An Online Peak Extraction Algorithm for Ion Mobility Spectrometry Data

Dominik Kopczynski[1] Sven Rahmann[1,2]

[1]Computer Science XI and Collaborative Research Center SFB 876, TU Dortmund, Germany
[2]Genome Informatics, Human Genetics, Faculty of Medicine, University of Duisburg-Essen, Essen, Germany

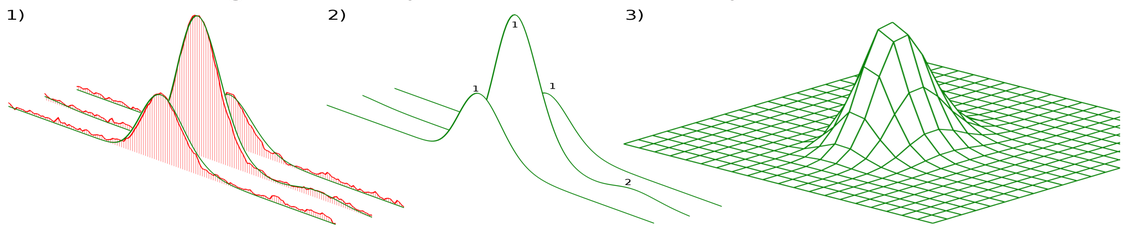
SFB 876 Verfügbarkeit von Information durch Analyse unter Ressourcenbeschränkung

technische universität dortmund

UNIVERSITÄT DUISBURG ESSEN

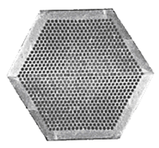
Abstract

Ion mobility (IM) spectrometry (IMS), coupled with multi-capillary columns (MCCs), has been gaining importance for biotechnological and medical applications because of its ability to measure volatile organic compounds (VOC) at extremely low concentrations in the air or exhaled breath at ambient pressure and temperature. Ongoing miniaturization of the devices creates the need for reliable data analysis on-the-fly in small embedded low-power devices e.g. the Raspberry Pi. We present the first fully automated online peak extraction method for MCC/IMS spectra [1]. Each individual spectrum is processed (with a time restriction of 100 ms) as it arrives, removing the need to store a whole measurement of several thousand spectra before starting the analysis, as is currently the state of the art.

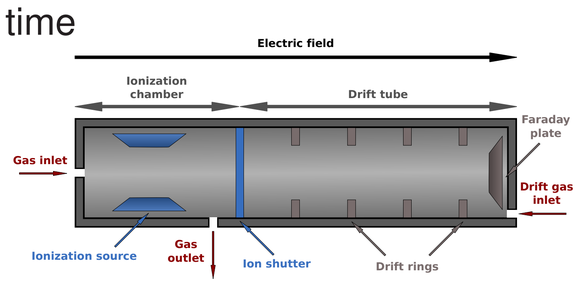


MCC/IMS Device and Data

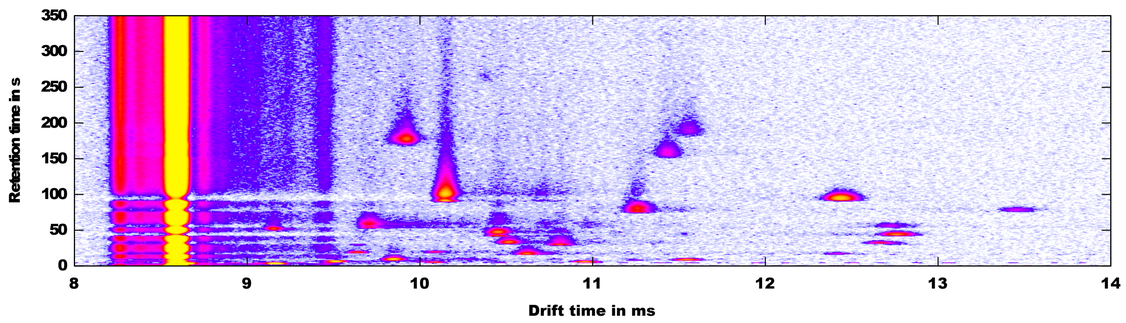
Multi-capillary Column (MCC)
Separates analytes according to time passing the column: retention time



Ion Mobility Spectrometer (IMS)
Separates analytes according to ion mobility (mass, shape): drift time



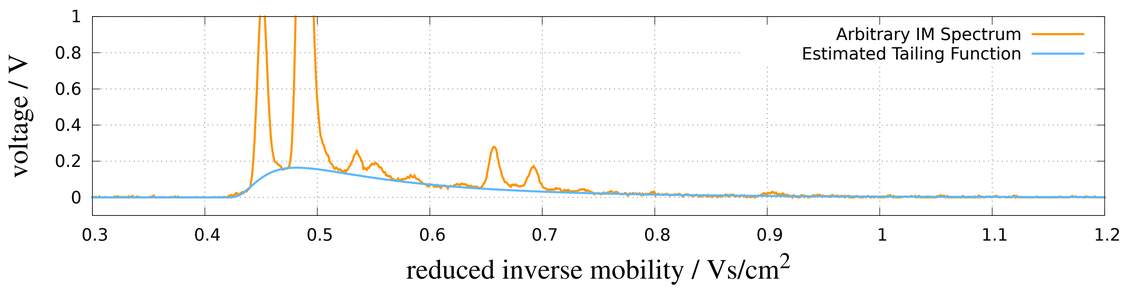
MCC/IMS Measurement
Measuring signal intensity at certain retention time R and drift time T provides $|R| \times |T|$ Matrix S.



Search for peak features (e.g. position, intensity). Peaks are potential biomarkers. *Challenge*: compute the features during the capturing without storing the whole matrix.

Step 1: Spectrum Reduction

Arbitrary single IMS Spectrum:



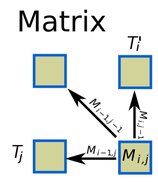
Describe spectrum as weighted mixture model of distributions:

$$f_{\omega,\theta}(t) = \frac{\omega_0}{|T|} + \sum_{j=1}^{c} \omega_j \cdot P_{\theta_j}(t)$$

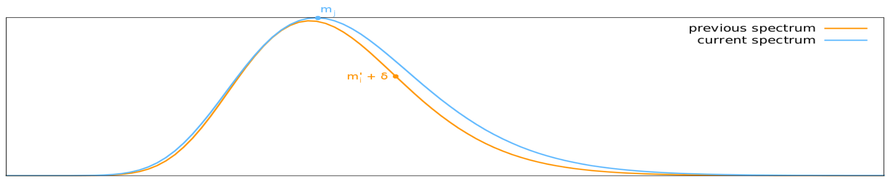
1. Estimate and subtract tailing function using linear least squares (NLLS) [2] with asymmetric error function
2. Use a moving window where size depends on physical properties
3. Fit a second order polynomial using NLLS within the window
4. Determine drift time parameters θ from polynomial for Inverse Gaussians $g_{\mu,\lambda,o}(x) := [x > o] \cdot \sqrt{\frac{\lambda}{2\pi(x-o)^3}} \cdot \exp\left(-\frac{\lambda\left((x-o)-\mu\right)^2}{2\mu^2(x-o)}\right)$
5. Subtract model from spectrum
6. Repeat 2, . . . , 5 until only noise remains

Step 2: Aligning consecutive Spectra

- Dynamic programming computation of best alignment score for sequences of individual modes up to indexes i and j of reduced spectrum T' and consecutive reduced spectrum T.
- Let $0 \leq i \leq |T'|$ and $0 \leq j \leq |T|$
- Score function $s(T'_i, T_j)$ for comparing Inverse Gaussians in drift times
- $M_{i,j}$ contains best alignment score up to indexes i and j
- $M_{i,j} := \max \begin{cases} M_{i-1,j-1} + s(T'_i, T_j) \\ M_{i-1,j} \\ M_{i,j-1} \end{cases}$



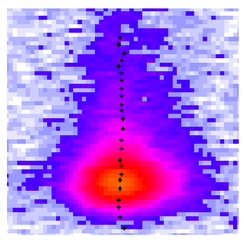
- Using log odds score function avoids additional computation
- Defining $s(T'_i, T_j) := \log \frac{g(m_j,\mu'_i,\lambda'_i,o'_i)}{g(m'_i+\delta,\mu'_i,\lambda'_i,o'_i)}$
- δ corresponds to the standard deviation of T'_i and depends on grid opening time, drift tube length and electric field strength

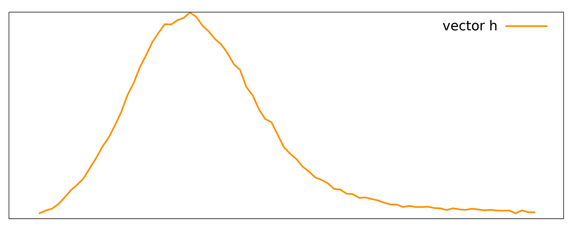


Step 3: Merging Peak Chain

Let $C = (P_1, \ldots, P_n)$ be a chain of one-dimensional Inverse Gaussian models:

1. Collect the peak height vector $h = (h_i)_{i=1,\ldots,n}$ at the individual modes





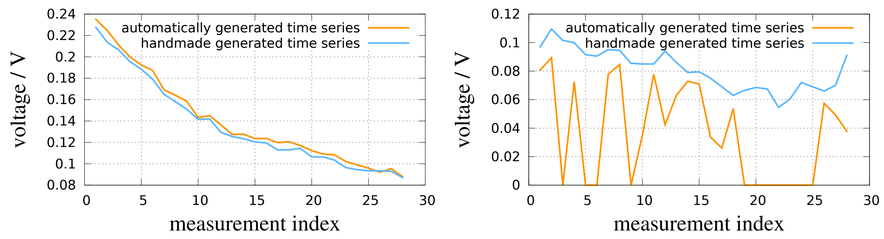
2. Scan vector h again by fitting a second order polynomial in a moving window
3. Estimate initial retention time parameters from polynomial using NLLS for Inverse Gaussian
4. Improve parameters by using EM algorithm [3]
5. Drift time parameters for Inverse Gaussian are the weighted average over all models within C
6. Reject model when not satisfying properties i.e. min peak height, width or model fitting score

Evaluation

Processing single spectrum:

Platform	*Average* 1	*Average* 2	*Average* 5
Desktop PC 2.8 GHz	7.79 ms	3.10 ms	1.52 ms
Raspberry Pi	211.90 ms	85.49 ms	37.82 ms

Time series of discovered intensities of two peaks:



Acknowledgements & References

DK, SR are supported by the Collaborative Research Center (Sonderforschungsbereich, SFB) 876 “Providing Information by Resource-Constrained Data Analysis” within project TB1, see http://sfb876.tu-dortmund.de.

[1] Kopczynski, D., Rahmann, S.: An online peak extraction algorithm for ion mobility spectrometry data. In: Algorithms in Bioinformatics, Lecture Notes in Computer Science, vol. 8701, pp. 232–246. Springer Berlin Heidelberg (2014)

[2] Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York, 2nd edn. (2006)

[3] Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological) pp. 1–38 (1977)