Paving the Way for Automated Clinical Breath Analysis and Biomarker Detection

Anne-Christin Hauschild[1,3], Jörg Ingo Baumbach[4], Jan Baumbach[1,3,2]

[1] *Computational Systems Biology Group, Max Planck Institute for Informatics, Saarbrücken, Germany*

[2] *Computational Biology group, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark*

[3] *Cluster of Excellence for Multimodal Computing and Interaction. Saarland University, Saarbrücken, Germany*

[4] *B & S Analytik, BioMedizinZentrum Dortmund, Germany*

a.hauschild@mpi-inf.mpg.de

Abstract: It is common knowledge that the human breath contains metabolites allowing to infer a patient's health status, especially for diseases related to the respiratory system. This information is encoded in the "volatolom", a combination of volatile organic compounds (VOCs) produced by the human metabolism and environmental perturbations. Nevertheless, due to a lack of alternative analytical techniques most of the traditional diagnostic methods are still based on invasive techniques, e.g. using blood or tissue samples.
During the last decade, the ion mobility spectrometer combined with a multi-capillary column (MCC/IMS) has become an established, inexpensive, and non-invasive bioanalytics technique for detecting VOCs with various potential applications in medical research. To pave the way for this technology towards daily usage in medical practice, different challenges still have to be solved. One of the main challenges is to establish an automated framework optimizing the processing algorithms in the pipeline yielding to an optimal performance for the final goals: Disease prediction and biomarker detection.
Although equivalent computational methods and standard procedures exist for other biomedical applications (e.g. sequence and microarray analysis) we still are lacking such a standard protocol in breath research. In four recently published papers presented here[HBJ12, HKD$^+$13, HSP$^+$12, SHBB13] we aimed at solving this challenge.

1 Introduction

Developed in the early 1970, the ion mobility spectrometers (IMS) has mainly been applied in the military domain, e.g. for detecting explosives. The resolution and precision was dramatically increased by combining the IMS with the multi capillary column for pre-separation. This technology leap opened up the horizon to many other applications. Furthermore, the IMS is inured to the moisture in exhaled air, comparably cheap, robust, and easy to use in every day practice, which makes it especially interesting for biotechnological and medical

applications [Bau09]. Due to these developments the IMS can now also be used for patient breath analysis and monitoring [WLM+11], identification of bacterial strains and fungi [BW06, PJV+11] and cancer sub-typing [WLM+11], just to name a few.

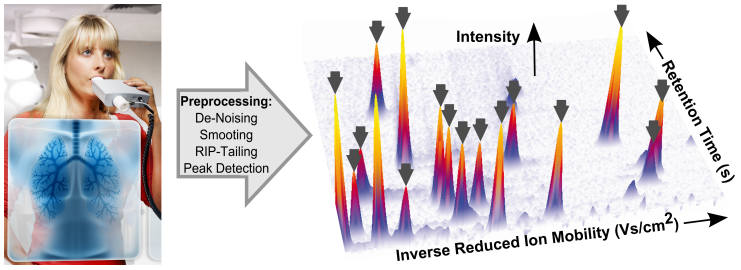

Figure 1: Result of the MCC/IMS measurement after preprocessing.

2 Computational Methods for MCC/IMS Data

Many challenges on the way to a comprehensive standard procedure in breath research have been solved in the last decade: the definition of a data format, the pre-processing (RIP detailing, smoothing, de-noising), the visualization and the evaluation using simple statistical techniques, see Figure 1. Nevertheless, to enable reliable disease prediction and biomarker detection we require an automated framework, following a standard protocol of carefully selected and adapted processing steps, see Figure 2.

Pre-processing and peak detection: With the increasing amount of performed measurements, a reliable and robust automated peak detection without manual intervention is an inevitable processing step. Although sophisticated peak detection approaches have been studied extensively in the last five to ten years, non of the proposed methods prevailed. This is most likely due to the fact that the assessment of the quality of these techniques has been done, if at all, solely by visual comparison with the manually selected peak lists[HSP+12]. Therefore, in a recent study we carefully evaluated four state-of-the-art approaches for automated MCC/IMS-based peak detection: local maxima search, watershed transformation, merged peak cluster localization, and peak model estimation (PME). We manually generated a gold standard with the aid of a domain expert (manual) and compared the performance of the automated methods with respect to two distinct criteria: (1) we systematically studied the classification performance of established machine learning methods (linear support vector machine and random forest) trained on the four peak detectors'

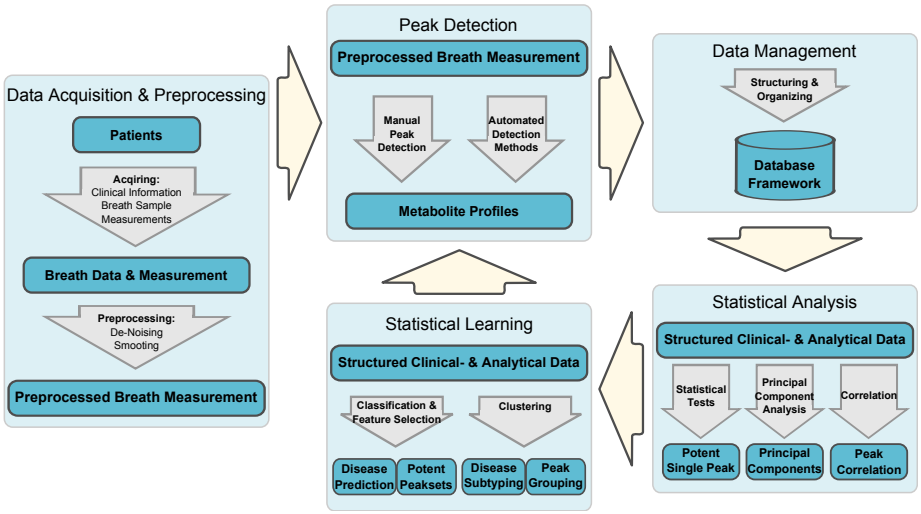

Figure 2: Automated framework for data analysis in breath research.

results and (2) we investigated the variance and robustness of the results, regarding overfitting and training set perturbations. In summary, all methods, though small differences exist, perform equally well in terms of classification error and are similarly robust against perturbations, while PME is most robust against overfitting. However, the trade off between a slightly higher accuracy (manual) and a huge increase in processing speed (automatic) has to be considered carefully [HKD+13].

Database and analysis platform: Another main part of an automated framework for MCC/IMS research is a flexible and comprehensive centralized data repository. To identify discriminating compounds as biomarkers, it is inevitable to have a clear understanding of the detailed composition of human breath and its potential confounders. Therefore, the challenge in MCC/IMS database development is not storing metabolic data in a fixed scheme but rather flexibly storing a huge set of heterogeneous clinical variables, varying for each study. To tackle this problem, we designed a comprehensive database application and analysis platform (IMSDB), combining metabolic maps with heterogeneous biomedical data in a well-structured manner. The model consists of a hybrid of the entity-attribute-value (EAV) model and the EAV-CR, incorporating the concepts of classes and relationships. Additionally, the IMSDB offers an intuitive user interface providing easy and quick access to the platform's functionalities: automated data integration and integrity validation, versioning and roll-back strategy, data retrieval as well as semi-automatic data mining

and basic machine learning capabilities. The standalone software system is based on the programming language Java and the WEKA learning package [HFH+09]. In conclusion, we are providing an intuitive software system for biologists, chemists, physicists and physicians conducting MCC/IMS-based research and diagnostics without demanding any computer skills [SHBB13].

Statistical learning: Finally, with respect to modern biomarker research and clinical diagnostics, one of the most important tasks is the automated classification of patient-specific data sets into different groups, e.g. healthy or not. In a recent pilot study, we investigated the potential of sophisticated statistical learning techniques for (1) VOC-based feature selection and (2) supervised classification on metabolic MCC/IMS data. Therefore, we utilized breath measurements of patients, either suffering from chronic obstructive pulmonary disease (COPD) or both, COPD and bronchial carcinoma (COPD+BC), as well as a healthy control group (CG). When distinguishing healthy from COPD all tested methods showed a reasonable performance between 84% (decision tree) and 94% (random forest) accuracy. However, the results indicated that further examination of the impact of bronchial carcinoma on COPD/no-COPD classification performance is necessary. The set of most important VOC features found by random forest coincided to a large extend with previous studies. Our findings demonstrate a generally high but improvable potential of statistical learning methods when applied to well-structured, medical MCC/IMS data. For more details see [HBJ12].

3 Summary

The MCC/IMS has become an established inexpensive, non-invasive bioanalytics technology for detecting VOCs with various potential applications in medical research. To pave the way for this analytical technique towards daily usage in medical practice, an automated framework following a standard protocol of carefully selected and adapted processing steps is needed. Therefore, our research focuses on the establishment of such a framework, with regard to the overall goal of disease prediction and biomarker detection. In this context, we evaluated different automated peak detection approaches and reported the peak model estimation as most robust [HKD+13]. The IMSDB combines two powerful computational tools, an extendible database as well as a machine learning toolkit, fully accessible through an intuitive graphical user interface, which will accelerate and expand the opportunities of clinical diagnostic research in the near future[SHBB13]. Finally, the application of sophisticated statistical learning methods for disease prediction and biomarker detection enables us to optimize the single steps of the proposed computational standard procedure for

breath research. The combination of the MCC/IMS methodology and the proposed automated standard procedure has the potential to successfully address a broad range of biomedical questions.

References

[Bau09] J. I. Baumbach. Ion mobility spectrometry coupled with multi-capillary columns for metabolic profiling of human breath. *Journal of Breath Research*, 3(3):1–16, 2009.

[BW06] J. I. Baumbach and M. Westhoff. Ion mobility spectrometry to detect lung cancer and airway infections. *Spectroscopy Europe*, 18(6):22–27, 2006.

[HBJ12] A.C. Hauschild, J. I. Baumbach, and Baumbach J. Integrated statistical learning of metabolic ion mobility spectrometry profiles for pulmonary disease identification. *Genet. Mol. Res.*, 11(3):2733–2744, 2012.

[HFH+09] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[HKD+13] A.C. Hauschild, D. Kopczynski, M. D'Addario, J. I. Baumbach, S. Rahmann, and J. Baumbach. Peak Detection Method Evaluation for Ion Mobility Spectrometry by Using Machine Learning Approaches. *Metabolites*, 3(2):277–293, 2013.

[HSP+12] A.C. Hauschild, T. Schneider, J. Pauling, K. Rupp, M. Jang, J. I. Baumbach, and J. Baumbach. Computational Methods for Metabolomic Data Analysis of Ion Mobility Spectrometry Data - Reviewing the State of the Art. *Metabolites*, 2(4):733–755, 2012.

[PJV+11] T. Perl, M. Juenger, W. Vautz, J. Nolte, M. Kuhns, B. Zepelin, and M. Quintel. Detection of characteristic metabolites of Aspergillus fumigatus and Candida species using ion mobility spectrometry - metabolic profiling by volatile organic compounds. *Mycoses*, 54(6):828–837, 2011.

[SHBB13] T. Schneider, A.C. Hauschild, J. I. Baumbach, and J. Baumbach. An Integrative Clinical Database and Diagnostics Platform for Biomarker Identification and Analysis in Ion Mobility Spectra of Human Exhaled Air. *Journal of Integrative Bioinformatics*, 10:733–755, 2013.

[WLM+11] M. Westhoff, P. Litterst, S. Maddula, B. Bödeker, and J. I. Baumbach. Statistical and bioinformatical methods to differentiate chronic obstructive pulmonary disease (COPD) including lung cancer from healthy control by breath analysis using ion mobility spectrometry. *International Journal for Ion Mobility Spectrometry*, pages 1–11, 2011.