Universität Bielefeld

Technische Fakultät
AG Genominformatik

BiPACE

a generic retention time alignment algorithm for gas-chromatography mass-spectrometry data

Nils Hoffmann[1], Mathias Wilhelm[2], Anja Doebbe[3], Karsten Niehaus[4], Jens Stoye[1]

[1] Genome Informatics, Faculty of Technology and CeBiTec, Bielefeld University, Germany; [2] Currently at Proteomics Research Group, Technical University Munich, Munich, Germany; [3] Department for Algae Biotechnology and Bioenergy, Faculty of Biology and CeBiTec, Bielefeld University, Germany; [4] Proteome and Metabolome Research Group, Faculty of Biology and CeBiTec, Bielefeld University, Germany

Introduction

GC×GC–MS is being used in the field of metabolomics due to its increased peak capacity compared to one-dimensional GC–MS, yielding improved chromatographic separation of chemically closely related analytes. However, this also increases the size of the resulting datasets drastically, hampering manual data analysis workflows. Thus we propose BiPACE 2D, based on BiPACE (bidirectional-best hits peak assignment and clique extension) [1] for automated matching and grouping of peaks and their associated mass spectra across many samples acquired using GC×GC–MS.

Peak Similarity Functions

For two peaks p and q, represented by their binned mass spectral intensity vectors with first column retention times $t_{1,p}$, $t_{1,q}$, second column retention times $t_{2,p}$, $t_{2,q}$, we define a similarity function for two-dimensional peaks in extension of [2] as:

2D Gaussian Product:

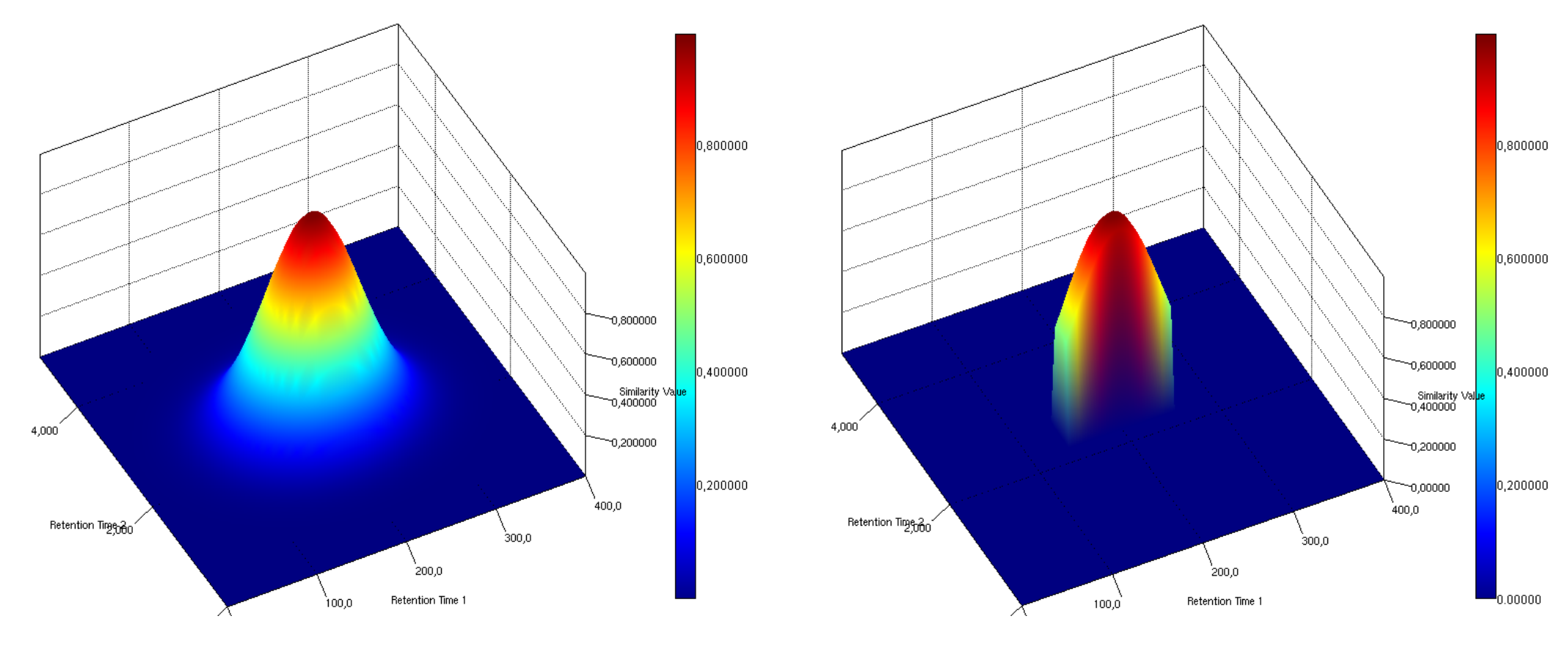
$$f_{2d}(p,q) = e^{\left(-\frac{(t_{1,p}-t_{1,q})^2}{2D_1^2}\right)} \cdot e^{\left(-\frac{(t_{2,p}-t_{2,q})^2}{2D_2^2}\right)} \cdot s(p,q), \quad (1)$$

where D_1 and D_2 are the retention time tolerances (σ) of the Gaussian distribution. $s(p,q)$ is an arbitrary similarity function between the mass spectral intensity vectors, such as the cosine, the dot product, Pearson's linear correlation coefficient, or Spearman's rank correlation coefficient.

Normalized 2D Inverse Gaussian Product:

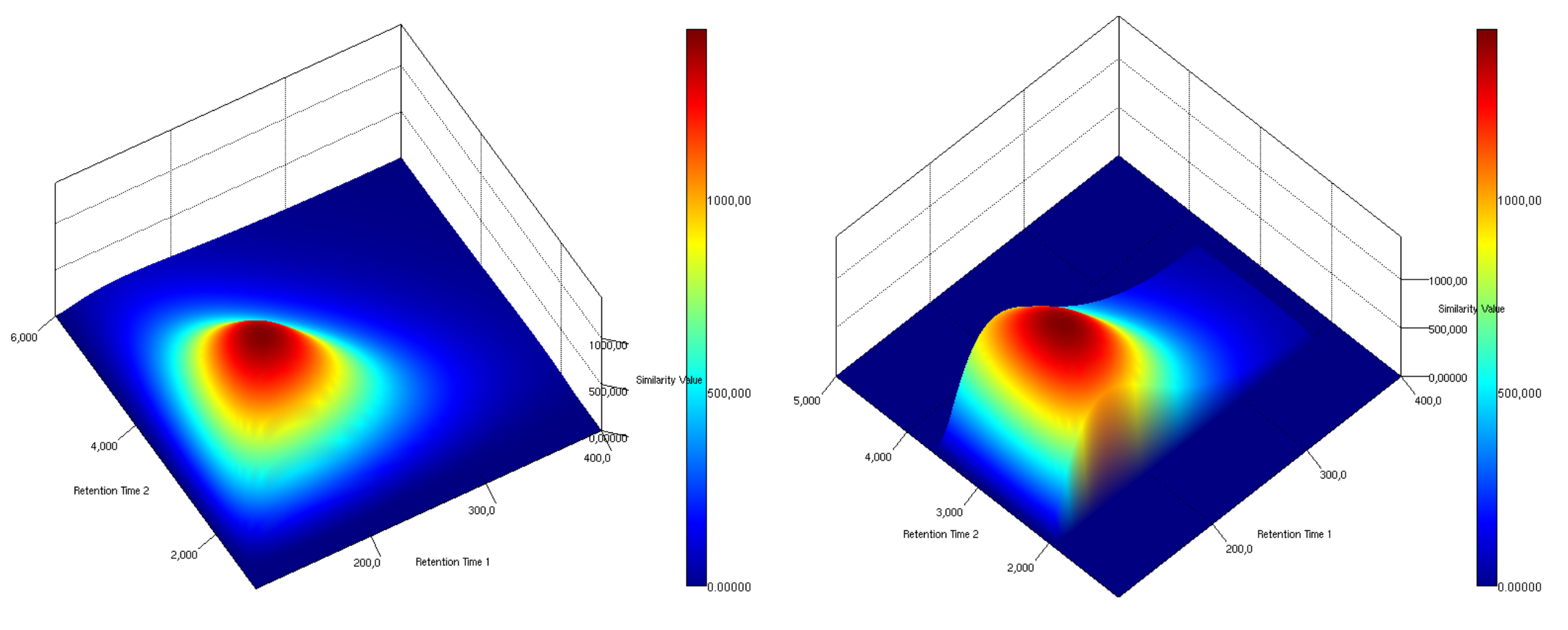
$$\begin{aligned} f_{2d}^{Inv}(p,q) = &\sqrt{\frac{\lambda_1}{2\pi(t_{1,p})^3}} \cdot e^{\left(-\frac{\lambda_1(t_{1,p}-t_{1,q})^2}{2(t_{1,q})^2 t_{1,p}}\right)} \\ &\cdot \sqrt{\frac{\lambda_2}{2\pi(t_{2,p})^3}} \cdot e^{\left(-\frac{\lambda_2(t_{2,p}-t_{2,q})^2}{2(t_{2,q})^2 t_{2,p}}\right)} \cdot s(p,q), \end{aligned} \quad (2)$$

where λ_1 and λ_2 are the shape parameters of the inverse Gaussian probability density function. Additional threshold parameters ($T1$, $T2$) can be set for each retention time similarity term separately for search space pruning (see Figures 1 and 2).



(a) $D_1 = 50$, $D_2 = 0.5$, $T_1 = 0$, $T_2 = 0$ (b) $D_1 = 50$, $D_2 = 0.5$, $T_1 = 0.5$, $T_2 = 0.9$

Figure 1: Plots of the 2D Gaussian product similarity (Eqn. 1) without threshold parameters (a) and with threshold parameters (b).



(a) $\lambda_1 = 5$, $\lambda_2 = 0.5$, $T_1 = 0$, $T_2 = 0$ (b) $\lambda_1 = 5$, $\lambda_2 = 0.5$, $T_1 = 20$, $T_2 = 2$

Figure 2: Plots of the 2D Inverse Gaussian product similarity (Eqn. 2) without threshold parameters (a) and with threshold parameters (b).

BiPACE

BiPACE is a generic algorithm for retention time alignment of multiple datasets from one and two-dimensional chromatography, coupled to MS or arbitrary detectors. It builds a k-partite graph based on the bidirectional-best hits found in the pairwise peak similarites (Eqns. 1 and 2) calculated between all input peak lists. The resulting multiple alignment is then constructed by enumerating all maximal cliques within that graph. Further details of the BiPACE algorithm are given in [1]. The time and space complexities of BiPACE 2D and BiPACE are $\mathcal{O}(K^2\ell^2)$ and $\mathcal{O}(K^2\ell)$, respectively, where K is the number of chromatograms and ℓ is the upper bound of the number of peaks in each chromatogram.

Evaluation

The evaluation was carried out on three different reference multiple alignments from a *Chlamydomonas reinhardtii* experiment, using the original method described in [3, 4] (GMA), the second created using a modified approach that tested GMA reference peak groups for temporal coherence (MGMA) and removed potentially bogus groups, and the third (MANUAL) based on a manual inspection of peaks by a domain expert.

We evaluated our methods against the mSPA [3] and SWPA [4] methods. TP, FP, FN, and TN values were counted against each reference multiple alignment, row by row. Peaks present in the reference but absent in the reported multiple alignment of a method were counted as additional FNs, normalizing the Recall value to the number of peaks contained in the reference, which explains the different numbers in comparison to those originally reported by Kim *et al.*. We used the F1 score, which is the harmonic mean of Precision and Recall for an overview comparison in Figure 3.

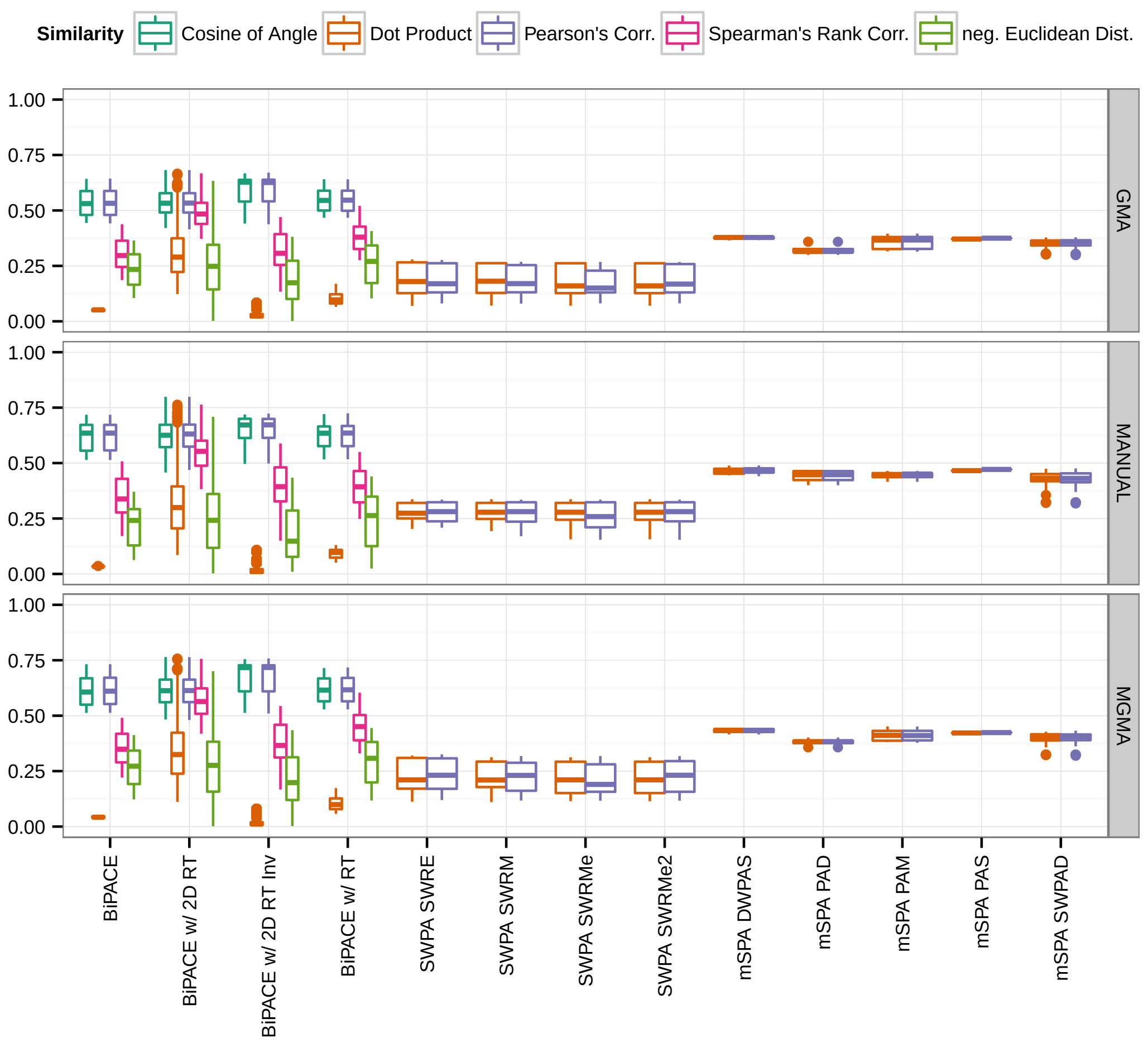


Figure 3: F1 score boxplots of the evaluated algorithms and their variants for the *Chlamydomonas reinhardtii* dataset.

Availability

BiPACE and BiPACE 2D are included in our framework MALTCMS which is freely available at:

`http://maltcms.sf.net`

C. reinhardtii Dataset

The metabolic difference resulting in H_2 production yield between the *Chlamydomonas reinhardtii* wild type stem cc406 (wt) and the high H_2-producing strain Stm6Glc4 (mut) at two different time points was compared before (t1) and during (t2) the H_2 production phase [5]:

- 12 samples, three for each factor combination
- acquisition with Leco Pegasus 4D GC×GC–MS
- total of 31695 peaks in raw peak reports
- GMA reference contained 2723 peaks, MGMA reference contained 1629 peaks, MANUAL reference contained 436 peaks
- raw data, peak lists, protocols, and manual reference alignment are available at:

`http://www.ebi.ac.uk/metabolights/MTBLS37`

Evaluation Framework

We developed an automatic evaluation framework for multiple alignment algorithms with the following features:

- easy integration of command-line tools
- combinatorial parameterization of tools
- parallel execution on computing grid environment
- local sql-database for bookkeeping and control
- allows dynamic, result-dependent workflows
- currently supports BiPACE 2D, BiPACE and CeMAPP-DTW [1], mSPA [3], SWPA [4]
- generation of joint evaluation table with classification performance, runtime and memory statistics
- plotting of results using GNU R and *ggplot2*

Results and Outlook

Using the two-dimensional inverse Gaussian product as a weight function for retention time deviations in the alignment of GC×GC–MS data resulted in a slightly increased Precision value when either the cosine or Pearson's correlation were used as mass-spectral similarities, in comparison to the standard two-dimensional Gaussian product weight function. However, this comes at the cost of a lower Recall and associated F1 value due to fewer reported TP values. Additional work will focus on a more intuitive parameterization of the inverse Gaussian weight function and an improved thresholding scheme for two dimensions.

References

[1] N. Hoffmann et al. "Combining peak- and chromatogram-based retention time alignment algorithms for multiple chromatography-mass spectrometry datasets". In: *BMC Bioinformatics* 13 (Aug. 2012), p. 214.

[2] M. Robinson et al. "A dynamic programming approach for the alignment of signal peaks in multiple gas chromatography-mass spectrometry experiments". In: *BMC Bioinformatics* 8 (Oct. 2007), p. 419.

[3] S. Kim et al. "An optimal peak alignment for comprehensive two-dimensional gas chromatography mass spectrometry using mixture similarity measure". In: *Bioinformatics* 27.12 (June 2011), p. 1660.

[4] S. Kim et al. "Smith-Waterman peak alignment for comprehensive two-dimensional gas chromatography-mass spectrometry". In: *BMC Bioinformatics* 12 (June 2011), p. 235.

[5] A. Doebbe et al. "The Interplay of Proton, Electron, and Metabolite Supply for Photosynthetic H_2 Production in *Chlamydomonas reinhardtii*". In: *Journal of Biological Chemistry* 285.39 (Sept. 2010), p. 30247.

Poster copy available at: `http://www.cebitec.uni-bielefeld.de/~hoffmann`