Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

A roadmap for the XCMS family of software solutions in metabolomics

Nathaniel G Mahieu[1,2], Jessica Lloyd Genenbacher[1,2] and Gary J Patti[1,2]

Global profiling of metabolites in biological samples by liquid chromatography/mass spectrometry results in datasets too large to evaluate manually. Fortunately, a variety of software programs are now available to automate the data analysis. Selection of the appropriate processing solution is dependent upon experimental design. Most metabolomic studies a decade ago had a relatively simple experimental design in which the intensities of compounds were compared between only two sample groups. More recently, however, increasingly sophisticated applications have been pursued. Examples include comparing compound intensities between multiple sample groups and unbiasedly tracking the fate of specific isotopic labels. The latter types of applications have necessitated the development of new software programs, which have introduced additional functionalities that facilitate data analysis. The objective of this review is to provide an overview of the freely available bioinformatic solutions that are either based upon or are compatible with the algorithms in XCMS, which we broadly refer to here as the 'XCMS family' of software. These include CAMERA, credentialing, Warpgroup, metaXCMS, X^{13}CMS, and XCMS Online. Together, these informatic technologies can accommodate most cutting-edge metabolomic applications and offer unique advantages when compared to the original XCMS program.

Addresses
[1] Department of Chemistry, Washington University in St. Louis, St. Louis, MO 63130, United States
[2] Department of Medicine, Washington University School of Medicine, St. Louis, MO 63110, United States

Corresponding author: Patti, Gary J (gjpattij@wustl.edu)

Current Opinion in Chemical Biology 2016, **30**:87–93

This review comes from a themed issue on **Omics**

Edited by **Daniel K Nomura**, **Alan Saghatelian** and **Eranthie Weerapana**

For a complete overview see the Issue and the Editorial

Available online 11th December 2015

http://dx.doi.org/10.1016/j.cbpa.2015.11.009

Introduction

Data from liquid chromatography/mass spectrometry (LC/MS)-based untargeted metabolomic experiments are highly complex. Therefore, bioinformatic software is typically required for processing of the results. At this time, there are many reliable software solutions available [1–9]. It is not the purpose of this review to comprehensively detail each, nor is it our intent to provide any type of comparative evaluation. Rather, we will exclusively focus on a selection of freely available software solutions which are interoperable with the XCMS program. Some of these software solutions bear variants of the XCMS name, while others do not. We broadly refer to the class as a whole as the 'XCMS family'.

Defining the needs: a general bioinformatic workflow

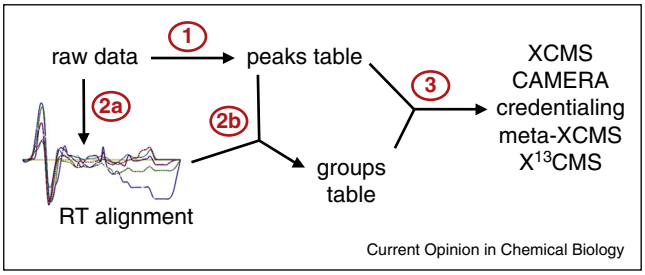
Historically, the bioinformatic workflow for processing untargeted metabolomic data has involved three general steps: feature detection, correspondence determination, and context-dependent analysis of the resulting measured values (Figures 1 and 2) [10,11••]. Each is briefly outlined below.

1. The first and perhaps most important step is feature detection (also known as peak detection or peak picking). The purpose of this step is to extract signals in the dataset that arise from real compounds, while attempting to exclude signals resulting from various noise sources [12]. Extracted signals with a unique mass-to-charge ratio, retention time, and peak shape are recorded as features (Figure 2a).

2. The second step in the workflow is establishing correspondence between the features detected from different sample runs. Correspondence refers to establishing which features from different analytical runs 'correspond' to the same analyte. Establishing correspondence is arguably the most challenging step in the processing of untargeted metabolomic data [11••]. Although the same analyte may be detected in multiple experimental runs, the measured mass-to-charge ratio and retention time of the analyte can vary in each run due to factors such as temperature fluctuation and column degradation (Figure 3a).

In practice, the majority of investigators performing LC/MS-based metabolomics currently assert correspondence by aligning the time domains of each run with time-warping techniques (Figure 2b). The objective is to correct for drift factors so that features can be grouped between samples by direct matching of retention time. Although the alignment approach for establishing correspondence has

Figure 1



The bioinformatic workflow for processing untargeted metabolomic data with XCMS. The workflow has three general steps: 1. Feature detection, 2. Correspondence determination, and 3. Additional context-dependent analysis. These steps are numbered in red on the schematic. After acquisition of LC/MS profiling data, feature detection is performed on the raw data to generate a peaks table (step 1). Next, retention time drift is corrected (step 2a). The OBI-warp algorithm implemented within XCMS operates on the raw data to determine retention time drift. This produces a retention time correction curve for each sample which, together with the peaks table, is used to establish correspondence and generate a groups table (step 2b). The peaks table and the groups table are the input for a variety of further analyses. The third step is dependent upon experimental objectives. In the standard XCMS workflow, step 3 is statistical analysis. The other programs listed use the peaks table and groups table to achieve different aims such as adduct and artifact annotation, multiple-factor analysis, and isotopic label tracking.

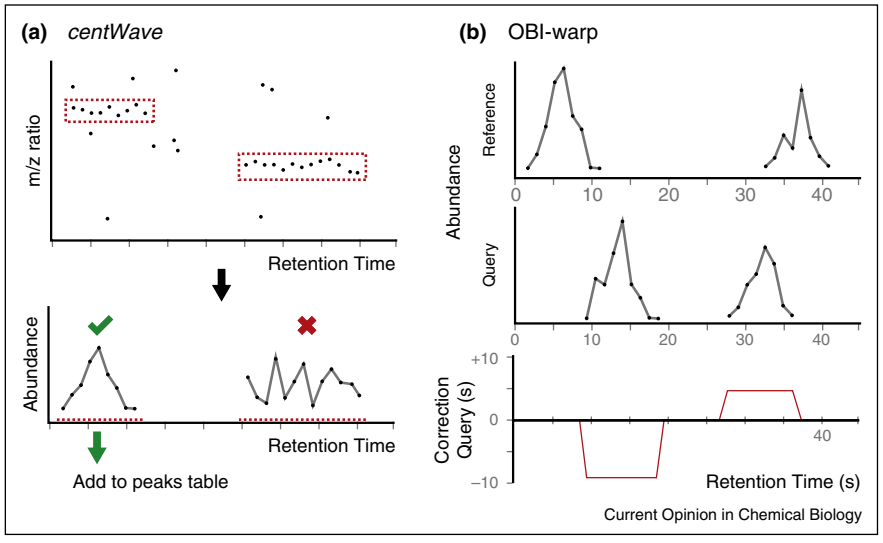
enabled many laboratories to successfully analyze untargeted metabolomic data, many drift factors are compound specific and therefore global-alignment techniques only reduce the total drift but do not eliminate it (Figure 3b). Accordingly, there remains a great need for robust correspondence determination algorithms and this is an active area of research interest [11••].

3. The last step of the workflow is context dependent. Analyses diverge, depending on experimental goals. In the simple cases when the objective is to compare sample classes, this step amounts to performing statistical analysis on the intensities of detected features. For more advanced objectives such as isotope tracing or tandem mass spectral analysis, additional algorithms are required.

Introducing XCMS

In 2006, the XCMS software was published as one of the first programs to provide a complete solution to the bioinformatic workflow outlined above for processing untargeted metabolomic data [13]. The 'X' in the XCMS acronym is used to denote that the software can be applied to any form of chromatography. To date, however, XCMS has been used the most to process LC/MS-based metabolomic data. The original XCMS software used the *matchedFilter* algorithm to accomplish feature detection, the *retcor.peakgroups* algorithm to perform alignment (an application of LOESS regression to well-behaved peak groups), and the *group.density* algorithm to group aligned features across samples on the basis of *m/z* bins. In recent years, a new algorithm for feature detection called *centWave* and a new algorithm for alignment called OBI-warp have been implemented within XCMS (Figure 2) [14,15]. It is worth noting that while these algorithms have led to better overall XCMS performance,

Figure 2



Schematic of the *centWave* and OBI-warp algorithms as implemented within XCMS. **(a)** The first step in centWave is to find consecutive scans in which peaks are detected within a specific mass error (top). These are referred to as regions of interest (ROIs). Two such ROIs are displayed here and boxed in red. Next, extracted ion chromatograms are created for each ROI (bottom). Extracted ion chromatograms that display a peak shape are then added to the peaks table, as illustrated by the green checkmark and arrow. **(b)** OBI-warp aligns a query sample to a reference sample. Here we illustrate a representative example in which two features are shifted in the query sample compared to the reference sample. Application of the correction curve to the query (bottom) brings the samples into alignment.