsei University was utilized for testing RGSA, noticing that the application was done with purpose for testing RGSA and illustrating application of RGSA, therefore not final analysis for the experiment. For use of this data, within-print tip, intensity-dependent normalization was performed before taking the steps of RGSA.

For the purpose of comparison of the sets before and after RGSA applied, a filtered set is chosen for set with reproducibility increased by 30% of CMRC data, in addition to the suggesting filtered set. Characteristics are compared before and after application of RGSA.

Simulation study shows the sensitivity of RGSA for detecting data with large variation reaches 32–79% at the suggesting filtered set, and it increases to 73–96% when a set of more reproducible genes is selected, according to the variation of the simulated data.

2. Materials and methods

2.1. cDNA microarray-based CGH

Thirty pairs of normal gastric mucosa and cancer tissues were obtained from gastric cancer patients who had undergone surgery at the Severance Hospital, Cancer Metastasis Research Center (CMRC), Yonsei University Health System, Seoul, Korea, from 1997 to 1999. The patients consisted of 27 males and 3 females with a median age of 65 years (41–78). The numbers of patients in each stage were 3, 9, 12 and 6 for stage I, stage II, stage III and stage IV, respectively.

Genomic DNA extraction was performed according to a conventional protocol using the phenol/chloroform/isoamyl alcohol method. The cDNA microarrays containing 17,000 sequence-verified human gene probes (CMRC-Genomictree, Korea) were used for CGH in a direct comparison design, where genomic DNAs from the normal and tumor tissues were labeled with fluorescent dyes Cy3 and Cy5, respectively, and cohybridized, following the standard protocol of CMRC, Yonsei University (Yang et al., 2005). The range of genomic copy-number in normal tissues was within ±0.3 of the $\log_2$ intensity ratios in autosomal genes (Park, Jeong, & Choi, 2006), and thus a gain in copy-number of the gene was identified if the $\log_2$ intensity ratio was over 0.3, and loss was identified if the ratio was below -0.3. The experiment was performed with direct comparisons (Churchill, 2002). The cDNA microarray-based CGH data for this experiment has been deposited into Array Express (http://www.ebi.ac.uk/arrayexpress/) Query:1283947172 E-TABM-171.

2.2. Data preparation

17K cDNA microarray contained the 15,723 unique genes with 17,664 spots and these unique genes were mapped for their chromosomal location using SOURCE (http://genome-www5.stanford.edu/cgi-bin/source/sourceSearch) and DAVID (http://apps1.niaid.nih.gov/david/).

Let R and G denote the fluorescent intensities of tumor and normal hybridizations, respectively. For the evaluation of relative intensity, $Y = \log_2(R/G)$ was used, and data were pre-processed with the following considerations: first, within-print tip, intensity-dependent normalization of Y was performed as described (Yang, Dudoit, & Luu, 2002); second, genes showing missing values for >20% of the total number of observations were deleted; third, the 10-nearest neighbor method was employed for imputation of missing values; and fourth, averaged values were used in cases with multiple spots. In this step, 10,514 genes were found in 30 microarrays, and this data set was designated *BF*. Reproducibility of the data among arrays in the initial data set was 17.74%.

2.3. Statistical method

2.3.1. Random effect model establishment

The random variable Y of $\log_2$ for the ratio of intensities is assumed to follow normal distribution. For the ith gene and jth array, the gene-based statistical model for $\log_2$ intensity ratio, y_{ij}, was as follows:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \tag{1}$$

where μ is the mean effect and α_i is the random effect of the ith gene, which explains the gene’s intrinsic variability. ε_{ij} is a random variable reflecting variation from sources other than those identified by the gene’s effect. The underlying mean for the ith gene is given by $\mu + \alpha_i$, where α_i is drawn from a normal distribution with mean 0 and variance σ_A^2. ε_{ij}, is assumed to be drawn from a normal distribution with mean 0 and variance σ^2. This model is referred to as a random effects one-way analysis of variance model.

2.3.2. Measurement of two types of variation and reproducibility

To measure variation between replicate measurements, variation is decomposed into two components. The first is the intrinsic variation of the genes, denoted by σ_A^2 which is the extent of the ‘between-gene’ variation. The second, denoted by σ^2, is the variation between replicates, including measurement error, which is the ‘within-gene’ variation. The ratio of σ_A^2 to $\sigma_A^2 + \sigma^2$, denoted by ρ, explains how closely the gene measurements of one array track the gene measurements of another. When σ_A^2 is relatively large