Information-based medicine

Christof Schütte and Tim Conrad

Tumor diseases rank among the most frequent causes of death in western countries coinciding with an incomplete understanding of the underlying pathogenic mechanisms and a lack of individual treatment options. Hence, early diagnosis of the disease and early relapse monitoring are currently the best available options to improve patient survival. In addition, it seems imperative to develop biological markers that can identify patients who are likely to benefit from a specific form of treatment. The progress in understanding molecular mechanisms underlying pathologies has started to revolutionize diagnostics. Most of these mechanisms are controlled by proteins (e.g., hormones) which can be detected in the blood stream using mass spectrometry technology. The entire set of all expressed proteins at a certain time is called the proteome. Monitoring and understanding changes in the proteome is going to bring the next wave of progress in diagnostics, since many changes can be linked directly to disease onset and progression. We call these disease-induced changes *disease fingerprints* since they represent a trace that a particular disease left in the proteome.

A mass spectrometer can be used to uncover the proteome from just a drop of blood. It produces a signal where every protein is represented by some peaks whose intensities are proportional to the protein concentration profile. Proteomics-based diagnostics means to find the fingerprint of a disease in this signal. Every increase in sensitivity and robustness of the fingerprint identification yields earlier and more robust disease detection and results in an increase in therapy success rates for most serious diseases, such as cancer.

Mathematical and algorithmic problems and their solutions. Our approach to fingerprint detection is via signal classification based on mass spectrometry data of large patient cohorts. These signals are extremely high-dimensional (typically 100.000 dimensions for a low-resolution spectrum and more than 150 million dimensions for high-resolution spectrum) and often show a bad signal to noise ratio. In close cooperation with physicians we developed a specific Standard Operating Procedure (SOP) under which the blood sample has to be processed in order to reduce the signal to noise level to below 25 %. Even for such high-

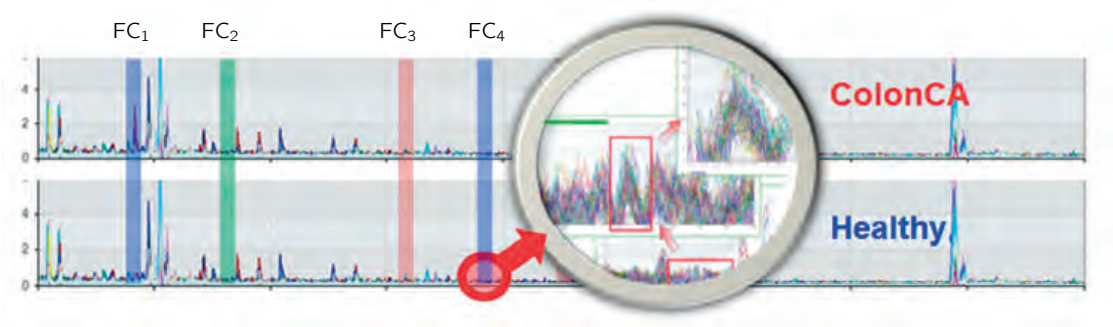

Figure 1. *This figure shows an illustrative example of mass spectra data and a fingerprint for Colon cancer with the exceptionally small number of four components ($FC_1 \ldots FC_4$). Note that only a very small fraction of the actual data set is shown. This data was acquired from two groups of individuals (after adequate preprocessing). The spectra in the top panel were created from blood serum of colon cancer patients. The lower spectra were created from healthy controls. The area inside the gray circle shows the magnified part of the red circled area to illustrate that our method allows detection of very small signals.*

quality mass spectrometry signals the high dimension renders standard signal classification infeasible. Therefore we invented a novel signal preprocessing technique that exploits knowledge about the physical processes underlying mass spectrometry, allows for peak detection across all available signals and results in peak detection with unprecedented accuracy. Based on the thus preprocessed signals we have designed novel sparse classification schemes. The idea behind these schemes is the following: the statistically significant differences between the classes ("healthy" and "different states of disease") results from a relatively small number of peaks that somehow reflect the proteins being characteristic for the disease in focus. This means that the fingerprint/classifier is sparse in comparison to the signal dimension even if the signals themselves are not sparse. Last but not least these preprocessing and classification techniques were implemented in a software environment able to handle this mass data (about 2.5 GB per patient, summing up to several TB for a typical patient population).

Impact and collaborations. The mathematical algorithms developed in this project were further improved towards real-world applicability in a subsequent BMBF-funded project within the ForMaT framework. In particular, components for handling very large medical data-sets from our clinical partners were added. This was done in very fruitful collaborations with our industrial partners IBM Germany and SAP Innovation. The pipeline has been applied to several data-sets and allowed to identify fingerprints for four different cancer types: lung, pancreas, colorectal, testicular (see, e.g., [1]). This was done together with our clinical partners from Helios Clinics, Charité – Berlin University Hospital, Leipzig University Hospital and Inselspital – Bern University Hospital. The resulting intellectual property has been patented [2] and is now been transferred into a spin-off company which will bring this to market. For their business plan for market entry the company won the second place in the *Berlin-Brandenburg business plan competition 2013*.

Further reading

[1] A. Leichtle, U. Ceglarek, P. Weinert, C. T. Nakas, J.-M. Nuoffer, J. Kase, T. O. F. Conrad, H. Witzigmann, J. Thiery, and G. M. Fiedler. Pancreatic carcinoma, pancreatitis, and healthy controls – metabolite models in a three-class diagnostic dilemma. *Metabolomics*, October 2012. URL: http://publications.mi.fu-berlin.de/1165/.

[2] M. von Kleist and C. Schütte. Patent no. de102010060311b3: Method for supporting planning, implementation and analysis of clinical studies, 2010.