A STRATEGY FOR IDENTIFYING PUTATIVE CAUSES OF GENE EXPRESSION VARIATION IN HUMAN CANCER

Sampsa Hautaniemi[1], *Markus Ringnér*[2], *Päivikki Kauraniemi*[3], *Anne Kallioniemi*[3], *Henrik Edgren*[4], *Olli Yli-Harja*[1], *Jaakko Astola*[1], *Olli-P. Kallioniemi*[2]

[1]Institute of Signal Processing, Tampere University of Technology, Finland, [2]Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, USA, [3]Laboratory of Cancer Genetics, Institute of Medical Technology, University of Tampere and Tampere University Hospital, Finland, [4]Biomedicum Biochip Center PL 63, Helsinki, Finland.

ABSTRACT

There is often a need to predict the impact of alterations in one variable on another variable. This is especially the case in cancer research, where much effort has been made to carry out large-scale gene expression screening by microarray techniques. However, the causes of this variability from one cancer to another and from one gene to another often remain unknown. In this study we present a systematic procedure for finding genes whose expression is altered by an intrinsic or extrinsic explanatory phenomenon. The procedure has three stages: preprocessing, data integration and statistical analysis. We tested and verified the utility of this approach in a study, where expression and copy number of 13,824 genes were determined in 14 breast cancer samples. The expression of 270 genes could be explained by the variability of gene copy number. These genes may represent an important set of primary, genetically "damaged" genes that drive cancer progression.

1. INTRODUCTION

Gene microarray experiments enable large-scale studies of gene expression. Gene expression patterns are useful for finding significant biological differences between samples from different kinds of tumors [1,7]. Most of the published studies on gene microarrays are descriptive. These studies typically describe genes and clusters of genes that are able to discriminate between two or more different tumor types, or between two or more biological treatments. Very little information is available on the underlying causes of the variability seen in gene expression patterns. We are interested in attributing the variability of expression levels of genes across multiple samples to either intrinsic (DNA sequence, biological role) or extrinsic features (measured with another method) of the genes. In this study we present a general and systematic procedure, which can be used in explaining gene expression variation across a set of experiments or samples.

The procedure consists of three stages: preprocessing, data integration and statistical analysis. Each stage is possible to implement according to purpose of the experiment. As preprocessing is very much dependent on the data, we are not discussing it in details. The heart of the procedure is data integration, where data from an explanatory phenomenon are combined with the gene expression data. In the statistical analysis stage the genes are ranked so that the top of the resulting list contains genes whose expression levels are very likely due to the explanatory phenomenon.

We assume that before the procedure is applied the phenomenon explaining the gene expression variation is defined. Furthermore, we assume that for each gene expression ratio 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 possibility to obtain an explanatory value. We are not making normal distribution assumptions in any stage of the procedure. Therefore our procedure is truly general in the sense that it is not restricted to process only normal distributed data.

As an example of usefulness of the procedure, we provide a case study where we determine the impact of gene copy-number on gene expression levels in breast cancer.

The order of the study is as follows. First we explain the principles of our procedure, i.e. how to determine the influence of an arbitrary phenomenon to gene expression levels. This theoretical discussion is followed by a case study where the procedure is used for finding genes that are overexpressed as a result of increased copy-number. In the end we discuss modifications of the procedure when underexpressed genes are studied.

2. PRINCIPLES OF THE PROCEDURE

The procedure has three main stages: Preprocessing, integration of the datasets and statistical analysis. These stages are illustrated in Figure 1.

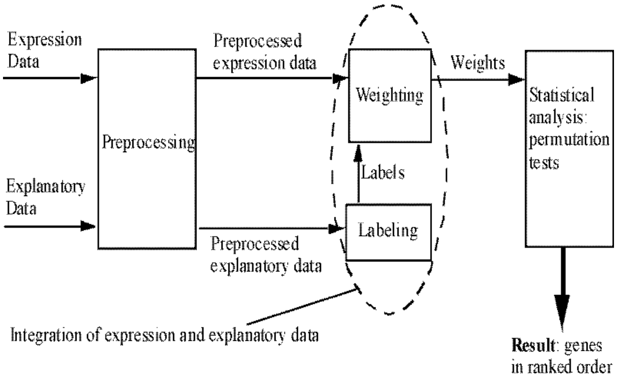
Expression Data
Explanatory Data
Preprocessing
Preprocessed expression data
Preprocessed explanatory data
Weighting
Labels
Labeling
Weights
Statistical analysis: permutation tests
Integration of expression and explanatory data
Result: genes in ranked order

Figure 1. Block diagram of the procedure.

The result of the procedure is a ranked list of genes. Top genes in the list are the ones having altered gene expression profile as a result of the explanatory phenomenon.

2.1 Preprocessing

In the preprocessing stage the data are modified so that they are comparable. Actual preprocessing algorithms are dependent on the data and the purpose of the experiment, so the scale of applied preprocessing and normalization methods is very large. There are no assumptions regarding to preprocessing, so any preprocessing technique is applicable.

2.2. Data integration

Data integration is done in two phases. The purpose of the first phase is to find data points from explanatory data set that may explain gene expression variation. This phase is referred as *labeling*. The second phase utilizes results from the labeling phase and associates them to gene expression data, and is referred to as *weighting*.

Labeling. In the labeling phase the explanatory data is divided into two groups. All data points in the first group (group 0) correspond to the situation where the following H_0 hypothesis holds: "this observation does not belong to the group that explains gene expression variation" and the data in the second group (group 1) to the H_1 hypothesis: "this observation belongs to the group that explains gene expression variation". The result of labeling is an index matrix, where entries are zeros and ones corresponding to the H_0 or H_1 hypothesis, respectively. The labeling matrix may contain missing values if there are such in the explanatory data set. Labeling is possible to execute with several statistical tests, clustering algorithms or *a priori* knowledge.

Weighting. In the weighting phase the index matrix obtained from the labeling phase is used to divide gene expression values into two groups. Then the goodness of separation (referred to as weight) for each gene is computed. The weight tells how well the expression values of one gene are separated into the two groups defined in the labeling phase using the explanatory data for that gene. Again, this phase can be utilized in almost arbitrary many ways. However, the algorithm should result in large weights when separation between the groups is good. One simple weighting method is to compute the mean of both groups and calculate their difference.

2.3. Statistical analysis

When a gene has large weight that 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 very likely to occur. Therefore, the final stage in our procedure is to analyze statistically the relevance of the weighting. One powerful statistical test class is permutation tests, which are used in this study. [4]

We used the permutation tests to test whether a large weight for a gene is really due to the explanatory phenomenon or not. The test is executed by permutating the label vector of the gene. The number of permutations is proportional to the number of samples in the groups 0 and 1. The number of data points belonging to group 0 and group 1 remains the same. In other words, permutation results in random groups whose sizes are the same as in the original grouping. The permutated labels are used for computing a new weight, which is compared to the old (original) weight. Pseudo-code for assessing α-value for one gene is illustrated below.

```
IN:   LabelVector (labels of gene x)
      cDNAData (cDNA values of gene x)

OUT: α-value

n := number of rounds, e.g. 10,000
counter := 0
$w_{old}$ = CompWeight(cDNAData, LabelVector)
```

```
repeat n times
    PermLabels := Permutate(LabelVector)
    $w_{new}$ := CompWeight(cDNAData, PermLabels)
    If $w_{new} > w_{old}$
        counter := counter +1
    end
end
$\alpha$ := counter/n
```

The result of the whole procedure is the probability that the null-hypothesis, i.e. “large weight is due to random event” is erroneously rejected. This probability is called the α-value.

Finally the genes are ranked according to their α-values so that the gene having the smallest α-value is ranked first, the gene having the second smallest α-value second etc.

3. CASE STUDY

In this section we illustrate how the procedure we have presented in earlier sections can be used for explaining the variation in gene expression profiles of breast cancer. We have elaborated the biological aspects of this case in [3].

Most functional genomic studies of cancer and other diseases are based on assessing steady-state expression levels of thousands of genes by cDNA microarrays. Our aim is to identify underlying causes of these patterns, a process that would eventually enable a mechanistic understanding of the dysregulation of gene expression in cancer. One important determinant of gene expression in cancer is variation in gene copy-number (by e.g. gene amplification), which can be measured by comparative genomic hybridization (CGH) [5,6]. We have used microarrays containing 13,824 genes to determine both the levels of gene expression (mRNA) and copy number (DNA) in 14 breast cancer specimens.

We arrange the expression data as a matrix with 13,824 rows and 14 columns. The explanatory phenomenon in this case is copy-number variation, which is observed by the CGH microarray experiment.

The first stage is preprocessing. We deleted measurements having low quality. In the cDNA data we discarded all ratios whose mean red (test sample) and green (reference sample) intensities were under 100 fluorescent units. Moreover, we discarded values whose area was less than 50 pixels. In the CGH data we deleted all ratios whose green intensity was below 100 fluorescent units. We did not exclude any gene from further analysis; only ratios of poor quality were discarded and treated as missing values. Then, CGH and cDNA calibrated intensity ratios were log-transformed and normalized using median centering of the values in each cell line. Furthermore, cDNA ratios for each gene across all 14 cell lines were median centered.

Preprocessing was followed by data integration. In labeling we assigned CGH values over 1.43 (5% of all CGH values) to group 1 (amplified) and the rest to group 0 (not amplified). In the weighting phase, we used the signal-to-noise statistic [2]:

$$w = \frac{m_1 - m_0}{\sigma_1 + \sigma_0},$$

where m_1, σ_1 and m_0, σ_0 denote the means and standard deviations for the expression levels for amplified and non-amplified cell lines, respectively.

Finally we applied a permutation test. For every gene we did 10,000 permutations and obtained an α-value for every gene. A low α-value indicates a strong association between gene expression and gene amplification. Figure 2 illustrates sorted α-values when $\alpha \in [0,0.1]$.

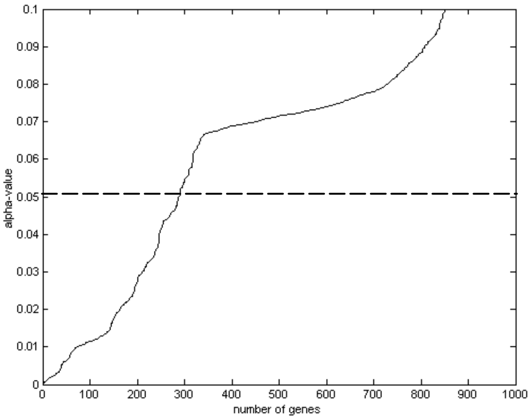

Figure 2. Sorted α-values (α< 0.1).

We defined our cut-point in α to be 0.05 (significant) resulting in 270 genes, as illustrated in Figure 2, that may be targets of gene amplification events in breast cancer. This set of genes includes several known oncogenes in breast cancer, such as the HER-2/ERBB2 oncogene as well as multiple novel genes included in amplicons at defined chromosomal loci, such as at 17q12, 17q23, 20q13 and 8q.

4. DISCUSSION

We have shown how this statistical analysis enabled us to quickly identify 270 genes whose expression levels potentially were due to an underlying gene amplification event in cancer. Since genes that undergo amplification or other “genetic damage” 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.

In the analysis of low-level expression changes it is very hard to distinguish whether the ratio is small due to biological reasons or noise. Signal-to-noise statistics consist of means and standard deviations, which are known to be sensitive to noise. Therefore, in the analysis of low-level expression changes, we need to assume more of the data and construct a model for the noise. After modeling the noise, we may try to filter the noise out from the measurements and use signal-to-noise statistics, or try an alternative weighting algorithm. For example, neural network based approaches such as support vector machines or learning vector quantization may turn out to be good choices.

In summary, we have developed a procedure that could be used in studies where the underlying causes of gene expression variations are examined. When we applied the procedure to explain overexpression in breast cancer study [3], the procedure was able to identify high-impact primary candidate gene targets for development of therapies and for sub-classification of breast cancer.

5. REFERENCES

[1] A. Alizadeh, M. Eisen, R. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson Jr, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, D. Botstein, P. Brown, L. Staudt. "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling", Nature, Vol.403, pp-503-511, 2000.

[2] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, E. Lander, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, Vol.286, pp. 531-537, 1999.

[3] E. Hyman, P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringnér, G. Sauter, O. Monni, A. Elkahloun, A. Kallioniemi, OP. Kallioniemi, "Impact of DNA amplification on gene expression patterns in breast cancer", Accepted to Cancer Research.

[4] J. Ludbrook, H. Dudley, "Why Permutation Tests are Superior to *t* and *F* Test in Biomedical Research", The American Statistician, American Statistical Association, Vol.52, No.2, pp. 127-132, 1998.

[5] O. Monni, M. Bärlund, S. Mousses, J. Kononen, G. Sauter, M. Heiskanen, Y. Chen, M. Bittner, A. Kallioniemi. "Comprehensive copy number and gene expression profiling of the 17q23 amplicon in human breast cancer", Proc Natl Acad Sci, USA, Vol. 98, pp.5711-5716, 2001.

[6] J. Pollack, C. Perou, A. Alizadeh, M. Eisen, A. Pergamenschikov, C. Williams, S. Jeffrey, D. Botstein, P. Brown. “Genome-wide analysis of DNA copy-number changes using cDNA microarrays”, Nature Genetics, 23, pp. 41-46, 1999.

[7] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, T. Golub. "Multiclass cancer diagnosis using tumor gene expression signatures", Proc Natl Acad Sci, USA, Vol. 98, no.26, pp.15149-15154, 2001.