A method for finding putative causes of gene expression variation

Autio, Reija; Hautaniemi, Sampsa; Ringnér, Markus; Kauraniemi, Päivikki; Edgren, Henrik; Yli-Harja, Olli; Astola, Jaakko; Kallioniemi, Anne; Kallioniemi, Olli-Pekka

Published in:
2nd TICSP Workshop on Computational Systems Biology, WCSB 2004

2004

Citation for published version (APA):

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 19. Dec. 2018
A METHOD FOR FINDING PUTATIVE CAUSES OF GENE EXPRESSION VARIATION

Reija Autio1, Sampsa Hautaniemi1, Markus Ringnér2, Päivikki Kauraniemi3, Henrik Edgren4,5, Olli Yli-Harja1, Jaakko Astola1, Anne Kallioniemi3, Olli-Pekka Kallioniemi4

1Institute of Signal Processing, Tampere University of Technology, P.O. Box 553, Tampere FIN-33101, Finland, 2Department of Theoretical Physics, Lund University, Lund, Sweden, 3Laboratory of Cancer Genetics, Institute of Medical Technology, Tampere FIN-33520, Finland, 4VTT Biotechnology, P.O. Box 106, Turku FIN-20521, Finland, 5Biomedicum Biochip Center, P.O. Box 63, FIN-00014, Helsinki, Finland

ABSTRACT

The majority of microarray studies evaluate gene expression differences between various specimens or conditions. However, the causes of this variability often remain unknown. Our aim is to identify underlying causes of these patterns, a process that would eventually enable a mechanistic understanding of the deregulation of gene expression in cancer. The procedure consists of three phases: pre-processing, data integration and statistical analysis. We have applied the strategy to identify genes that are overexpressed due to amplification in breast cancer. The data were obtained from 14 breast cancer cell lines, which were subjected to cDNA microarray based copy number and expression experiments. The result of the analysis was a list that consisted of 92 genes. This set includes several genes that are known to be both overexpressed and amplified in breast cancer. The complete study was published in Journal of the Franklin Institute 2004 [1], and in this paper we focus on the main issues of the study.

1. INTRODUCTION

We are interested in attributing the variability of expression levels of genes across multiple samples to either intrinsic or extrinsic features. In this study we present a method for identifying putative causes of gene expression variation.

We assume that for each gene expression value there is a corresponding explanatory value. The explanatory value could be another microarray measurement, gene ontology term, promoter sequence etc. The procedure allows missing values, so actually we assume that for each gene expression value, there is the possibility to obtain an explanatory value.

2. SYSTEMATIC PROCEDURE FOR EXPLAINING GENE EXPRESSIONS

In order to identify the impact of an explanatory variable on gene expression we present a general and systematic procedure to be used in explaining the gene expression variation across a set of experiments or samples. The strategy consists of three stages: pre-processing, data integration and statistical analysis. The schematic is illustrated in Figure 1. The heart of the strategy is the data integration, which consists of labeling and weighting.

2.1. Preprocessing

Preprocessing includes both within slide and between slide normalizations. Since preprocessing is strongly dependent on the quality of the data and the purpose of the experiment we make no assumptions about regarding the applied preprocessing method and any sensible preprocessing method is applicable.

2.2. Data integration

The core of the procedure is the data integration stage, in which explanatory data and expression data are first preprocessed and then integrated. Finally, statistical significance is computed using weights, labels and gene expression data.
where $m_1$, $\sigma_1$ and $m_0$, $\sigma_0$ denote the sample means and sample standard deviations for the expression levels for amplified and non-amplified samples, respectively. Signal-to-noise statistics results in a large weight if the means of the groups are far away from each other and standard deviations within the groups are small.

2.3. Statistical analysis

A large W does not necessarily mean that the gene’s expression variation can be explained by the explanatory phenomenon, since, depending on the algorithm chosen in the labeling and weighting phases, some misclassifications are likely to occur. Therefore, the final stage in our procedure is to compute statistical significance for the weighting. In this study we used permutation tests to test if a large weight for a gene is really due to the explanatory phenomenon.

3. CASE STUDY

We have applied the strategy to identify genes that are overexpressed due to amplification in breast cancer. The data were obtained from 14 breast cancer cell lines, which were subjected to cDNA microarray based copy number and expression experiments. The materials and the methods for the CGH and the gene expression experiments are given in [3, 4, 5].

The quality of the data included in to this study was good and therefore we performed within-slide normalization for both cDNA and CGH experiments. We also filtered out all spots whose mean red and mean green intensities were under 100 fluorescent units. Furthermore, the spots with area smaller than 50 pixels were discarded.

We have integrated CGH-Plotter to this procedure and we label the data with CGH-Plotter which is a MATLAB toolbox for identifying the copy number changes in microarray data [2]. CGH-Plotter identifies amplicons and deletions from microarray based copy number data and the resulted amplified genes are labeled with ‘1’ and the rest of the genes with ‘0’. For each gene, the expression levels of the amplified and non-amplified groups are calculated and a weight value for the difference between these two groups is computed with the signal-to-noise statistics.

Finally, statistical significance of the weight values is assessed with permutation tests [6]. We carried out 10,000 permutations and obtained an $\alpha$-value for each gene. The low $\alpha$-value indicates a strong association between gene expression and amplification.

The result of the analysis was a list that consisted of 92 genes. This set includes several genes that are known to be both overexpressed and amplified in breast cancer such as EGFR and ERBB2, as well as novel ones, such as HOXB7 gene, which was validated with RT-PCR and FISH.

4. CONCLUSION

We have shown a systematic approach for identifying genes whose expression levels are significantly influenced by an explanatory phenomenon. Since genes that undergo amplification or other ‘‘genetic change’’ in cancer may be the primary ‘‘driver genes’’ of cancer development and progression, the procedure enabled us to quickly identify a small subset of genes for further analysis. This approach is therefore highly valuable in trying to prioritize and simplify the most essential gene expression information in cancer. The crucial phase in our strategy is the labeling. If the labels in it are erroneous, they cannot be compensated in $\alpha$-value computation. However, permutation tests could be used in assessing statistical significance to labels.

5. REFERENCES


