Analytica Chimica Acta 841 (2014) 10–16

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

Pairwise alignment of chromatograms using an extended Fisher–Rao metric

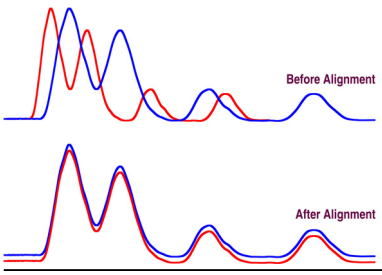
W.E. Wallace [a,*], A. Srivastava [b,d], K.H. Telu [a], Y. Simón-Manso [c]

[a] Chemical Sciences Division, National Institute of Standards and Technology, 100 Bureau Drive Stop 8320, Gaithersburg, MD 20899-8320, USA
[b] Statistical Engineering Division, National Institute of Standards and Technology, 100 Bureau Drive Stop 8980, Gaithersburg, MD 20899-8980, USA
[c] Biomolecular Measurement Division, National Institute of Standards and Technology, 100 Bureau Drive Stop 8362, Gaithersburg, MD 20899-8362, USA
[d] Department of Statistics, Florida State University, Tallahassee, FL, USA

HIGHLIGHTS

- A new approach to the alignment of chromatograms is presented.
- The entire chromatogram is aligned without priority given to any user-selected features.
- This eliminates operator bias and allows for unattended alignment of chromatograms.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 7 May 2014
Received in revised form 7 July 2014
Accepted 8 July 2014
Available online 10 July 2014

Keywords:
Alignment
Chromatography
Metabolomics
Registration
Retention time
Warping

ABSTRACT

A conceptually new approach for aligning chromatograms is introduced and applied to examples of metabolite identification in human blood plasma by liquid chromatography–mass spectrometry (LC–MS). A square-root representation of the chromatogram's derivative coupled with an extended Fisher–Rao metric enables the computation of relative differences between chromatograms. Minimization of these differences using a common dynamic programming algorithm brings the chromatograms into alignment. Application to a complex sample, National Institute of Standards and Technology (NIST) Standard Reference Material 1950, Metabolites in Human Plasma, analyzed by two different LC–MS methods having significantly different ranges of elution time is described.

Published by Elsevier B.V.

1. Introduction

In most types of chromatographic separation [1], irreproducibility in run-to-run retention time arising from instrument instability is a common occurrence. In liquid chromatography, changes in separation column temperature, mobile phase composition, mobile phase flow rate, stationary phase age, and instrument usage history are each sources of unintentional measurement variability. Retention time variability may also arise intentionally as when instrument parameters are varied systematically to find the conditions under which the best separation occurs [2]. Additionally, variability can occur when samples are analyzed on different instruments to compare separation performance or to find as many components in a mixture as may be identified by a suite of methods [3]. In any circumstance chromatograms must be retention-time aligned in order to allow run-to-run comparisons to be made [4,5].

Broadly considered, there are two approaches to chromatogram alignment [6]. One approach identifies common features in the chromatograms to be aligned, forces alignment of these features, and then interpolates an alignment function between these fiducial features. The popular XCMS program [7] uses this method.

* Corresponding author. Tel.: +1 301 975 5886; fax: +1 301 975 3670.
E-mail address: William.Wallace@nist.gov (W.E. Wallace).

http://dx.doi.org/10.1016/j.aca.2014.07.004
0003-2670/Published by Elsevier B.V.

These features can be individual peaks, clusters of peaks, or entire segments of the chromatogram. The difficulty in this approach lies in defining and detecting features with consistency in real data. For instance, a single missed peak or a false positive identification can alter the alignment of the remaining spectrum. A second approach seeks to align all points in the chromatogram without any added importance given to chromatographic features by finding a warping function that minimizes the discrepancy between two chromatograms across the entire range of retention time. An example of this is parametric time warping [8]. The algorithm introduced here uses this second full-chromatogram approach but relies on the derivative of the chromatogram to create a warping function. Use of the derivative enhances the sensitivity of the alignment to subtle features in the chromatograms.

In this work a new approach for alignment is presented for consideration by the chromatography community. A purely geometric framework for separating the phase and the amplitude variability [9] based on an extension of the Fisher–Rao metric is described and applied to the problem of chromatogram alignment. The chromatograms are treated as mathematical functions f_i without regard to any specific features or details as to how the data was taken. The user must select the beginning and ending points of each chromatogram that will remain fixed during the alignment procedure. For highly misaligned chromatogram pairs it has been observed that the method functions best when the chromatograms to be aligned have the same, or similar, number of points. This may entail pruning or re-sampling of the chromatograms as will be discussed below. Alignment examples are given on liquid chromatography–mass spectrometry data (LC–MS) on National Institute of Standards and Technology (NIST) Standard Reference Material 1950, Metabolites in Human Plasma [3,10], a sample that is chromatographically complex.

2. Mathematical framework

We treat individual chromatograms as functions on an interval $[a, b]$ and consider the issues and challenges that arise in aligning such functions. Given two functions f_1 and f_2, our goal is to find a warping function $\gamma:[a, b] \to [a, b]$ such that f_1 is optimally aligned to $f_1 \circ \gamma$. The most basic idea in alignment is to solve a problem of the type:

$$\inf_\gamma \|f_1 - (f_2 \circ \gamma)\| \tag{1}$$

where inf means the infimum (the greatest lower bound), $\|\cdot\|$ denotes the standard Euclidean norm and quantifies the difference between f_1 and the warped f_2. However, there are some known issues with this formulation, including the possibility of a degenerate solution known as the pinching effect. This occurs when one can pinch the whole of f_2 into a single point, via warping, and consequently reduce the alignment cost to zero despite f_1 and f_2 being very different. A common solution that prevents pinching is to regularize γ by including its roughness in the optimization, according to:

$$\inf_\gamma(\|f_1 - (f_2 \circ \gamma)\| + \lambda \mathcal{R}(\gamma)) \tag{2}$$

here $\mathcal{R}(\gamma)$ is a measure of roughness associated with the warping function γ, and λ is a positive constant. While this is a commonly-used solution, it suffers from the problem of asymmetry. That is, the optimal registration of f_1 to f_2 can be different from that of f_2 to f_1. This asymmetry makes it difficult to provide a meaningful interpretation to the alignment. An additional issue is the choice of λ that provides a balance between the matching term and the roughness term. Different values of λ can lead to very different solutions.

More recently there has been a large interest in developing alignment criteria that result in proper distances between aligned functions. The advantages of this approach are: (1) the solutions are symmetric, i.e. the optimal alignment of f_1 to f_2 is the same as that of f_2 to f_1, resulting in a better interpretability of alignments, (2) the regularization term is already included in these distances, i.e. one does not need any explicit roughness penalty term and, thus, avoiding the tricky issue of selecting λ, and (3) the distance can be used for ensuing statistical analysis such as PCA or classification. The last item is important because the same distance can be used for both the alignment and the accompanying analysis, rather than using two different distances for these steps.

In this paper a specific distance formulation, termed elastic functional data analysis [11] is used. It is based on extending the traditional Fisher–Rao metric used in statistics [13–17] to include more general functions such as the chromatograms. The details of this construction are provided in papers [9,11,12,18,19] and, therefore, here we simply state the alignment solution and apply it to chromatography data.

This new approach is based on a mapping that takes the original functions and transforms them into new functions. This new function, $q : [a, b] \to \mathbb{R}$, is called the square-root slope function (SRSF) of the original function f, and is defined as follows:

$$q(t) = \text{sign}(\dot{f}(t))\sqrt{|\dot{f}(t)|}. \tag{3}$$

As described in earlier papers [18,19], there are several reasons for choosing SRSFs for the alignment problem. The main reason is that at the Fisher–Rao metric, which has the requisite mathematical properties to facilitate alignment, but is difficult to work with, becomes a standard Euclidean norm of the difference, when we use SRSFs instead of the original functions. The use of Euclidean norm naturally simplifies the alignment solution as it is a familiar quantity. While the Fisher–Rao metric has been used for studying probability density functions and cumulative density functions in the past, the SRSFs allow an extension to more general functions. Notice that we do not require the chromatograms to be positive or have positive derivatives; we allow general functions. For every SRSF $q(t)$, the original function f can be obtained precisely using the equation $f(t) = f(0) + \int_0^t q(s)|q(s)|\text{ds}$, since $q(s)|q(s)| = \dot{f}(s)$. If we warp the function f by γ, the SRSF of the warped function $f_1 \circ \gamma$ is given by $\tilde{q}(t) = (q, \gamma)(t) = q(\gamma(t))\sqrt{\dot{\gamma}(t)}$. With this expression it can be shown that for any f_1, f_2 and a warping γ, we have:

$$\|q_1 - q_2\| = \|(q_1, \gamma) - (q_2, \gamma)\|, \tag{4}$$

where q_1, q_2 are SRSFs of f_1, f_2, respectively. This is called the isometry property, and implies preservation of distances under identical warping. Readers may be familiar with more common isometries such as the preservation of Euclidean distance between vectors under identical rotation and/or translation. SRSF allows a similar behavior for warping. This property is central in suggesting a new cost term for pairwise registration of functions: $\inf_{\gamma\in\Gamma}\|q_1 - (q_2, \gamma)\|$. Therefore, in this framework, one aligns the SRSFs of any two functions first and then maps them back to the original function space to obtain the registered functions. We point out that due to the presence of the square-root of the derivative of γ inside the norm, this cost function has a built in regularization term and does not require any additional penalty term. In the case one wants to further control the amount of warping this can be done by using an additional penalty term.

The pairwise alignment problem can now be solved using the optimization:

$$\gamma^* = \inf_{\gamma\in\Gamma}\left\|q_1 - (q_2 \circ \gamma)\sqrt{\dot{\gamma}}\right\| = \inf_{\gamma\in\Gamma}\left\|q_2 - (q_1 \circ \gamma)\sqrt{\dot{\gamma}}\right\|. \tag{5}$$