UNIVERSITY OF AMSTERDAM

Innovative Methods for Data Analysis in Analytical Chemistry Using Bayesian Statistics and Machine Learning.

M.T. Woldegebriel

Innovative Methods for Data Analysis in Analytical Chemistry Using Bayesian Statistics and Machine Learning

Michael Woldegebriel

Summary

Compound screening is a process aimed at identification of known or un-known compounds from any given biological/non-biological sample. As such, it is routinely applied in most analytical laboratories around the world. In general, the compound identification process can be approached either as targeted or untargeted screening. For both approaches, in most laboratories, highly sensitive separation techniques such as liquid chromatography coupled to high resolution mass spectrometer (LC-HRMS) is commonly used as the state-of-the-art technique. In recent years, there has been a growing interest to analyze compounds beyond target compound lists, and therefore a shift towards non-targeted screening is on its way.

In general, the process of both targeted and untargeted compound screening for LC-HRMS can be summarized into two main steps: (i) feature detection, and (ii) feature matching. The main difference between targeted and untargeted screening is that for targeted case, the first step (i) can be performed by looking into only predefined regions. These regions can be defined based on prior knowledge (previous experimentation), that is assumed to be the most probable location for compound of interest. Any feature present in those regions can then be assessed for a match with a reference profile of compound of interest. That way, the feature detection and matching can be performed simultaneously. On the other hand, for the case of untargeted screening, the first step (i) refers to efficient detection of all existing peak features within the raw-data, taking into account the entire multi-dimensional separation space. That way, regions consisting of signals generated by truly existing compounds can be detected, followed by exhaustive probabilistic assessment of features configuration for a match with a pattern of any candidate compound.

Due to the complexity and large size of high resolution datasets, almost all currently existing algorithms approach the above mentioned steps in overly simplistic manner; by first reducing the data-size using centroiding approach, and/or using a high cut-off value for the intensity (threshold). Such a stringent approach is vulnerable to loss of important chromatographic profiles (i.e. centroiding results in loose of peak shape in the mass/charge domain), as well as discards low abundant peak features with chemical relevance, too early in the data analysis pipeline. In addition, the non-exhaustiveness of the chemical formula databases normally utilized for feature matching is a limiting factor. In this thesis, one of the biggest challenges in separation science, i.e. algorithms incapability to fully utilize datasets generated by high resolution instruments, has been addressed in the context of forensic, food-safety and material science. However, the algorithmic frameworks can as easily be utilized in other application areas with similar scientific problems.

Chapter one of this thesis, gives a general introduction and a brief history of analytical chemistry with more emphasis in compound screening. In addition, the interdisciplinary nature of the discipline from earlier stages of its development, up to the era of 'big data' is also presented. Here, the benefit of Bayesian statistics and machine learning for handling the computationally challenging large datasets nowadays produced by high resolution instruments is also briefly discussed.

Chapter two of this thesis introduces a Bayesian untargeted peak detection algorithm developed for high resolution mass-spectrometry data. This chapter highlights the benefit of probabilistic feature detection opposed to the conventional binary (deterministic) algorithms. At this stage, the foundation for a probabilistic data analysis for high resolution mass-spectrometry data was laid.

Chapter three of this thesis presents an algorithm developed for the purpose of targeted compound screening in toxicology. In this chapter, a novel and robust Bayesian algorithm for

xenobiotic screening in forensic toxicology, that utilizes the concept developed and discussed in *chapter one*, is introduced. Further discussion on how this new method implies a paradigm shift in the way data is treated in the laboratory is also discussed. This approach, opposed to a frequentist method based on a threshold, applies a Bayesian framework to make use of all the evidence extracted from the chromatogram in a sequential state-of-knowledge updating process. Such a method requires no decision at a sub-step level preventing error propagation.

Chapter four is a natural extension of *chapter three.* In this chapter a Bayesian framework developed for utilizing additional evidence for targeted compound screening is presented. For this, the Bayesian framework presented in *chapter three* was further expanded to accommodate additional evidence from the fragment ions of an LC-MS/MS data, opposed to just being limited to precursor information. In addition, contrary to most existing open-source/commercial software for targeted-compound screening purposes, this method applies machine learning approach for sequentially updating the model utilized from every dataset.

Chapter five of this thesis presents a new method for compound identification in LC-HRMS data. As such, this novel method differs from the classical methods in the way the hypothesis itself is defined; 'the compound is present', opposed to answer the question 'the compound feature is present' by taking into account the probability of interfering compounds (i.e. isomers and isobaric compounds).

Chapter six of this thesis describes a novel solution to one of the most common, but unaddressed problem in chromatographic systems i.e. retention time shift. In this chapter, as an extension of *chapter five,* a novel method that tackles the retention time ambiguity problem in an elegant way is introduced. Taking into account the retention time shift can result in the presence of more than one potential peak with similar m/z value within the search region of compound of interest, the probabilistic framework presented in this chapter approach the

problem probabilistically, giving an improved result, opposed to the conventional methods usually applying the ‘nearest’, or ‘tallest’ peak criteria.

Chapter seven of this thesis discusses on the application of probabilistic peak detection for forensic DNA analysis. In this chapter, the Bayesian framework for peak detection originally developed and introduced for chromatography and mass-spectrometric data in *chapter two* to *six* proved to be ideally suited for peak detection in electropherogram produced by laser induced fluorescence multi-capillary electrophoresis (CE-LIF) for DNA fragments. As such, the benefit of Bayesian approach for forensic DNA analysis is also further discussed.