Analytica Chimica Acta 719 (2012) 8–15

Contents lists available at SciVerse ScienceDirect

Analytica Chimica Acta

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

Global test for metabolic pathway differences between conditions

Diana M. Hendrickx[a,b,c], Huub C.J. Hoefsloot[a,c,*], Margriet M.W.B. Hendriks[b,c], André B. Canelas[d], Age K. Smilde[a,c]

[a] Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
[b] Department of Metabolic Diseases, University Medical Centre Utrecht, Lundlaan 6, 3584 EA Utrecht, The Netherlands
[c] Netherlands Metabolomics Centre, Einsteinweg 55, 2333 CC Leiden, The Netherlands
[d] Department of Biotechnology, Kluyver Centre for Genomics of Industrial Fermentation, Delft University of Technology, Julianalaan 67, 2628 BC Delft, The Netherlands

ARTICLE INFO

Article history:
Received 22 August 2011
Received in revised form
30 November 2011
Accepted 20 December 2011
Available online 4 January 2012

Keywords:
Goeman's global test
Group statistic
Metabolomics
Score test
Pathway statistic

ABSTRACT

In many metabolomics applications there is a need to compare metabolite levels between different conditions, e.g., case versus control. There exist many statistical methods to perform such comparisons but only few of these explicitly take into account the fact that metabolites are connected in pathways or modules. Such a priori information on pathway structure can alleviate problems in, e.g., testing on individual metabolite level. In gene-expression analysis, Goeman's global test is used to this extent to determine whether a group of genes has a different expression pattern under changed conditions. We examined if this test can be generalized to metabolomics data. The goal is to determine if the behavior of a group of metabolites, belonging to the same pathway, is significantly related to a particular outcome of interest, e.g., case/control or environmental conditions. The results show that the global test can indeed be used in such situations. This is illustrated with extensive intracellular metabolomics data from *Escherichia coli* and *Saccharomyces cerevisiae* under different environmental conditions.

1. Introduction

Many current problems in metabolomics can be summarized as finding differences between conditions. The prototypical metabolomics biomarker study is an example: diseased versus control individuals are subjected to urine or serum metabolomics measurements and subsequently statistical methods are used to find the differences. This is mostly done using multivariate data analysis tools such as PLS-DA (Partial Least Squares Discriminant Analysis) [1,2], but also univariate tools are used [3]. Both tools have drawbacks, e.g., in univariate methods the multiple testing problem is present and in multivariate analysis model interpretation can be difficult. Shortcuts have been proposed, such as simplivariate models [4] that try to find groups of similarly behaving metabolites. Another route to tackle the problem is to use a priori biological information, such as the knowledge of pathways or modules.

Cellular processes arise as the result of many reactions between metabolic intermediates [5]. These reactions are functionally organized in pathways, which together form a large network. Most studies focused on relating changes in pathways to different conditions by using RNA micro-array data [6–9]. Here we describe the extension of a statistical tool, previously developed for analysis of RNA micro-array data, to the analysis of metabolomics data.

Studying statistics for a whole group of genes or metabolites avoids the often time consuming task of multiple testing for each gene or metabolite separately [10]. For metabolomics, predefined groups of pathways [5,11,12] or functional modules can be used in this approach. For example, in lipidomics, the test can be performed per lipid class instead of per lipid. Another advantage of group testing is that it can detect differences between conditions that are caused by subtle changes in several metabolites, which are difficult to discover by single metabolite testing [13].

Nam and Kim [14] distinguished three types of methods for testing pathways, depending on the hypothesis that is tested. The first kind of algorithms test if under particular conditions, a group of genes belonging to a certain pathway is differentially expressed compared with the rest of the genes in the data set (= H1 hypothesis), e.g. T-profiler [15] and PAGE (Parametric Analysis of Gene Set Enrichment) [16]. The second type of methods examines if a selected group of genes from the same pathway has a different behavior under a first condition, compared to a second condition (= H2 hypothesis), e.g. Goeman's global test [7] and SAM-GS (Significance Analysis of Microarray for Gene Sets) [17]. The third kind of methods, known as Gene Set Enrichment Analysis (GSEA), test the hypothesis that none of the predefined groups of genes in the data set is different between two conditions (= H3 hypothesis).

* Corresponding author at: Biosystems Data Analysis, Swammerdam Institute for Life Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands.
E-mail address: H.C.J.Hoefsloot@uva.nl (H.C.J. Hoefsloot).

0003-2670/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.12.051

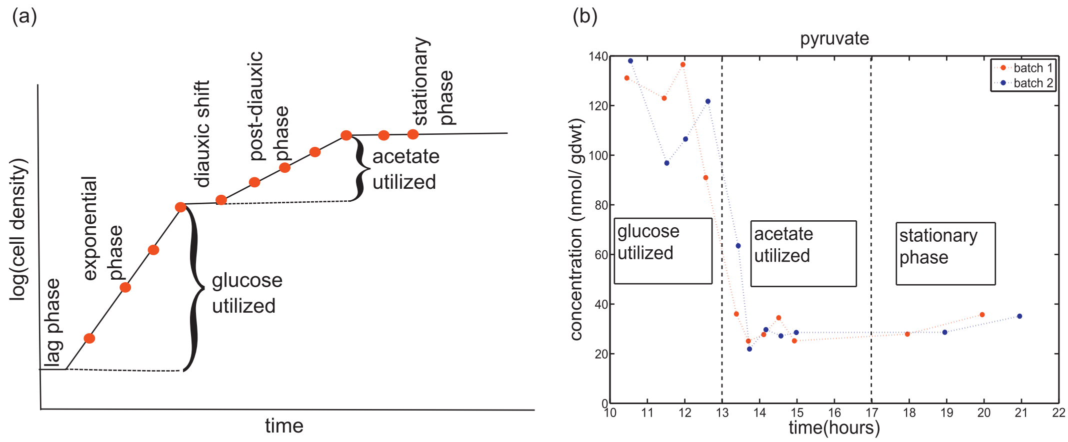

Fig. 1. (a) Diauxic growth curve. The red points indicate in which phases the measurements were taken. (b) An example of a metabolite concentration profile (pyruvate) under diauxic growth. The different growth phases are indicated on the graph. Abbreviations: nmol, nanomoles; gdwt, gram dry weight. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Two types of GSEA are developed: simple GSEA [18,19] and GSEA using linear models [20,21]. The tested groups of genes can be predefined groups from e.g. Gene Ontology or KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [5,11,12,18,13,19,21] or can be defined based on chromosome location [20,19]. Extensions of GSEA for metabolomics data have been implemented in the web-based tools MSEA (Metabolite Set Enrichment Analysis) [22], MPEA (Metabolite Pathway Enrichment Analysis) [23] and MBRole (Metabolite Biological Role) [24]. In Quantitative Enrichment Analysis (QEA), which is part of MSEA, the *Q*-statistic from Goeman's global test was used [22], but the method was not described in the literature about MSEA.

In the current paper, we explain the working of Goeman's global test for metabolomics in full detail. We discuss the usefulness of this test for establishing significant differences between conditions at the pathway (or module) level. We critically evaluate the validity of the method by using two worked out examples and studying the biological relevance of the test results. For the *Escherichia coli* data set, the test is applied to find pathways that are different under glucose growth compared to acetate growth. With the *Saccharomyces cerevisiae* data set, the behavior of glycolysis and the tricarboxylic acid (TCA) cycle under three sets of conditions is examined: aerobic versus anaerobic; glucose pulse versus short-term glucose deprivation (feed off); larger versus smaller glucose pulse. The results show that Goeman's global test can indeed be used in situations where one wants to know if a metabolic pathway is significantly related to a change in conditions.

2. Materials and methods

2.1. *E. coli* data set

GC–MS (gas chromatography–mass spectrometry) and LC–MS (liquid chromatography–mass spectrometry) data [25] of batch cultures on glucose of *E. coli* were obtained from TNO Quality of Life (Zeist, The Netherlands). During growth on glucose, acetate is produced. After depletion of glucose, there is a diauxic shift to acetate growth [26]. Sampling of two fermentation processes at eleven time points was performed: four time points in the exponential phase during growth on glucose, five in the post-diauxic phase (growth on acetate), and two in the stationary phase (all carbon sources exhausted) (see Fig. 1(a)). The data set consists of absolute concentrations (in nanomoles per gram dry weight) of metabolites from glycolysis, the tricarboxylic acid (TCA) cycle and biosynthesis of amino acids, nucleotides and nucleosides. The data are not equidistantly sampled: the time between two subsequent samples ranges from 0.5 to 2 h. The window of observation is from 10.5 to 20.5 h elapsed fermentation time (see example for pyruvate, Fig. 1(b)).

2.2. *S. cerevisiae* data set

LC–MS data [27–29] of continuous cultures[1] of *S. cerevisiae* were obtained from the Kluyver Centre for Genomics of Industrial Fermentation (Biotechnology Department, TU Delft, The Netherlands). The cells were cultivated to steady-state in glucose-limited chemostats under aerobic ($D = 0.1\,\mathrm{h}^{-1}$) or anaerobic ($D = 0.05\,\mathrm{h}^{-1}$) conditions. Furthermore, each steady-state was used to perform a short-term perturbation response experiment, by rapid addition of a concentrated pulse solution and withdrawing samples within a short time frame. Eleven aerobic and four anaerobic experiments were performed. Different perturbations were obtained depending on the composition of the glucose pulse solution. An overview is given in Table 1.

The data set consists of measurements of absolute metabolite concentrations (in micromoles per gram dry weight) from glycolysis and some of its branches and from the tricarboxylic acid cycle (TCA cycle). The data are not equidistantly sampled: in most experiments the sampling frequency is higher immediately after the pulse and decreases throughout the rest of the time series. The window of observation also differs between experiments.

2.3. Data pre-treatment

The intracellular concentrations of different metabolites can differ by more than five orders of magnitude [30]. Furthermore, the abundance of a given compound is not necessarily related to its biological importance [33]. Therefore, the data sets were autoscaled, so that all metabolite levels have zero mean and unit variance. In this way, all compounds are put on the same scale [32].

2.4. Goeman's global test

Assume that n samples of p metabolites are measured, of which m metabolites belonging to the same pathway are selected. Our selection of pathway metabolites is based on the Kyoto Encyclopedia of Genes and Genomes (KEGG) [5,11,12]. Let i be the index for

[1] Chemostat cultures, continuous inflow and outflow.