Universität Bielefeld

Technische Fakultät
AG Genominformatik

Maltcms an Application Framework for Processing of Metabolomics-Data

Nils Hoffmann[1,2], Mathias Wilhelm[1], Matthias Keck[3], Anja Döbbe[3], Karsten Niehaus[3], Jens Stoye[1]

[1] AG Genominformatik, Bielefeld University, [2] International NRW Graduate School in Bioinformatics and Genome Research
[3] Department of Proteome and Metabolome Research, Faculty of Biology, Bielefeld University

Introduction

We present the current state of our modular application toolkit for chromatography-mass spectrometry (Maltcms), and two derived applications: ChromA [1], which is applicable to gas-chromatography (GC) and liquid-chromatography (LC) data with single-dimension detectors (FID, FL) or multi-dimension detectors (MS), and ChromA4D, which is applicable to data from GCxGC-MS experiments. The framework Maltcms allows to setup and configure individual processing components with few effort. All processing steps center around the pipeline paradigm, where each step can define dependencies on previous steps. Individual processing components are easily implemented within the JAVA language. Maltcms is freely available under the L-GPL v3 license at `http://maltcms.sourceforge.net`

ChromA

ChromA is a configuration of Maltcms, which includes preprocessing, in the form of mass binning, time-scale alignment and annotation of signal peaks found within the data, as well as visualizations of unaligned and aligned data.

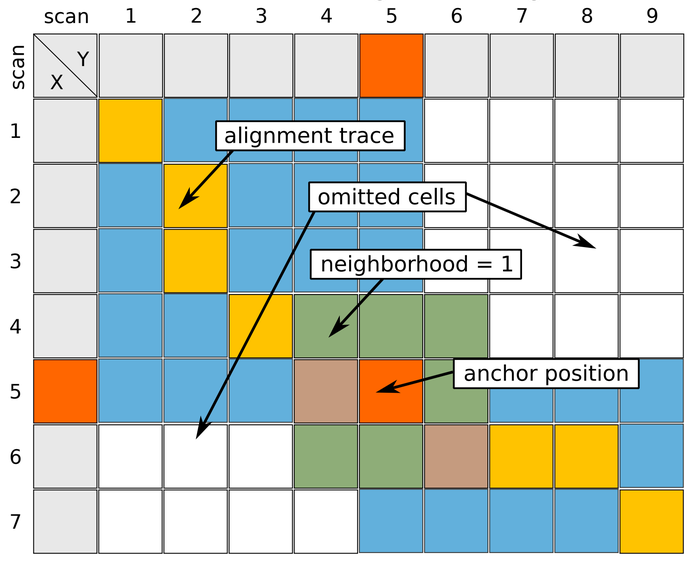

Figure 1: Schematic of the constrained pairwise alignment matrix with anchors. Omitted cells are never evaluated and do not occupy any memory.

The alignment is based on a star-wise or tree-based application of an enhanced variant of pairwise dynamic time warping (DTW) as shown in Figure 3. To reduce both runtime and space requirements, we identify conserved signals throughout the data, constraining the search space of DTW, which is illustrated by Figure 1. These alignment anchors can be augmented or overwritten by user-defined anchors, such as previously identified compounds, characteristic mass or MS/MS identifications. Then, the candidates are paired by means of a bidirectional best-hits (BBH) criterion [2], which can compare different aspects of the candidates for similarity [3], which is illustrated in Figure 2. Paired anchors are then extended to k-cliques with configurable k, which help to determine the conservation or absence of signals across measurements.

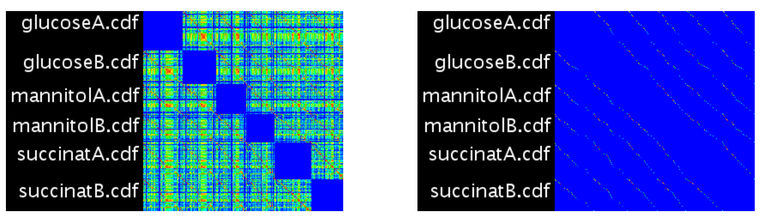

Figure 2: Peak similarities for all peaks before (left) and for retained peaks after (right) peak matching.

ChromA visualizes alignment results including paired anchors as shown in Figure 3 *b*). Additionally, absolute and relative differential charts are provided, which allow easy spotting of quantitative differences.

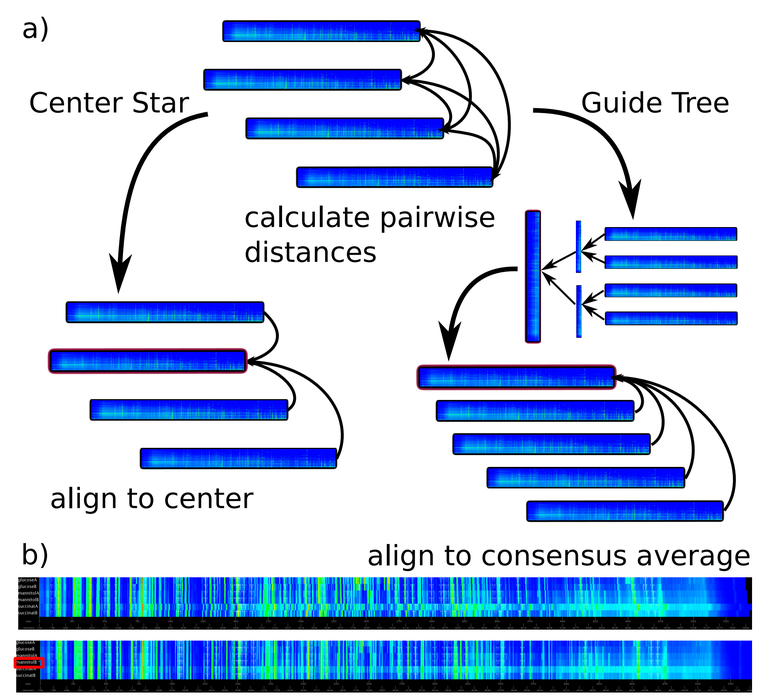

Figure 3: *a*) The two implementations of multiple alignment in ChromA. Based on pairwise distance calculation between chromatograms, a center sequence is selected as the reference for alignment by the center-star method. Alternatively, the pairwise distances are used to derive a guide tree, which is used to progressively align and merge the chromatograms. The chromatogram at the root is then a consensus average chromatogram, which is used as the reference for alignment. *b*) The result of aligning six chromatograms with the center-star method.

ChromA4D

ChromA4D applies DTW for alignment of GCxGC-MS chromatograms, but in this case comparing slices of the 2D-TIC instead of the binned intensities of mass spectra.

Peak areas are found by a modified seeded region growing algorithm [4] (Figure 4). All local maxima of the TIC representation which exceed a threshold are selected as initial seeds. Then, the peak area is determined by using the distance of the seed mass spectrum to all neighbour mass spectra as a measure of the peak's coherence. The area is extended until the distance exceeds a given threshold. No information about the expected peak shape is needed. The peak integration is based on the sum of TICs of the peak area. An identification of the area's average mass spectrum or the seed mass spectrum is possible by using our in-house MetaboliteDB.

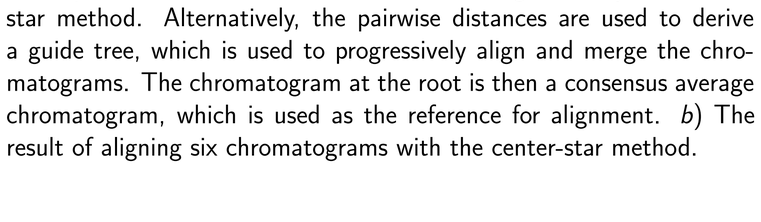

Figure 4: Left: Image of a 2D-TIC of a GCxGC-MS chromatogram from a standard FAME mix. Right: Peak boundary overlay after peak detection and integration with seeded region growing.

To represent the similarities and differences between different chromatograms, we used bidirectional best hits to find co-occurring peaks. BBHs are found by using a distance which exponentially penalizes differences in the first and second retention times of the peaks to be compared. To avoid a full computation of all pairs of peaks, only those peaks within a defined window of retention times are evaluated.

ChromA4D's visualizations represent aligned chromatograms as color overlay images, as shown in Figure 5. This allows a direct visual comparison of signals present in one sample, but not present in another sample.

Common Features

Maltcms has been designed to require only small effort to implement new functionality. It is accompanied by many libraries for different purposes, such as the *JFreeChart* library for 2D-plotting or, such as the *Colt* library, for BLAS compatible linear algebra, math and statistics implementations. Building upon the base library *Cross*, which defines the commonly available interfaces and default implementations, Maltcms provides the domain dependent data structures and specializations for processing of chromatographic data. However, the modularity allows for easy adaptation of the framework to other application domains, which involve similar, multi instance time series data.

Currently, the following data formats are supported:

- Input in common formats: *netcdf*, *mzxml* and *mzdata* (*mzml* in next release)
- Output of warped chromatograms as *netcdf*
- Peak matching, corrected retention times and integration results are saved in *csv* format

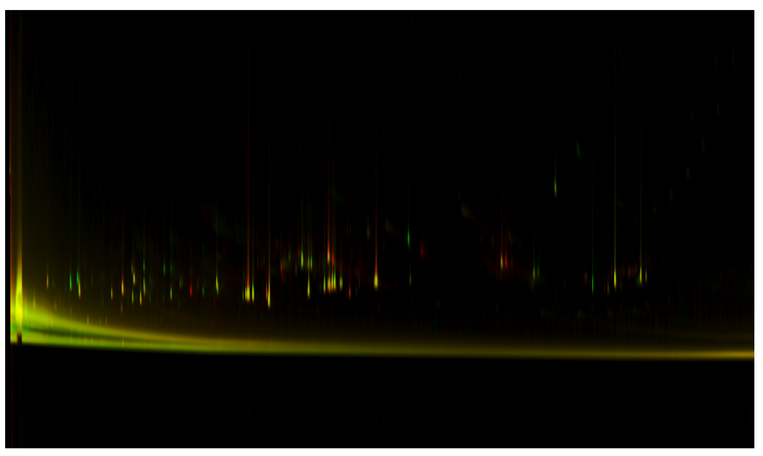

Figure 5: Overlay of GCxGC-MS chromatograms after alignment with DTW. Yellow peaks are present in both chromatograms while red ones are only present in the left chromatogram and green ones only in the right one.

Outlook

We currently develop an in-house database API based on the object database *db4o*, MetaboliteDB, which will be included as a seperate package within the next major release of Maltcms. It can parse data in *msp* compatible format and is used within a pipeline to identify metabolites by their mass spectrum with user definable similarity measure and matching threshold. The alignment with DTW in ChromA4D is not yet optimal and also does not cover the multiple alignment case yet. Thus, our work will focus on the development of an improved alignment algorithm for GCxGC-MS data.

References

[1] N Hoffmann and J Stoye. Chroma: signal-based retention time alignment for chromatography-mass spectrometry data. *Bioinformatics*, 25(16):2080–1, 2009.

[2] M P Styczynski, J F Moxley, L V Tong, J L Walther, K L Jensen, and G N Stephanopoulos. Systematic identification of conserved metabolites in gc/ms data for metabolomics and biomarker discovery. *Anal. Chem.*, 79(1):966–973, 2007.

[3] M D Robinson, D P De Souza, W Keen, E C Saunders, M J Mcconville, T P Speed, and V A Likić. A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments. *BMC Bioinformatics*, 8:419, 2007.

[4] R Adams and L Bischof. Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):641 – 647, 1994.