Organic Geochemistry 79 (2015) 44–48

Contents lists available at ScienceDirect

Organic Geochemistry

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

TEXPRESS v1.0: A MATLAB toolbox for efficient processing of GDGT LC–MS data

James T. Dillon [*], Yongsong Huang [*]

Department of Geological Sciences, Brown University, Providence, RI 02912, USA

ARTICLE INFO

Article history:
Received 3 July 2014
Received in revised form 7 November 2014
Accepted 24 November 2014
Available online 16 December 2014

Keywords:
Glycerol dialkyl glycerol tetraether
GDGT
TEX_{86}
MBT
CBT
MATLAB

ABSTRACT

Liquid chromatography–mass spectrometry (LC–MS) analysis of glycerol dialkyl glycerol tetraethers (GDGTs) from sediment and soil samples has become a widely adopted approach for reconstructing past ocean and continental climate variables such as temperature and pH. The LC–MS data used for constructing these GDGT climate proxies are taken directly from the peak area values of individual GDGT $[M+H]^+$ ions, often determined from manual peak integration due to unreliable computer integrators in commercial software, particularly in cases of complex baselines, asymmetric peak shapes and peak coelution. Manual integration is not only time consuming, but also prone to user induced inconsistency when individuals utilize different criteria for peak/baseline definition. To overcome these problems, we have developed a user friendly, graphical user interface (GUI) programmed in the MATLAB environment, allowing users to efficiently and reproducibly perform batch processing and peak integration of LC–MS data. The program, "TEXPRESS" v1.0 ("tetraether index express"), incorporates modern chemometric based techniques for baseline definition and deconvolution of complex chromatographic peaks and we show that LC–MS data processed using the TEXPRESS toolbox are in strong agreement with results obtained from manual peak integration. We provide a general overview of the concepts and architecture of the TEXPRESS toolbox and discuss the advantages of chemometric based peak integration methods for processing branched and isoprenoid GDGT LC–MS data.

1. Introduction

Glycerol dialkyl glycerol tetraethers (GDGTs) have attracted attention in recent years as potential proxies for studying past changes in paleoenvironments (Schouten et al., 2013). For example, the TEX_{86} index, based on isoprenoid GDGTs (isoGDGTs) with 86 carbons, is shown to correlate with annual mean sea surface temperature (SST; Schouten et al., 2002) and lake surface temperature (LST; Powers et al., 2004, 2010) spanning locations across the globe (Kim et al., 2008, 2010; Liu et al., 2009). Ratios based on branched GDGT (brGDGT) distributions, such as the methylation (MBT) or cyclization (CBT) branched tetraether indices, provide information about mean air temperature (MAT) and soil pH (Weijers et al., 2007; Peterse et al., 2012). The widespread occurrence of brGDGTs in, for example, lacustrine sediments (Pearson et al., 2011), loess deposits (Gao et al., 2012) and speleothems (Blyth and Schouten, 2013), has the potential to provide valuable insights into past continental climate change.

A standard approach for quantifying the distribution of brGDGTs and isoGDGTs generally includes some form of high performance liquid chromatography (HPLC) combined with mass spectrometry (MS) for the detection of individual GDGT $[M+H]^+$ ions (Schouten et al., 2007; Escala et al., 2009; Becker et al., 2013). Regardless of instrument model or separation technique, the amount of MS data produced in a typical batch analysis often requires a substantial amount of processing time, particularly when tasks such as peak integration are performed manually. Automated peak detection and integration features available in most commercial chromatography software are unreliable in cases of substantial baseline noise, asymmetric peak shape (peak fronting/tailing) and coeluting peaks. These troublesome characteristics, common in many GDGT containing samples, leave many to rely on manual peak integration for a substantial number of peaks in a given dataset. In addition to the time requirements, the inherent subjectivity in manually defining the start and end points of a chromatographic peak is an often overlooked source of error, particularly when there is a lack of consistency in the manner in which the points are defined.

Modern chemometric approaches to chromatographic data analysis provide an effective way for automating consistent

---

[*] Tel.: +1 401 863 3822; fax: +1 401 863 2058.
E-mail addresses: james_dillon@brown.edu (J.T. Dillon), yongsong_huang@brown.edu (Y. Huang).

http://dx.doi.org/10.1016/j.orggeochem.2014.11.009
0146-6380/

measurement of peak areas, while maintaining rigorous mathematical and statistical treatment of the data (Amigo et al., 2010). When performed within a numerical computing environment, such as MATLAB, these advanced techniques can be performed on standard personal computing hardware and yield excellent results (Skov and Bro, 2008). The "TEXPRESS" v1.0 ("tetraether index express") toolbox presented here, integrates these approaches into a user friendly LC–MS data processing environment using the MATLAB programming language, designed for automating the measurement of peak areas in brGDGT and isoGDGT ion chromatograms. Described here are the basic concepts and architecture of the TEXPRESS toolbox, a brief look at potential sources of variance from manual integration using commercial software, and a comparison of results obtained with the TEXPRESS toolbox with those obtained from manual integration.

2. Experimental

2.1. System requirements

The software was developed in MATLAB (R2014a) and tested on Windows 7 and OSX 10.9 operating systems. A general description of the software architecture may be found in the Supplementary information. Annotated source code and installation instructions are available online at the MATLAB File Exchange website and Github.com.

2.2. LC–MS datasets

Datasets of previously analyzed GDGT samples were reprocessed in MATLAB using the TEXPRESS toolbox. All the samples had been analyzed using HPLC–MS with atmospheric chemical ionization (APCI) and selected ion monitoring (SIM) of iso and brGDGT $[M+H]^+$ ions (Figs. S1 and S2). The HPLC system comprised a binary pump, thermostated column compartment and autosampler (Agilent 1200 series), coupled to a quadrupole mass spectrometer (Agilent 6130) with an APCI probe. GDGTs were separated using a Prevail Cyano HPLC column (3 µm, 150 × 2.1 mm i.d.) with a mobile phase gradient of hexane and isopropanol (Schouten et al., 2007). Agilent OpenLab CDS ChemStation software was used for instrument control and manual processing of peak data. Further details about GDGT analysis may be found in the Supplementary information.

3. Results and discussion

3.1. Drawbacks of manual integration

Measurement of peak area using most commercial software is generally accomplished by some form of trapezoidal integration (Dyson, 1998; Poole, 2003), where a peak is defined by its start and end points (represented by a straight line connecting the two), which are defined by the user or a computer peak detection algorithm. The area is calculated by segmenting the peak into a number of trapezoids, drawn from the baseline up to the signal, and the summed area of these trapezoids represents the approximate peak area (Fig. 1). For an isolated peak with a stable baseline, a high degree of accuracy and precision in the peak area can be expected from manual or computer integration methods. As peaks begin to coelute or noise in the baseline increases, the accuracy and precision of the area obtained from either manual or computer integration methods typically declines, introducing an often overlooked source of error. Although trapezoidal based integration methods are computationally fast, they are particularly prone to error when the perpendicular drop method is used to compensate for peak coelution (Poole, 2003), where the larger peak is under quantified and the smaller peak over quantified.

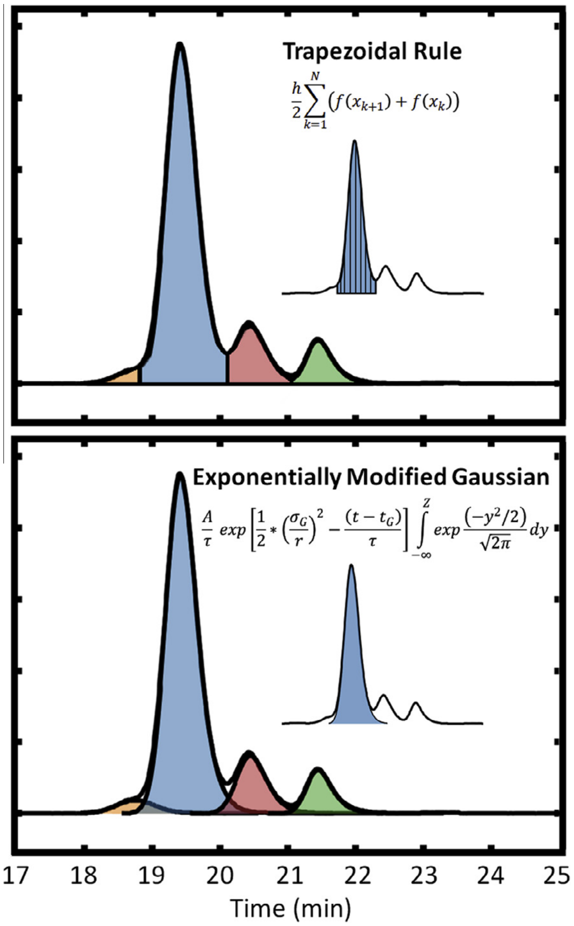

Fig. 1. Depiction of integration methods used for determining peak area. Top panel depicts trapezoidal integration commonly used in commercial chromatography software. Bottom panel depicts a curve fitting peak integration model using the exponentially modified Gaussian equation, which is used for peak area determination in TEXPRESS.

To assess potential variance in peak area associated with manual integration, five participants with a range of experience in processing GDGT LC–MS data participated in a 'blind' integration of brGDGT LC–MS data using the manual integration features provided within Agilent ChemStation software. Participants were asked to integrate five samples and one replicate sample, each with nine extracted ion chromatograms (brGDGTs I–III, a, b, c) of varying complexity (Table S1). The consistency of each user was assessed from the replicate sample, where each participant integrated the same file twice during the course of this 'blind' test. Participants were given no prior knowledge of sample information included in the dataset (sample name, source, location, etc.) in order to prevent bias in the integration of the replicate sample.

Overall, the majority of peak areas obtained by all participants were in good agreement, generally within a ± 1% margin of absolute area (Figs. S3–S7). The area of coeluting peaks differed, however, among participants by on average ± 4–5%. The greater variance among participants arose from differences in the peak start and end points defined by each participant. Samples with numerous coeluting peaks ultimately showed the greatest variance among MBT/CBT derived pH and MAT values (Fig. 2), where pH values varied up to ± 0.50 units and MAT values up to ± 4.0 °C. On average though, pH varied by ± 0.01 units from the sample mean, while MAT values varied by ± 0.50 °C among participants.