Genomics Proteomics Bioinformatics 11 (2013) 219–229

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ORIGINAL RESEARCH

In silico Proteome-wide Amino aCid and Elemental Composition (PACE) Analysis of Expression Proteomics Data Provides A Fingerprint of Dominant Metabolic Processes

David M. Good [1,#], Anwer Mamdoh [1], Harshavardhan Budamgunta [1], Roman A. Zubarev [1,2,*]

[1] *Division of Physiological Chemistry I, Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE 171 77 Stockholm, Sweden*
[2] *Science for Life Laboratory, SE 171 21 Solna, Sweden*

Received 22 February 2013; revised 29 May 2013; accepted 6 June 2013
Available online 3 August 2013

KEYWORDS

Shotgun proteomics;
Mass spectrometry;
LC–MS/MS;
Data reduction;
Cyanobacterium;
Arginine deprivation

Abstract Proteome-wide Amino aCid and Elemental composition (PACE) analysis is a novel and informative way of interrogating the proteome. The PACE approach consists of *in silico* decomposition of proteins detected and quantified in a proteomics experiment into 20 amino acids and five elements (C, H, N, O and S), with protein abundances converted to relative abundances of amino acids and elements. The method is robust and very sensitive; it provides statistically reliable differentiation between very similar proteomes. In addition, PACE provides novel insights into proteome-wide metabolic processes, occurring, *e.g.*, during cell starvation. For instance, both *Escherichia coli* and *Synechocystis* down-regulate sulfur-rich proteins upon sulfur deprivation, but *E. coli* preferentially down-regulates cysteine-rich proteins while *Synechocystis* mainly down-regulates methionine-rich proteins. Due to its relative simplicity, flexibility, generality and wide applicability, PACE analysis has the potential of becoming a standard analytical tool in proteomics.

Introduction

Modern proteomics analysis provides the identities and the relative abundance changes for thousands of proteins per a single LC–MS/MS experiment [1,2]. However, since many proteins have multiple functions and the exact function of many proteins is not yet known, this information is not always easy to rationalize. Pathway analysis [3,4] provides mapping of the proteome onto more than 160 known signaling pathways and dozens of metabolic pathways. Nonetheless, molecular

* Corresponding author.
E-mail: Roman.Zubarev@ki.se (Zubarev RA).
Current address: Department of Medicine, University of Wisconsin – Madison, Madison, WI 53706, USA.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

ELSEVIER Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.gpb.2013.07.002

pathways are often overlapping and inter-related, such a mapping is rarely unequivocal. A similar problem plagues the popular gene ontology (GO) mapping. Ideally, an aggregate analysis of the proteome state would involve mapping onto a reasonably small number orthogonal, *i.e.*, non-overlapping and mutually independent, classification factors that have clear physico-chemical interpretations. Although mutually orthogonal ("extreme") pathways have been constructed for microorganisms [5,6], such constructs are usually artificial, *i.e.*, do not have clear counterparts at the molecular level.

However, methods to reduce the proteome to a manageable number of orthogonal entities do exist. For example, proteins can be broken down into their constituent amino acids (AAs). Since amino acids in protein sequences are, in general, not mutually interchangeable (the evidence for which is their survival of the evolutionary pressure), they represent an orthogonal set for global proteome analysis. And since all organisms try to minimize the "cost" of protein synthesis by adjusting their AA content to specific growth conditions [7], it is reasonable to assume that changes in these conditions will be reflected in the abundances of the component AAs. Thus, a proteome-wide AA composition analysis can provide an aggregate fingerprint characterizing the specific state of a given organism.

Unfortunately, the current methods for AA analysis all possess significant drawbacks. Edman degradation [8], for instance, is limited with regard to the size of polypeptide which can be interrogated. Meanwhile, acid hydrolysis [9,10] followed by quantification with either ninhydrin [11–13] or mass spectrometry (MS) [14–17] is limited by exposing proteins to harsh chemical treatment, which in turn completely destroys unstable AAs, *e.g.*, tryptophan. Even a short hydrolysis duration leads to deamidation of asparagine and glutamine to aspartic acid and glutamic acid, respectively [10,18].

As will be shown below, the AA and element analyses of whole proteomes can provide valuable information on the ongoing metabolic processes. Here, we present a novel, non-destructive method of performing such analysis on quantitative data obtained in expression proteomics experiments. The entire Proteome-wide Amino aCid and Elemental composition (PACE) analysis is performed *in silico*, and as it can be applied to previously acquired data, it can provide fresh insights from earlier results without a requirement of new experiments. In addition, this method is platform-independent, *i.e.*, can be used for data generated with any mass spectrometric, and even non-mass-spectrometric (*e.g.*, laser fluorescence or antibody-based) quantitative proteomics platforms.

What relevant biological insights can PACE mapping provide? At a very basic level, it can answer the question of whether two given proteomes are different better than any other known statistical method while providing a quantitative estimate of this difference and associated *P* value. PACE mapping also yields a fingerprint of the dominant metabolic processes and, in some cases, even reveals their character. For instance, PACE analysis confirms that single-cell organisms deprived of a single element (*e.g.*, sulfur) during growth exhibit depletion of this element in their proteins [7]. Analyzing both our own and published data with PACE, we investigated the question of whether this depletion is proteome-wide or is instead concentrated in a few highly abundant proteins. We also used PACE to reveal which AA residues get depleted and to what degree. Processes not involving nutrient depletion (*e.g.*, cold or heat stress) also leave a specific mark in the PACE domain, which subsequently can be used as a fingerprint for their recognition. As a novel and informative way of interrogating the proteome, which combines relative simplicity, flexibility and wide applicability, PACE has the potential of becoming a standard analytical tool in proteomics.

Results

Distribution of PACE signal in the proteome

Until very recently, proteomics analyses were unable to reveal the entire expressed proteome due to the high dynamic range of protein expression. Thus, in any real-life experiment, a subset of the total expressed proteome is sampled, representing the most abundant part of the proteome. To investigate whether the partial nature of the proteomics data affects the PACE diagram, we analyzed a "deep proteomics" (> 50% of the expressed proteome) literature dataset of the model cyanobacterium *Synechocystis* sp. PCC 6803 [19]. The total list of ~2000 quantified proteins was randomly split into two halves, and a PACE AA (**Figure 1**) and elemental histogram (Figure S1) were produced for each of the half-proteomes. The visual similarity between the two histograms is confirmed by correlation analysis (**Figure 2**; $R^2 \geqslant 0.8$ for both correlations). This example demonstrates that the PACE signal is distributed throughout the whole proteome, and the partial nature of real-life proteomics data does not affect the PACE analysis fatally.

Detection of small differences between proteomes

To answer the question as to whether the observed proteome differences between two cellular states are statistically significant, one typically needs to use principal component analysis (PCA) or a similar statistical method to differentiate two groups, each consisting of multiple replicate analyses. In the absence of *a priori* knowledge of statistics associated with protein abundances (each protein being, strictly speaking, a separate statistical entity), there is no easy method to assign statistical significance to a difference, if only two proteomics datasets are available. However, this task becomes solvable with PACE analysis, as the following example demonstrates. In this example, a pair of measured proteomes (lists of ~500 protein identities and respective abundances; T1 and T2) represents two technical replicates of the same proteome B1, while a third measured proteome (B2) represents a separate biological replicate. The protein abundances of the same proteome analyzed repeatedly (technical replicates) are affected by random, statistically independent errors in the measured abundances of individual proteins, while non-identical but biologically similar proteomes (biological replicates) vary in a fundamentally different way, where abundances of the proteins within the same pathway are statistically linked. A simple comparison through the correlation coefficient *R* gives similar values when T1 and T2 are compared ($R^2 = 0.9999$) as well as for the similarity between T2 and B2 ($R^2 = 0.9989$), and provides no estimate for *P* values of the differences (Figure 2A). The failure of standard approaches to robustly differentiate between the biologically unique samples as compared to technical replicates of the same sample is further demonstrated by unsupervised PCA of the data (Figure 2A). Here, the PCA model yields a nonsensical negative Q2 value, illustrating the inability to separate these datasets from each other.