UNIVERSITY OF AMSTERDAM

Chromatographic Profiling: From Samples to Information
S. Peters

Summary

Chromatography exists for over a hundred years and has become an important part of analytical chemistry. With the development of new instrumentations and columns that can measure more analytes at higher sensitivities simultaneously a new tool has been made possible: chromatographic profiling. Here, samples are measured untargeted and their entire profiles are then correlated to their properties, the chemical process or whatever type of research question needed to be answered. One special application of chromatographic profiling are the so-called –omics applications such as e.g. metabolomics.

In chromatographic profiling, huge quantities of data are obtained, and hypothetically, a tremendous increase in information can be obtained from profiling studies. Unfortunately, methods that can deal with the data in a timely manner are still lacking. To date, automated data pre-processing and analysis have become one of the major bottlenecks of large-scale sample analyses. In this thesis, we have tried to deal with some of the issues in a practical yet sound way. We have focused on developing new methods that either result in better data (*i.e.* more robust and reliable), better pre-processing or better analysis of the data, all with the aim to improve the link from analytical data to information.

In Chapter 1, a general introduction to the set-up of large-scale profiling studies is presented. The analytical workflow from experimental design to biological interpretation is discussed and all methods developed in this thesis are linked to the workflow. The importance of automation, whether it is of instrumental or data-analytical nature is highlighted. Only when all or most parts of the workflow are automated, reliable and robust data is obtained in a timely manner in large-scale experiments. Automating sample preparation is hereby seemingly one of the simplest steps, provided the appropriate instrumentation is available. In Chapter 2, an automated sample preparation method based on Micro-extraction in a Packed Sorbent (MEPS) for the analysis of plasma by gas chromatography is presented. As the micro-extraction material is incorporated into the gas chromatographic system, no manual handling of the samples is required. As the extraction material is not very specific (C18 in our case), broad chromatographic profiles could be obtained. Additionally, compounds in very low abundances could be measured in the presence of other high-abundant compounds.

All data analysis methods are either based on peak tables or on the raw (chromatographic) data. Alignment of the signal over all samples is hereby a necessity. While this is relatively simple for detected peaks it becomes more challenging when aligning the raw signal. Luckily, software (commercial or freeware)

exists that performs signal alignment; however, setting all parameter correctly can be difficult and time-consuming. In Chapter 3, we show how with little analytical effort, the parameter selection for alignment using any given software package can be optimised. The method is based on using quality control samples, often anyhow included in the analysis plan.

(Automatic) peak detection can immensely simplify further data analysis as the proceeding method does not have to deal with the parts of the data that are irrelevant (i.e. noise), providing it has been detected as such. In Chapter 4, a new peak detection algorithm for chromatograms obtained from two-dimensional gas chromatography is presented. Peaks are hereby detected as "mountains", i.e. a peaks' start and end point in the first and second chromatographic dimension are identified. In this way, quantitative data for the peaks are obtained.

In profiling studies less than perfect (chromatographic) data is often obtained, and depending on the complexity of the samples not all compounds are usually separated chromatographically. In such cases mathematical deconvolution presents a powerful alternative. In Chapter 5 a new method is presented that automatically determines the number of components to use in a mathematical deconvolution model. It is based on a cross-validation procedure that makes use of the fact that the overall error of cross-validation will first decrease when adding more components and then increase once the optimum number of components has been reached. In this way, a general method has been developed that does not require any manual intervention.

In many profiling studies it is not known what changes in the data can be expected. Therefore, in addition to the instrumental side that should be as versatile as possible, any proceeding data-analytical method should be non-specific and include the entire information available. However, if information on the samples is present *a priori* it should be included in the data analysis. This will not only simplify the method but also should result in more relevant information. The obtained results from broad, multivariate methods often require a high understanding of the data and biological knowledge and often the results are still subject for interpretation. Incorporating pre-knowledge in the method will force the model in a certain direction thus excluding results not relevant for the present data. In Chapter 6 and 7, two methods for the analysis of data obtained from kinetic studies are presented. The data sets that these methods were based on differ in one respect: in Chapter 6, the kinetic data were taken from several volunteers that were exposed to one particular dose and in Chapter 7, the kinetic data were taken from one volunteer at several doses. Thus, in both cases pre-knowledge on the structure of the data was present. In Chapter 7, an additional knowledge was that the responses of the several doses should follow the same trend. For both studies, two methods were developed that even though they differ inherently

in their mathematical nature they both aim at reducing the thousands of (often redundant) variables present in such data sets to a reasonable number.

In Chapter 8 a new method to express resolution in two-dimensional chromatography has been developed. Resolution is a very important measure when comparing sample chromatograms and it is often used to determine the goodness of a given chromatogram. In one-dimensional chromatography resolution is well-defined, but its definition cannot be directly transferred to data obtained from two-dimensional chromatography. We have developed a new resolution function based on the valley-to-peak ratios between peaks. We have defined a new measurement of “vicinity”, i.e. a measurement to check between which peaks in a two-dimensional (2D) chromatogram a resolution minimum occurs. Finally, the 2D resolution method developed can be used to determine the overall quality of a 2D chromatogram.