Available online at www.sciencedirect.com

SciVerse ScienceDirect

Genomics Proteomics Bioinformatics 10 (2012) 127–135

GENOMICS PROTEOMICS & BIOINFORMATICS

www.elsevier.com/locate/gpb

Original Research

CDS: A Fold-change Based Statistical Test for Concomitant Identification of Distinctness and Similarity in Gene Expression Analysis

Nicolas Tchitchek [1], José Felipe Golib Dzib [1], Brice Targat [1], Sebastian Noth [1], Arndt Benecke [1,2,*], Annick Lesne [1,3]

[1] *Institut des Hautes Etudes Scientifiques, Bures-sur-Yvette 91440, France*
[2] *Centre National de la Recherche Scientifique, USR3078, Bures-sur-Yvette 91440, France*
[3] *Laboratoire de Physique Théorique de la Matière Condensée, CNRS UMR7600, Université Pierre et Marie Curie-Paris 6, Paris 75005, France*

Received 9 May 2012; revised 8 June 2012; accepted 10 June 2012
Available online 25 June 2012

Abstract

The problem of identifying differential activity such as in gene expression is a major defeat in biostatistics and bioinformatics. Equally important, however much less frequently studied, is the question of similar activity from one biological condition to another. The fold-change, or ratio, is usually considered a relevant criterion for stating difference and similarity between measurements. Importantly, no statistical method for concomitant evaluation of similarity and distinctness currently exists for biological applications. Modern microarray, digital PCR (dPCR), and Next-Generation Sequencing (NGS) technologies frequently provide a means of coefficient of variation estimation for individual measurements. Using fold-change, and by making the assumption that measurements are normally distributed with known variances, we designed a novel statistical test that allows us to detect concomitantly, thus using the same formalism, differentially and similarly expressed genes (http://cds.ihes.fr). Given two sets of gene measurements in different biological conditions, the probabilities of making type I and type II errors in stating that a gene is differentially or similarly expressed from one condition to the other can be calculated. Furthermore, a confidence interval for the fold-change can be delineated. Finally, we demonstrate that the assumption of normality can be relaxed to consider arbitrary distributions numerically. The Concomitant evaluation of Distinctness and Similarity (CDS) statistical test correctly estimates similarities and differences between measurements of gene expression. The implementation, being time and memory efficient, allows the use of the CDS test in high-throughput data analysis such as microarray, dPCR, and NGS experiments. Importantly, the CDS test can be applied to the comparison of single measurements ($N = 1$) provided the variance (or coefficient of variation) of the signals is known, making CDS a valuable tool also in biomedical analysis where typically a single measurement per subject is available.

Keywords: Statistical test; Fold-change; Distinctness; Similarity; Gene expression; Single measurement; Patient study

Introduction

The problem of identifying differentially expressed genes has been widely studied [1]. Considering two different biological conditions, one aims to decide which genes are differentially expressed from one biological condition to the other, each composed of one or several gene expression measurements. RNA quantification, which is being used in transcriptome analysis here will serve as an instance representative of any type of high-throughput quantification of cellular components such as DNA, RNA, protein, or metabolites, as the underlying problem of identifying statistically significant changes remains similar independent of the nature of the experiment. Therefore, all of what follows similarly applies to proteome or other measurements.

[*] Corresponding author.
E-mail: arndt@ihesfr (Benecke A).

http://dx.doi.org/10.1016/j.gpb.2012.06.002

For the sake of simplicity, we will only continue to discuss the case of gene expression investigations. First attempts to tackle the question of differential quantities did not involve statistics and genes having expression levels differing by more than an arbitrary cut-off fold-change value were considered to be differentially expressed [2,3]. Although the identification of statistically differentially expressed genes has been widely covered [1], the identification of similarly expressed genes has been far less studied. This is surprising, for several reasons. (i) Statistical measures for similarity are an important tool in establishing reproducibility and thus track technical and biological variation. (ii) In relative quantification, such as microarray experiments, where no absolute numbers of, *e.g.*, transcripts is established, a defining procedure for what is considered similar, or unchanged, expression would in turn also provide a sound basis for defining what is to be considered different. (iii) Finally, especially in the case of biomedical studies on human subjects and patients, the question of genes with conserved expression across different biological conditions is of similar importance to the one of change [4].

When reasoning in a statistical manner, assumptions can generally be made that gene expression measurements are normally distributed. The simplest statistical method for detecting differentially expressed genes is the two-sample *t*-test [5]. The two-sample *t*-test allows us to formulate statements concerning the difference between the means of two normally distributed variables with the assumption that the variances are unknown. On the other hand, the two-sample *z*-test allows us to formulate statements concerning the difference between the means of two normally distributed variables with the assumption that the variances are known. However as this assumption can only be made with a large sample of independent records or with additional information about the variances, the two-sample *t*-test is more often used in the identification of differentially expressed genes. Different variants of the two-sample *t*-test can be classified in two groups: (i) methods such as the two-sample *t*-test with relative thresholds [6] carrying out local adjustments to account for biologically meaningful differences, and the Significance Analysis of Microarrays method [7] that uses a gene-specific correction; and (ii) jointly global and local methods such as the B-statistic [8] and the regularized two-sample *t*-test [9]. In addition to simple fold-change or *t*-test-like methods, another approach is to consider the statistical properties of the ratio of means of the two biological conditions sampled. Based on the previous work [10], Chapman [11] proposed for the first time a statistical test in this direction. Recent methods (*e.g.*, [12,13]) extended this approach by considering confidence intervals for the statistic of the ratio of the two means used in hypothesis testing. When comparing different methods for differential expression detection, among the desirable characteristics that a method should have are reproducibility and control of type I and type II errors. Not all of the existing methods necessarily combine both characteristics [14]. Another way of comparing different methods is to measure their false positive and false negative rates [15].

Assume two sets of gene expression measurements obtained from two different biological conditions (**Figure 1**). By initially making the assumptions that the gene measurements are normally distributed with known variances, we represent the fold-change as the tangent of θ in Figure 1A. Having two biological conditions we can expect different scenarios. If both biological conditions have a small variance within biological replicates and then show differential expression, then methods should detect them as signifi-

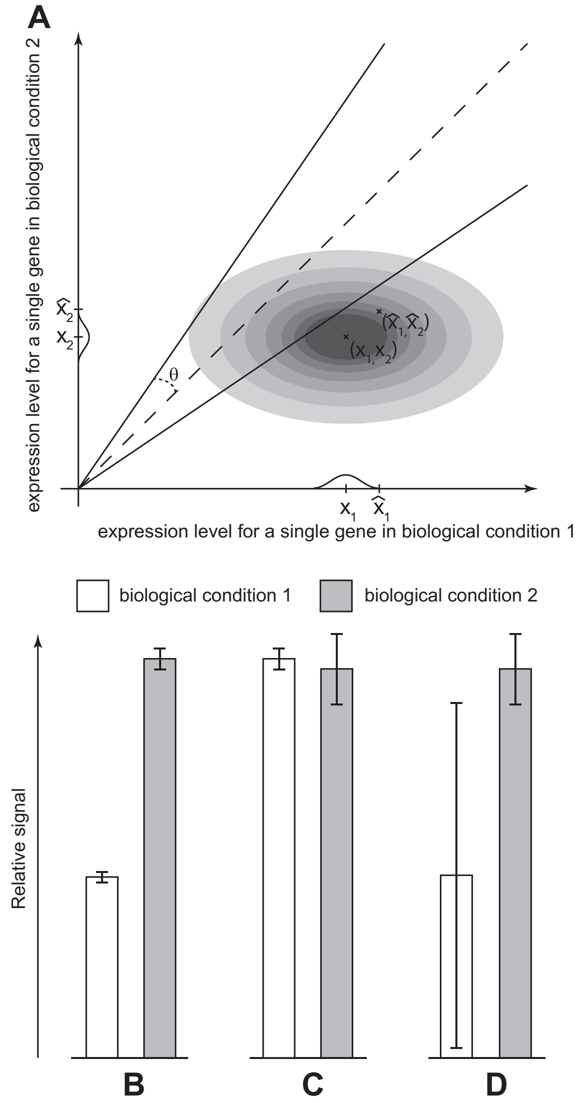
A
expression level for a single gene in biological condition 2
$\widehat{x}_2$
x_2
θ
$(\widehat{x}_1,\widehat{x}_2)$
(x_1,x_2)
x_1
$\widehat{x}_1$
expression level for a single gene in biological condition 1
□ biological condition 1
▨ biological condition 2
Relative signal
B
C
D

Figure 1 Graphical representation of the problematic and encountered scenarios

A. Expression signals of a single gene in two different biological conditions, with normal distributions having the parameters x_1 and x_2 (mean values) and σ_1 and σ_2 (variances). The fold-change criteria defining the difference or similarly is represented with a conic section defined by parameter θ. The problem is to determine the value of (x_1,x_2) having the values of estimators $(\widehat{x_1},\widehat{x_2})$. **B.** Potential scenario for the statistical test for differential expression and low variability. **C.** Potential scenario of having low variability and similarly expressed genes. **D.** Potential scenario for the statistical test for no statistical significance and high variabilities.