Mining of High-Resolution Mass Spectrometry Data to Monitor Organic Pollutant Dynamics in Aquatic Systems

Diss. ETH No. 23098

Mining of High-Resolution Mass Spectrometry Data to Monitor Organic Pollutant Dynamics in Aquatic Systems

A dissertation submitted to
ETH Zürich

for the degree of
Doctor of Sciences
(Dr. sc. ETH Zürich)

presented by
MARTIN JÜRGEN LOOS
MSc. Environmental Science, ETH Zürich
Dipl. Geoecology, University of Potsdam
born 25 October 1980
citizen of Germany

accepted on the recommendation of
Prof. Dr. Juliane Hollender, examiner
Heinz Singer, co-examiner
Dr. Steffen Neumann, co-examiner
Prof. Dr. Kristopher McNeill, co-examiner

2015

Summary

The ecological balance and human usage of aquatic environments are frequently affected by the widespread intentional or accidental release of anthropogenic compounds. As a particular case, the Rhine River has long been strained by a variety of emissions, whether from point sources like sewage treatment effluents and spill accidents or ubiquitous diffuse inputs from agriculture and horticulture. A long-term monitoring of the resultant pollutant levels is therefore crucial. Herein, especially the polar and therefore mobile portion of emitted organic compounds has come into focus. Termed micropollutants for their trace-level occurrence, these compounds can exert toxicological effects even at low concentrations and are composed of a complex range of pharmaceuticals, pesticides, surfactants or illicit drugs, amongst others. Many of the anthropogenic compounds are not suspected or even known to occur in the Rhine network, e.g., the sometimes hard-to-predict set of transformation products or unregistered industrial intermediates. Consequently, these compounds cannot be approached in a targeted manner. A relatively new method of choice for the chemical monitoring of such known and unknown micropollutants is thus the detection via high-performance liquid chromatography (LC) coupled to high-resolution mass spectrometry (HRMS). In this setup, electrospray ionization (ESI) offers a soft technique to transfer analytes from LC to HRMS without much fragmentation.

Among the few institutions responsible for detecting micropollutants in the named river network, the Swiss Rhine monitoring station (RÜS) in Basel has been equipped with LC-HRMS. Using a fully developed sampling strategy, the RÜS has acquired a long sequence of daily LC-HRMS measurements over the last years, comprising several hundred river samples. However, certain issues in the post-acquisition analysis have not been fully resolved yet and concern the simulation, reduction and automated trend analysis for LC-HRMS data. The aim of this thesis was to resolve these data mining issues in four steps.

A first part elaborated on the simulation of isotopic fine structures, necessary to compare the measured and theoretical mass spectra of compounds at high instrument resolution. Even for small compounds, dominating fractions of low-probable isotopologues exist, which must be efficiently pruned without

loss in computational performance. To this end, a novel transition tree approach was introduced to organize non-redundant, single-isotopic changes between isotopologues into separate tree branches. The latter can be efficiently pruned, with tree growth directed to find a relative instead of absolute pruning threshold first. The method was consequently able to outcompete existing approaches both in memory usage and computation time during an extensive performance comparison.

Benefitting from these transition trees, a second step improved the nontargeted isotopologue and adduct grouping of measured LC-HRMS signals. First, a large set of organic compounds from a public database was used to simulate isotopologue pairs and to sample defining characteristics between them, some of which are not covered by available low-resolution approaches. These characteristics were thereupon discretized in a recursive partitioning procedure and used to group coeluting measurement signals of isotopologues. A high recall and precision could therein be attested, based on a complementary evaluation with external simulation data and a targeted screening. When combined with a grouping for main ESI adducts, large fractions of measured LC-HRMS signals could thus be assorted into their chemical components.

A third step developed a first algorithm for a direct and unsupervised recognition of homologue series pattern in LC-HRMS data; that is, sets of signal series with constant shifts in mass and smooth changes in retention time. By introducing a specialized data structure, this novel approach swiftly revealed the patterns of numerous signal series even from very crowded spectra, despite the combinatorial complexity of this task. The detected series information could then be annotated to the grouped nontarget components. Further investigation also revealed multiple assignments of measured signals to different series, indicative of homologue series with different reoccurring chemical units.

Finally, the fourth step approached a primary goal of automatized trend detection to reveal riverine micropollutant spills from big data LC-HRMS sequences, incorporating the achievements in simulation and grouping. An automatized workflow for the fast extraction of chromatograms, picked signal peaks, time-intensity profiles, target and nontarget components, and trends was therefore implemented and equipped with a user interface. When tested for the routine monitoring at the RÜS, the workflow successfully prioritized numerous spill events, whereupon international alarms could be issued and the responsible emission sources of partly unknown micropollutants uncovered.

It can be concluded that environmental long-term monitoring of river systems with LC-HRMS results in data amounts which can neither be fully identified to appoint all the measured analytes nor sufficiently inspected by manual analysis alone. When fused with data mining strategies tailored to high-resolution acquisition and nontarget analysis, however, the emitted universe of polar aquatic pollutants can be routinely assessed for critical trends first and for selected chemical identification afterwards. A suite of five software tools for this former task has been made publicly available as part of this thesis (R packages *enviPat*, *enviPick*, *nontarget*, *nontargetData* and *enviMass*).

Zusammenfassung

Die umfangreiche und allgegenwärtige Emission von anthropogenen Stoffen in aquatische Systeme steht häufig im Konflikt mit der menschlichen Nutzung oder ökologischen Funktion selbiger Systeme. Ein beispielhafter Fall ist der Rhein, welcher kontinuierlich durch solche Emissionen beeinflusst wird, entweder in Form von Punktquellen (z.B. Kläranlagenausflüsse, Unfälle) oder durch diffuse Einträge aus der Landwirtschaft und dem Gartenbau. Eine langfristige Überwachung der resultierenden Schadstoffbelastungen ist daher unerlässlich, wobei der Eintrag organischer polarer Chemikalien verstärkt in den Fokus gerückt ist, zumal diese Chemikalien gut wassergängig und somit mobil sind. Letztere Substanzklasse wird aufgrund ihrer häufig niedrigen Konzentration auch als Mikroschadstoffe bezeichnet; einzelne Vertreter können zudem selbst bei starker Verdünnung noch toxische Beeinträchtigungen verursachen. Das zugrundeliegende Substanzspektrum wiederum ist vielfältig und betrifft unter anderem Stoffe wie Pharmazeutika, Pestizide, Tenside oder auch Drogen. Viele der im Rhein auftretenden Substanzen sind ausserdem noch nicht gänzlich aufgeklärt, da sie z.B. schwer hervorsagbare Transformationsprodukte oder intermediäre Industriechemikalien darstellen. Einerseits können solch unbekannten Substanzen nicht vorab eingegrenzt werden, da oft keine ausreichende Erwartungshaltung definierbar ist. Andererseits können polare Substanzen aber – wenngleich nicht identifizier – so dennoch häufig detektiert werden. In der chemischen Spurenstoffanalytik ist hierbei vor allem eine Kopplung aus (a) Hochleistungsflüssigkeitschromatographie, (b) Elektrospray-Ionisierung und (c) hochauflösender Massenspektrometrie zum Einsatz gekommen (abgekürzt LC-ESI-HRMS oder LC-HRMS).

Neben den verschiedenen mit der Detektion von Mikroschadstoffen im Rhein betrauten Einrichtungen betreibt auch die Rheinüberwachungsstation Basel (RÜS) eine solche LC-HRMS Analytik. Durch eine tägliche automatisierte Probennahmestelle konnte die RÜS hierbei über die letzten Jahre eine beachtliche LC-HRMS Messreihe aufbauen. Obgleich die analytische Seite stark ausgereift ist, ist in der Datenauswertung hinsichtlich Simulation von Messergebnissen, Datenreduktion und der automatisierten Trenddetektion jedoch Nachholbedarf ersichtlich. Die hier präsentierte Doktorarbeit behandelt diese Schwachstellen in insgesamt vier Schritten.

Ein erster Schwerpunkt zielte auf die Simulation von Isotopenfeinmustern ab, welche zum Vergleich von theoretischen und gemessenen Massenspektren benötigt werden. Da selbst kleine Moleküle einen Hauptanteil von Isotopologen mit vernachlässigbarer Auftretwahrscheinlichkeit aufweisen, sollten letztere effizient in solchen Berechnungen umgangen werden. Zu diesem Zweck wurde eine neuartige Berechnungsmethode erarbeitet, welche die Übergänge mit einzelnen Isotopen zwischen paarweisen Isotopologen in sogenannten Transitions-Bäume organisiert. Dabei wurden nicht nur redundante Übergänge vermieden, sondern auch einzelne Bereiche dieser Bäume entweder abgegrenzt oder vorranging auf wahrscheinliche Isotopologe hin überprüft. Der einhergehende Gewinn in Berechnungsgeschwindigkeit und Speicherverbrauch ist dabei besser als derjenige existierender Methoden, was durch umfangreiche Simulationsvergleiche gezeigt werden konnte.

Ein zweiter Schwerpunkt befasste sich mit einer Verbesserung der Isotopologen- und Adduktgruppierung für LC-HRMS Messdaten von unbekannten Substanzen. Unter Zuzug einer öffentlich zugänglichen Substanzdatenbank wurden hierfür zunächst umfassende Simulationen von Isotopologenpaaren berechnet. Die beobachteten Eigenschaften dieser Paare - die überdies deutlich von denjenigen aus niedrig aufgelöste Massenspektren abwichen – wurden mithilfe einer rekursiven Partitionierung dann diskretisiert und zur Sortierung von gemessenen Isotopensignalen herangezogen. Validierungsschritte mit externen simulierten Isotopologenpaaren sowie mit solchen aus gemessenen und a priori aufgeklärten Paaren wiesen dabei sowohl eine hohe Wiederfindungrate als auch eine zufriedenstellende Klassifizierungsgenauigkeit auf. In Kombination mit einer Gruppierung der wichtigsten ESI-Addukte konnten hernach grosse Signalanteile in einzelne chemische Messkomponenten zusammengeführt werden.

Ein nachfolgender dritter Teilschritt entwickelte einen ersten Algorithmus zur direkten und flexiblen Erkennung von Messmustern, welche das Vorhandensein von homologen Reihen in LC-HRMS Datensätzen nahelegen. Diese Muster charakterisierten sich einerseits durch konstante Massenabstände zwischen mehreren verketteten Messsignalen, andererseits aber auch durch gleichmässige Veränderungen in der Retentionszeit. Durch Ausarbeitung einer angepassten Datenstruktur konnten diese Muster nunmehr mit hoher Rechengeschwindigkeit detektiert werden, selbst in mit Messsignalen stark überfrachteten Messungen und trotz der Vielzahl an kombinatorischen Möglichkeiten für diese Mustererkennung.

In einem letzten übergeordneten Aufgabenbereich wurde eine automatisierte Trendüberwachung von Mikroschadstoffen durch das Filtern von LC-HRMS Datensätzen und unter Zuzug der bisher erarbeiteten Teilschritte (Simulation, Signalgruppierung und Mustererkennung von Homologen) erstellt. Der resultierende Workflow umfasste u.a. die Extraktion von Chromatogrammen, die Detektion von Signalpeaks, die Erstellung von Intensitäts-Zeitprofilen, die oben erwähnte Komponentenbildung und die Priorisierung von Intensitätsanstiegen um rasch auf plötzlich ansteigende Intensitätsveränderungen hinweisen zu können. Weitergehend wurde eine Benutzeroberfläche generiert um an der RÜS angewendet werden zu können. Dort konnten infolgedessen zahlreiche Intensitätstrends im Rhein nachgewiesen werden – und im Rahmen von internationalen Alarmfällen nicht nur etliche Verursacher, sondern auch einige bis dato unbekannte Mikroverunreinigungen zur Aufklärung gebracht werden.

Zusammengefasst lässt sich festhalten, dass eine umweltspezifische und langfristige Überwachung von Flüssen wie dem Rhein ohne LC-HRMS kaum machbar ist, jedoch die durch ausgedehnte Generierung von Messdaten anfallenden Anforderungen der Datenanalyse manuell nicht mehr hinreichend erschöpfbar sind. Neue Möglichkeiten ergeben sich daher aus der Erstellung von automatisierten Workflows. Hier können aus der Vielzahl der durch das breite Substanzspektrum hervorgerufenen Messsignale einzelne Trends herausgefiltert werden – ohne dass eine völlige Identifizierung der gemessenen Matrix nötig wäre, welche stattdessen vereinzelt nachfolgen kann. Die im Rahmen der Doktorarbeit erarbeiteten Methoden wurden in fünf Softwarepaketen implementiert und öffentlich verfügbar gemacht (R Pakete *enviPat*, *enviPick*, *nontarget*, *nontargetData* und *enviMass*).