Talanta 83 (2011) 1289–1294

Contents lists available at ScienceDirect

Talanta

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

Application of parallel computing to speed up chemometrics for GC × GC–TOFMS based metabolic fingerprinting

Thomas Gröger[a,b], Ralf Zimmermann[b,a,*]

[a] Institute of Ecological Chemistry, Cooperation Group "Analysis of complex molecular systems", Helmholtz Zentrum München, 85764 Neuherberg, Germany
[b] Chair for Analytical Chemistry/Mass Spectrometry Centre – Institute for Chemistry, University of Rostock, 18051 Rostock, Germany

ARTICLE INFO

Article history:
Available online 16 September 2010

Keywords:
Parallel computing
Chemometrics
Comprehensive two-dimensional gas chromatography
Metabolic fingerprinting

ABSTRACT

Parallel computing was tested regarding its ability to speed up chemometric operations for data analysis. A set of metabolic samples from a second hand smoke (SHS) experiment was analyzed with comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC × GC–TOFMS). Data was further preprocessed and analyzed. The preprocessing step comprises background correction, smoothing and alignment of the chromatographic signal. Data analysis was performed by applying *t*-test and partial least squares projection to latent structures discriminant analysis (PLS-DA). The optimization of the algorithm for parallel computing led to a substantial increase in performance. Metabolic fingerprinting showed a discrimination of the samples and indicates a metabolic effect of SHS.

1. Introduction

The aim of metabolomics is a comprehensive quantitative and qualitative characterization of the metabolome of a biological system and its dynamics [1,2]. However, due to the large qualitative and quantitative diversity not all components and processes of the metabolome can be analyzed at the same time on one analytical platform. Therefore, different strategies have been established focusing on different biological tasks. Metabolic fingerprinting is focused on a relative comparison of biological systems based on their metabolomic patterns which could be addressed by one experiment or one analytical platform without optimizing the system for a certain small subset of metabolites. The strength of metabolic fingerprinting is its ability to screen and classify huge numbers of samples in short progression. Very common are hyphenated techniques like gas chromatographic–mass spectrometric (GC–MS) or liquid chromatographic-mass spectrometric (LC–MS) couplings [3–6]. The aim of such hyphenation is the separation of different metabolites and matrix before they enter the MS. Considering the complexity of metabolic samples much effort has been made to further increase the separation power of the analytical platforms or to adapt them for a special purpose. Basically, these approaches could be divided into two efforts.

The first one focuses on the analytical platform itself and tries to further enhance the selectivity or separation power of the hardware. With regard to the enhancement of the chromatographic side, higher dimensional separation techniques, like comprehensive two-dimensional gas chromatography (GC × GC) [7–9] in combination with a fast time-of-flight MS, have become very popular over the last years. Due to the introduction of a second orthogonal separation direction, the metabolites become separated over a plane. The increased selectivity of such systems leads to a higher separation power and offers also additional opportunities for data analysis of metabolomic data [10–15].

The second attempt concentrates on the application of chemometrics to further improve the physical/chemical separation. Chemometrics can be applied during data acquisition, data processing and/or data analysis [16–22]. Former attempts for application of chemometrics to GC or MS during data acquisition were utilized e.g. by Phillips [23]. Nowadays, the application of chemometrics for the preprocessing of chromatographic and/or mass spectrometric data [24–27] is more common. The main objectives are the enhancement of the analytical signal and its isolation from interfering signals [12,24,28]. A further field of chemometrics in GC/MS based metabolomics is the statistical analysis of the data [29–31]. The principle objects here are the classification of the different samples according to their metabolite pattern, (semi-)quantification and the identification of discriminating metabolites [32]. In any case, chemometrics has to be applied with care, since complex

* Corresponding author at: Institute of Ecological Chemistry, Cooperation Group "Analysis of complex molecular systems", Helmholtz Zentrum München, D-85764 Neuherberg, Germany. Tel.: +49 089 3187 4544; fax: +49 089 3187 3510.
E-mail address: ralf.zimmermann@helmholtz-muenchen.de (R. Zimmermann).

0039-9140/$ – see front matter © 2010 Published by Elsevier B.V.
doi:10.1016/j.talanta.2010.09.015

issues require a careful selection and interpretation of chemometric tools [33,34].

Both attempts can only be realized at the expense of data size and computationally intensive processing. While higher dimensional separation in combination with fast MS systems produce very large data sets as a consequence of high sample throughput and fast repetition rates of MS detection, chemometrical operations on these data sets can become very intense in resource demands and time, if the complete data set of many samples should be considered.

At the moment only a few vendors offer commercial software for GC × GC (–MS). These packages (e.g. Pegasus, LECO Corporation or GC-Image, Zoex Corporation) provide basic processing or analyzing tools and are highly suitable for target analysis due to a user-friendly and sophisticated graphical user interface. Yet state-of-the-art chemometric operations like proper alignment or multivariate statistics for a comprehensive non-targeted analysis are lacking. In addition, this software does not support state-of-the-art architectures like 64-bit, multi-core processing or emerging techniques like general purpose graphic processing units (GPGPU's). One opportunity could be the application of software packages developed for closely related data sets like two-dimensional gel electrophoresis. Since this sector has a larger sales volume the software is in most cases further developed and it would meet the requirements for the processing GC × GC data files. Recently published work [35] looks very promising but currently the adaptation is not ready for end user application. Appropriate algorithms can also be programmed based on popular programming languages like MatLab, R, and others.

This paper will focus on the implementation of parallel computing [36,37] for analysis of GC × GC–TOFMS data from metabolic fingerprinting to speed up chemometric operations [22,38] based on MatLab.

The main purpose of parallel computing is the ability to either distribute one large data block to smaller blocks or speed up a computer algorithm by distributing different data sets (e.g. from GC × GC–TOFMS) on different workers. Nowadays, the first approach is only relevant for 32-bit Windows systems in which a single application can address only about 3 GB. While data is often collected and stored in lower precision like integer, data processing is often based on double precision operations which increase the space needed in memory size of the data dramatically. Data sets from GC × GC–TOFMS often reach this boarder, at least if multivariate operations are part of the processing. With the introduction of 64-bit architecture and the adaptation of the software, the 3GB boarder has vanished. Now the limitation is the physically available memory of the computer system.

Of much more interest is the ability to speed up data processing by distributing the processing of data to different workers. A requirement for parallel computing is the feasibility to distribute the original data set and to do all further operations on such a distributed set. While the first necessity depends on the used software the second one depends on the structure of the data and the kind of data operation. A problem could be an algorithm which has to access data from the memory of another worker. Such inter worker processes would slow down the overall process due to excessive data transfer. Therefore, as a rule of thumb, the data has to be distributed in such a manner, that all workers can operate on their own. For that reason it could be applicable to redistribute the data set during operation to meet the requirements of each programming step. An example is the alignment of different sample chromatograms to a target chromatogram. Popular algorithms are based on piecewise shifting a small section of a sample chromatogram along a target chromatogram within predefined limits until some quality criterion is optimized. In case of GC × GC such an alignment has to be done in two dimensions. If the chromatogram is distributed among multiple workers it could be necessary to shift a part of the chromatogram from one worker to another, which would break the mentioned rule. For this example a total chromatogram has to be stored on one worker. Still, the whole processing can benefit from distribution, if different workers processed different chromatograms. While such a distribution scheme can be suitable for alignment, it can become a problem, if statistics should be applied to the data set. In such a case quantitative data from different chromatograms but the same time index has to be processed from one worker. In that case, the data has to be redistributed prior to statistics.

Technically, parallel computing is based on multi-core technology. Multi-core processors consist of two or more in most cases identical individual processors. These cores are normally placed within one central processing unit (CPU) and share some of the architecture of the hosting chip. Up to date, dual-core CPU's have come up to a standard in personal computers (PC) and quad- or octo-core CPU's are now commercially available. The gain in performance depends mainly on the used software. In order to take advantage from multi-core architecture, the used software has to divide pending work into different threats which can be processed by different cores. A limitating factor is the ability to divide a task into different threats and the transfer time. The maximum achievable speed-up is described by Amdahl's law [39]. MatLab introduced a parallel computing toolbox to take advantage of local multi-core architecture. However, scripts and data structure have to be modified and optimized for the application of parallel computing.

2. Experimental

2.1. Sample material

Prepared sample material was obtained from Fiehn Labs, Genome Center, UC Davis, CA, USA and had already been analyzed there by GC–MS and FT-ICR-MS and subsequent statistical analysis [40].

Male Sprague–Dawley rats were exposed with aged and diluted side stream cigarette smoke at a concentration of 1 mg/m^3 total suspended particulates for 6 h/d for one (group one, 7 individuals) or 21 days (group two, 7 individuals). There was also a control group with 8 and 7 individuals for each group. (The original experiment includes additional groups from 3 and 7 days exposure).

An aliquot of 30 μL rat plasma was transferred into clean microcentrifuge vial and 400 μL of solvent (isopropanol:acetonitrile:water = 3:3:2) were added. The mixture was vortexed for 10 s and then mechanically shacken for 5 min at 4 °C. After centrifugation at 13,000 × *g* for 2.5 min the supernatant was transferred to new centrifuge tubes and taken to dryness under vacuum and centrifugation. Vials were filled with nitrogen and stored at room temperature until derivatization.

Methyl oxime derivatives were produced by dissolving the dry extracts in 50 μL of freshly prepared *O*-methylhydroxylamine·HCl (40 mg/mL in pyridine). Incubation was done at 37 °C for 90 min under continuous shaking. Subsequent trimethyl silylation was achieved by the addition of 50 μL of *N*-methyl-*N*-trimethylsilyltrifluoroacetamide, followed by continuous shaking for 30 min at 60 °C.

The analysis of variance of the original GC–MS data set by Fiehn Labs from plasma and lung samples, showed that several metabolites were significant at the 0.05 level, including palmitoleic, palmitic and arachidic and cis-2-octadecanoic acid.

2.2. GC × GC–TOFMS

GC × GC–TOFMS analysis was performed on a Pegasus III GC × GC–TOFMS instrument (LECO Corporation, St. Joseph, MI,