Universität Bielefeld

Technische Fakultät
AG Genominformatik

ChromA4D

an integrated OpenSource Application Framework for comprehensive GCxGC-MS-based Metabolomics

Nils Hoffmann[1,2], **Mathias Wilhelm**[1], **Matthias Keck**[3], **Anja Döbbe**[4], **Karsten Niehaus**[3], **Olaf Kruse**[4], **Jens Stoye**[1]

[1] Genome Informatics Group, Faculty of Technology, Postfach 10 01 31, Bielefeld University, Germany
[2] International NRW Graduate School in Bioinformatics and Genome Research
[3] Proteome and Metabolome Research Group, [4] Algae Biotech and Bioenergy Group, CeBiTeC, Bielefeld University

Introduction

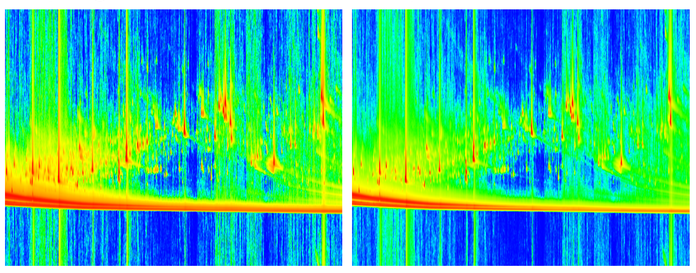
Two-dimensional gas-chromatography mass-spectrometry (GCxGC-MS) has developed into an important technique for the analysis of complex biological samples. It is applied by a growing number of researchers in the field of metabolomics due to its increased peak capacity compared to one-dimensional GC-MS. Yet, this has lead to larger amounts of data, which are even harder to inspect and analyze manually than one-dimensional GC-MS data. Currently, only a few commercial software solutions exist, which cover most parts of the workflow from raw data preprocessing to statistical data analysis, such as ChromaTOF (LECO Corp.) and GC Image (GC Image, LLC).

However, there are no OpenSource solutions available yet, which integrate all steps of processing, analysis and visualization of GCxGC-MS data into a complete solution. We present our software ChromA4D, which captures the typical workflow from data preprocessing to feature detection, grouping and output, as well as comprehesive visualizations. All processing steps generate data formats compatible with OpenSource statistics software such as R or general spreadsheet programs such as OpenOffice, so that they can be integrated easily into existing workflows. ChromA4D is based on our framework Maltcms (see Poster 374) [1].

To show the applicability of our pipeline, we compared four measurements of the TMS derivatized extract from Chlamydomonas reinhardtii wildtpye (406) and mutant (*glc*4) before ($t0$) and during ($t4$) H_2 production.

Preprocessing

The samples were recorded on a LECO Pegasus® 4D GCxGC-TOFMS machine and were converted to netcdf ANDIMS format using the ChromaTOF® software after resampling to 100 Hz MS scan acqusition rate.

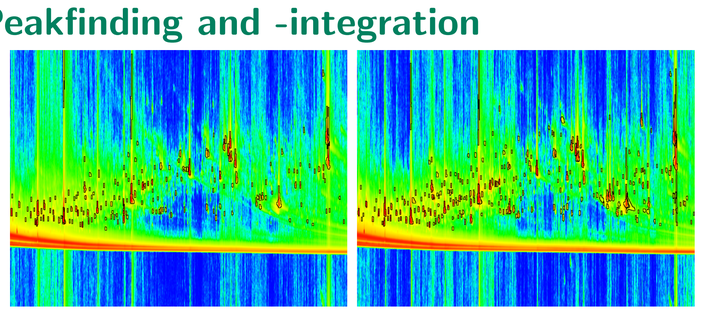


(a) Chromatogram before removal of mass channels within lowest 0.5% of standard deviation of the intensities.

(b) Chromatogram after removal of mass channels within lowest 0.5% of standard deviation of the intensities.

Figure 1: Application of the background noise reduction filter. (a) shows the raw chromatogram, before application of the filter, (b) shows the chromatogram after mass channel intensities which fall below a user defined threshold of the total standard deviation have been removed. ChromA4D also provides another preprocessing method to remove noisy mass channel intensities with very high background and low variation based on the coefficient of variation for each individual mass channel (not shown).

Peakfinding and -integration

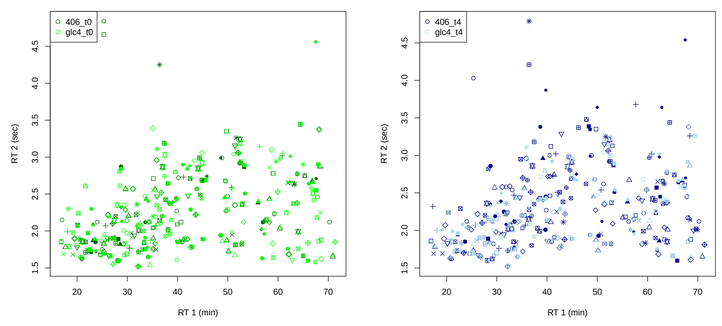


(a) Chromatogram heatmap visualization of 406_$t0$ with highlighted peak areas.

(b) Chromatogram heatmap visualization of 406_$t4$ with highlighted peak areas.

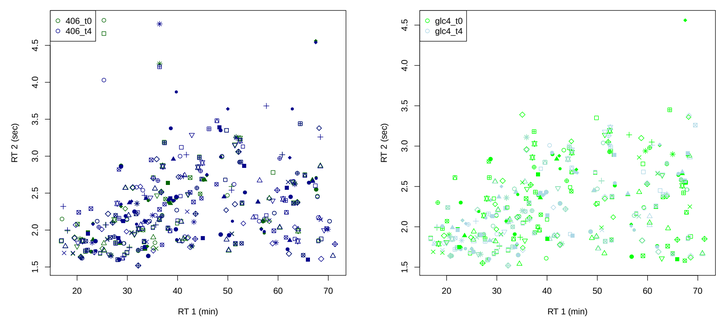
Figure 2: In order to find coherent peak areas, ChromA4D currently uses a variant of seeded region growing [2], based on mass spectral similarities between adjacent 2D-TIC signals.

Peakmatching



(a) Wildtype (406) vs. mutant (*glc*4) at time $t0$, before H^2 production.

(b) Wildtype (406) vs. mutant (*glc*4) at time $t4$, during H^2 production.



(c) Wildtype (406) at times $t0$ and $t4$.

(d) Mutant (*glc*4) at times $t0$ and $t4$.

Figure 3: Result of bidirectional best hits peak matching, based on the similarity of apex mass spectra of peaks, penalized by exponentially weighted time deviation on each separation dimension. Identical symbols of neighboring peaks correspond to the same peak group. In (c), a total of 141 peaks were paired between 406 and *glc*4 at time $t0$ and in (c) 135 peaks were paired for time $t4$. Within groups, (c) 151 peaks were paired between wildtype 406 at time points $t0$ and $t4$ and (d) 145 peaks were paired for the mutant *glc*4 at time points $t0$ and $t4$. In total, 100 cliques were found, which contained peaks from all four chromatograms. Looking at the files individually, our peakfinder reported 178 peaks for *glc*4_$t0$, 258 peaks for *glc*4_$t4$, 224 peaks for *glc*4_$t0$ and 168 peaks for *glc*_$t4$.

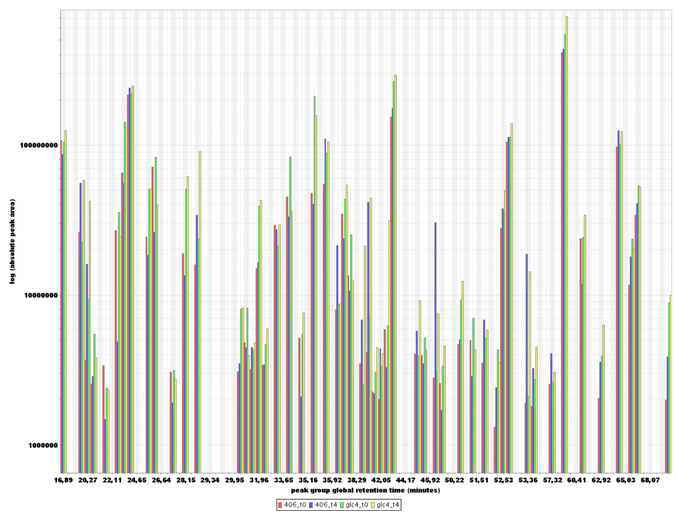


Figure 4: Boxplot of global retention time (rt1 + rt2) of peak groups versus $\log_{10}$ of absolute peak areas for all peaks, which were present in all four samples.

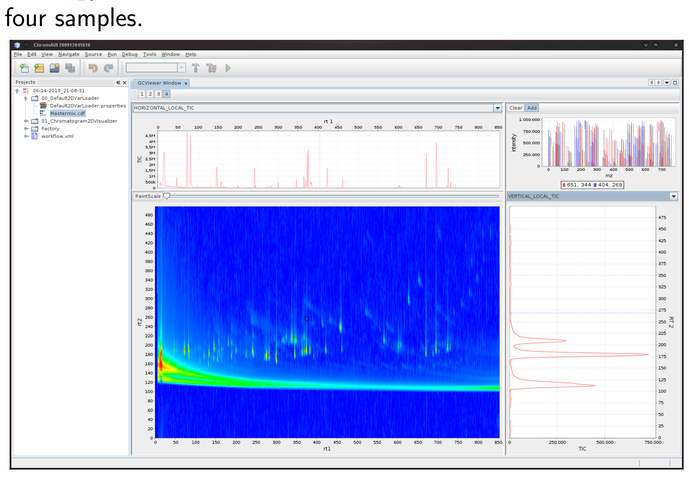
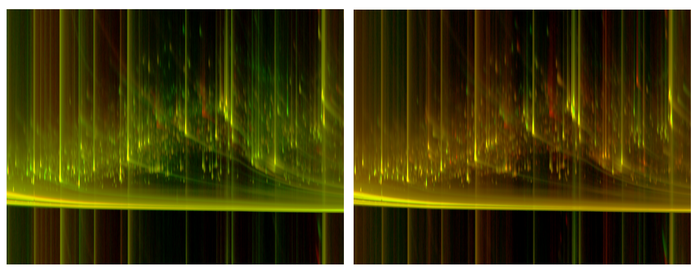


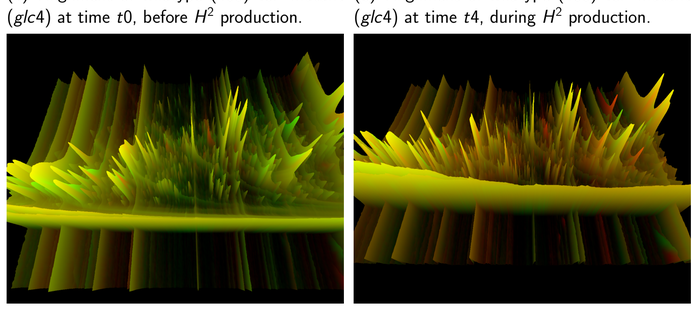
Figure 5: Screenshot of the Netbeans rich client platform based ChromAUI application, showing the GCxGC-MS heatmap viewer component.

Alignment



(a) Alignment of wildtype (406) vs. mutant (*glc*4) at time $t0$, before H^2 production.

(b) Alignment of wildtype (406) vs. mutant (*glc*4) at time $t4$, during H^2 production.



(c) Alignment of wildtype (406) vs. mutant (*glc*4) at time $t0$, before H^2 production visualized as a three dimensional heat map surface.

(d) Alignment of wildtype (406) vs. mutant (*glc*4) at time $t4$, during H^2 production visualized as a three dimensional heat map surface.

Figure 6: Result of the application of dynamic time warping using the total ion count of each modulation as a feature vector for pairwise similarity calculation. Yellow peaks are present in both chromatograms at the same intensity level while red ones are only present in the left (406) chromatogram and green ones only in the right one (*glc*4). It is also possible to use other data for the alignment, e.g. derived from the mass spectra. It is also possible to generate a differential image of the aligned chromatograms, visualizing the integrated peak areas only as a schematic image (not shown).

Preliminary Results and Outlook

We presented first results applying our GCxGC-MS workflow ChromA4D to study wildtype versus mutant samples from Chlamydomonas reinhardtii under two different conditions. Figure and the differential views in Figure 6 show, that there are some peaks, which are immediately differentially expressed. However, we have not yet taken replicates into account, which will allow for a more statistically sound analysis of variances of peak areas between the sample groups. Finally, no peaks have yet been identified against a mass spectral database, which is one of the next steps to be undertaken, before a comparison against the manual domain expert annotation and the automatic ChromaTOF (Leco Corp.) annotation will be performed.

Further work is focused on the integration of the ChromA4D workflow into our application ChromAUI, which will soon be available from the same website as Maltcms/ChromA4D. Furthermore, we are integrating ChromAUI with MeltDB [3] to allow access to experiments which were defined within MeltDB as well as the generation and editing of peak annotations.

Availability

ChromA4D is freely available under the L-GPL v3 license at `http://maltcms.sourceforge.net` within our framework Maltcms. It runs under all personal computer operating systems for which a JAVA Runtime Environment is available and typically requires $1 - 2$ GBytes of RAM.

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

[3] H Neuweger, S Albaum, M Dondrup, and M Persicke. Meltdb: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics*, Jan 2008.