

Strategies for high resolution profiling of natural extracts: UHPLC-MS and physicochemical approaches for early metabolite identification

EUGSTER, Philippe

Abstract

Natural products (NPs) play a key role in chemical biology and drug development thanks to their high diversity in chemical space. This diversity however renders their separation and universal detection challenging. Recently, ultra-high pressure liquid chromatography (UHPLC) systems have been recognised as the most versatile technique for the efficient separation of NPs in crude complex mixtures. This work firstly investigated the separation of small molecules and peptides in complex natural mixtures by UHPLC and ion mobility spectrometry (IMS). Secondly, a dereplication (rapid identification of known compounds) method was developed for the identification of NPs present in natural extracts by high resolution mass spectrometry (HR-MS) with the combined use of heuristic filters, chemotaxonomic information and a new and original retention prediction method. This thesis showed that the new developments in UHPLC and HR-MS applied to high resolution metabolite profiling of complex natural matrices give promising perspectives in dereplication and metabolomics.

Reference

EUGSTER, Philippe. *Strategies for high resolution profiling of natural extracts: UHPLC-MS and physicochemical approaches for early metabolite identification*. Thèse de doctorat : Univ. Genève, 2013, no. Sc. 4589

URN : [urn:nbn:ch:unige-343735](http://nbn-resolving.org/urn:nbn:ch:unige-343735)

DOI : [10.13097/archive-ouverte/unige:34373](http://dx.doi.org/10.13097/archive-ouverte/unige:34373)

Available at:

<http://archive-ouverte.unige.ch/unige:34373>

Disclaimer: layout of this document may differ from the published version.



UNIVERSITÉ
DE GENÈVE

Strategies for High Resolution Profiling of Natural Extracts:

UHPLC-MS and Physicochemical Approaches for Early Metabolite Identification

THESE

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention sciences pharmaceutiques

par

Philippe Eugster

de

Speicher (AR)

Thèse N° 4589



**UNIVERSITÉ
DE GENÈVE**

FACULTÉ DES SCIENCES

**Doctorat ès sciences
Mention sciences pharmaceutiques**

Thèse de *Monsieur Philippe EUGSTER*

intitulée :

**" Strategies for High Resolution Profiling of Natural Extracts :
UHPLC-MS and Physicochemical Approaches for Early
Metabolite Identification "**

La Faculté des sciences, sur le préavis de Messieurs J.-L. WOLFENDER, professeur ordinaire et directeur de thèse (Section des sciences pharmaceutiques), P.-A. CARRUPT, professeur ordinaire et codirecteur de thèse (Section des sciences pharmaceutiques), E. ALLÉMANN, professeur ordinaire (Section des sciences pharmaceutiques), C. BICCHI, professeur (Dipartimento di Scienza e Tecnologia del Farmaco, Torino, Italia) et M. AFFOLTER, docteur (Proteins, Carbohydrates & Tracers Group, Analytical Sciences Pillar Nestec Ltd, Nestlé Research Center, Lausanne, Suisse), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 30 août 2013

Thèse - 4589 -


Le Doyen, Jean-Marc TRISCONE

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

À mes parents.

Foreword

Scientific research has become nowadays a multidisciplinary approach of the questions to be answered. Projects often require various knowledge, and, as an example, many projects in pharmaceutical sciences involve chemists, biologists, and statisticians. This is the reason why so many projects in the School of Pharmacy Geneva-Lausanne (EPGL) are close collaborations between laboratories, or with external partners.

Thesis works follow this evolution, and this one is no exception. It deals with the analysis of natural products in complex natural samples by liquid chromatography coupled with mass spectrometry, to quickly determine their composition using as few experiments and data processing as possible. It is composed of two main parts. The first part aims at optimising the high resolution separation of complex natural samples, in both theoretical and experimental viewpoints. The second part aims at identifying the components of these natural samples from the online data obtained, and using new tools based on chemotaxonomy knowledge and physicochemical properties of the natural products.

This thesis work required knowledge in various domains such as phytochemical analysis, natural products chemistry, fundamentals of liquid chromatography, molecular physicochemical properties, statistical analysis, biological assays, and taxonomy. Since it is not possible to deal with all these areas alone, it is thus of prime importance to collaborate with experts in these domains. Therefore, this thesis work is co-directed by Prof. Jean-Luc Wolfender, head of the laboratory of phytochemistry and bioactive natural products, and by Prof. Pierre-Alain Carrupt, head of the laboratory of pharmacology, and involved many collaborations, for example with the laboratory of analytical pharmaceutical chemistry, and with external partners such as venom experts from Atheris Laboratories in Geneva, or Brazilian phytochemists of the São Paulo University.

Such diversity is a chance for a PhD thesis, and I hope that the reader will find as much interest as I had doing this research.

Most of the chapters of this thesis are based on research articles, reviews, and book chapters published or submitted during the thesis work.

Abstract

Natural products (NPs) are known to possess a very high diversity in chemical space, and play a key role in chemical biology and drug development. NPs are extracted from various natural organisms and include proteins, peptides and small molecular weight molecules. The whole array of small molecular weight metabolites found in a given organism, known as the metabolome, can be extremely large and has been estimated to contain a few thousand constituents. The high chemical diversity of secondary metabolites is directly linked to a high variability of the intrinsic physicochemical properties of NPs, which render the separation and universal detection of NPs extremely challenging in complex biological matrices. In this respect, ultra-high pressure liquid chromatography (UHPLC) systems, using sub-2 μm packing columns, and developed in the early 2000s, have been recognised as the most versatile technique for the efficient separation of NPs in crude mixtures without the need for complex sample preparation. This new technology has allowed a remarkable decrease in analysis time and increase in peak capacity, sensitivity and reproducibility compared to conventional HPLC. UHPLC is nowadays used more and more in dereplication and metabolomics applications, in conjunction with both photodiode array (PDA) and MS detection.

The first part of the thesis is dedicated to the optimisation of NP separation. Indeed, high resolution separations are indispensable for obtaining high quality spectrometric information in NP research. Hence, fundamental chromatographic

parameters were studied to optimise the UHPLC profiling of complex samples, such as a cone snail venom and a plant extract, containing peptides and small molecules respectively. The optimised method provided high peak capacities for high resolution metabolite profiling and proved its applicability in peptidomic or metabolomic studies for early metabolite identification or peptide deconvolution.

Despite its high resolution, UHPLC was not able to separate closely related isomers, frequently contained in natural matrices. The capacity of ion mobility spectrometry to separate the constituents of complex natural samples was then evaluated. This technique separates analytes based on mechanisms different from LC and provides a high number of detected features, and, most importantly, an efficient separation of closely related isomers that were not separated using LC.

The second part of the thesis deals with the dereplication of NPs present in natural extracts. Dereplication is the process of rapid identification of known compounds present in a mixture without classical isolation steps. This identification step was performed based on high resolution MS (HR-MS) data by the combined use of heuristic filters, chemotaxonomic information and retention information, and was applied to a chemotaxonomic study of several *Lippia* species. In this study, the high quality of the profiling data and the applied multivariate data analysis expanded the knowledge of the chemical relationships existing between the various *Lippia* species investigated.

Finally, a method for the retention time prediction of NPs was developed, which is challenging because of the high chemical diversity of NPs. While still limited, the

prediction of retention times may however be extremely useful in many NP-related applications, e.g. metabolomics, or as an additional tool for dereplication.

These various studies, based on high resolution metabolite profiling of complex natural matrices and performed on both LC and MS dimensions, showed promising perspectives offered by the new development in UHPLC and HR-MS for dereplication and metabolomics. However, to fully exploit the possibilities offered by these huge instrumental advances in high-resolution profiling, automated software tools are required to deal with the increasing amount of data acquired. The recent developments in this field are promising but insufficient. It is therefore important in the future to integrate bioinformatics tools in natural product analysis procedures to automatically extract relevant information online and quickly deconvolute complex biological matrices.

Still, these new advances are welcome since pharmaceutical companies had gradually abandoned NP research over the last decades, while embracing the development of combinatorial chemistry and modern high throughput screening techniques. Indeed, if the techniques used to discover new bioactive natural compounds become more efficient in terms of throughput and efficiency, pharmaceutical companies will probably turn back toward NP research.

Résumé

Les produits naturels (NPs) sont caractérisés par une grande diversité dans l'espace chimique et jouent un rôle-clé en biologie chimique et dans le développement de nouveaux médicaments. Ces NPs sont extraits d'organismes naturels très divers et comportent différents types de métabolites, tels que des protéines, peptides et composés de faible poids moléculaire. L'ensemble de ces derniers, le métabolome, peut être très grand et comporter plusieurs milliers de composés. L'importante chimiodiversité des NPs se traduit par une grande diversité de leurs propriétés physicochimiques, et implique que leur séparation et détection sont complexes, en particulier dans des matrices complexes. La chromatographie liquide à ultra haute pression (UHPLC) s'est imposée dès son introduction au début du millénaire comme la technique de référence pour l'analyse de NPs dans des mélanges complexes. L'UHPLC est caractérisée par des temps d'analyse faibles, des capacités de pics élevées ainsi que de grandes sensibilité et reproductibilité. C'est pourquoi cette technique, généralement couplée à un détecteur à photodiodes (PDA) ou à un spectromètre de masse (MS), est de plus en plus utilisée en déréplication et en métabolomique.

La première partie de la thèse est consacrée à l'optimisation de la séparation des NPs dans des matrices complexes. En effet, une séparation à haute résolution permet l'acquisition de données spectrométriques de bonne qualité, c'est à dire sans pic interférents. Les paramètres chromatographiques fondamentaux ont été

étudiés en vue de leur optimisation, lors du profilage de deux échantillons naturels types, à savoir un venin de cône marin et un extrait de plante, contenant majoritairement des peptides et des composés de faible poids moléculaire, respectivement. La méthode optimisée a fourni une capacité de pic élevée pour le profilage à haute résolution et s'est montrée adaptée aux études métabolomiques ou peptidomiques.

Malgré sa haute résolution, l'UHPLC n'est pas capable de séparer des isomères proches, fréquemment rencontrés dans les matrices naturelles. Dès lors, la capacité de la spectrométrie à mobilité ionique (IMS) à séparer les composants des matrices naturelles a été évaluée. Cette technique sépare les analytes par de des mécanismes différents de ceux à la base de l'UHPLC, détecte un grand nombre de composants, et, contrairement à la LC, sépare facilement des isomères proches.

La deuxième partie de cette thèse traite de la déréplication de composés présents dans des extraits naturels. La déréplication est le processus qui vise à identifier rapidement des composés connus dans un mélange sans avoir à effectuer un isolement classique. Ce processus d'identification a été réalisé à partir des données MS à haute résolution (HR-MS) en utilisant des filtres heuristiques, chimiotaxonomiques et l'information tirée de la rétention. Il a été appliqué lors d'une étude chimiotaxonomique de plusieurs espèces du genre *Lippia*, qui a permis d'améliorer les connaissances des liens entre différentes espèces de ce genre.

La dernière étude s'est attachée au développement d'une méthode de prédiction de rétention de NPs en UHPLC. Bien qu'elle soit encore limitée, la prédiction de

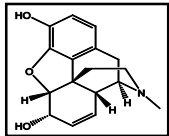
rétection est extrêmement utile dans de nombreuses applications liées aux NPs, telles qu'en métabolomique ou comme outils supplémentaire de dérégulation.

Ces différentes études, basées sur le profilage à haute résolution de mélanges naturels, et qui s'appuient sur les dimensions chromatographique et spectrométrique, ont montré les perspectives prometteuses offertes par les nouveaux développements en UHPLC et HR-MS pour les études de dérégulation et en métabolomique. Cependant, des outils efficaces sont nécessaires pour faire face à la quantité croissante de données acquises par le profilage à haute résolution. Les développements récents dans ce domaine sont prometteurs, mais insuffisants. Il est donc important de mettre au point des outils bio-informatiques pour extraire automatiquement les informations pertinentes acquises en ligne.

Enfin, ces nouvelles avancées dans l'analyse de NPs sont bienvenues, car elles permettront probablement d'améliorer le processus de découverte de nouveaux composés bioactifs naturels en termes de débit et d'efficacité. Ainsi, la recherche de NPs retrouvera grâce auprès des compagnies pharmaceutiques qui l'ont peu à peu abandonnée durant les dernières décennies au profit de la chimie combinatoire et des techniques de criblage à haut débit modernes.

I. General Introduction

Natural Products

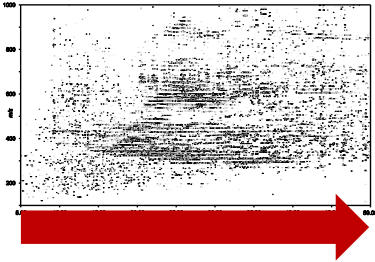


II. Introduction to UHPLC in Natural Products Analysis

III. Optimisation of UHPLC Resolution

IV. Ion Mobility Spectrometry: an Additional Separation Dimension

Separation

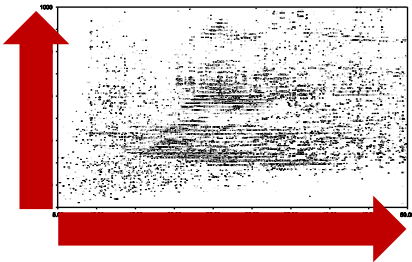


V. Rational Approach for LC-MS Online Dereplication

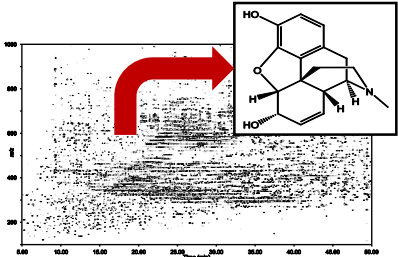
VI. LC-MS Online Dereplication - Practical Application to a Crude Plant Extract

VII. Retention Prediction: an Additional Tool for Dereplication

Identification



VIII. Concluding Remarks



Contents

Foreword	5
Abstract	7
Résumé	10
Contents	16
Abbreviations.....	23
Chapter I - General Introduction.....	27
1. Interest of natural products.....	29
1.1. The golden age of NP-based drug discovery.....	29
1.2. The pharmaceutical R&D NCE productivity decline	31
1.3. The role of natural products in modern drug discovery	32
1.4. Sources of natural products.....	37
2. Classical chemical investigation of samples of natural origin	39
2.1. General procedure for the chemical investigation of natural extracts	39
3. Modern techniques for natural product discovery.....	41
3.1. Miniaturised and integrated setup for bioactivity-guided isolation.....	41
3.2. Metabolomic strategy and peak annotation.....	43
4. LC-MS high resolution metabolite profiling	45
4.1. High resolution separation.....	45
4.2. High resolution MS detection	47
4.3. High resolution profiling for dereplication and metabolite identification	47
5. Other uses of natural products	51
6. Aim of thesis	53
7. References	54
Chapter II – Introduction to UHPLC in Natural Products Analysis	59
Foreword	61
Abstract	64
1. Introduction.....	65
1.1. Implementation of UHPLC in NP Analysis	66
2. Multiple Facets of UHPLC in NP research	68

2.1.	UHPLC Detectors used for NP Analysis	68
2.2.	Targeted vs Untargeted Analyses of NPs	69
2.3.	Column Phase Chemistries for NP Analysis.....	72
3.	Fast Targeted Analysis	75
3.1.	UHPLC-UV	76
3.2.	UHPLC-MS.....	77
4.	Fast Non-targeted Analysis, Fingerprinting and Metabolomics	83
4.1.	UHPLC-MS for Plant Metabolomics	83
4.2.	UHPLC-MS/MS-based Targeted Metabolomics	87
4.3.	UHPLC Fingerprinting for QC	88
4.4.	Chemotaxonomic Studies	88
5.	High-resolution Profiling and Metabolite Identification	89
5.1.	Very High-resolution Profiling.....	90
5.2.	LC x LC for Improved Resolution	90
5.3.	Metabolite Identification and Dereplication	91
6.	Conclusions.....	93
	Acknowledgments	94
	References	95
	Chapter III - Optimisation of UHPLC Resolution.....	103
	Foreword	105
	Abstract	110
1.	Introduction.....	112
2.	Experimental.....	114
2.1.	Experimental design	114
2.2.	Chemicals.....	114
2.3.	Sample preparation	114
2.4.	Instrumentation and analytical conditions	116
2.5.	S value determination.....	117
2.6.	System and column characterisation.....	117
2.7.	Experimental peak capacity determination	119
2.8.	Theoretical peak capacity calculation	120
2.9.	Evaluation of the number of resolved peaks	120
2.10.	<i>Conus</i> venom temperature stability study.....	120
3.	Results and discussion	122

3.1.	Peak capacity for the evaluation of profiling performance.....	122
3.2.	Comparison of the experimental and calculated peak capacity values	124
3.3.	Effect of the nature of the analytes on experimental peak capacity	125
3.4.	Effect of the flow rate on experimental peak capacity	127
3.5.	Effect of the temperature on experimental peak capacity	128
3.6.	Effect of gradient time on the calculated peak capacity.....	130
3.7.	Effect of the particle diameter and of the column geometry on the calculated peak capacity 132	
3.8.	Effect of the detection mode on peak capacity	136
3.9.	MS/MS deconvolution applications	136
4.	Conclusion	138
	Acknowledgements	141
	References	142
	Chapter IV - Ion Mobility Spectrometry: an Additional Separation Dimension.....	147
1.	Introduction	149
2.	Ion mobility spectrometry	150
3.	Evaluation of drift time IMS for the metabolite profiling of complex mixtures	153
3.1.	Metabolite profiling of a <i>Ginkgo biloba</i> extract by IMS- and UHPLC-TOF-MS.....	153
3.2.	Separation of closely related stereoisomers by drift time IMS and UHPLC	155
4.	Conclusion	158
5.	References	160
	Chapter V – Rational Approach for LC-MS Online Dereplication	163
	Foreword	165
	Abstract	177
1.	Introduction	178
2.	Materials.....	180
2.1.	Solvents and reagents.....	180
2.2.	Equipment	180
3.	Method	182
3.1.	Sample preparation for dereplication based on UHPLC-QTOF-MS.....	182
3.2.	3.1.2 Sample preparation	182
3.3.	UHPLC-QTOF-MS analysis	183
3.4.	Data processing	184
3.5.	Targeted LC-MS isolation	190

3.6. Micro-flow NMR analysis	193
4. Notes	194
Acknowledgments	198
References	199
Chapter VI - LC-MS Online Dereplication - Practical Application to a Crude Plant Extract.....	203
Foreword	205
Abstract	208
1. Introduction	210
2. Experimental.....	212
2.1. Chemicals.....	212
2.2. Plant material.....	212
2.3. Extraction and concentration	212
2.4. Sample preparation	214
2.5. HPLC fractionation of isomeric flavanone glucosides	214
2.6. UHPLC-PDA-ESI-TOF-MS experiments	214
2.7. UHPLC-PDA-ESI-TOF-MS data processing and analysis.....	215
2.8. Hierarchical clustering analyses.....	215
3. Results and discussion	216
3.1. Optimisation of the UHPLC-PDA-TOF-MS conditions.....	216
3.2. Study of the interconversions of some flavanones	217
3.3. Comparison of the phenolic profiles of all extracts	218
3.4. Dereplication procedure.....	218
3.5. Phytochemical and chemotaxonomic considerations.....	227
4. Conclusion	230
Acknowledgements	232
References	233
Chapter VII – Retention Prediction: an Additional Tool for Dereplication	237
Foreword	239
Abstract	243
1. Introduction	244
2. Experimental Section	246
2.1. Chemicals and Sample Preparation	246
2.2. UHPLC-TOF-MS Experiments	246
2.3. Calculation of the Physicochemical Parameters	246

2.4.	Clustering and PLS Regressions.....	247
2.5.	ANN Models.....	248
3.	Results and Discussion.....	249
3.1.	NP Database of RT and Physicochemical Parameters.....	249
3.2.	Development of Preliminary Models.....	250
3.3.	Development of Refined Models based on NP Clusters.....	250
3.4.	Development of the ANN Model.....	257
3.5.	Applicability of the Prediction Models.....	258
4.	Conclusion.....	263
5.	Supporting information.....	264
	Acknowledgements.....	281
	References.....	282
	Chapter VIII – Concluding Remarks.....	287
1.	High resolution metabolite profiling and online dereplication.....	289
2.	Maturity of LC-MS instrumentation.....	291
2.1.	The LC dimension.....	291
2.2.	The MS dimension.....	291
3.	Constraints in post-LC-MS analysis steps.....	293
3.1.	Automated molecular formulae annotation.....	294
3.2.	Online metabolite identification.....	294
4.	Non-universality of analytical techniques used in NP research.....	296
5.	The future of the natural product research.....	297
	References.....	299
	Acknowledgements.....	303
	Appendices.....	309

Abbreviations

CapNMR	capillary NMR
CID	collision-induced dissociation
DAD	photodiode array detection
DB	database
EI	electron impact (ionisation)
ESI	electrospray ionisation
GC	gas chromatography
HCA	hierarchical clustering analysis
HILIC	hydrophilic interaction chromatography
HPLC	high performance liquid chromatography
HR-MS	high resolution mass spectrometry
HTS	high throughput screening
LC	liquid chromatography
LC-MS	liquid chromatography coupled to mass spectrometry
MS	mass spectrometry (or mass spectrometer)
MS/MS	tandem mass spectrometry
MS ⁿ	multi-stage mass spectrometry
MW	molecular weight

NI	negative ionisation
NMR	nuclear magnetic resonance
NP	natural product
PCA	principal component analysis
PI	positive ionisation
PLS	partial least square (regression)
QTOF	quadrupole-time-of-flight
R&D	research and development
RP-LC	reversed-phase liquid chromatography
RT	retention time
RT _{exp}	experimental retention time
RT _{pred}	predicted retention time
S	parameter describing the chromatographic behaviour or an analyte
SPE	solid-phase extraction
TLC	thin-layer chromatography
TOF	time-of-flight
UHPLC	ultra-high pressure liquid chromatography
UV	ultraviolet

Chapter I - General Introduction

1. Interest of natural products

1.1. The golden age of NP-based drug discovery

For thousands of years, natural products (NPs) played – and still play – an important role in medicine and drug discovery. Until the beginning of the 20th century, nature was the only source of drugs, and medical documents from the Egyptians, Chinese, Greeks, and later Arabs and European monks detail the use of herbal medicines or other medicines of natural sources [1]. Morphine was the first isolated NP, in the beginning of the 19th century: Friedrich Sertürner, a pharmacist's apprentice, isolated morphine from opium produced by cut seed pods of *Papaver somniferum* L. (Figure I.1) [2]. From this time, many NPs from plants were isolated, purified, clinically studied and administered [3]. Later, after the Second World War, modern techniques expanded and NPs were often used as scaffolds for the development of new drugs [4, 5].

Over the last three decades, however, important revolutions have occurred in research and development (R&D) in the pharmaceutical industry, such as the development of high-throughput screening (HTS) techniques, genomics and combinatorial chemistry. Many pharmaceutical companies have dramatically reduced their investments in NP research, considered a slow and expensive technique of drug discovery for several reasons, including the complexity of isolation procedures and the supply problems, as well as the intellectual property concern and the difficulties with the collection of biomaterials which are consequences of the 1992 Rio Convention on Biological Diversity [3, 5-7]. On the other hand, the large libraries of compounds obtained through combinatorial chemistry and their subsequent testing by HTS techniques have brought the highest hopes of efficient and low-cost discovery of active lead compounds. The pharmaceutical R&D focused on these promising techniques.

Natural products (NPs) are chemical compounds from natural sources such as plants, animals or microorganisms. They have usually complex 3D structures since they are biosynthesised by specific enzymes and hold chiral centers with well-defined configuration.

Combinatorial chemistry is the generation of large collections, or 'libraries', of compounds by synthesising combinations of a set of smaller chemical structures.

The **Convention on Biological Diversity** signed in Rio de Janeiro in 1992 aims at developing strategies for the conservation and sustainable use of biological diversity.

A **lead compound** is a bioactive compound whose chemical structure is used as a starting point to develop new drugs with improved properties.

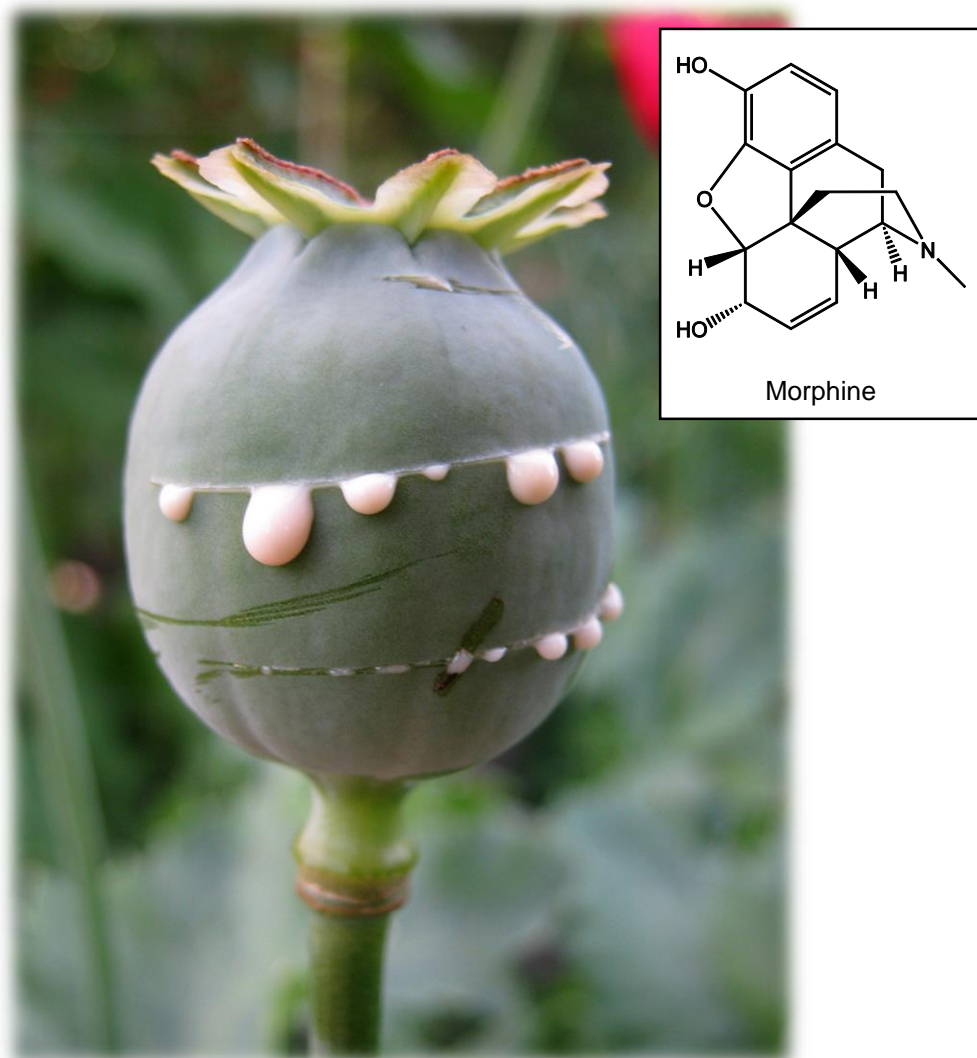


Figure I.1. Seed pod of *Papaver somniferum* L. with latex. Inset: Structure of morphine. Morphine, the widely used and potent opiate analgesic drug is extracted from the latex of unripe seed pods of *Papaver somniferum* (Papaveraceae). The use of opium (which is the dried latex) as a postoperative analgesic was first mentioned by James Moore in 1784. This alkaloid was isolated for the first time by Friedrich Sertürner in 1804 and commercialised by Merck 25 years later. Its structure was elucidated only in the beginning of the 20th century [2]. Photo: Nigel.

1.2. The pharmaceutical R&D NCE productivity decline

Unfortunately, the modern R&D techniques introduced in the last decades such as combinatorial chemistry and high throughput screening didn't provide the promised results [8]. As an example, only one drug, resulting exclusively from combinatorial chemistry was approved: the antitumor compound known as sorafenib (Nexavar, from Bayer) [4].

This example demonstrates the paradox of today's pharmaceutical R&D: although huge technological advances have been made over the last 50 years, the number of drugs arriving on the market is continuously decreasing [9]. Moreover, a recent review [10] showed a decrease in the number of new drugs approved by the American FDA per billion US\$ (inflation-adjusted) spent on R&D, more precisely, this number halved roughly every 9 years, as illustrated in Figure I.2.

Several authors have tried to explain this decrease in number of new drugs approved. The most cited causes are (1) regulatory pressure, (2) the "has to be better" issue, and (3) drug-likeness and target-based problems.

(1) Regulatory pressure is probably the most cited cause of the decrease of the number of new approved drugs. Indeed, the tendency today is to increase the number of clinical trials, with larger populations treated in each case. Moreover, rare side effects have to be better investigated, partly because of media pressure (this was recently the case with oral contraceptives). Because of this,

the development time and cost of a new drug are largely increased [9, 10].

(2) The "has to be better" issue is linked to the high number of efficient drugs on the market for a given indication. To get approval and reimbursement for a new drug, it has to be better than the existing treatments in at least one domain such as efficacy, cost, or safety, and equal in the other domains [9, 10]. Because of this, there is no more (or less) research in some therapeutic areas while R&D activity focuses on other axes that are usually more complex resulting in less financial profit. As an example, since statins are largely accepted, widely used, well-tolerated, efficient, and because of the presence of cheap generic drugs, there is probably very little R&D activity to provide new hypocholesterolemic agents.

(3) As mentioned above, the modern pharmaceutical R&D strategy is partly based on methods providing a high number of drug candidates, such as combinatorial chemistry and high throughput screening methods. Models or filters were thus developed to quickly reduce the number of candidates in the pipelines, such as the drug-likeness concept. Drug-likeness describes how a molecule is 'drugable' *i.e.* relate molecular properties such as molecular weight or simple physicochemical properties to its ability to become a drug, with respect to activity, bioavailability and safety, among others. For example, Lipinski's famous 'Rule of Five' aims at predicting the oral absorption of substances [11]. Unfortunately, by the strict application of such rules there is a risk of missing valuable discoveries – indeed, several approved drugs are known to

A **new chemical entity** (NCE) is an active drug that has not been previously approved for marketing in any form.

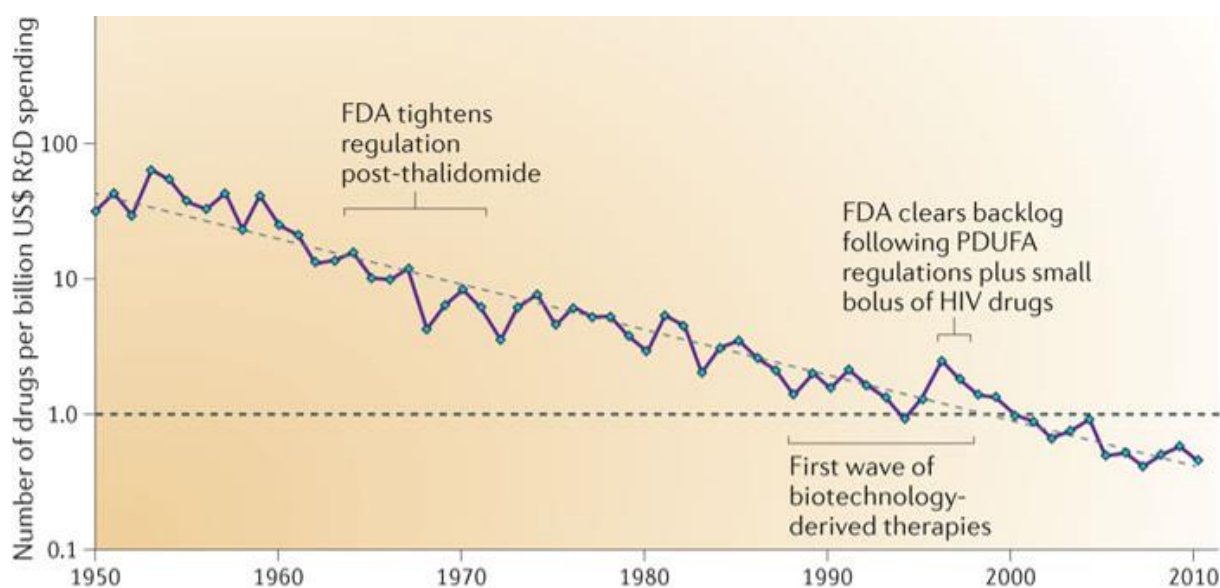


Figure I.2. The number of new drugs approved by the American FDA per billion US\$ (inflation-adjusted) spent on R&D has halved roughly every 9 years. Adapted with permission from Macmillan Publishers Ltd: Nature Reviews Drug Discovery [10], copyright 2012.

break the rules. Moreover, being drug-like does not mean that the candidate is more likely to become an approved drug [12].

To conclude, the decline of productivity of the pharmaceutical R&D seems to be multifactorial and probably won't end in the near future. In this context, NP research can be again considered as a valuable and efficient strategy [5, 7]. This regain of interest may be explained partly as NPs bring structural novelty and bioactive scaffolds (see below) that are needed by pharmaceutical research to provide diversified lead compounds, and partly because new methods in analytical chemistry and statistics to NP research (see subchapters 3 and 4) allow a remarkable increase in efficiency and throughput, that was considered before as too low.

1.3. The role of natural products in modern drug discovery

As mentioned in the previous subchapters, as new drug discovery techniques emerged, the pharmaceutical R&D has lost much interest in NP research during the last decades. NPs have, however, played a key role as lead compounds in the recent years [1, 5]. Indeed, NPs were involved in approximately 50% of all small molecules officially approved in the years 2001–2010 [4, 6], as illustrated in Figure I.3. These NPs may be approved without any structural modifications, with semisynthetic modifications, or may be used as lead compound, providing natural pharmacophores for synthetic drugs.

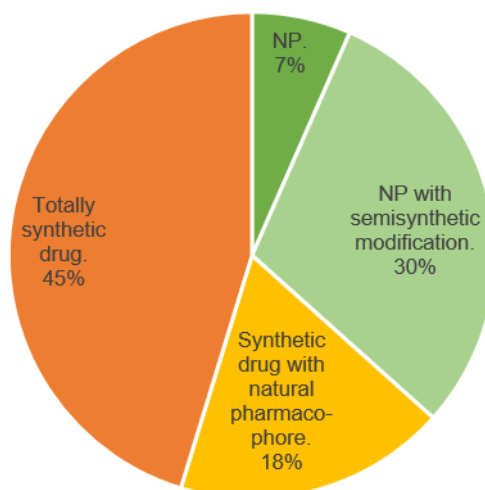


Figure I.3. Sources of approved drugs in the period 2001 to 2010. Only the small molecules are displayed. NPs were involved in approximately 50% of all small molecules officially approved in this period. Adapted from data of [4].

The importance of NPs in modern drug discovery is illustrated below by three examples from various natural sources, detailed below: (1) romidepsin which is isolated from a bacterium, (2) ombrabulin which was found in a tree bark, and (3) statins which originated from fungal strains. Romidepsin was approved in 2009, while ombrabulin is still in phase III clinical trials, showing that NPs play an important role in drug discovery today. The statins, in particular, show that NPs may lead to the development of blockbusters.

(1) Romidepsin (Figure I.4) is a depsipeptidic NP isolated in 1994 by a Japanese researcher's team from the Gram-negative *Chromobacterium violaceum* bacteria for its antibacterial activity [13]. However its cytotoxicity against several human cancer cell lines raised interest and finally

this potent inhibitor of histone deacetylase (HDAC) enzyme was approved without any structural modification by the American FDA in 2009 with the indication for cutaneous T-cell lymphoma and commercialised by the Celgene company (brandname Istodax) [14]. Clinical trials are currently being conducted for several additional indications.

(2) Ombrabulin (AVE8062, Sanofi-Aventis, Figure I.5) is a novel vascular-disrupting agent that might be a future anticancer drug [15]. This agent was derived from combretastatin A-4 (Figure I.5), a stilbenoid isolated from the South African medicinal tree *Combretum caffrum* Kuntze (Combretaceae). It is being investigated in a phase III trial in patients with advanced-stage soft tissue sarcoma, and in phase I and phase II trials for other indications [16].

A **pharmacophore** is the ensemble of steric and electronic properties that provides optimal interactions with a specific biological target to trigger or block its biological response.

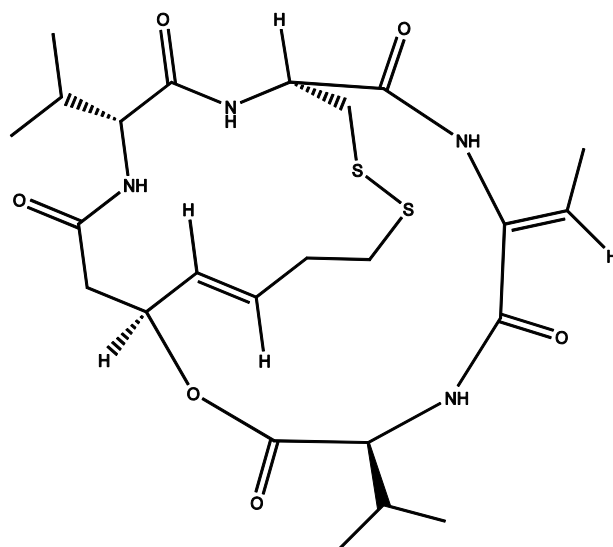


Figure I.4. Structure of romidepsin. This natural product has been approved without any structural modification by the FDA in 2009 for the treatment of cutaneous T-cell lymphoma.

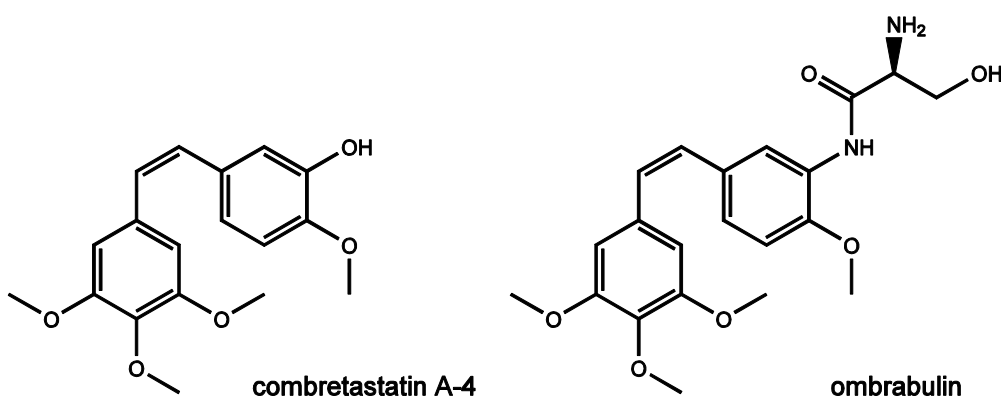


Figure I.5. Combretastatin A-4, a stilbenoid isolated from *Combretum caffrum* Kuntze, a South African tree, and its derived analogue ombrabulin, a vascular-disrupting agent in phase III trial for advanced stage soft tissue sarcoma.

(3) The statins are extensively prescribed hypocholesterolemic agents that inhibit the HMG-CoA reductase enzyme [17]. It is generally agreed that this class of agents brings long-term cardiotoxic benefits and represents a very large commercial value. The first two statins,

mevastatin and lovastatin (Figure I.6), were isolated from the fungi *Penicillium brevicompactum* and *Aspergillus terreus*, respectively [18]. Lovastatin was the first statin approved by the FDA in 1987 and was commercialised by Merck (Mevacor in the USA)

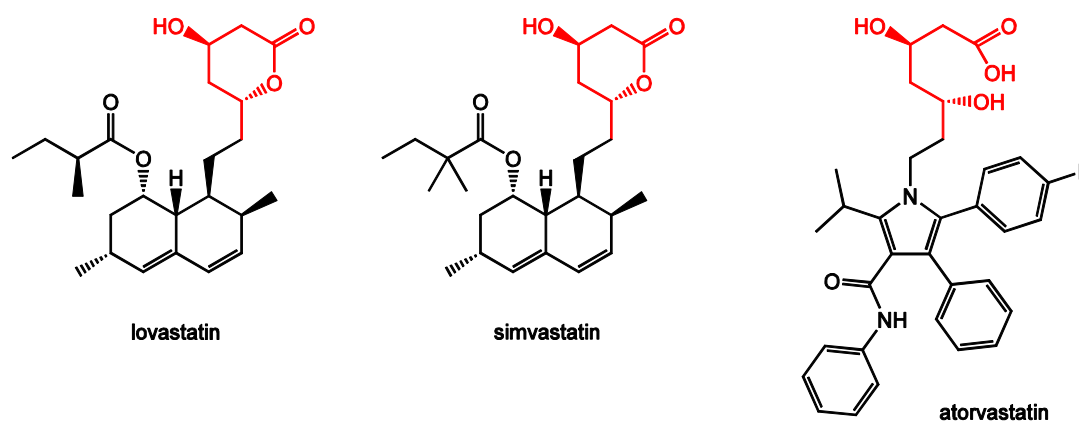


Figure I.6. Structures of lovastatin, simvastatin and atorvastatin, inhibitors of the HMG-CoA reductase. The statin pharmacophore is highlighted in red.

[14]. Many semi-synthetic derivatives were developed based on the pharmacophore of these NPs. Among them, simvastatin incorporates only one additional methyl group (Figure I.6). This derivative was approved in 1990 in Switzerland and was commercialised by Merck with the brandname Zocor [19]. Atorvastatin [17], which has a pharmacophore very similar to lovastatin's (Figure I.6) [18], was approved in 1997 in Switzerland and commercialised by Pfizer with the brandname Sortis [19] and became the world's bestselling drug of all time, with a 130 billion US\$ profit in the 1997-2010 period [20].

Based on these three recent successful examples, and on the high number of recently approved drugs that are NP-related (Figure I.3), it is no wonder that NPs have had and will continue to have a high impact on drug discovery.

Several authors have linked this success to the high *drugability* of NPs, *i.e.* their high ability to become bioactive drugs related to their structures and ADMET properties [21, 22]. There are three main explanations for this high *drugability*. The first reason is the high chemical diversity of NPs compared to synthetic compounds [7]. This is illustrated in Figure I.7, where synthetic compounds and NPs (Figures I.7A and B, respectively) were placed in a chemical space defined by the first two principal components of a PCA analysis built on several molecular parameters [22, 23]. The second reason is that NPs have been naturally optimised by evolutionary pressure to create biologically active molecules, *i.e.* ligands adapted to their targets [7], which is illustrated by Figure I.7, where NPs (Figure I.7B) cover a similar chemical space as active drugs (Figure I.7C). Moreover, this is the case not only for the bioactivity of NPs, but also for their ADME properties since NPs are, on average, better absorbed than

The **ADME** or **ADMET properties** (absorption, distribution, metabolism, elimination and toxicity) are pharmacokinetic properties describing the disposition of a compound in an organism.

synthetic drugs [5]. Finally, the third reason is linked to the use of natural extracts or NPs in traditional medicine, which brings valuable information on their bioactivity, sample preparation, and, last but not least, on toxicity, due to their long-term use. Indeed, almost 75% of the currently used drugs originating from plants were previously used in traditional medicine [1].

Based on all these considerations, one can consider that NPs still represent a valuable source of core scaffolds or drugs for modern drug discovery. This is also demonstrated by the important number of NPs-related drugs that are currently under clinical trials [4].

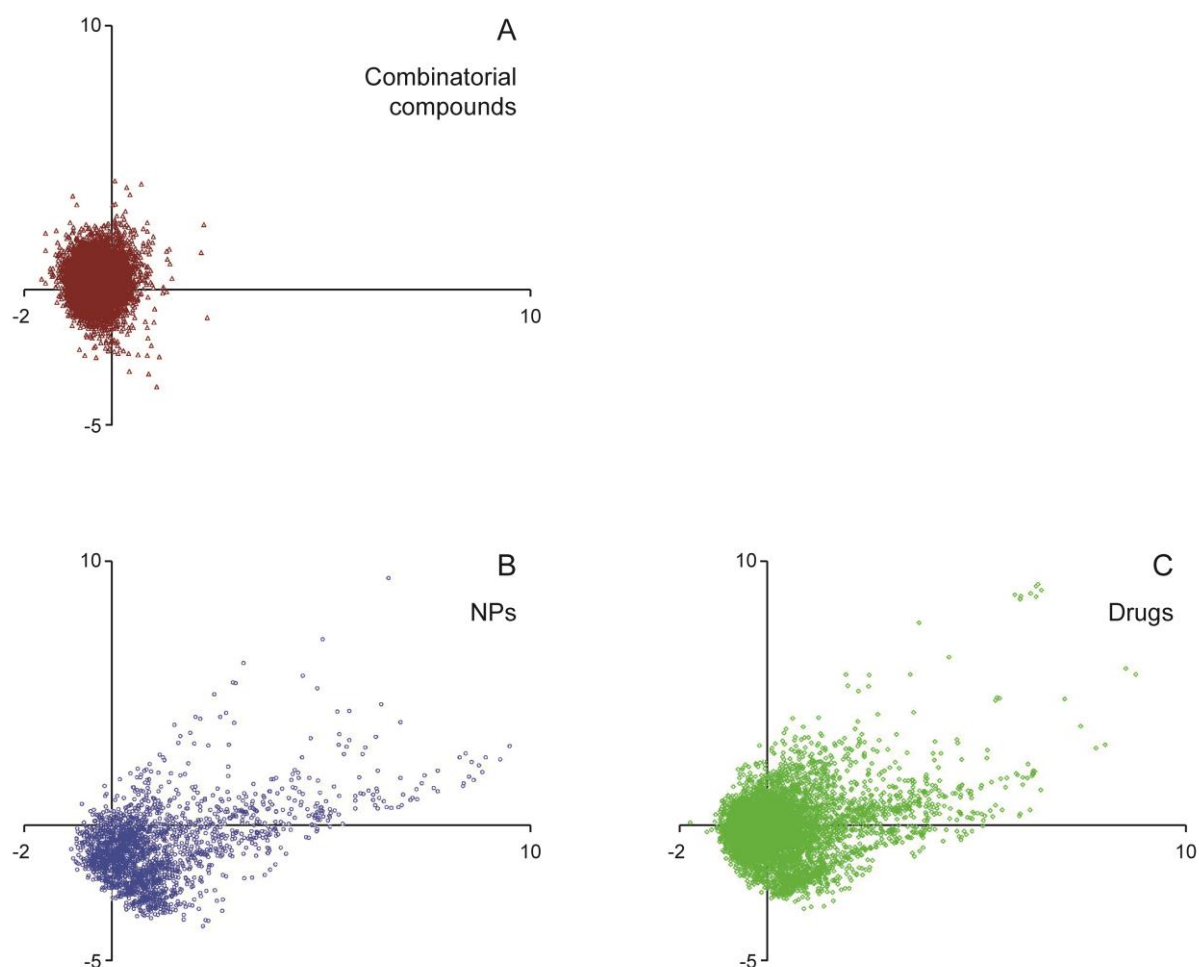


Figure I.7. The plot of the first two principal components (PC) obtained from (A) a random selection of combinatorial compounds, (B) NPs, and (C) bioactive drugs. Principal components (PC) are derived from simple molecular properties such as the number of chiral and rotatable bonds, of C-N, C-O, C-S bonds, etc. The first two PC explain about 54% of the variance. The figure shows that NPs cover a larger area of the chemical space defined by the two PC than compounds originating from combinatorial chemistry. Moreover, NPs and drugs have the same coverage of this space. Adapted with permission from [22]. Copyright 2003 American Chemical Society.

1.4. Sources of natural products

As mentioned above, NPs are interesting scaffolds for bioactivity screening because of their high chemical diversity. This diversity is linked to the high number of NPs that may be found in a given organism and to the almost infinite number of living organisms. This will be presented in this subchapter.

NPs are, by definition, extracted from natural organisms. The whole array of small metabolites found in a given organism is known as *the metabolome*, and includes primary and secondary metabolites. The primary metabolites are directly involved in essential processes such as normal growth, development and reproduction of the organism and include molecules such as carbohydrates, vitamins, amino acids. The secondary metabolites, however, are not essential for the organism but possess important ecological functions such as defence and provide most of the NPs used as drugs [24], and comprises various molecules such as alkaloids and terpenoids. The metabolome of a given organism may be extremely complex, because of the huge number of constituents, their physicochemical diversity, and their extreme variation in concentration [25]. Because of this, the extraction, isolation and detection of all metabolites of a given metabolome in one single analysis is extremely challenging [24]. This will be discussed in details in the next chapters.

The number of constituents of the metabolome may be extremely large and has been estimated to be a few thousand [26]. For example, more

than 2'500 metabolites have been identified in the tobacco plant, *Nicotiana tabacum* L. [27]. As another example, the number of features detected in the venom of the marine snail *Conus consors* is higher than 1600 [28] (this number was obtained after removing adducts and MS data cleaning to provide only the real features, which may differ from the number of metabolites). Besides the difference between LC-MS features and metabolites, it has to be noted that the number of detected metabolites may differ from the total number of constituents of the organism. This difference may be explained by the lack of universality and sensitivity of the analytical techniques used. Moreover, some key metabolites are not constitutively present in a given organism, or only in low concentrations, and are induced by a specific stress or an interaction with another organism. This is typically the case of defence compounds such as phytoalexins in plants [29] or stress related signaling molecules such as jasmonates [30, 31] or β -lactam antibiotics produced by *Penicillium* fungi when their growth is inhibited.

For decades, plants were traditionally the major source of NPs for drug discovery because of the links that could be easily established with traditional medicine. Today, the interest in alternative sources such as algae and other marine organisms, as well as microorganisms (fungi, bacteria and viruses), is growing, and the development of new technologies (e.g. genetic approaches) with higher sensitivity offers new perspectives in this direction, although there are probably millions of microbes in the environment that are still untouchable for science. Table I.1 lists some sources of NPs.

An LC-MS **feature** is the content of a cell in a matrix for multivariate data analysis after data cleaning (e.g. deisotoping). More than one LC-MS feature may correspond to one metabolite.

Table I.1. The numbers of known and possible living species. Adapted from [32] with permission of Nature Publishing Group.

Group	Number of described species	Number of estimated / supposed species
Bacteria	~ 6'000	500'000 / 1'500'000
Actinomycetes	~ 4'000	35'000 / 50 - 80'000
Fungi	~ 8'000	1'500'000 / several millions
Viruses	~ 5'000	~ 50'000
Algae	~ 2'500	~ 50'000/ ~ 40'000
Higher plants	~ 35'000	500 – 600'000 / 1'500'000
Insects	> 1'000'000	several millions / 8 -10'000'000
Marine invertebrates	20 - 25'000	150 – 200'000 / several millions
Vertebrates	~ 50'000	50 - 55'000

A highly interesting review based on statistical data [32] raised the question of the number of NPs yet unknown and yet discoverable, which is related to both the total number of known/described species and the number of known/described metabolites for a given species. Table I.1 displays the number of described, estimated and supposed existing species for several group of natural organisms. The author considers that the higher the number of unknown species, the higher the probability of finding new metabolites, although there is always some redundancy in the metabolome of different species. According to this table, microorganisms are largely unexplored, and the biggest potential is in fungal species. Moreover, methods based on DNA analysis have proven that the number of microorganisms in the soil is much higher than formerly thought [32].

The total number of NPs recognised until now, including both bioactive and inactive compounds, is around one million [32]. Considering the number of species yet to be discovered, and the number of species not yet investigated, one cannot see any limit to the number of NPs that may be still discovered in the future.

To conclude, NPs represent a valuable source of lead compounds in drug discovery, thanks to their high physicochemical diversity, their natural *drugability*, their high probability to be bioactive, and the huge number of NPs yet to be discovered. NP research, however, is often considered a slow and costly technique because of the difficulties to get the pure compound using either isolation or synthesis, and because of the problems related to intellectual property, and collection and supply concerns.

2. Classical chemical investigation of samples of natural origin

The traditional – and oldest – drug discovery strategy consists in bioactivity guided isolation that aims at providing pure compound in milligram amounts for structural elucidation and biological testing. This procedure may be divided in three steps - extraction, isolation and structural elucidation. A biological monitoring is performed at all steps to efficiently target the bioactive compound(s) [33]. The whole procedure is shortly presented below and illustrated by Figure I.8.

2.1. General procedure for the chemical investigation of natural extracts

Extraction is the first step of any investigation of the metabolome composition of a natural organism. The technique has to be adapted to the amount of the biological matrix and to its physical properties. There is no comprehensive and total extraction procedure, and the choice of solvents and techniques will determine the nature of the extracted metabolites. The more specific the extraction is, the less unwanted compounds will be present in the extract, with possibly a higher concentration of the desired metabolites. For example, two different extraction procedures will be required to obtain a crude extract containing as many metabolites as possible or to specifically extract the alkaloids from a plant. Solid-phase extraction (SPE), liquid-liquid extraction (LLE), sonication, maceration, percolation, supercritical fluid extraction (SFE), microwave-assisted extraction (MAE) and pressurised solvent

extraction (PSE) are the mainly used extraction techniques. A recent book comprehensively reviewed the subject [34].

Separation of the extract is the second step and aims at isolating the pure targeted compound(s). The bioactive extract is separated in tens or hundreds of fractions that are systematically biologically tested to highlight the fraction(s) containing the compound(s) responsible for the desired activity. This procedure is repeated on the active fraction in a dichotomic way by repeated fractionation steps to finally isolate the pure and active compound. Such a method has been used for a collaborative work during this thesis to study the oestrogenic components of *Salvia officinalis* (see Annexe IV). Separation is performed by chromatographic methods adapted to the nature of the metabolite and their amounts. Liquid chromatography at the preparative scale or in open columns is the most frequently used separation technique and reversed phase, normal phase and steric exclusion are among the most used separation mechanisms [34]. A combination of various chromatographic approaches ideally leads to the isolation of a pure NP that can then be fully characterised.

The structure of the isolated pure compound is elucidated using spectroscopic methods, and its activity and toxicity may be tested. Structural elucidation is usually performed by 2D-NMR, MS, UV and IR, and sometimes MSⁿ, and circular dichroism [35, 36].

Such fractionation procedures have been - and are still - performed successfully for the isolation of NPs for decades and enabled the isolation, identification and biological testing of many NPs. However, the classical investigation of a natural extract possesses two main drawbacks that make it unsuitable for lead discovery [6]. Firstly, it is a complex, slow and costly procedure, and secondly, the huge efforts put in for isolating an NP may provide a compound that has already been studied.

Some solutions to overcome these problems were recently provided thanks to the new developments in analytical chemistry. Firstly, the whole isolation procedure may be accelerated by its miniaturisation to the microfractionation setup described below in subchapter 3.1. Secondly, the application of dereplication procedures based on online techniques such as LC-MS (Figure I.8) represent an efficient way to detect and identify already known compounds and avoid their unnecessary isolation, as described in subchapter 4.

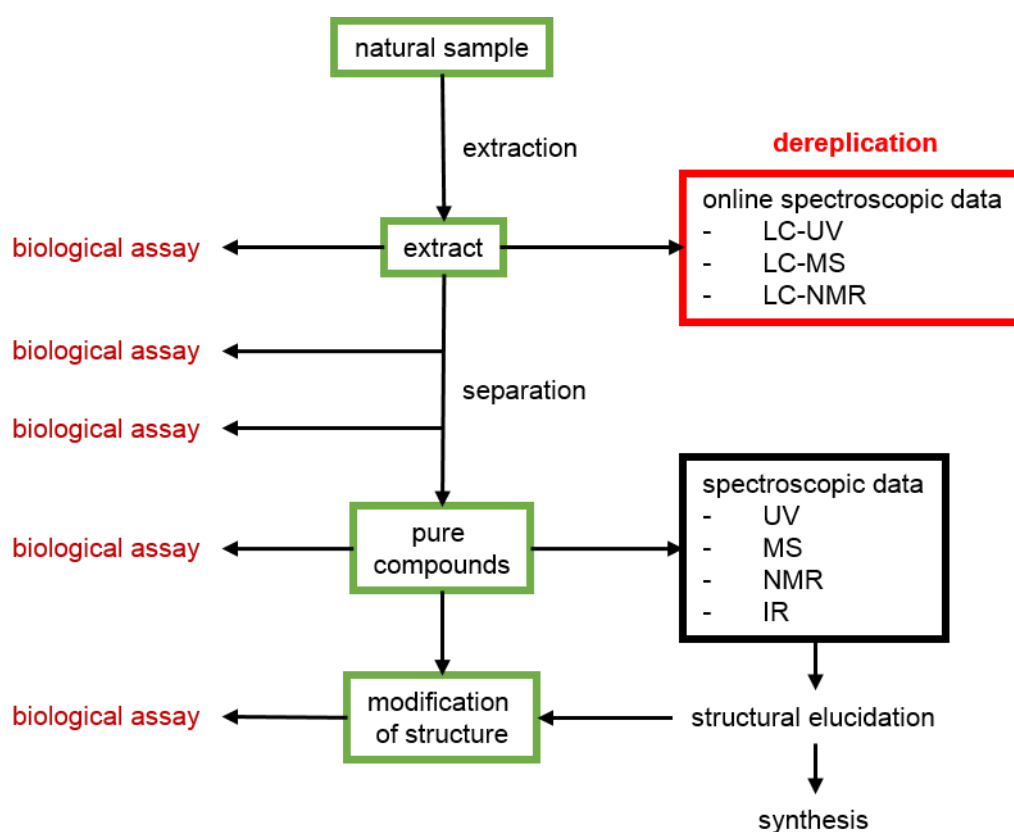


Figure I.8. General procedure for the isolation of a natural product.

Dereplication is the process of identifying known metabolites in a sample from online data, to avoid focusing on compounds that were already studied.

3. Modern techniques for natural product discovery

The classical investigation of natural extracts is a slow and costly technique that pharmaceutical companies almost stopped using because new promising high throughput techniques emerged [6]. Since NPs remain an invaluable source of interesting bioactive molecules, new approaches for NP discovery were proposed.

Firstly, the development and application of online detection techniques such as LC-PDA and LC-MS permitted the efficient application of dereplication to bioguided isolation procedures. Dereplication is the process of highlighting known compounds in the studied extract to avoid their unnecessary isolation. These hyphenated techniques were then used not only for dereplication, but also for the *de novo* online metabolite identification during crude extract profiling. This so-called 'high resolution metabolite profiling' will be presented in detail in subchapter 4.

Secondly, the whole bioactivity-guided isolation process can be miniaturised using HPLC at the semi-preparative scale, increasing its throughput and decreasing the amount of sample used. This method is described in subchapter 3.1 below.

Finally, metabolomics approaches play an important role in NP research. Indeed, they aim at identifying metabolites in an untargeted way and also highlight dynamic changes in biological systems [37]. Such strategies dramatically improved the efficiency of biomarker discovery. Several metabolomics studies were successfully conducted in our laboratory and one of them is

described in subchapter 3.2 below. Large scale metabolomics studies however highlighted the lack of adapted online identification techniques.

3.1. Miniaturised and integrated setup for bioactivity-guided isolation

The whole procedure of bioactivity-guided isolation described in subchapter 2 may be downscaled from the milligram to the microgram scale using semi-preparative HPLC and MS monitoring to provide high throughput and miniaturised fractions [38]. Such microfractions contain almost pure compounds thanks to the high chromatographic resolution of the HPLC separation [39]. Because of the low amount of sample injected, however, detection and bioactivity testing represent big challenges. PDA and MS detectors are well adapted to enable the detection of analytes at such concentrations, but CapNMR is required instead of classical NMR for structural elucidation because of the lower amount of compound in the microfraction compared to classical isolation [39, 40].

Moreover, the use of semi-preparative HPLC columns that possess the same phase chemistry as analytical columns allows the geometrical transfer of a previous method at the analytical scale to the microisolation at the semi-preparative scale [41]. This process that is presented in details in Chapter II allows the transfer of precious information from metabolite profiling to microisolation.

Biological testing of microfractions is a difficult task because of the low concentration of the analytes. Two strategies were reported for HPLC biological profiling, based on online and at-line testing [42].

Online bioassays are usually based on ligand-receptor interactions or enzymatic reactions. The bioactivity is monitored by a change of fluorescence of the substrate or ligand in presence or absence of the inhibitor. The bioassay is performed post-column, where a split directs part of the eluent to the fluorescence detector after the addition of the protein and the ligand or substrate(s) by make-up pumps. Such a setup was used for the detection of acetyl cholinesterase inhibitors [43] and angiotensin-converting enzyme inhibitors [44], among others.

The at-line monitoring of the microfractions collected is performed in microplates and is then related to the corresponding peak(s) of the chromatogram. Biological testing is usually based on *in vitro* chemical or enzymatic reactions. Recently, however, papers have reported the use of cell-based assays, for example in a calcium uptake assay for highlighting Ca^{2+} uptake inhibitors [45]. The use of cell-based assays is very interesting since it provides a better predictability of the *in vivo* activity of molecules

than classical enzymatic reactions. Another study showed that the amount of compounds obtained in microfractions is also adapted to antibacterial testing [40]. Finally, some recently developed *in vivo* miniaturised assays were efficiently applied to microfractions, with the clear advantage that they take into account the whole organism instead of the target or one single cell only. For example, the *in vivo* biological high-throughput assays based on the embryos and larvae of zebrafish (*Danio rerio*) provide a wide number of miniaturised bioassays compatible with the low concentrations of pure compounds obtained by microfractionation [46, 47]. Such a miniaturised setup coupling the microfractionation with the zebrafish bioassay has been successfully used on a Fabaceae species for the search of NPs inhibiting angiogenesis in zebrafish and proved that traditional bioactivity-guided isolation may be downscaled to semi-preparative scale [48]. To conclude, the miniaturisation of traditional bioactivity-guided isolation to HPLC-based microfractionation provides an efficient solution to increase the throughput of the method, providing almost pure microfractions that may be further tested biologically using online or at-line methods. Moreover, the hyphenation of the HPLC system to PDA or MS detectors provides useful online information. Still, there is a need for an efficient LC-MS dereplication method to avoid the isolation of known compounds.

The geometrical transfer of a method aims at transferring an LC separation performed in given chromatographic conditions to new ones, with identical thermodynamic parameters (mobile and stationary phases, temperature) but with the possibility to change all kinetic parameters such as column geometry (length, internal diameter, particle size), flow rate, or gradient, to keep the resolution and selectivity constant. For more details, see Chapter II and [41, 48].

3.2. Metabolomic strategy and peak annotation

Metabolomic approaches were recently applied to NP discovery [24] and, depending on the aim of the study, may highlight an active biomarker. The main analytical techniques used for metabolite profiling are NMR [50] and mass spectrometry [51]. In a second step, biomarkers are highlighted by data mining performed by multivariate analysis to link bioactivity and features [52]. Metabolomics aims at identifying and quantifying all metabolites present in a biological system, usually using short analysis time [53] to get a 'fingerprint' of the metabolome, while differential metabolomics aims at quantifying the response of an organism to a stimulus or an interaction [54] and usually consists of a statistical comparison of the metabolomes of two different populations.

The number of MS-based metabolomic applications exploded with the use of high throughput LC techniques in routine analysis and with the miniaturisation of instruments and methods (e.g. 96/384-well plates or highly-sensitive detection techniques) [51, 55, 56].

There are typically two main steps in such large scale metabolomic studies, *i.e.* data collection and data treatment. Data are usually collected using an ultra-rapid sample preparation procedure (e.g. using a ball mill extractor providing efficient extraction in ca. 2 minutes) prior to a short gradient 'fingerprint' analysis by LC-MS [54]. Induced metabolites (called biomarkers) are then highlighted by multivariate analysis such as PCA and HCA [52]. Metabolomics has been successfully used in our laboratory in several projects related to the dynamic induction of metabolites as a consequence of different types of stress applied to plants or microorganisms. For example, this approach was applied to highlight induced metabolites in the confrontational zone of fungal co-cultures on Petri dishes. Given that fungal growth is usually inhibited in this zone, metabolites responsible for this inhibition may be present [57, 58]. Metabolomics is presented more in detail in Chapter II.

To conclude, both bioactivity-guided microfractionation and metabolomics are efficient strategies used in NP discovery, aiming at highlighting bioactive compounds or

Multivariate analysis comprises of statistical tools designed to deal with more than one variable at the same time. They are divided in two groups, unsupervised (HCA, PCA) and supervised (PLS).

Hierarchical cluster analysis (HCA) groups objects based on their similarity (in terms of characteristics found in the data).

Principal component analysis (PCA) is an exploratory method aiming at summarising a dataset of high dimensionality with a small number of factors (the principal components, PC) by using orthogonal coordinate systems.

Partial least square regression (PLS) is a supervised regression method that aims at maximising the covariance between linear combinations of variables and observations.

biomarkers in samples from natural sources. For all these approaches there is a need for efficient tools for online dereplication or for *de novo* structural identification of metabolites. In this respect, hyphenated techniques such as LC-MS and LC-NMR may be considered. Online LC-NMR was introduced twenty years ago and evolved towards at-line techniques (LC-SPE-NMR) allowing the pre-concentration of the LC peaks and their subsequent analysis using microflow NMR probes [59]. However LC-MS has been considered for 2-3 decades now as the gold standard for the characterisation of natural samples [60], thanks to its high separation power,

sensitivity and quasi-universal detection possibilities, and is already routinely and successfully used as an identification tool in protein and peptide analysis by LC-QTOF-MS platforms [61]. Although its use for the identification of small molecular weight NPs has only emerged more recently, probably because of the lack of efficient and adapted tools, it is nowadays also recognised as a valuable tool for the *de novo* structure identification of NPs. Based on these considerations, the LC-MS profiling of natural extracts is introduced in the next subchapter.

4. LC-MS high resolution metabolite profiling

Natural extracts are valuable but complex samples because of their high number of compounds characterised by a high variability in term of physicochemical properties. The hyphenation of liquid chromatography with mass spectrometry (LC-MS) is thus the method of choice for the analysis of these complex samples, and has been routinely used in pharmaceutical research for a few decades now (e.g. for quantification purposes). LC-MS has recently benefited from significant advances in both chromatographic and spectrometric dimensions, and has brought new perspectives in NP research for the online identification of small molecular weight metabolites.

4.1. High resolution separation

Like other chromatographic techniques, LC has been used for decades in NP research, since it

allows the separation of the metabolites contained in a natural samples [33, 62]. The recent commercialisation by Waters in 2004 of the first UHPLC system operating at very high pressures and using sub-2 μm packing columns have allowed a remarkable decrease in analysis time and increase in peak capacity compared to conventional HPLC as illustrated in Figure 1.9 (see Chapter II for more details) [63]. For example, a classical HPLC separation of a *Ginkgo biloba* extract (Figure 1.9A) may be transferred with equivalent analysis time to a high resolution UHPLC separation (Figure 1.9B) where the efficiency is multiplied by a factor of 3, and provides a very high peak capacity, which is of particular interest for the separation of complex matrices. On the other hand, the same HPLC separation (Figure 1.9A) may be also transferred to a UHPLC separation with the same efficiency in a shorter analysis time (gradient time divided by a factor of 9), and provides satisfactory

Fingerprint analysis is a fast and global analysis of a sample that does not aim at identifying compounds, but at providing a « fingerprint » of the sample. It is often subjected to multivariate analysis in metabolomic studies.

High resolution metabolite profiling is a high resolution analysis aiming at separating all compounds in a sample to provide a complete and quantitative picture of all the metabolites present.

Peak annotation is the extraction of MW and/or molecular formulae from online data.

De novo structural identification is the complete identification of a metabolite based on online data and on further experiments on the pure isolated compound.

separations in very short times which is adapted to fingerprint analysis. Moreover, UHPLC separations are clearly advantageous in terms of sensitivity and reproducibility, and of solvent and sample consumption, and allows easy geometrical method transfer thanks to the

availability of the same column chemistries in HPLC, UHPLC and semi-preparative scales [41, 49]. For all these reasons, UHPLC is gradually replacing HPLC separations in NP research. Chapters II and III present this technology in detail.

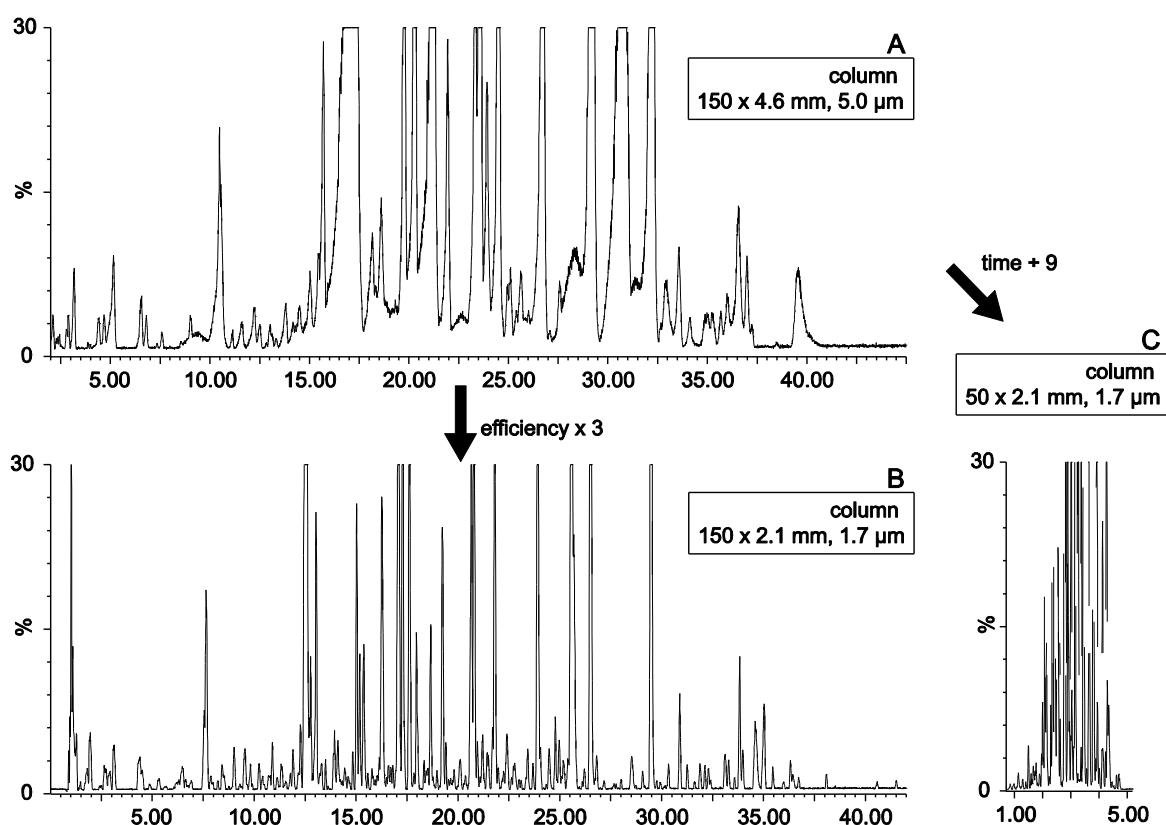


Figure I.9. Chromatograms of three different analyses of a *Ginkgo biloba* extract using a UHPLC-TOF-MS system in negative ionisation mode. (A) Conventional HPLC separation in 60 minutes using a 5-40% ACN gradient on a 150 x 4.6 mm, 5.0 µm HPLC C18 column. (B) High resolution profiling in 60 minutes using a 5-40% ACN gradient on a 150 x 2.1 mm, 1.7 µm UHPLC C18 column. (C) Fingerprint analysis in 6.76 minutes using a 5-40% ACN gradient on a 50 x 2.1 mm, 1.7 µm UHPLC C18 column. Compared to the conventional HPLC separation, the high resolution UHPLC profiling provided an efficiency multiplied by a factor 3 for the same analysis time, while the fingerprint analysis time was divided by a factor 9 for the same efficiency (see Chapter II for more details). UHPLC technology is thus able to provide both high resolution separation needed for the high resolution profiling and high throughput fingerprint required for metabolomics applications.

4.2. High resolution MS detection

Mass spectrometry has been used for decades in NP research, mainly since the 80s with the development of hyphenation of LC with MS [56, 64]. Indeed, the use of an MS detector hyphenated with an HPLC system provides a high sensitivity and selectivity in the analysis of NPs in complex biological matrices as well as important online structural information, such as the molecular mass and diagnostic fragments, which are crucial for dereplication and rapid online characterisation [36, 55, 60, 65]. MS analyser may be divided in two categories, *i.e.* low and high resolution instruments.

On the one hand, low resolution instruments such as single quadrupole, triple quadrupole and ion trap instruments are among the most used in NP research. Single quadrupole are the simplest and least expensive MS instruments, and are able to provide the nominal mass of ions and to specifically monitor a selected mass. Triple quadrupole instruments allow the specific selection of a given ion prior to its fragmentation when operated in multiple reaction monitoring (MRM) mode and are typically used for quantitative analysis, while ion trap MS are able to produce multiple stage fragmentation (MS^n) to get structural information based on the study of the fragments.

On the other hand, high resolution instruments, able to provide the molecular mass with a high accuracy (1-5 ppm) [66] are more and more frequently used in NP research in the last decade. Among them, time-of-flight (TOF) instruments are the most used in hyphenation with LC [55]. They provide sensitive detection and high MS resolution on the entire m/z range with a high acquisition rate compatible with both high throughput and high resolution separations [65], and are mainly used to get the molecular formula

of unknown compounds in fingerprinting and dereplication analysis. The recently introduced hybrid QTOF instruments that are a combination of both triple quadrupole and TOF technologies are probably today the most versatile analysis technique in NP research.

In summary, modern LC-MS systems based on UHPLC and (Q)TOF-MS provide a high resolution in both chromatographic and spectroscopic dimensions, which is defined as high resolution profiling. Such platform provide a baseline separation of the analytes and high quality spectrometric information for both dereplication and *de novo* structural identification of the metabolites present in a natural sample [67-69].

4.3. High resolution profiling for dereplication and metabolite identification

As mentioned above, high resolution LC-MS profiling based on UHPLC and (Q)TOF-MS provides a baseline separation of analytes. This is important for the analysis of complex mixtures such as natural extracts to get high quality MS data, *i.e.* clean spectra displaying the spectroscopic information of ideally a single metabolite, without other interfering peaks. Moreover, it enables the detection of minor metabolites and of isomers that cannot be separated by the MS dimension and reduces the ionisation suppression problems that often occur with electrospray ionisation [65]. Besides the high resolution obtained by modern LC systems, today's HR-MS instruments provide a high mass accuracy, usually below 5 ppm. Such mass accuracy is sufficient to provide few or even one single molecular formula for a given ion, which is the basis of the dereplication strategy [70]. Figure I.10 illustrates the UHPLC-TOF-MS high resolution profiling of an extract of *Viola tricolor*, an herbal drug traditionally used for its anti-inflammatory properties.

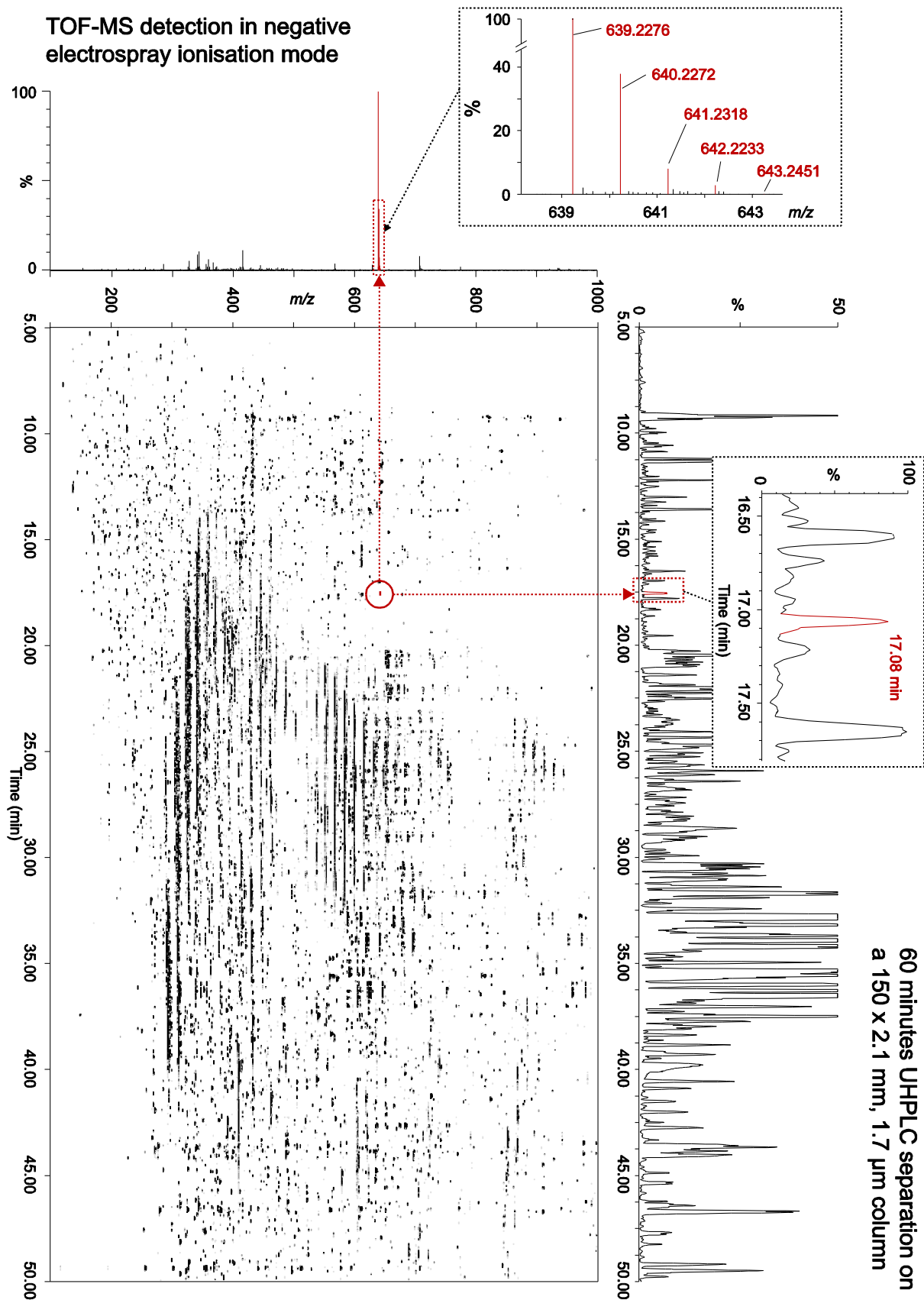


Figure I.10. Two-dimensional map of a *Viola tricolor* UHPLC-TOF-MS high resolution metabolite profiling. No blank subtraction was performed (see Chapter V for more details on data processing). For more details on high resolution profiling, see comments in subchapter 4.3 and Chapters II and III.

The extract was analysed in 60 min on a BEH C18 column (150 x 2.1 mm, 1.7 μm), in negative ionisation mode. As shown, a good separation of most of the metabolites was obtained in the LC dimension, and minor compounds were also detected thanks to the sensitivity of the TOF-MS system. The UHPLC-TOF-MS system provided high resolution two-dimensional information, *i.e.* exact molecular weights (below 5 ppm) and retention time information for all of the compounds detected, showing that such a platform is adapted for dereplication studies. As a result, the number of applications of metabolite profiling in NP research significantly increased over the past ten years with the emergence of this new LC-MS instrumentation. The number of publications for the last 13 years on the topic of LC-MS metabolite profiling of non-human samples is illustrated in Figure I.11.

However, high resolution in both LC and MS dimensions is not sufficient to provide efficient metabolite identification. Indeed, a recent review on plant metabolomic and metabolite profiling

studies involving LC-MS, GC-MS, CE-MS and $^1\text{H-NMR}$ showed that only approximately 10% of the detected peaks in LC-MS studies are identified [71]. This ratio is however much higher with GC-MS analysis, where the reproducible EI-MS fragmentation allows specific spectra database search for fast and automated identification. Similar approaches are however scarcely used for LC-ESI-MS/MS analyses because of the instrument-dependent fragmentation pattern of the electrospray ionisation that makes the database instrument-specific. Only large companies may consider the construction of in-house MS/MS databases. Finally, this 10% ratio may be explained by the fact that HR-MS instruments only provide the molecular formula, but no identification. There is thus a need for efficient and generic dereplication methods to fully exploit the high resolution profiling and to improve the number of identified peaks in complex mixtures. Several methods or tools were recently published for online metabolite identification based on these high resolution LC-HR-MS data (see Chapter V for a small review of existing methods).

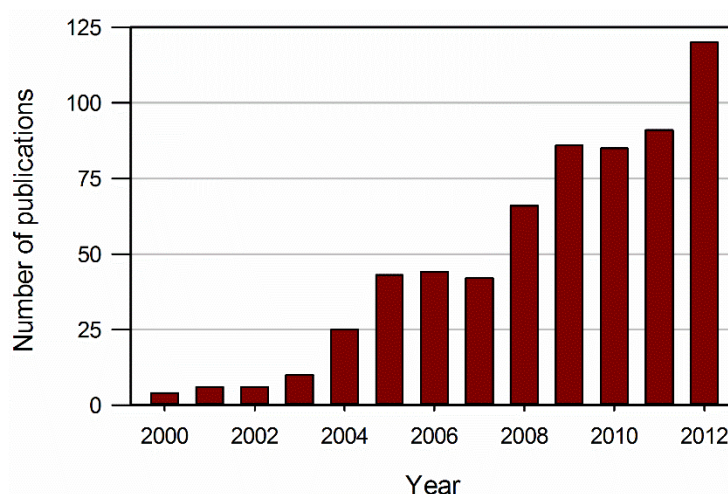


Figure I.11. Number of publications related to the LC-MS metabolite profiling of non-human samples. Retrieved in Web of Knowledge using the following keywords: "metabolite profiling" OR "metabolic profiling", AND "LC-MS" OR "mass spectrometry", NOT "plasma", NOT "urine", NOT "blood", NOT "serum", NOT "cell", NOT "human". Data collected on the 5th of April 2013.

Usually metabolite identification is a multistep procedure based on both the m/z and the isotopic information from HR-MS spectra (see MS spectra in Figure I.10). It involves, among others, the search for the pseudomolecular ion, subsequent calculation of possible molecular formulae, matching of theoretical and experimental isotopic patterns, application of successive filters based on heuristic rules to reduce possibilities and ascertain molecular formula, and database queries [70].

The identification of metabolites present in a natural sample based on HR-MS data is however still not fully automatized and not implemented in routine work, for two main reasons. Firstly, the existing approaches are able to identify a small number of the metabolites present in a natural sample only [71]. Secondly, none of the tools are fully automated, although this is mandatory to process the huge amount of data obtained from large scale metabolomic studies or other LC-MS analyses. There is today an urgent need for efficient dereplication methods of NPs based on LC-MS data.

5. Other uses of natural products

Although the main topic of the present thesis is related to NP research for therapeutic applications, it is worth mentioning that NPs are widely exploited since centuries in many other domains such as foods and perfumery.

As described above, products from nature have always been used as source of drugs. Besides pure NPs used in modern medicine, phytomedicines, *i.e.* extracts from plants or from parts of a plant, are still widely used [72], such as *Ginkgo biloba*, *Echinacea* species, and *Hypericum perforatum* extracts. There are several reasons that explain this success. First, phytomedicines represent the main available treatments in some countries. Second, many people think that phytomedicines are totally inoffensive, because of their natural origin. Third, there is a growing number of scientific evidences of their efficacy. Since they are plant extracts, phytomedicines are complex mixtures of metabolites. Their standardisation and quality control is today required and present a great challenge because of the complexity of the sample. Moreover the active principle is often unknown, and/or the activity is due to the synergic effect of more than one metabolite. Because of this, LC-MS based

fingerprint is the most efficient method to get a comprehensive picture of the composition of the preparation, and is often linked with a metabolomic approach [73]. There is thus a need for efficient LC-MS methods in this domain.

Natural products are not only used for medical reasons, but they are also important sources of food, perfumes, spices and materials for humans. The perfume industry, for example, strongly depends on nature for the creation of new flavours and fragrances, even if industrial production is often synthetic. Most of the techniques used for research and production of perfumes are similar to those used for the NP research for drugs, including extraction procedures and analytical tools such as GC-MS and LC-MS profiling, or multivariate analysis [74]. Nutraceuticals are other examples of products from natural origin that possess a huge economical and scientific potential. The search for active ingredients in food or in medicinal plants is very similar because of the complexity of the matrices studied, thus both nutraceuticals and NPs research require similar methods for chemical profiling [75].

To conclude, many domains related to NPs represent a huge economic value and a continuous scientific interest. Because the

production, quality control and research techniques related to these areas are similar,



they may be transferred from one to another and possess the same needs.

6. Aim of thesis

Based on these considerations, this thesis focuses on the LC-MS metabolite profiling in the frame of the analyses of complex natural samples and may be divided in two main parts.

The first part is aimed at optimising the chromatographic conditions of HPLC metabolite profiling. Chapter II introduces the UHPLC technology and presents its theoretical aspects and practical applications in NP analysis. Chapter III provides solutions to increase the LC resolution of metabolite profiling of complex natural samples containing small MW metabolites and peptides. In Chapter IV, the potential of ion mobility spectrometry as an additional separation dimension for complex mixtures

analysis is investigated. These high resolution separations approaches provide well-resolved peaks and high quality HR-MS spectra for further peak annotation.

The second part explores dereplication and online metabolite identification. Chapter V details a comprehensive LC-MS methodology for the dereplication of NPs based on the high resolution profiling and using heuristic filters and database search. This procedure is used in Chapter VI for a chemotaxonomic study of Brazilian *Lippia* species. Chapter VII presents a method for the LC retention prediction of NPs that may be used as an additional filter for their dereplication.

7. References

- [1] D.J. Newman, G.M. Cragg, K.M. Snader. The influence of natural products upon drug discovery. *Natural Product Reports*, **2000**. 17: 215-234.
- [2] G.R. Hamilton. In the arms of Morpheus: the development of morphine for postoperative pain relief. *Canadian Journal of Anesthesia*, **2000**. 47: 367-374.
- [3] J.W.-H. Li, J.C. Vederas. Drug Discovery and Natural Products: End of an Era or an Endless Frontier? *Science*, **2009**. 325: 161-165.
- [4] D.J. Newman, G.M. Cragg. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. *Journal of Natural Products*, **2012**. 75: 311-335.
- [5] A.L. Harvey. Natural products in drug discovery. *Drug Discovery Today*, **2008**. 13: 894-901.
- [6] F.E. Koehn, G.T. Carter. The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery*, **2005**. 4: 206-220.
- [7] J. Larsson, J. Gottfries, S. Muresan, A. Backlund. ChemGPS-NP: Tuned for navigation in biologically relevant chemical space. *Journal of Natural Products*, **2007**. 70: 789-794.
- [8] R. Macarron. Critical review of the role of HTS in drug discovery. *Drug discovery today*, **2006**. 11: 277-279.
- [9] F. Pammolli, L. Magazzini, M. Riccaboni. The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*, **2011**. 10: 428-438.
- [10] J.W. Scannell, A. Blanckley, H. Boldon, B. Warrington. Diagnosing the decline in pharmaceutical R&D efficiency. *Nature reviews Drug discovery*, **2012**. 11: 191-200.
- [11] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, **1997**. 23: 3-25.
- [12] I. Yusof, M.D. Segall. Considering the impact drug-like properties have on the chance of success. *Drug Discovery Today*, **2013**. 18: 659-666.
- [13] H. Ueda, H. Nakajima, Y. Hori, T. Fujita, M. Nishimura, T. Goto, M. Okuhara. FR901228, A novel antitumor bicyclic depsipeptide produced by *Chromobacterium violaceum* no 968.1. Taxonomy, fermentation, isolation, physicochemical and biological properties and antitumor activity. *Journal of Antibiotics*, **1994**. 47: 301-310.
- [14] FDA. FDA Approved Drug Products. [Access April 18, 2013]; Available from: http://www.accessdata.fda.gov/scripts/cder/drugsatfda/index.cfm?fuseaction=Search.Search_Drug_Name.
- [15] A. Delmonte, C. Sessa. AVE8062: a new combretastatin derivative vascular disrupting agent. *Expert Opinion on Investigational Drugs*, **2009**. 18: 1541-1548.
- [16] S.-A. Oncology. Ombrabulin (AVE8062). [Access May 8, 2013]; Available from: <https://www.sanofioncology.com/pipeline/ombrabulin.aspx>.
- [17] A. Lea, D. McTavish. Atorvastatin. *Drugs*, **1997**. 53: 828-847.
- [18] R.M. Wilson, S.J. Danishefsky. Small Molecule Natural Products in the Discovery of Therapeutic Agents: The Synthesis Connection†. *The Journal of Organic Chemistry*, **2006**. 71: 8329-8351.
- [19] Swissmedic. Swiss Authorised Therapeutic Drugs. [Access April 18, 2013]; Available from: <http://www.swissmedic.ch/daten/00080/index.html?lang=en>.
- [20] T. Lancet. Lessons from Lipitor and the broken blockbuster drug model. *The Lancet*, **2011**. 378: 1976.
- [21] C.A. Lipinski. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods*, **2000**. 44: 235-249.

- [22] M. Feher, J.M. Schmidt. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences*, **2002**. 43: 218-227.
- [23] C.M. Dobson. Chemical space and biology. *Nature*, **2004**. 432: 824-828.
- [24] L.W. Sumner, P. Mendes, R.A. Dixon. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **2003**. 62: 817-836.
- [25] R. Verpoorte. Secondary Metabolism, in *Metabolic Engineering of Plant Secondary Metabolism*, R. Verpoorte and A.W. Alfermann, Editors. **2000**, Springer Netherlands. p. 1-29.
- [26] J.L. Wolfender, G. Glauser, J. Bocard, S. Rudaz. MS-based Plant Metabolomic Approaches for Biomarker Discovery. *Natural Product Communications*, **2009**. 4: 1417-1430.
- [27] I. Wahlberg, C.R. Enzell. Tobacco isoprenoids. *Natural Product Reports*, **1987**. 4: 237-276.
- [28] D. Biass, S. Dutertre, A. Gerbault, J.L. Menou, R. Offord, P. Favreau, R. Stocklin. Comparative proteomic study of the venom of the piscivorous cone snail *Conus consors*. *Journal of Proteomics*, **2009**. 72: 210-218.
- [29] R. Bari, J.G. Jones. Role of plant hormones in plant defence responses. *Plant Molecular Biology*, **2009**. 69: 473-488.
- [30] E. Grata, J. Bocard, G. Glauser, P.A. Carrupt, E.E. Farmer, J.L. Wolfender, S. Rudaz. Development of a two-step screening ESI-TOF-MS method for rapid determination of significant stress-induced metabolome modifications in plant leaf extracts: The wound response in *Arabidopsis thaliana* as a case study. *Journal of Separation Science*, **2007**. 30: 2268-2278.
- [31] G. Glauser, E. Grata, L. Dubugnon, S. Rudaz, E.E. Farmer, J.L. Wolfender. Spatial and temporal dynamics of jasmonate synthesis and accumulation in *Arabidopsis* in response to wounding. *Journal of Biological Chemistry*, **2008**. 283: 16400-16407.
- [32] J. Berdy. Bioactive microbial metabolites. *The Journal of antibiotics*, **2005**. 58: 1-26.
- [33] K. Hostettmann, J.-L. Wolfender, C. Terreaux. Modern screening techniques for plant extracts. *Pharmaceutical biology*, **2001**. 39: 18-32.
- [34] S.D. Sarker, L. Nahar. *Natural Products Isolation*. Methods in Molecular Biology. Vol. 864. **2012**, Humana Press.
- [35] K. Hostettmann, A. Marston, M. Hostettmann. *Preparative Chromatography Techniques: Applications in Natural Product Isolation*. 2nd ed. **1997**, Berlin, Springer.
- [36] J.L. Wolfender. HPLC in Natural Product Analysis: The Detection Issue. *Planta Medica*, **2009**. 75: 719-734.
- [37] W. Weckwerth. Metabolomics in systems biology. *Annual Review of Plant Biology*, **2003**. 54: 669-689.
- [38] G. Glauser, D. Guillarme, E. Grata, J. Bocard, A. Thiocone, P.-A. Carrupt, J.-L. Veuthey, S. Rudaz, J.-L. Wolfender. Optimized liquid chromatography-mass spectrometry approach for the isolation of minor stress biomarkers in plant extracts and their identification by capillary nuclear magnetic resonance. *Journal of Chromatography A*, **2008**. 1180: 90-98.
- [39] J.F. Hu, E. Garo, H.D. Yoo, P.A. Cremin, L. Zeng, M.G. Goering, M. O'Neil-Johnson, G.R. Eldridge. Application of capillary-scale NMR for the structure determination of phytochemicals. *Phytochemical Analysis*, **2005**. 16: 127-133.
- [40] J.-F. Hu, H.-D. Yoo, C.T. Williams, E. Garo, P.A. Cremin, L. Zeng, H.C. Vervoort, C.M. Lee, S.M. Hart, M.G. Goering, M. O'Neil-Johnson, G.R. Eldridge. Miniaturization of the Structure Elucidation of Novel Natural Products—Two Trace Antibacterial Acylated Caprylic Alcohol Glycosides from *Arctostaphylos pumila*. *Planta Medica*, **2005**. 71: 176-180.
- [41] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *European Journal of Pharmaceutics and Biopharmaceutics*, **2008**. 68: 430-440.
- [42] O. Potterat, M. Hamburger. Concepts and technologies for tracking bioactive compounds in natural product extracts: generation of libraries, and hyphenation of analytical processes with bioassays. *Natural Product Reports*, **2013**. 30: 546-564.

- [43] L.A. Marques, J. Kool, F. de Kanter, H. Lingeman, W. Niessen, H. Irth. Production and on-line acetylcholinesterase bioactivity profiling of chemical and biological degradation products of tacrine. *Journal of Pharmaceutical and Biomedical Analysis*, **2010**. 53: 609-616.
- [44] D.A. van Elswijk, O. Diefenbach, S. van der Berg, H. Irth, U.R. Tjaden, J. van der Greef. Rapid detection and identification of angiotensin-converting enzyme inhibitors by on-line liquid chromatography–biochemical detection, coupled to electrospray mass spectrometry. *Journal of Chromatography A*, **2003**. 1020: 45-58.
- [45] P. Tammela, T. Wennberg, H. Vuorela, P. Vuorela. HPLC micro-fractionation coupled to a cell-based assay for automated on-line primary screening of calcium antagonistic components in plant extracts. *Analytical and Bioanalytical Chemistry*, **2004**. 380: 614-618.
- [46] L.I. Zon, R.T. Peterson. *In vivo* drug discovery in the zebrafish. *Nature Reviews Drug Discovery*, **2005**. 4: 35-44.
- [47] S. Basu, C. Sachidanandan. Zebrafish: A Multifaceted Tool for Chemical Biologists. *Chemical Reviews*, **2013**. In press.
- [48] S. Challal, N. Bohni, O.E. Buenafe, C.V. Esguerra, P.A.M. de Witte, J.-L. Wolfender, A.D. Crawford. Zebrafish Bioassay-guided Microfractionation for the Rapid *in vivo* Identification of Pharmacologically Active Natural Products. *Chimia*, **2012**. 66: 229-232.
- [49] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part I: Isocratic separation. *European Journal of Pharmaceutics and Biopharmaceutics*, **2007**. 66: 475-482.
- [50] H.K. Kim, Y.H. Choi, R. Verpoorte. NMR-based metabolomic analysis of plants. *Nature Protocols*, **2010**. 5: 536-549.
- [51] K. Dettmer, P.A. Aronov, B.D. Hammock. Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, **2007**. 26: 51-78.
- [52] J. Boccard, J.L. Veuthey, S. Rudaz. Knowledge discovery in metabolomics: An overview of MS data handling. *Journal of Separation Science*, **2010**. 33: 290-304.
- [53] N.E. Madala, F. Tugizimana, P.A. Steenkamp, L.A. Piater, I.A. Dubery. The Short and Long of it: Shorter Chromatographic Analysis Suffice for Sample Classification During UHPLC-MS-Based Metabolic Fingerprinting. *Chromatographia*, **2013**. 76: 279-285.
- [54] W.B. Dunn, D.I. Ellis. Metabolomics: Current analytical platforms and methodologies. *TrAC, Trends in Analytical Chemistry*, **2005**. 24: 285-294.
- [55] S. Forcisi, F. Moritz, B. Kanawati, D. Tziotis, R. Lehmann, P. Schmitt-Kopplin. Liquid chromatography–mass spectrometry in metabolomics research: Mass analyzers in ultra high pressure liquid chromatography coupling. *Journal of Chromatography A*, **2013**. 1292: 51-65.
- [56] J.W. Allwood, R. Goodacre. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis*, **2010**. 21: 33-47.
- [57] S. Bertrand, O. Schumpp, N. Bohni, A. Bujard, A. Azzollini, M. Monod, K. Gindro, J.-L. Wolfender. Detection of metabolite induction in fungal co-cultures on solid media by high-throughput differential ultra-high pressure liquid chromatography–time-of-flight mass spectrometry fingerprinting. *Journal of Chromatography A*, **2013**. 1292: 219-228.
- [58] S. Bertrand, O. Schumpp, N. Bohni, M. Monod, K. Gindro, J.-L. Wolfender. *De novo* production of metabolites by fungal co-culture of *Trichophyton rubrum* and *Bionectria ochroleuca*. *Journal of Natural Products*, **2013**. 76: 1157-1165.
- [59] K.A. Leiss, Y.H. Choi, R. Verpoorte, P.G. Klinkhamer. An overview of NMR-based metabolomics to identify secondary plant compounds involved in host plant resistance. *Phytochemistry Reviews*, **2011**. 10: 205-216.
- [60] K.W. Cheng, F. Cheng, M. Wang. Liquid chromatography-mass spectrometry in natural product research, in *Bioactive Natural Products: Detection, Isolation, and Structural Determination, Second Edition*, S.M. Colegate and R.J. Molyneux, Editors. **2008**, CRC press: London. p. 245-266.

- [61] J.R. Yates, C.I. Ruse, A. Nakorchevsky. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications, in *Annual Review of Biomedical Engineering*, **2009**. p. 49-79.
- [62] A. Marston. Role of advances in chromatographic techniques in phytochemistry. *Phytochemistry*, **2007**. 68: 2786-2798.
- [63] D.T.T. Nguyen, D. Guillarme, S. Rudaz, J.L. Veuthey. Fast analysis in liquid chromatography using small particle size and high pressure. *Journal of Separation Science*, **2006**. 29: 1836-1848.
- [64] J.-L. Wolfender, E.F. Queiroz, K. Hostettmann. The importance of hyphenated techniques in the discovery of new lead compounds from nature. *Expert Opinion on Drug Discovery*, **2006**. 1: 237-260.
- [65] D. Guillarme, J. Schappler, S. Rudaz, J.-L. Veuthey. Coupling ultra-high-pressure liquid chromatography with mass spectrometry. *TrAC, Trends in Analytical Chemistry*, **2010**. 29: 15-27.
- [66] A.G. Marshall, C.L. Hendrickson. High-Resolution Mass Spectrometers. *Annual Review of Analytical Chemistry*, **2008**. 1: 579-599.
- [67] G.A. Theodoridis, H.G. Gika, E.J. Want, I.D. Wilson. Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Analytica Chimica Acta*, **2012**. 711: 7-16.
- [68] X.-F. Chen, H.-T. Wu, G.-G. Tan, Z.-Y. Zhu, Y.-F. Chai. Liquid chromatography coupled with time-of-flight and ion trap mass spectrometry for qualitative analysis of herbal medicines. *Journal of Pharmaceutical Analysis*, **2011**. 1: 235-245.
- [69] J.-L. Zhou, L.-W. Qi, P. Li. Herbal medicine analysis by liquid chromatography/time-of-flight mass spectrometry. *Journal of Chromatography A*, **2009**. 1216: 7582-7594.
- [70] W.F. Smyth, T.J.P. Smyth, V.N. Ramachandran, F. O'Donnell, P. Brooks. Dereplication of phytochemicals in plants by LC-ESI-MS and ESI-MSn. *TrAC Trends in Analytical Chemistry*, **2012**. 33: 46-54.
- [71] G. Marti, M. Erb, S. Rudaz, T. Turlings, J.-L. Wolfender. Search for Low-Molecular-Weight Biomarkers in Plant Tissues and Seeds Using Metabolomics: Tools, Strategies, and Applications, in *Seed Development: OMICS Technologies toward Improvement of Seed Quality and Crop Yield*, G.K. Agrawal and R. Rakwal, Editors. **2012**, Springer Netherlands. p. 305-341.
- [72] P.M. Barnes, B. Bloom, R.L. Nahin. Complementary and alternative medicine use among adults and children: United States, 2007, **2008**. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
- [73] L. Mattoli, F. Cangì, C. Ghiara, M. Burico, A. Maidecchi, E. Bianchi, E. Ragazzi, L. Bellotto, R. Seraglia, P. Traldi. A metabolite fingerprinting for the characterization of commercial botanical dietary supplements. *Metabolomics*, **2011**. 7: 437-445.
- [74] P.Q. Tranchida, I. Bonaccorsi, P. Dugo, L. Mondello, G. Dugo. Analysis of *Citrus* essential oils: state of the art and future perspectives. A review. *Flavour and Fragrance Journal*, **2012**. 27: 98-123.
- [75] J.L. Wolfender, P.J. Eugster, N. Bohni, M. Cuendet. Advanced Methods for Natural Product Drug Discovery in the Field of Nutraceuticals. *Chimia*, **2011**. 65: 400-406.

Chapter II – Introduction to UHPLC in Natural Products Analysis

This chapter is based on a book chapter entitled *UHPLC in Natural Products Analysis*, published in *UHPLC in Life Sciences*.

Foreword

As mentioned in the introduction, there is a need for high resolution LC separation techniques for the high resolution profiling of complex natural samples. Several parameters may be optimised to increase the resolution, such as particle diameter, nature of the stationary phase, or mobile phase temperature [1, 2]. Recent developments in liquid chromatography provide solutions to increase chromatographic resolution, that are based on the usage of (1) silica-based monolithic supports, (2) sub-2 μm particles columns with their dedicated UHPLC systems and (3) fused-core columns packed with sub-3 μm superficially porous particles.

(1) Monolithic silica columns (or 'monoliths') introduced at the end of the 90's are constituted of a single piece of porous silica. They are characterised by a higher permeability compared to conventional packed columns, which enable the use of high flow rate and makes these columns well-adapted to rapid separations and high resolution analyses [3]. Monoliths are however not frequently used for the separation of small molecules for several reasons, including the low number of commercialised columns due to patent exclusivity, and the low physicochemical resistance of the support.

(2) Sub-2 μm particles columns, based on UHPLC technology introduced in 2004, are based on the reduction of the conventional HPLC particles to a diameter below 2 μm . The simultaneous development of systems and phase chemistries able to withstand the high backpressure generated provide very high chromatographic efficiencies [1, 4]. Thanks to this, the analysis time

of existing HPLC separations may be dramatically reduced using a simple method transfer. Moreover, the UHPLC technology provide very high peak capacities in high resolution separations.

(3) Finally, fused-core columns packed with sub-3 μm superficially porous particles were developed in the 90's but were only commercialised in 2007. Their particles are composed of a 1.7 μm solid core surrounded by a 0.5 μm porous silica layer (providing a total particle diameter of 2.7 μm) to reduce the mass transfer, *i.e.* the C-term of the Van Deemter curve [1]. This is especially useful for the separation of macromolecules that possess a low diffusivity [5]. Moreover, the efficiency of these columns is almost as high as the one of conventional HPLC columns, while the generated backpressure is reduced to 50% [6]. This allows working with higher flow rates or coupling columns in series, providing fast or high resolution separations [7, 8]. In addition, fused-core columns present the advantage of providing higher resolutions than conventional columns without the need to update the HPLC instrumentation, contrary to UHPLC technology. However, the fused core columns have not yet been widely used in NP analysis due to their recent commercialisation (its introduction after the adoption of UHPLC by many laboratories), and the limited number of available phase chemistries due to patent exclusivity.

Of these three strategies, UHPLC is clearly the most frequently used for NPs analysis, as displayed in Figure II.1 (red bars). While fused-

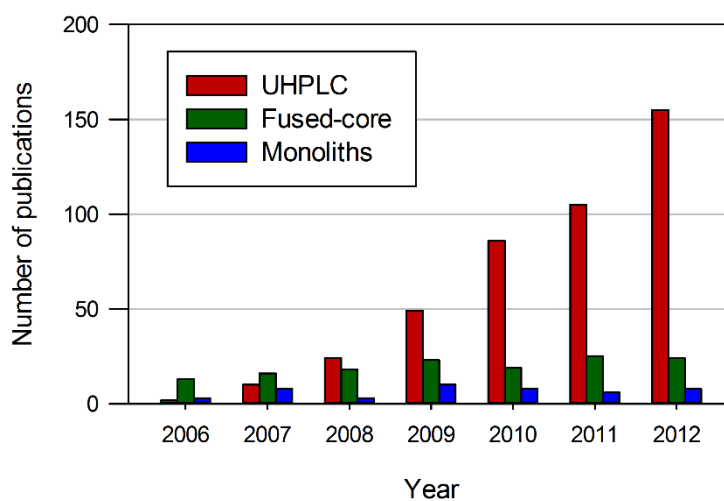


Figure II.1. Number of publications retrieved in Web of Knowledge using the following keywords: "natural product", NOT "plasma", NOT "urine", NOT "blood", NOT "serum", NOT "cell", NOT "wastewater", as well as "UHPLC" OR "UPLC" for the red bars plot, "fused-core" OR "core-shell" for the green bars plot, and "monolithic column" for the blue bars plot. Data collected on the 13th of June 2013.

core (green bars) and monoliths (blue bars) columns were used or mentioned in 10 to 20 papers per year for the last 7 years, this number dramatically increased for UHPLC.

This approach is probably the most adapted to high resolution profiling of natural samples in the

frame of NP research. Therefore, our laboratory has extensively studied the possibilities offered by UHPLC technology in NP analysis since its first years of commercialisation. This chapter introduces the fundamentals of LC to understand the advantages and constraints of reducing the particle size to 1.7 μm and its many possible applications in NP research.

UHPLC in Natural Products Analysis

Philippe J. Eugster

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Jean-Luc Wolfender

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Book chapter published in *UHPLC in Life Sciences*, Editors D. Guillarme and J.-L. Veuthey, RSC Publishing.

Available on 8th of June 2012.

Abstract

HPLC is an efficient analytical chromatographic technique that has been used for the direct separation of natural products (NPs) in complex crude extracts. Ultra-high pressure liquid chromatography (UHPLC) has been recently introduced in NP research and has demonstrated that it can advantageously replace existing HPLC methods for many applications, including quality control, profiling and fingerprinting, dereplication, and metabolomics. The development of sub-2 μm packing columns has allowed for a remarkable decrease in analysis time and an increase in peak capacity, sensitivity, and reproducibility. The resulting excellent chromatographic performance also opens new research possibilities, notably for the detailed profiling of metabolite crude extracts and for metabolomics. In this chapter, an overview of the latest applications of this technology to NP analysis is provided. Several new trends involving UHPLC in this field of research are discussed.

1. Introduction

Natural products (NPs) are known to possess a very high diversity in chemical space [9], and, as a result, they have a profound impact on chemical biology and drug development [10]. Bioactive NPs can be found in many different biological matrices, such as plants, marine organisms, micro-organisms and animals. In many cases, each organism produces a huge variety of these NPs. Plants, for example, are known to produce NPs that either are essential for their life (primary metabolites) or are not directly involved in their normal growth, development or reproduction, but are necessary for survivability, fecundity or aesthetics (secondary metabolites). The complete composition of a given organism, known as the metabolome, can be extremely large and has been estimated to contain a few thousand constituents; however, the exact size of a plant or fungal metabolome is still unknown [11].

The high chemical diversity of secondary metabolites can probably be explained by the effects of evolutionary pressure, which provided an impetus for organisms to create biologically active molecules, and/or by the structural similarity of protein targets across many species. This large chemical diversity [12] is also directly linked to a high variability of the intrinsic physicochemical properties of NPs, which causes the separation and universal detection of NPs to be extremely challenging.

The analysis of individual NPs in a complex crude extract requires the efficient separation of individual components before detection. In this respect, high-performance liquid chromatography (HPLC) has been recognised

since the early 1980s as the most versatile technique for the efficient separation of NPs in crude mixtures without the need for complex sample preparation [13]. HPLC has been greatly developed through the years in terms of its convenience, speed, choice of stationary phases, sensitivity, applicability to a broad variety of sample matrices and its ability to couple to spectroscopic detection methods [14]. The development of HPLC columns with different phase chemistries (especially reversed-phase) enabled the separation of almost any type of NPs.

HPLC is thus widely used and has been adapted to the analysis of a broad range of NPs, generally without the need for complex sample preparation. Because in many cases, the NPs of interest must be isolated from their original biological matrix, other liquid chromatography (LC) preparative techniques that use similar HPLC phase chemistries can be used to isolate milligram amounts of pure NPs. These techniques include low pressure LC (LPLC), medium pressure LC (MPLC), semi-preparative and preparative HPLC [15].

Crude extracts of natural origin can be separated either by using the raw mixtures or by using samples that are enriched by extraction via solid phase extraction (SPE) or liquid–liquid extraction (LLE). These separations are usually performed by reversed-phase chromatography on C₁₈ material with the acetonitrile/water (ACN/H₂O) or methanol/water (MeOH/H₂O) solvent systems in the gradient elution mode. To improve the separation efficiency, various modifiers or buffers can be added to the mobile phase to tune the

selectivity of the separation or the sensitivity of detection.

However, the choice of the appropriate HPLC detector is crucial because of the diversity of NPs, and thus there is no universal technique for NP detection. Simple detectors, such as ultraviolet (UV), evaporative light scattering detection (ELSD), fluorescence detection (FD), electrochemical detection (ECD), refractive index detection (RID), flame ionisation detection (FID), chemiluminescence (CL) and charged aerosol detection (CAD), can be used, with UV and ELSD being the most widespread [16]. In addition, the coupling of HPLC with photodiode array (PDA), mass spectrometry (MS), nuclear magnetic resonance (NMR) and infrared (IR) is often of key importance in the dereplication process to collect online preliminary structural information during the HPLC separation.

The latest developments in HPLC technologies, including the recent introduction of phase chemistries that are stable at a wide range of pH values, fully porous sub-2 μm and core-shell sub-3 μm packing particles [17] or monolith columns, have considerably improved the performance of HPLC systems in terms of their resolution, speed and reproducibility. Efficiencies exceeding 100 000 plates and peak capacities over 900 can be attained by coupling columns together [18]. In this Chapter, we will review all aspects related to the introduction of ultra-high-pressure liquid chromatography (UHPLC) for NP analysis and the transition of conventional HPLC profiling methods to this new technology in different fields of plant research, including quality control (QC), metabolite profiling/fingerprinting and dereplication, and metabolomics.

1.1. Implementation of UHPLC in NP Analysis

As in other fields of analytical science, the introduction of UHPLC systems that operate at very high pressures and use porous sub-2 μm packing columns in NP research has allowed for a remarkable decrease in analysis time and increases in peak capacity, sensitivity and reproducibility as compared to conventional HPLC. This technology has started being implemented in many laboratories that work in NP research, and it has not only replaced conventional HPLC but also opened new fields of research, such as metabolomics and high-resolution profiling.

Thus, NP chemists have used UHPLC to considerably enhance the throughput of their targeted analyses using very rapid gradients on short columns (*e.g.*, for QC or for crude extract standardisation). In addition, they have also pushed the UHPLC technology to its limit for performing very high resolution profiling of complex mixtures using slow gradients on long columns (*e.g.*, for detailed metabolite profiling and dereplication). For many applications, important improvements in the overall performance have been reported.

Thus, there is a growing interest in UHPLC in plant science and in other aspects of NP research, such as fungal or bacterial metabolite studies and the standardisation of herbal products. The number of reported UHPLC-related applications for NP analysis has constantly increased since its introduction in 2006, while the reported applications of conventional HPLC methods remain relatively stable [19]. In 2010, 171 articles were published on the applications of UHPLC in NP analysis, while only one paper was reported in 2005 [20]. For comparison, in 2010, the number of reported HPLC applications exceeded 3500 [number of papers by year of publication

retrieved from SciFinder Scholar (Chemical Abstracts) using the keywords “UHPLC” or “UPLC” or “RRLC” and “plants” or “phytochemistry” or “natural products” in September 2011]. The main factor hindering a

faster implementation of UHPLC in NP research laboratories is the need for specific instrumentation, as conventional HPLC systems cannot tolerate the high pressure generated by the sub-2 μm columns

2. Multiple Facets of UHPLC in NP research

As mentioned above, HPLC has been extensively used in many aspects of NP research, and UHPLC is advantageously replacing HPLC for both high-throughput and high-resolution applications. However, to use UHPLC, the acquisition speed of the detectors must be adapted for monitoring thinner LC peaks and the sample extracts must be prepared in a way that meets the requirements of the sub-2 μm columns.

2.1. UHPLC Detectors used for NP Analysis

As described for HPLC, different detectors have been used with UHPLC to analyse NPs. Spectroscopic methods are often used in hyphenation with UHPLC, which is important for dereplication in metabolite profiling studies. Compared to HPLC, however, UHPLC imposes some limitations in the choice of the detector, both in terms of the acquisition rate and the loading capacities of the column. For these reasons, UHPLC at present is not compatible with spectroscopic detectors such as NMR (LC-NMR hyphenation) and IR (LC-IR), which is a disadvantage, as these types of detectors are important for the *de novo* structure determination of NPs online or at-line [21]. However, UHPLC can be efficiently hyphenated to MS which is the most useful detector for NP analysis [16, 22]. Indeed, although it is expensive, the use of a mass spectrometer as detector for LC systems provides excellent sensitivity and selectivity in the analysis of NPs in complex biological matrices. Furthermore, MS detection provides important online structural information,

such as the molecular mass, molecular formula and diagnostic fragments, which are crucial for dereplication and rapid online characterisations of NPs [21].

Single or triple quadrupole systems have been coupled to UHPLC but have mainly been used for the specific detection of NPs through single ion monitoring (SIM) [23] or multiple reaction monitoring (MRM) [24, 25] experiments rather than for full scan acquisitions of MS or tandem MS (MS/MS) spectra. With quadrupole-based analysers, the sampling rate can be problematic, and modern instruments that possess improved acquisition rates should be selected for coupling with UHPLC. Dwell times and inter-channel delays have, however, been reduced down to less than 5 ms in the SIM mode with new analyser generations, for example [26].

The use of a high acquisition-rate mass analyser, such as the time-of-flight MS (TOF-MS) detector, has considerably boosted the use of UHPLC, providing a powerful UHPLC-TOF-MS platform with high sensitivity and specificity of detection and accurate mass detection is used [27]. Indeed, TOF-MS instruments are well adapted to record and store data over a broad mass range without compromising sensitivity, with high resolving power [generally $\cdot 10\ 000$ full width at half maximum (FWMH)] to be attained in routine analysis at speeds up to $40\ \text{spectra s}^{-1}$ [19]. Even higher resolving power, up to $50\ 000$ for the latest generation of TOF instruments have been reported, reducing the risk of false negative

results when complex biological matrices are to be analysed [28].

Furthermore, with hybrid systems, such as hybrid quadrupole TOF mass spectrometer (Q-TOF-MS/MS), acquisition of MS/MS spectra at high frequency provides more online structural information or more specific detection.

Trap systems have given very useful structural information on conventional LC-MS system through MSⁿ experiments. High-resolution Orbitrap Fourier transform (FT) MS, with a resolving power up to 100 000, is very useful for structural identification and provides high quality spectra when used with infusions of pure NPs [29]. Trap systems have rarely been used in conjunction with UHPLC [30] because the lower peak duration in UHPLC does not match the time needed by the spectrometer to acquire different MSⁿ spectra with sufficient ion statistics, especially when high-resolution measurements are required. The most recent generation of Orbitrap systems are, however, able to work at up to a 5 to 10Hz acquisition frequency with a reasonable loss in resolution [31]. First attempts of hyphenation of such analysers with UHPLC were successful in proteomics [32] and are also promising for the high-resolution profiling of low-molecular-mass compounds such as NPs.

In conclusion, MS analysers are the most useful detectors coupled to UHPLC for NPs analysis. For specific and sensitive detection, the high acquisition rate of triple quadrupole instruments should be preferred, while non-targeted analysis with full scan spectra acquisition can be efficiently performed on TOF-MS systems. In the years to come, faster Orbitrap systems may compete with the high resolving power of TOF-MS and also provide MSⁿ structural information online [33].

2.2. Targeted vs Untargeted Analyses of NPs

UV detection has often been used for targeted, quantitative UHPLC analyses of NPs (ca. 60% of applications). In addition to simple UV detectors, PDA detectors have also been applied for online UV spectra acquisition [34]. Indeed, UV is the most simple and the most widely used among all LC detectors [35]. It is quite easy to optimise UV-visible (UV-Vis) and UV-diode array detector (UV-DAD) detectors to meet the requirements of UHPLC in terms of the sampling rate. The UV cell volume should be reduced to avoid peak dispersion in UHPLC, while maintaining a sufficient path length of light passing through the UV cell, as the absorbance is directly proportional according to the Beer-Lambert law. Generally, the UV cell in conventional HPLC systems has a volume between 10 and 25 μL and a path length of 10 mm. The UV cell in UHPLC, however, is reduced to 0.25–3 μL with a path length of 3–10 mm, depending on the provider. Although UV suffers from some limitations, particularly for NPs that do not possess UV chromophores, this detection method has the best combination of sensitivity, linearity, versatility and reliability of all of the LC detectors that have been developed. Most NPs adsorb UV light in the range of 200–550 nm, including all substances having one or more double bonds and all substances that have unshared electrons. Thus, even compounds with weak chromophores, such as triterpene glycosides, can be successfully detected by UV at short wavelengths (203 nm) [36]. However, in this system, several mobile-phase constituents that exhibit high UV cut-offs should be avoided, as they might inhibit the detection of NPs with weak chromophores [37]. The use of PDA detection provides UV spectra directly online and is particularly useful for the detection of natural products with characteristic chromophores [38]. For example, polyphenols can be efficiently localised by this method because they possess characteristic chromophores. With this type of

compound, PDA-UV spectral libraries can be built and used for dereplication [34].

ELSD is also compatible with UHPLC. ELSD is a quasi-universal detector for LC, as it can detect any analyte that is less volatile than the mobile phase, regardless of its optical, electrochemical or other properties [39, 40]. ELSD is a mass-dependent detector (in contrast to UV, which is a concentration-dependent detector), and the generated response does not depend on the spectral or physicochemical properties of the analyte. Because the detection is based on the measurement of light scattering (using a photomultiplier or a photodiode) produced by the non-volatile residual particles after the evaporation of the mobile phase, the sampling rate is generally not critical (equal to at least 50 Hz in any commercial devices) and thus is sufficient for even ultra-fast experiments. As was recently reported [41, 42], the coupling of UHPLC with ELSD is possible, but the latter remains a non-negligible source of additional dispersion that increases with higher mobile-phase flow rates.

In NPs analysis, HPLC-ELSD has been mainly used for the detection of compounds with weak chromophores, such as terpenes, in both aglycone and glycosidic forms, saponins and some alkaloids [16]. For example, coupling ELSD to UHPLC has been used to quantify triterpenoids in phytopharmaceuticals containing black cohosh (*Actaea racemosa*) [42].

Figure II.2 shows the complementarity of UHPLC-PDA-ELSD and -MS for profiling a plant extract (crude isopropanol extract of *Arabidopsis thaliana*). The UV detection at 350 nm is rather selective, showing mainly peaks that are related to flavonoids. The trace at 254 nm displays most of the NPs containing an aromatic chromophore

or conjugated double bonds. The ELSD provides more peaks than the UV detector, especially for the detection of non-polar compounds (mainly lipids, in this example). Finally, the MS trace [base peak intensity (BPI) positive ionisation (PI) mode] demonstrates that almost all constituents can be ionised, with the exception of the very polar constituents that are detected by ELSD. With LC-MS, however, the response cannot be linked to the quantity of NP detected, as it is compound-dependent [16]. Figure II.2 demonstrates the use of TOF-MS detection for the selective detection and rapid online characterisation of natural products, which is not possible with UV or ELSD detectors. As shown in the display of the UHPLC-ESI-PI-TOF-MS trace of the ion at m/z 741, the TOF-MS system provided a selective detection of this compound in the crude extract of *A. thaliana*. In addition, the corresponding high-resolution spectrum of compound F enabled the precise determination of the molecular formula ($C_{33}H_{41}O_{19}$) of its protonated molecule $[M+H]^+$ (m/z 741.2242), allowing for the identification of this compound as a flavonol triglycoside. More details on this dereplication procedure are provided in Section 5.3.

As previously mentioned, not all detectors that have been used for HPLC are fully compatible with UHPLC. However, those detectors that can be adapted to this technology, mainly by increasing their acquisition rates, generally demonstrate important improvements in data quality and throughput as compared to their HPLC counterparts. The use of UV, ELSD or MS enables very fast, targeted analysis to be performed, mainly on a quantitative viewpoint, while PDA and high-resolution MS or MS/MS provide high-quality online spectroscopic information in non-targeted analysis, which is especially useful for metabolite profiling, chemical screening or dereplication applications.

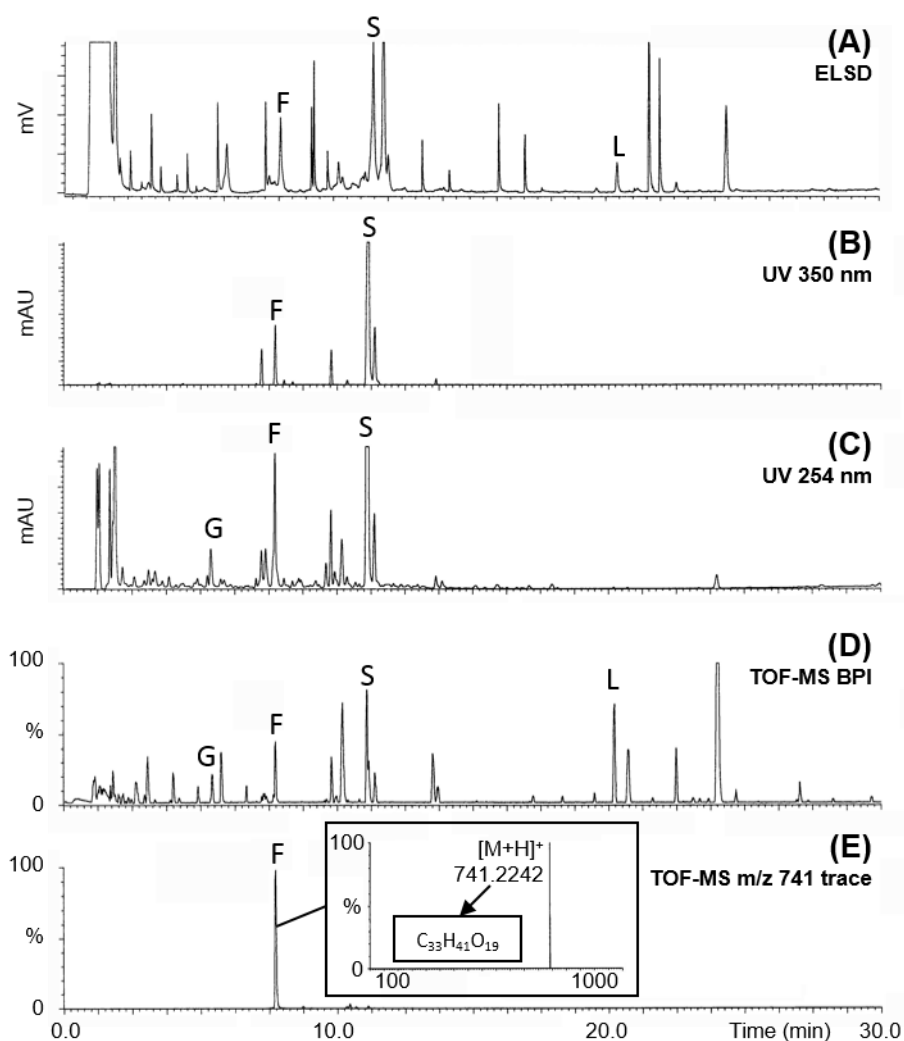


Figure II.2. Chromatograms of *A. thaliana* (crude leaves isopropanol extract) obtained with different detection techniques: (A) ELSD, (B) PDA trace at 350 nm, (C) PDA trace at 254 nm, (D) TOF-MS BPI in PI mode and (E) extracted ion trace of m/z 741.2, TOF-MS in PI mode. Separation was carried out on an Acquity BEH C₁₈ column (150 x 2.1 mm; 1.7 μ m), with a 5–98% ACN gradient in 45 min; both water and ACN contained 0.1% FA. Compound F is a flavonol glycoside, G is a glucosinolate, S is a synapoyl derivative and L is a galactolipid. Inset: TOF-MS spectra of F. Adapted from [16].

UHPLC has thus conquered domains related to the QC of plants or food extracts, especially for the standardisation and safety assessment of medicinal plants, phytopharmaceuticals or dietary supplements. In this respect, standard HPLC procedures are gradually being replaced by high-throughput, targeted and quantitative

UHPLC methods [43]. UHPLC is also being used more commonly for dereplication purposes in drug discovery programmes in conjunction with both PDAs and MS detection. Dereplication is the process of differentiating NP extracts that contain known secondary metabolites from those that contain novel compounds of interest

[44]. Here, the high resolving power of UHPLC is required for the deconvolution of closely related metabolites, such as isomers, to obtain high-quality online spectra without interference, which can then be used for database searching or for spectral interpretation. Such a process represents an important step in drug discovery programmes, as the early structural determination of known NPs avoids the time-consuming processes required for their isolation and enables the optimisation of bioactive-guided isolation procedures [45].

In order to illustrate both the high throughput and the high chromatographic resolution that can be obtained by UHPLC as compared to standard HPLC, the metabolite profiling of a representative crude plant extract (the widely used phytomedicine *Ginkgo biloba*) is displayed in Figure II.3. By using UHPLC conditions for the profiling of this standardised extract, a 9-fold reduction in analysis time could be obtained by transferring the 60 min gradient from the HPLC column (150 x 4.6 mm; 5 μ m) to a short gradient on a 50 mm UHPLC column (50 x 2.1 mm; 1.7 μ m). It should be noted that even if chromatographic calculations indicate that the peak capacity should remain constant with such a method transfer, a significant decrease is measured, mainly due to peak broadening in the MS source. The use of the same gradient time on a longer UHPLC column (150 x 2.1 mm; 1.7 μ m) provided a notable increase in resolution.

In addition, UHPLC, and especially UHPLC-TOF-MS, is beginning to play an important role in new research fields, such as metabolomics [11]. This holistic approach has recently emerged with other 'omics' technologies in biological research [46] and concerns the large-scale analysis of metabolites in given organisms at different physiological states. Profiling the metabolome

has the potential to provide the most "functional" information among the 'omics' technologies that are used in systems biology [47]. Currently, UHPLC-TOF-MS represents a key method for both metabolite fingerprinting and for metabolite profiling from crude extracts. Metabolic fingerprinting consists of high-throughput separations that are used not with the intention of identifying each observed metabolite, but rather to compare patterns or "fingerprints" of metabolites that change in response to disease, nutrition, toxin exposure or environmental or genetic alterations. In contrast, metabolic profiling focuses on the analysis of a group of metabolites that are either related to a specific metabolic pathway or to a class of compounds. In most cases, metabolic profiling is a hypothesis-driven approach rather than a hypothesis-generating one [48].

Metabolic fingerprinting is also becoming frequently used to assess phytoequivalence in untargeted QC methods [43, 49].

2.3. Column Phase Chemistries for NP Analysis

As mentioned previously, most separations of crude extracts with complex NP compositions are carried out using a gradient mode on reversed-phase columns. However, as discussed in the Introduction, the chemical space occupied by NPs is very broad, and the analysis of both very polar and very lipophilic compounds is important. The large variety of sub-2 μ m phase chemistries that are available can resolve almost all analytical issues: C₈ and C₁₈ are used for plant extracts of average polarity; C₄ and cyano for the most apolar fractions; hydrophilic interaction liquid chromatography (HILIC) with bare silica or diol, amino bonding for the most polar fractions;

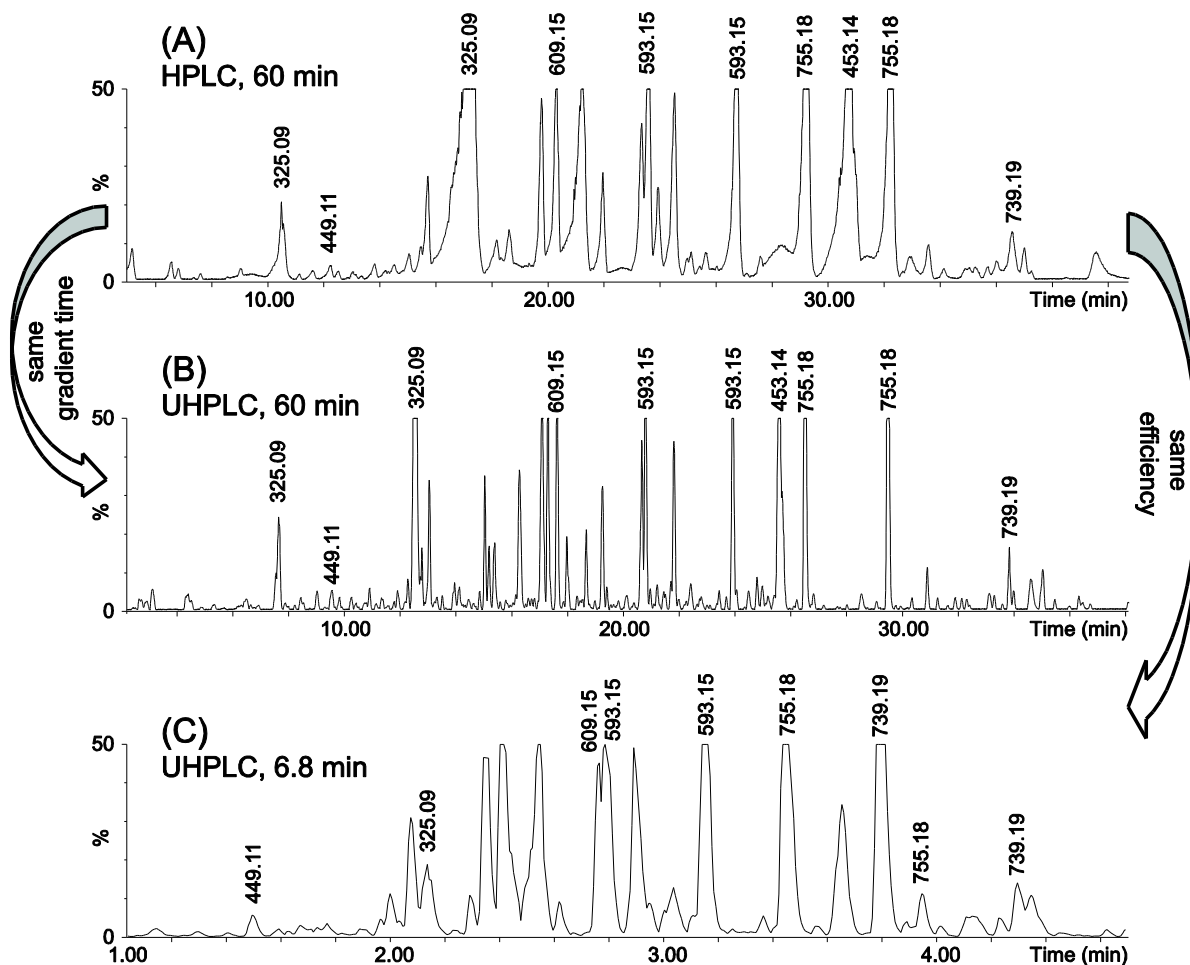


Figure II.3. Example of an HPLC–UHPLC transfer: chromatograms of a standardised *G. biloba* extract with a 5–40% ACN gradient using (A) a classic HPLC column (150 x 4.6 mm; 5 μ m) in 60 min at 1.0 mL min^{-1} , (B) a UHPLC column (150 x 2.1 mm; 1.7 μ m) in 60 min at 0.35 mL min^{-1} and (C) a UHPLC column (50 x 2.1 mm; 1.7 μ m) in 6.8 min at 0.60 mL min^{-1} . Detection was carried out by an ESI-TOF-MS analyser in the NI mode in the 100–1000 m/z range.

and biphenyl, pentafluorophenyl (PFP) or zirconia for alternative selectivity [19]. Despite this wide variety, relatively few applications have reported the use of reversed-phase columns other than C_{18} (see Figure II.4).

Concerning column geometries, in Figure II.4, it can be also noted that more than 50% of all

separations are performed on a 100 mm column, 13% are performed on a 150 mm column, and approximately 30% are performed on a 50 mm or shorter column. The coupling of up to three 150 mm column in series (total column length of 450 mm) has also been reported for very high-resolution profiling of NP [50].

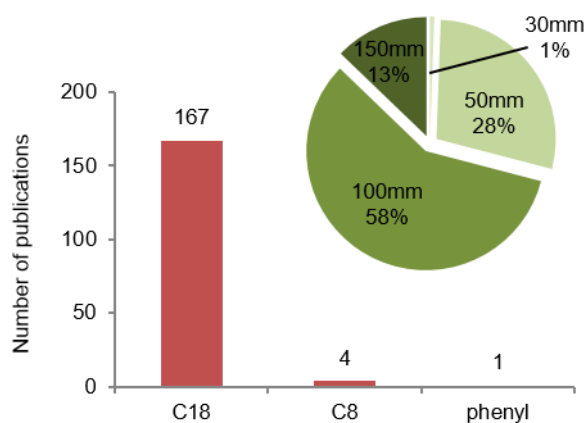


Figure II.4. Classification of all papers retrieved from a SciFinder Scholar (Chemical Abstracts) search using “UPLC OR UHPLC” and “natural products” or “plants” or “phytochemistry” keywords as related to column phase chemistries (C₁₈, C₈ and phenyl). The inset shows the column length distribution. (Compiled in August 2011.)

3. Fast Targeted Analysis

Fast targeted analyses are typically used to check the purity of an isolated NP, to quantify a compound in a complex mixture or as QC in the pharmaceutical field. With the introduction of UHPLC, the targeted QC of plant extracts consists of high-throughput methods where only a few constituents that are representative of the plant sample must be evaluated. As illustrated in Figure II.3, it is theoretically possible to obtain a 9-fold reduction of the analysis time as compared to standard HPLC methods, while maintaining an equivalent performance with UHPLC. However, it should be noted that method transfers in plant analysis are seldom purely geometric due to the use of different stationary phase chemistries in many applications. Thus, analysis times are often slightly shorter or longer than the predicted 9-fold reduced time.

A targeted analysis using UHPLC can be carried out to check the purity of a given NP during its isolation from a complex extract through a rapid analysis of the collected LC fractions. In this case, the analysis is very similar to that performed for any simple mixture of organic compounds, and the separation is optimised for the target compound. The high throughput of UHPLC is an appreciable advantage for bioactivity-guided fractionation approaches due to the high number of simplified fractions that are generated and the need for these fractions to be individually analysed for pooling and for the final purification of the compound of interest.

In the large majority of literature reports using targeted analysis by UHPLC (mainly UHPLC-UV or UHPLC-MS/MS), the targeted applications of UHPLC have been developed for the specific

detection and quantification of a given NP or a set of NPs in a complex mixture. This was applied, for example, for monitoring the biosynthesis of microbial products in fermentation broths or in plant cell cultures, determining phytohormones in model or crop plants, quantifying specific markers in herbal drugs or food, standardising nutraceuticals or phytopharmaceuticals, or detecting NPs with toxic properties.

Using UHPLC for the QC of herbal medicine is of particular interest. Indeed, a suitable standardisation and QC procedure is required to guarantee the botanical identity of the raw material and the quality, safety and efficacy of the final phytopharmaceutical product. Furthermore, NP amounts in a given extract are strongly dependent on the season, time, place of harvest, and extraction method; thus, the NP amounts need to be quantified to guarantee efficacy. The same is also needed for some dietary supplements or functional foods, although to a lesser extent because of less strict regulations. In general, most of the standardisation methods consist of the quantification of one marker (a secondary metabolite characteristic of the plant of interest) and the verification that possible toxic constituents are absent or below a given limit. In an ideal case, the quantified marker is the NP holding the bioactivity, but, in many cases, this compound is unknown, or the activity is the result of several compounds acting synergistically [49]. For the standardisation of such preparations, a complete fingerprinting or the selective quantification of many markers might be required. Because of these characteristics, the

QC of plant extracts is difficult, but it nonetheless remains mandatory [51].

3.1. UHPLC-UV

Several plant extracts, mainly phytopharmaceuticals and food, have been analysed by high-throughput UHPLC-UV methods using simple UV or PDA detection. For example, the polyphenols of green tea have been extensively studied for their potential health benefits [52]. In a recent study, a comprehensive profiling of 29 phenolic compounds comprising caffeine in tea preparations, infusions and extracts has been carried out quantitatively. The

phenolic compounds were separated in less than 20 min on a C₁₈ column (100 x 2.1 mm; 1.7 μm) using a gradient elution mode with a 0.1% formic acid and methanol mobile phase [53]. The high-throughput capability of UHPLC was demonstrated in another study on green tea with the baseline separation of seven catechin standards in only 30 s using a short 50 mm column (50 x 2.1 mm; 1.7 μm), although the analysis of enriched extracts required longer columns [25]. In Figure II.5, we demonstrate that with specific detection at 280 nm, caffeine and the main catechin of green tea, epigallocatechin gallate (EGCG), could be separated and quantified using the same 50 mm column in

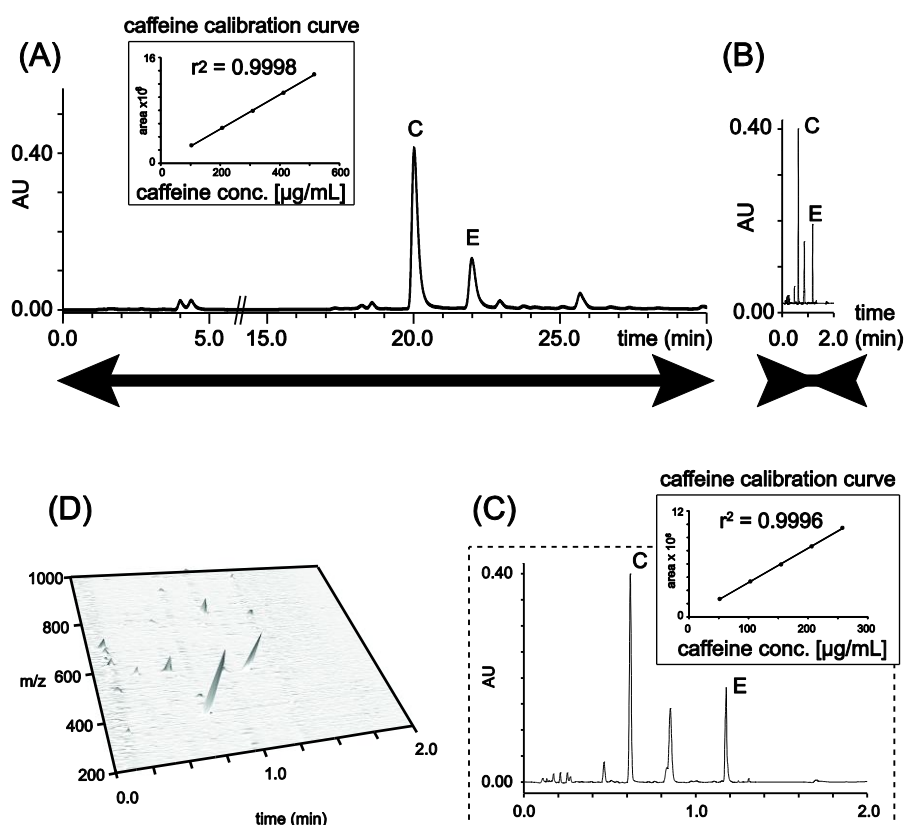


Figure II.5. Quantitative caffeine analysis of a green tea infusion. (A) 30 min HPLC separation on a C₁₈ column (150 x 4.6 mm; 5 μm) with UV detection at 280 nm. The calibration curve is shown in the inset. (B) and (C) are the same analysis after transfer of the HPLC method on to a short UHPLC column (50 x 2.1 mm; 1.7 μm) with further optimisation. The final analysis time is 2.0 min. (D) A three-dimensional ion map of the same UHPLC separation with TOF-MS detection in PI mode.

2 min (including the reconditioning step) by transferring and re-optimising an existing HPLC method in which the gradient time exceeded 30 min. The linearity of both the HPLC-UV and the UHPLC-UV methods were similar, and the selectivity of the UHPLC method was demonstrated by further UHPLC-TOF-MS analyses.

For the standardisation of phytopharmaceuticals containing anthraquinones, an UHPLC-PDA method enabled the simultaneous determination of five anthraquinone derivatives in three *Rheum* species [54]. The method was fully validated in terms of precision, accuracy, and linearity according to ICH guidelines, and UHPLC analysis was performed in only 3 min after optimisation using a 50 mm column (50 x 2.1 mm; 1.7 μm). A more recent study also incorporated the ionic liquid-based ultrasonic/microwave-assisted extraction (IL-UMAE) of five anthraquinones (physcion, chrysophanol, emodin, rhein and aloemodin) from rhubarb prior to their UHPLC-UV determination [55]. Another example of the quantification of NPs in phytopharmaceuticals is the standardisation of black cohosh by UHPLC-UV and UHPLC-ELSD. In this case, triterpenoids and isoflavonoids were identified and quantified in the rhizomes [42]. The analysis time of the extracts was reduced to 7 min with UHPLC using a 45–65% gradient of ACN/MeOH (7:3). Finally, faster separations have also been developed. For example, a powerful 1 min UHPLC-PDA quantification method of N-acyl-D/L-homoserine lactones in *Hordeum vulgare* and in *Pachyrhizus erosus* plants was proposed using a specific sample preparation [56].

3.2. UHPLC-MS

MS or MS/MS analysers providing more specific detection have been used for the quantitation of NPs in various matrices. A good example of fast standardisation of phytopharmaceuticals and dietary supplements is the specific detection of

the terpenes (ginkgolides and bilobalide) responsible for the antiplatelet activity of *Ginkgo biloba* by LC-MS on both a simple quadrupole and a TOF system. These methods involve the use of the $[\text{M}+\text{NH}_4]^+$ and $[\text{M}+\text{H}]^+$ ions of *Ginkgo* terpene in the PI mode with extractive ion monitoring by HPLC-TOF-MS and selected ion monitoring by UHPLC-MS using a single quadrupole analyser. The limit of detection (LOD) values for ginkgolide J, ginkgolide C, ginkgolide B, and ginkgolide A were in the range of 1–10 ng mL^{-1} with both methods. The LOD for bilobalide was 200 ng mL^{-1} by HPLC-TOF-MS and 35 ng mL^{-1} by UHPLC-MS. The gradient analysis in UHPLC was performed in 7 min [57]. For an even more specific detection, UHPLC-MS/MS was used on a triple quadrupole system operated in the MRM mode for the quantification of eight major alkaloids in extracts of *Coptis chinensis*, a commonly used herbal drug in traditional Chinese medicine (TCM). In this example, the mobile phase comprised an ammonium acetate buffer to optimise the peak shape and the separation of the alkaloids, and the complete analysis was performed in 5 min. This method was used for a rapid authentication and quality evaluation of this TCM obtained from various locations [58].

Very specific detection methods are also important for the determination of minor plant constituents that have key hormonal effects. In this respect, a validated method has been used for the simultaneous analysis of different phytohormones (auxins, cytokinins, and gibberellins) in vegetables in less than 7 min. UHPLC-MS/MS was performed in both PI and negative ionisation (NI) modes. The sample preparation was reduced to a minimum using a simple and fast extraction procedure in which all extractions and sample preparations are performed in the same tube (QuEChERS-based method). The method was validated, and mean recoveries were evaluated at three concentration levels (50, 100 and 250 mg kg^{-1}), ranged from 75 to 110% at the three levels assayed. Intra-day and

inter-day precisions, expressed as relative standard deviations (RSDs), were lower than 20 and 25%, respectively. Limits of quantification (LOQs) were equal to or lower than 10 mg kg⁻¹. The developed procedure was applied to different courgette samples, and naphthylacetic acid, naphthylacetamide, and benzyladenine were found in several of the analysed samples [24].

QC in routine analysis must be fast and well optimised to ensure high reproducibility. Because plant samples are often complex and several homologous compounds need to be separated in a fast gradient run, chromatographic modelling software has been used to optimise separations.

For example, the UHPLC conditions for the QC of *Rhizoma coptidis*, a plant containing different NPs with overlapping LC peaks, were calculated based on the retention time (RT) and peak shape parameters of the target peaks. The calculated chromatograms proved to be well correlated to the experimental ones, and the calculated method was found to be very helpful in obtaining satisfactory separation conditions of target compounds that were rapid and efficient [59].

Several additional, recent characteristic applications are summarised in Table II.1. Other applications reported especially prior to 2009 have been previously reviewed [19].

Table II.1. Summary of some published NPs UHPLC analyses classified according to the type of studies performed.

Ref.	Year	Compounds analysed or aim of the study	Plant or organism studied	Stationary phase chemistry	Column size (mm x mm)	Analysis time (a) (min)	Detection
			QC (targeted)				
[60]	2008	Simultaneous quantitative determination of 10 diterpenes	<i>Salvia miltiorrhiza</i>	BEH C18	50 x 2.1	15 / 10	PDA
[54]	2008	Quantitative determination of 5 anthraquinone derivatives	<i>Rheum</i> spp	BEH C18	50 x 2.1	isocratic : 3	PDA
[42]	2009	Quantitative analysis of formononetin and triterpenoid glycosides	<i>Actaea racemosa</i> L.	BEH C18	100 x 2.1	12.5 / 7	ELSD + MS
[57]	2009	Quantitative analysis of ginkgolides and bilobalides	<i>Ginkgo biloba</i> L.	BEH Shield RP18	50 x 2.1	12.5 / 7	MS
[23]	2010	Quantification of steviol (st) and its glycosides (g); 2 conditions with high temperature of mobile phase	<i>Stevia rebaudiana</i>	HSS C18	150 x 2.1 (g), 100 x 2.1 (st)	? / 4 (g), 4.4 isocr. (st)	MS/MS
[25]	2010	Qualitative analysis of 7 polyphenols in tea samples	<i>Camellia sinensis</i> L.	BEH Shield RP18	50 x 1.0 + 100 x 2.1	? / 0.5 and 7.2	UV + MS/MS
[53]	2010	Quantitative analysis of 29 phenolics in tea infusions or extracts	<i>Camellia sinensis</i> L.	BEH C18	100 x 2.1	29 / 20	PDA
[24]	2011	Quantitative analysis of phytohormones; use of a QuEChERS-based extraction method	<i>Cucurbita pepo</i> , courgette Elena variety	BEH C18	100 x 2.1	6 / 4	MS/MS
[55]	2011	Quantitative determination of 5 anthraquinones; use of ionic liquid-based	<i>Rheum</i> spp	BEH C18	100 x 2.1	? / 33	PDA

ultrasonic/ microwave-assisted extraction (IL-UMAE)							
QC (untargeted)							
[61]	2009	Metabolite fingerprinting for untargeted standardised QC; use of chemometric tools	<i>Angelica acutiloba</i>	BEH C18	150 x 2.1	? / 10.1	TOF-MS
[62]	2010	Metabolic profiling for the evaluation of raw and steamed <i>P. notoginseng</i>	<i>Panax notoginseng</i>	BEH C18	100 x 2.1	10 / 8.3	TOF-MS
[63]	2010	Quality evaluation of the Radix Linderae TCM by fingerprint analysis	<i>Lindera aggregata</i> Sims.	HSS T3	150 x 2.1	? / 37	PDA + MS/MS
Metabolomics or signal/biomarker study							
[56]	2007	Study of bacterial signal molecules in plants	<i>Hordeum vulgare</i> + <i>Pachyrhizus erosus</i>	BEH C18	100 x 2.1	? / 1	PDA
[64]	2008	Metabolomic study of oxylipins induced by wounding	<i>Arabidopsis thaliana</i> L.	BEH C18	150 x 2.1	? / 107	TOF-MS
[65]	2008	Metabolomic study of oxylipins induced by wounding in a two-step strategy	<i>Arabidopsis thaliana</i> L.	BEH C18	50 x 1.0 + 150 x 2.1	? / 10 and 119.7 or 325.8	TOF-MS
[66]	2009	Study of metabolic changes in potato tissues after pulsed electrical field stress	<i>Solanum tuberosum</i> L. cv. Bintje	BEH C18	100 x 2.1	19 / 12.5	TOF-MS
[67]	2009	Signal propagation study in wounded plant leading to jasmonic acid accumulation	<i>Arabidopsis thaliana</i> L.	BEH C18	100 x 2.1	13 / 6	TOF-MS
[68]	2010	Study of the metabolic changes after flower opening	<i>Brunfelsia calycina</i>	BEH C18	100 x 2.1	26 / 22	UV + QTOF-MS

[69]	2011	Study of the lipid metabolism in plant under light or temperature stress	<i>Arabidopsis thaliana</i> L.	C8	150 x 2.1	25 / 17	QTOF-MS
[70]	2011	Monitoring of the dynamic network of benzoxazinoids at the plant-insect interface; 2 methods: (A) metabolomic study, (B) quantification	<i>Zea mays</i>	BEH C18 (A,B)	50 x 1.0 (A), 50 x 2.1 (B)	8.0 / 4.9 (A), 5.0 / 2.9 (B)	TOF-MS (A), QTOF-MS (B)
Profiling / dereplication / screening							
[2]	2009	High resolution profiling optimisation; use of high and low temperature	<i>Arabidopsis thaliana</i> L. + <i>Ginkgo biloba</i> L.	BEH C18	150 x 2.1 + 2x 150 x 2.1	many, up to 240min	TOF-MS
[71]	2009	Identification of 10 alkaloids	3 Lycopodiaceae spp.	BEH C18	100 x 2.1	28 / 12	TOF-MS
[59]	2009	Separation optimisation of a TCM; use of computer target optimisation	<i>Coptis chinensis</i> Franch.	BEH Shield RP18	100 x 2.1	? / 9	PDA
[72]	2010	Study of composition of pine needles in organic acids and antibacterial activity	<i>Pinus massoniana</i> Lamb.	BEH C18	100 x 2.1	isocratic	MS/MS
[73]	2010	Identification of 39 bufadienolides	<i>Bufo bufo gargarizans</i> Cantor	HSS T3	100 x 2.1	? / 30	QTOF-MS
[74]	2011	Fast and comprehensive profiling using a 2D LC (NP x RP) strategy ; UHPLC as the 2nd dimension	<i>Stevia rebaudiana</i>	Zorbax SB C18	30 x 2.1	0.33 / 0.27	PDA
[75]	2011	Identification of 28 triterpenoid saponins	<i>Albizia julibrissin</i> Durazz.	RRHD SB-C18	100 x 2.1	? / 60	PDA + QTOF/MS
[30]	2011	Chemical screening of micro-organisms associated with marine invertebrate, in 96-well plates	<i>Erythropodium caribaeorum</i>	?	?	? / 6	ELSD + IT-MS

[34]	2011	Identification of caffeoylquinic acids and flavonoids based on retention time and PDA spectra	<i>Hemerocallis fulva</i>	BEH C18	100 x 2.1	? / 9	PDA
[76]	2011	Analysis of anthocynins in red wine	Red wine	BEH C18	2x 100 x 2.1	? / 98	PDA + QTOF/MS
Chemotaxonomy							
[77]	2009	Metabolic profiling of Gentiana and Gentianella spp for chemotaxonomic study (flavonoids and xanthones)	Gentianaceae spp.	BEH C18	150 x 2.1	? / 16.1	TOF-MS

^(a) Total analysis time with / without wash and equilibration. Unless specified, all analyses are performed in gradient mode.

^(b) Isocratic mode.

^(c) Abbreviations: UV, single trace UV; PDA, photo diode array detector; Q-MS, simple quadrupole MS; QqQ-MS/MS, triple quadrupole MS in MS/MS mode; TOF-MS, time-of-flight MS; QTOF-MS/MS, quadrupole-time-of-flight MS in MS/MS mode; IT-MSⁿ, ion trap MS in MSⁿ mode; ELSD, evaporative light scattering detector.

4. Fast Non-targeted Analysis, Fingerprinting and Metabolomics

The high-throughput capacities of UHPLC are not only interesting for targeted quantification methods of NPs but also for non-targeted qualitative analyses. Indeed, coupling UHPLC with a mass spectrometer such as a TOFMS provides a sensitive MS detection in full-scan mode with a high acquisition rate, resulting in a powerful analytical platform for the non-targeted metabolite fingerprinting of crude plant extracts or other natural matrices [2, 78]. When performed in a given series of analytical runs, UHPLC-TOF-MS fingerprinting data are reproducible and accurate. The separation of the different features (m/z vs. RT) related to each NP generates ion maps where metabolites are resolved in both the m/z scale and RT dimensions. This approach generates metabolite fingerprints that can be used for various purposes such as extract standardisation based on fingerprinting, metabolomics and chemotaxonomic studies.

In the field of metabolomics, UHPLC-MS has been largely used for biomarker discovery in human and animal samples (see related Chapter 14). In applications related to natural products analysis, several studies have demonstrated the usefulness of UHPLC fingerprinting as a rich and valuable source of analytical data for differential metabolomic studies [11].

4.1. UHPLC-MS for Plant Metabolomics

For differential metabolomic studies, in our group, we use both (i) the high-throughput

capabilities of UHPLC on short columns to acquire rapid UHPLC-TOF-MS fingerprints of numerous replicates and (ii) slow high-resolution profiling on long columns of pool representative samples for the localisation and determination of biomarkers (see below). The high-throughput analysis of many biological replicates improves the reproducibility of the LC-MS detection, allowing large series of samples to be analysed over a short time period and thereby avoiding drift of the MS detection. The increase in the number of biological replicates gives more significant weight to metabolome variations in relation to a given physiological modification *versus* the natural biological variation of the samples. The data mining of such fingerprinting data is thus notably improved [79].

Such a strategy was used to study the stress caused by wounding in the model plant *A. thaliana*. Indeed, wounding is known to mimic the attack of herbivores, and metabolomics has the potential to provide a global picture of all chemical events triggered by this stress for the discovery of new wound biomarkers [67]. In these experiments, UHPLC-MS analyses were performed in a two-step strategy to detect induced metabolites and precisely localise these compounds among the numerous constitutive metabolites from a leaf of *A. thaliana*. In a first step, rapid UHPLC-TOF-MS fingerprints of the isopropanol leaf extracts were acquired on a 50 mm column (50 X 1.0 mm; 1.7 μ m) with a rapid gradient (see Figure II.6A) and were submitted to multivariate analysis.

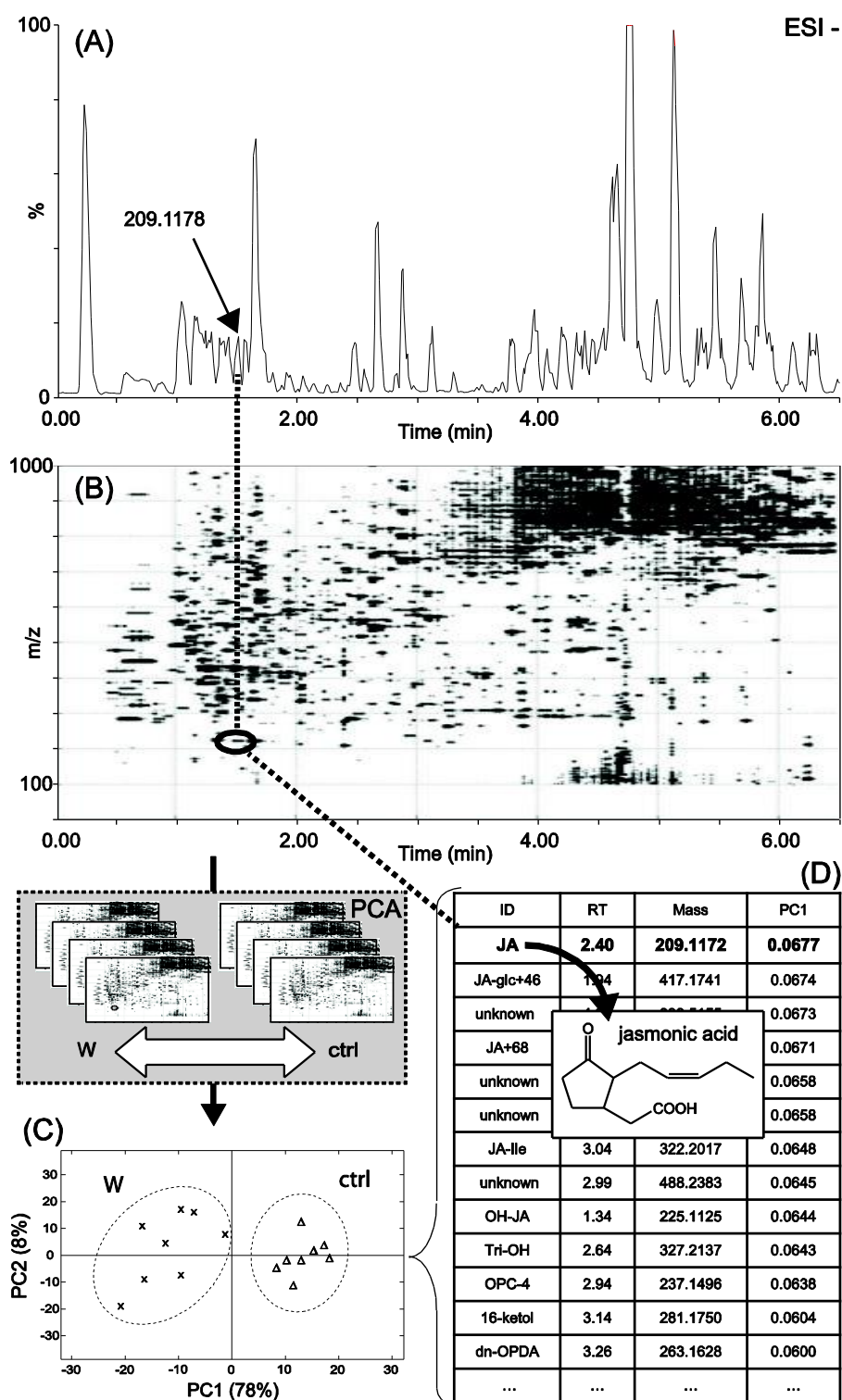


Figure II.6. UHPLC-TOF-MS based metabolomic analysis of the wound response in *A. thaliana*. (A) Metabolite fingerprint of a wounded *A. thaliana* leaf extract using a fast 6.0 min gradient with TOF-MS detection in NI mode. (B) 2D ion map of the fingerprint (m/z vs. RT). (C) PCA score plot of nine wounded (W) and control (ctrl) leaf extracts. (D) Loading of the PCA analysis indicating that jasmonic acid (JA) was the most significant biomarker induced upon wounding of the leaves. All other features (m/z vs. RT) ranked according to the PC1 score are related to additional wound biomarkers.

For each sample, UHPLC-TOF-MS produces a large amount of three-dimensional information (retention RT x m/z x intensity) that can be displayed in the form of two-dimensional (2D) ion maps (see Figures II.6B and II.6C). Pre-processing of the data was required for data mining [65]. In the first step, noise filtering, peak detection and matching were concomitantly performed, making use of both the UHPLC high peak capacity and resolving chromatographic power and the high mass accuracy of TOF-MS detection. Due to the high reproducibility of the data sets obtained, no alignment of the LC-MS was required. After completing the integration parameters, a report of peaks based on areas was generated for each sample and a comprehensive list of the detected components was created. The final data table consisted of retention times and positive or negative m/z data pairs as labels; these data were then exported to perform multivariate analyses.

The data were then used to produce interpretable projections of samples in a reduced dimensionality (score plots) (see Figure II.6D) and to highlight putative biomarkers responsible for the group separation (loading plots) (see Figure II.6E). Statistical methods, such as principal components analysis (PCA), were initially used (see Figure II.6D) and provided an unsupervised data reduction without using class information. Complementary analysis tools and supervised methods were also used for further in-depth investigations of subtle metabolome modifications that occurred at different times after wounding [80].

After PCA, a clear clustering of plant specimens was demonstrated (control vs. wounded plant after 90 min, Figure II.6D), and the highest discriminating ions given by the complete data analysis were selected, leading to the specific detection of discrete-induced ions (m/z values) (see Figure II.6E).

In the generated loading plots, jasmonic acid was found to be the most significant wound biomarker responsible for the PCA clustering between wounded and unwounded plants (see Figure II.6E). Jasmonic acid is a well-known phytohormone involved in the wound response, and its detection by this non-targeted approach validated the model used. The majority of the other biomarkers highlighted in the loading were then characterised either based on the formula that was detected or in a second, confirmatory step.^{57,74,75} For biomarker identification, high-resolution LC profiling was performed on pooled samples by UHPLC-TOF-MS. An example of the type of high resolution profile obtained is illustrated in Figure II.7. This strategy allowed for a precise localising of the putative biological markers induced by wounding through the specific extraction of accurate m/z values. The localised markers could then be isolated using semi-preparative LC after method transfer to allow for their subsequent characterisation by capillary NMR [64].

In addition to our own investigations, the model plant *A. thaliana* has been the topic of many metabolomics studies involving several MS-based metabolomic approaches [11]. Most of these studies are related to the evaluation of responses to different type of biotic or abiotic stresses. For example, nonpolar lipids were efficiently analysed by UHPLC-TOF-MS in *A. thaliana* using a C_8 column at 60 °C with a relatively long (25 min) aqueous gradient with MeOH and isopropanol as organic modifiers. The aim of this study was to analyse the short-term changes in the *A. thaliana* glycerolipidome in response to temperature and light in a time-resolved manner. This UHPLC-TOF-MS lipidomic approach enabled the monitoring of several glycerolipid species that have been reported in *Arabidopsis* leaves. The exposure of these plants to various light and temperature regimes resulted in two major effects. The first effect was the dependence of the saturation level of phosphatidylcholine and

monogalactosyldiacylglycerol pools on light intensity, probably arising from light regulation of *de novo* fatty acid synthesis. The second effect concerns an immediate decrease in unsaturated species of phosphatidylglycerol at high temperatures (32 °C), which could mark the first stages of adaptation to heat stress conditions [69].

Interactions between plant and insects were also studied from a metabolomic perspective. In this case, a comprehensive study of the interaction between feral cabbage (*Brassica oleracea*) and small caterpillars (*Pieris rapae*) was conducted based on a 15 min UHPLC-TOF-MS fingerprinting of the extracts of both participating organisms in this plant–insect

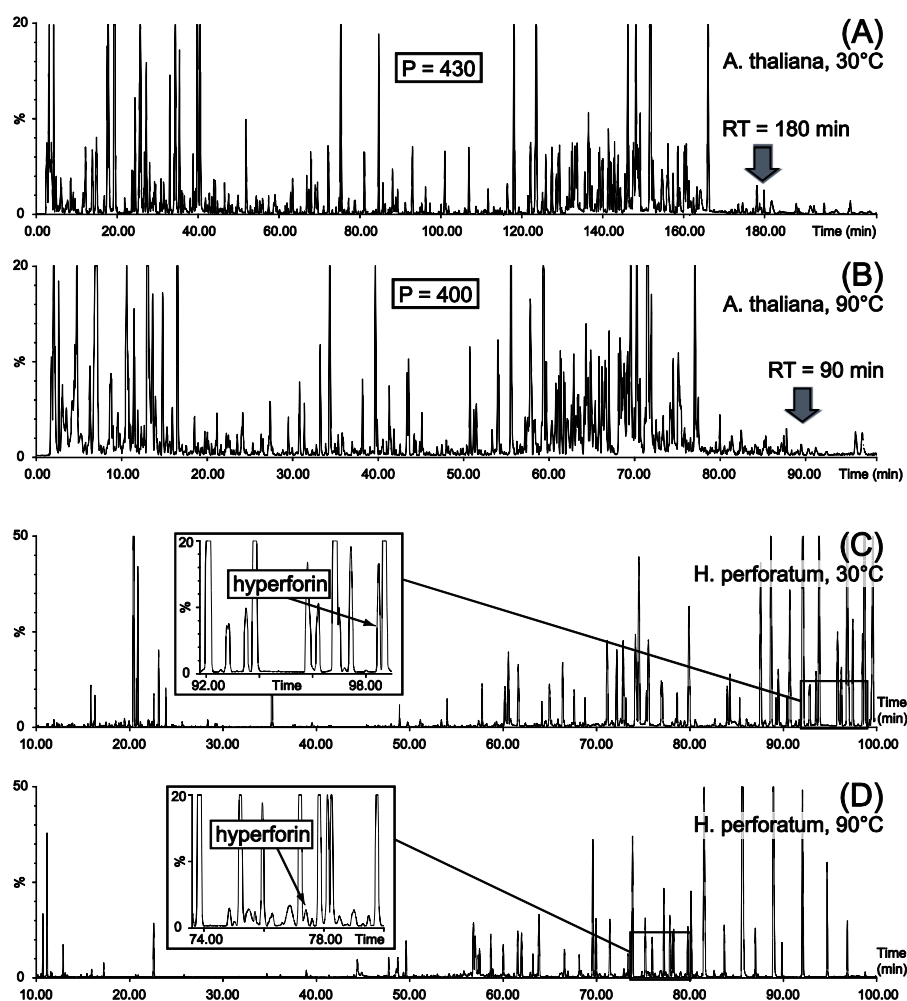


Figure II.7. (A) *A. thaliana* extract analysed on two Acquity BEH C₁₈ columns (150 x 2.1 mm; 1.7 μm) coupled in series at 30 °C with a 240 min gradient. (B) The same separation carried out at 90 °C, with a higher flow rate providing the same pressure with a 120 min gradient. (C) *H. perforatum* extract analysed on an Acquity BEH C₁₈ (150 x 2.1 mm; 1.7 μm) column at 30 °C with a 93 min gradient. (D) The same separation carried out at 90 °C with a higher flow rate providing the same pressure. The selectivity is changed, and the peaks are thinner [see inset in (C) and (D)].

herbivore interaction. The metabolomic results provided more insight into the metabolites that were possibly involved in such interactions, and it was finally concluded that the attack history of the plants affects a specific part of the herbivore's metabolome [81]. Similarly, the plant–insect interface [*Spodoptera* spp. (maize)] was also investigated from a metabolic viewpoint to highlight modifications of bioactive plant secondary metabolites by insect herbivores for understanding animal detoxification processes and plant–insect interactions [70].

The effect of artificial stresses that are known to induce non-thermal permeabilisation of cell membranes, such as those generated by pulsed electric fields (PEFs), were also assessed based on a UHPLC-TOF-MS metabolomic approach on potato tissue. In this case, the UHPLC-TOF-MS fingerprinting data were complemented by gas chromatography (GC) using a GC-TOF-MS system to obtain a more comprehensive survey of the potato metabolites. Clustering analysis showed that 24 h after the application of PEFs, the potato metabolism shows PEF-specific responses characterised by changes in the hexose pool that may involve starch and ascorbic acid degradation [66].

UHPLC-QTOF-MS in complement to headspace-solid phase micro extraction-GC-MS (HS-SPME-GC-MS) and 2D gel electrophoresis was also used for an extensive characterisation of the metabolic changes occurring in *Brunfelsia calycina* petals after the flower's opening. In particular, the anthocyanin degradation products were profiled and characterised based on the UHPLC-MS/MS analyses performed. Globally, this multi-level metabolomic study resulted in the identification of nine main anthocyanins in *Brunfelsia* flowers, 146 up-regulated genes, 19 volatiles, seven proteins and 17 metabolites that increased during anthocyanin degradation, suggesting an induction of the shikimate pathway [68].

In addition to its application in fundamental plant sciences, UHPLC-TOF-MS studies based on metabolomics have been applied to the study of metabolic variations that occur in plants of medicinal value. For example, the effects of the duration of steaming on the metabolome composition of *Panax notoginseng* were monitored by UHPLC-TOF-MS using a 10 min generic water/ACN gradient. A qualitative profiling of multi-parametric metabolic changes of raw *P. notoginseng* during the steaming process was thus obtained. Both the unsupervised and supervised data mining on the fingerprinting results demonstrated strong classification and clear trajectory patterns with regard to the duration of steaming. Using this tool, the minimum duration of steaming for the maximum production of bioactive ginsenosides could be predicted [62]. Such a methodology can be used for fundamental research and for quality assessment for commercial preparations. For other recent applications, see Table II.1.

4.2. UHPLC-MS/MS-based Targeted Metabolomics

Metabolomics studies can also be performed in a semi-targeted manner by UHPLC-MS/MS when hundreds of previously selected constituents are included in the data sets. This metabolomics methodology has been established to quantify hundreds of targeted plant metabolites by MRM in a high-throughput manner in 14 plant accessions from Brassicaceae, Gramineae and Fabaceae. As mentioned, the use of MRM after high-throughput UHPLC separation is a well-established method for the targeted analysis of specific NPs. In this study [82], however, the inclusion of a high number of metabolites provides a rich data set that can be investigated with similar data mining tools as those used for non-targeted metabolomics. Thus, approximately 100 metabolites were quantified in each of the plant extracts investigated, and five transitions were monitored in each 3 min UHPLC

gradient run. A hierarchical cluster analysis based on the metabolite accumulation patterns clearly showed differences among the plant families, and family-specific metabolites could be predicted using a batch-learning, self-organising map analysis. Such an automated, widely targeted metabolomics approach represents an interesting alternative method for elucidating metabolite accumulation patterns in plants. It also represents an elegant way to combine the high-throughput potential of UHPLC to the performance of MS/MS for quantitation and appropriate data mining to achieve a comprehensive evaluation of the results obtained.

4.3. UHPLC Fingerprinting for QC

All of the examples previously described demonstrate the potential of UHPLC-TOF-MS-based metabolite fingerprinting to obtain a fast overview of an extract metabolic content. Thus, QC procedures are progressively adopting such strategies for identification, categorisation or standardisation purposes. UHPLC is especially useful in these situations because plant extracts are complex and consist, among other things, of numerous metabolites acting synergistically that could not be accurately considered separately. TCM preparations, which often consist of several herbs, require even more extensive rational approaches. Moreover, the identification of plants based on fingerprints is more valuable than identification based on one or few constituents (targeted analysis) [83]. NMR metabolomics [84] is well established for this type of global metabolite fingerprinting, but only

provides the detection of the main NPs in a given extract. UHPLC-MS is now also starting to be more extensively used for detailed composition comparisons. For example, the QC in commercial preparations of angelica roots (*Angelica acutiloba*) was performed by comparison of high-throughput (10 min gradient) PI and NI UHPLC-TOF-MS fingerprinting using chemometric tools. Partitioning of root samples was effectively achieved by PCA, showing that the cultivation area was one of the most significant parameters for quality determination. This method proved to be an efficient and rapid QC method that can be used on a routine basis [61]. For other recent applications, see Table II.1.

4.4. Chemotaxonomic Studies

As mentioned above, fingerprinting can be used as a chemotaxonomic tool to discriminate plant species based on their secondary metabolite composition. The *Gentiana* and *Gentianella* genera were distinguished among the Gentianaceae family based on their UHPLC-TOF-MS fingerprints. Separations were carried out on a UHPLC column (150 x 2.1 mm; 1.7 μ m) in 15 min with a 5–55% aqueous/ACN gradient. The fingerprints of three *Gentianella* species were strikingly similar. In contrast, fingerprints of the *Gentiana* species were very different from those of *Gentianella* species and from each other. Several compounds were determined as unique to each genus and, therefore, could be used as biomarkers. This result was helpful for an unambiguous classification of plants belonging to these genera [77]. Another study enabled the classification of different Brazilian species of the *Lippia* genus [85].

5. High-resolution Profiling and Metabolite Identification

In most of the fingerprinting studies discussed above, the UHPLC conditions were mainly optimised for a high-throughput comparison of many crude extract replicates for data mining with either very fast separation or with fingerprinting methods, providing peak capacities equivalent to conventional HPLC methods.

UHPLC can also be used to extend the achievable resolution for the separation of NPs in crude extracts with complex compositions. In this case, the goal is to provide the best possible separation of closely related NPs, which often occur as positional isomers or diastereoisomers. Furthermore, the high chromatographic resolution of UHPLC reduced the ionisation suppression problems that often occur with electrospray ionisation in UHPLC-MS; thus, a better detection of minor constituents is also obtained. When coupled with MS or PDA, the high peak purity obtained by this method provides a better deconvolution of the MS or UV spectra recorded for online structural determination or dereplication purposes [65].

However, the price of this enhancement in the quality of data is a longer analysis time. With similar run times as in conventional HPLC profiling methods, an enhancement of peak capacity of approximately a factor of 3 can be expected with UHPLC systems [50]. This improvement in resolution also depends on the column length and the number of theoretical plates and optimised conditions obtained. Because the profiling of crude extracts metabolites is generally performed in gradient

mode and because peak capacity is related both to the plate number and to the column dead time, the improvement in peak capacity is not dependent just on the column length. Thus, an optimum for the column length and the gradient time has to be found. An accepted compromise is that a 150 mm, 1.7 μm column should be preferentially selected for gradient lengths up to 60 min at 30 °C, while the columns coupled in series (3 x 150 mm, 1.7 μm) are attractive only for a gradient time higher than 250 min [50].

The enhancement of peak capacity that can be obtained on crude plant extracts has been well demonstrated by the chromatograms obtained for the extract of *G. biloba* (see Figures 13.2A and 13.2B). In this case, the gradient time was kept similar to the original HPLC method, and the UHPLC columns and conditions were optimised to achieve the maximum peak capacity.

In the case of UHPLC-TOF-MS coupling, this enhancement of chromatographic resolution also provides a much more detailed localisation of the different NPs that constitute the metabolome of a given organism. The ion maps obtained with a high-resolution profiling method transferred from the fast fingerprinting method were used for the metabolomic study of the wound response in *A. thaliana* (see the corresponding high-resolution profile in Figure II.7A). With such a high peak capacity measurement, several isomers were well separated, and, for example, for this plant, a peak at m/z 225 that appeared as a single wound biomarker in the metabolomic study based on the rapid fingerprinting was

found to correspond to four different isomers that could be resolved using a high-resolution metabolite profiling method [64].

5.1. Very High-resolution Profiling

In order to push forward the quest for high resolution, it is also possible to increase the column length of the UHPLC column and increase the peak capacity by using gradient times exceeding 60 min [50]. In the case of the metabolite profiling of *A. thaliana*, the use of two 150 mm columns coupled in a series provides an increase of 40% in peak capacity as compared to the separation obtained from one 150 mm column. The gradient transfer on this 300 mm column was, however, performed in 240 min, as compared to 60 min on the 150 mm column [2].

One possible way to decrease this very long gradient time is to perform the separation at high temperature. Indeed, most of the new hybrid silica-based columns are stable at high temperatures, rendering this type of analysis possible. In the case of *A. thaliana*, the gradient time could be reduced from 240 to 120 min while maintaining approximately the same high peak capacity when the separation was performed at 90 °C instead of 30 °C [2]. On the 300 mm column, the maximum flow rate at 30 °C was 200 mL min⁻¹, but this could be increased to 350 mL min⁻¹ at higher temperatures. Figures 13.6(A) and 13.6(B) show metabolite profiles obtained at 30 °C and 90 °C, respectively. The baseline separation of more than 300 metabolites could be practically achieved by this means. The potential degradation of NPs during separation was examined, but no apparent degradation was observed for even the longest separations at 90 °C [2].

As described for short gradients, to a lesser extent, however, a non-negligible loss of

resolution may occur due to the extra-column volume and is related to the type of detector used. This parameter has to be taken into account in addition to the gradient time and column length optimisation in order to improve the performance of the analytical platform. The source of some MS detectors, for example, may generate a loss of more than 20% in peak capacity as compared to UV detectors [2].

The use of high temperatures provides a significant increase in throughput; however, temperature modifications also affect the polarity of the mobile phase and the selectivity of the separation. For the separation of non-polar NPs, the use of HT-UHPLC can represent an advantage, and compounds that would be difficult to elute from C₁₈ columns even with a high percentage of organic solvent may elute much faster in these conditions. For example, this result was seen for hyperforine, a non-polar phloroglucinol derivative found in the standardised extract of *Hypericum perforatum* that is involved in the antidepressant effect of this phytopharmaceutical. As shown in Figures 13.6(C) and 13.6(D), the use of high temperature significantly affects the selectivity and above all the retention for the different *Hypericum* constituents. Hyperforine was found to elute at 78 min at 90 °C, while it did not elute until 98 min at 30 °C.

5.2. LC x LC for Improved Resolution

For complex plant extracts, the use of reversed-phase separation alone, even with very high peak capacity, might not be sufficient for the separation of all metabolites in a single profiling analysis. In this case, the use of an orthogonal separation using a column with different phase chemistry might be needed. In this respect, a very recent 2D LC application based on UHPLC has been described for separating all of the

components of interest contained in *Stevia rebaudiana*, a plant from Paraguay that is currently used worldwide as a sweetener. For the profiling of this plant, neither RP-HPLC nor NP-HPLC alone has been capable of separating all of the components of interest. A combination of 2D LC (LC x LC) for the profiling of this extract was used. The first dimension used a classical polyamine HPLC (250 x 1.0 mm; 5 μm) column in normal-phase mode at ambient temperature with a 100 min gradient at 20 mL min^{-1} . UHPLC was employed for the fast second dimension: the eluate was divided into fractions by a 20 μL loop and then injected online in triplicate in the second dimension. This second dimension consisted of a Zorbax RRHD SB-C₁₈ UHPLC column (30 x 2.1 mm; 1.7 μm) operating in reversed-phase mode at 70 °C in a fast gradient of 20 s (with re-equilibration) at 3.4 mL min^{-1} . This high flow rate allowed a very short wash and reequilibration times. Thanks to the high throughput of the UHPLC separation, the reduced cycle time allowed 3 to 12 samplings for each peak eluted by the first dimension. Polyphenolic and stevioside compounds were thus efficiently identified by combining the information coming from the position of the compounds in the 2D plot and the UV spectra with that of reference materials [74].

5.3. Metabolite Identification and Dereplication

The high-resolution profiling of UHPLC provides a good separation of NPs in complex mixtures. This baseline separation of analytes is important for quantification if a simple detector, such as UV or ELSD, is used. The deconvolution of LC peaks based on chromatography is also important for recording online UV-PDA and MS spectra of good quality for facilitating the dereplication process. An example of the type of online spectral information that can be obtained for the dereplication of natural products is illustrated in Figure II.8 for the profiling of *Viola tricolor*, a

herbal drug used traditionally for its anti-inflammatory properties. The extract of *V. tricolor* was analysed with an optimised gradient on a BEH C₁₈ column (150 x 2.1 mm; 1.7 μm) in 60 min. As shown, a good separation of most of the metabolites was obtained, and minor compounds were also detected in both PI and NI modes thanks to the sensitivity of the TOF-MS system. The TOF-MS detection provided exact molecular weights (< 5 ppm) and retention time information for all of the compounds detected. Different successive filters were applied to extract and ascertain molecular formulas in order to reduce the number of structural possibilities. This filtering is derived from heuristic rules: (1) restrictions for the number of elements; (2) LEWIS and SENIOR chemical rules; (3) isotopic patterns; (4) hydrogen/carbon ratios; (5) element ratio of nitrogen, oxygen, phosphorus and sulphur versus carbon; and (6) element ratio probabilities [86]. For the most abundant unknown compounds, PDA-UV spectra were recorded and used as a complement to the MS data in the dereplication process. Chemotaxonomic information was then also added for the final selection of putative structures. Based on these structural hypotheses, correlations between retention time, lipophilicity and elution behaviour in a series of related compounds were performed.

As shown in Figure II.8A, the peak at RT 11.33 min displayed a molecular ion at m/z 577.1543 $[\text{M}-\text{H}]^-$ in NI mode and at m/z 579.1741 NI $[\text{M}+\text{H}]^+$ in PI mode (Figure II.8C). This information confirmed that the MW was 578 Da. With a 15 ppm tolerance in NI mode, this exact online mass determination gave five possible formulas (Figure II.8B). The application of the heuristic filter confirmed that the only valid possibility for was C₂₇H₃₀O₁₄. A cross-search of this formula with chemotaxonomic information on the *Viola* genus found in a NP database [87] revealed that the only possible hit corresponded to violanthin (see Figure II.8C). This peak annotation was confirmed

by the UV-PDA spectrum that was recorded online, confirming that both maxima matched well with this flavonoid diglycoside, which could be efficiently dereplicated by this means. The

same procedure was applied to all LC peaks that were efficiently separated by this high-resolution profiling [88].

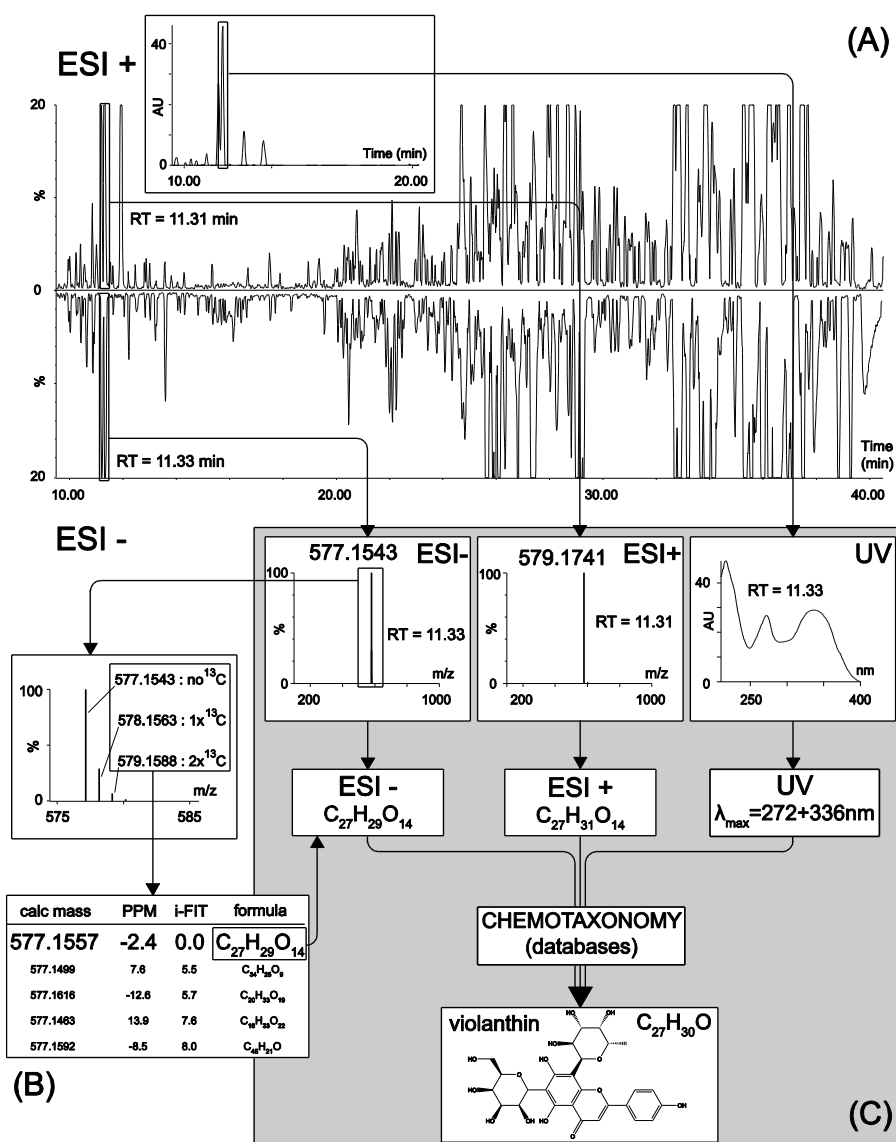


Figure II.8. Peak annotation procedure for dereplication based on a high resolution *V. tricolor* profiling on a C_{18} UHPLC column (150 x 2.1 mm; 1.7 μ m). (A) PI (upper trace) and NI (lower trace) ESI-TOF-MS BPI chromatograms obtained with a slow gradient (5–95% ACN in 50 min). A UV trace (366 nm) is displayed in the inset. (B) Putative molecular formulas assignment based on the precision (ppm) and isotopic pattern (iFIT) obtained from the NI ESI-TOF-MS spectrum of the LC peak at RT 11.33 min. (C) Annotation of the LC peak at RT 11.33 min based on PI and NI molecular formula assignment and the UV PDA spectrum. Final structural assignment based on a cross search with chemotaxonomic information.

6. Conclusions

As shown here, UHPLC presents many advantages for analysing NPs in complex biological matrices, such as crude plant extracts. Indeed, for all of the examples discussed, the efficiency of UHPLC either in terms of its high-throughput (QC, fingerprinting) or in terms of its high-resolution (dereplication, profiling) is very advantageous as compared to classical HPLC. Also, because the diameter of the column used is smaller, there is a significant reduction in solvent and sample consumption. For metabolomics, this technique provides clear advantages in terms of its reproducibility, resolution and throughput, yielding data that could not be attained by conventional HPLC methods in practically achievable analysis times. Such characteristics are essential for the satisfactory comparison of fingerprints with data mining methods.

In NPs research, compounds often must be isolated for either *de novo* structure determination or for bioactivity assessment. In this respect, it is still difficult to find semi-preparative columns with similar phase chemistries as those developed for UHPLC, which might hinder the possibility of performing efficient gradient transfers for all types of applications. However, this problem will likely be solved when the technique spreads more widely to research groups involved in NP research. The number of applications of UHPLC in crude extract analysis is still scarce compared to HPLC, probably due to the required dedicated instrumentation that is needed to work with the high pressures generated by the use of columns with sub-2 μm particles.

The development of highly efficient, sub-2 μm columns has also stimulated the development of other columns that share similar characteristics and are compatible with a conventional HPLC system. An example of this is columns with core-shell particles consisting of a 1.7 μm solid core surrounded by a 0.5 μm porous silica shell, as discussed in Chapter 5. This type of particle shares a similar chromatographic performance to the sub-2 μm particles; however, their backpressure is much lower [6]. These types of columns have already been used with success for profiling crude extracts [89, 90]. Core-shell particles can thus be a good low-pressure alternative to columns packed with the sub-2 μm particles for the separation of complex mixtures with only a small sacrifice in peak efficiency.

For detection and dereplication, an MS analyser is the optimal detector to be coupled with UHPLC. Although TOF-MS detectors have an adequate acquisition frequency to cope with the LC peak width obtained by UHPLC, this type of MS detector still requires improvement to achieve faster acquisition rates. Indeed, an ideal system should be able to provide both MS and MS/MS spectra in a single run at high resolution and in both PI and NI mode simultaneously. However, this process is a very demanding one for a MS analyser and further improvements are expected to appear in the coming years.

UHPLC thus represents a very valuable tool for NP chemists. With the increasing requirements for QC, profiling and fingerprinting, dereplication and metabolomics, it is very likely that UHPLC will gradually replace most of the HPLC applications developed for NP research in years to come.

Acknowledgments

The authors are indebted to the Swiss National Foundation for supporting their plant metabolomic studies (grant no. 205320-124667/1 to J.-L. W.).

References

- [1] D. Guillarme, J. Ruta, S. Rudaz, J.L. Veuthey. New trends in fast and high-resolution liquid chromatography: a critical comparison of existing approaches. *Analytical and Bioanalytical Chemistry*, **2010**. 397: 1069-1082.
- [2] E. Grata, D. Guillarme, G. Glauser, J. Boccard, P.A. Carrupt, J.L. Veuthey, S. Rudaz, J.L. Wolfender. Metabolite profiling of plant extracts by ultra-high-pressure liquid chromatography at elevated temperature coupled to time-of-flight mass spectrometry. *Journal of Chromatography A*, **2009**. 1216: 5660-5668.
- [3] K. Cabrera. Applications of silica-based monolithic HPLC columns. *Journal of Separation Science*, **2004**. 27: 843-852.
- [4] J.R. Mazzeo, U.D. Neue, M. Kele, R.S. Plumb. A new separation technique takes advantage of sub-2- μm porous particles. *Analytical Chemistry*, **2005**. 77: 460A-467A.
- [5] J. Ruta, D. Guillarme, S. Rudaz, J.L. Veuthey. Comparison of columns packed with porous sub-2 μm particles and superficially porous sub-3 μm particles for peptide analysis at ambient and high temperature. *Journal of Separation Science*, **2010**. 33: 2465-2477.
- [6] J.M. Cunliffe, T.D. Maloney. Fused-core particle technology as an alternative to sub-2- μm particles to achieve high separation efficiency with low backpressure. *Journal of Separation Science*, **2007**. 30: 3104-3109.
- [7] J. Ruta, D. Zurlino, C. Grivel, S. Heinisch, J.-L. Veuthey, D. Guillarme. Evaluation of columns packed with shell particles with compounds of pharmaceutical interest. *Journal of Chromatography A*, **2012**. 1228: 221-231.
- [8] D.V. McCalley. Some practical comparisons of the efficiency and overloading behaviour of sub-2 μm porous and sub-3 μm shell particles in reversed-phase liquid chromatography. *Journal of Chromatography A*, **2011**. 1218: 2887-2897.
- [9] C.M. Dobson. Chemical space and biology. *Nature*, **2004**. 432: 824-828.
- [10] J. Hong. Role of natural product diversity in chemical biology. *Current Opinion in Chemical Biology*, **2011**. 15: 350-354.
- [11] J.L. Wolfender, G. Glauser, J. Boccard, S. Rudaz. MS-based Plant Metabolomic Approaches for Biomarker Discovery. *Natural Product Communications*, **2009**. 4: 1417-1430.
- [12] J. Larsson, J. Gottfries, S. Muresan, A. Backlund. ChemGPS-NP: Tuned for navigation in biologically relevant chemical space. *Journal of Natural Products*, **2007**. 70: 789-794.
- [13] D.G.I. Kingston. High-Performance Liquid-Chromatography of Natural-Products. *Journal of Natural Products*, **1979**. 42: 237-260.
- [14] T.K. Natishan. Recent developments of achiral HPLC methods in pharmaceuticals using various detection modes. *Journal of Liquid Chromatography & Related Technologies*, **2004**. 27: 1237-1316.
- [15] K. Hostettmann, A. Marston, M. Hostettmann. *Preparative Chromatography Techniques: Applications in Natural Product Isolation*. 2nd ed. **1997**, Berlin, Springer.
- [16] J.L. Wolfender. HPLC in Natural Product Analysis: The Detection Issue. *Planta Medica*, **2009**. 75: 719-734.
- [17] D.T.T. Nguyen, D. Guillarme, S. Rudaz, J.L. Veuthey. Fast analysis in liquid chromatography using small particle size and high pressure. *Journal of Separation Science*, **2006**. 29: 1836-1848.
- [18] F. David, G. Vanhoenacker, B. Tienpont, I. Francois, P. Sandra. Coupling columns and multidimensional configurations to increase peak capacity in liquid chromatography. *Lc Gc Europe*, **2007**. 20: 154-158.

- [19] P.J. Eugster, D. Guillaume, S. Rudaz, J.L. Veuthey, P.A. Carrupt, J.L. Wolfender. Ultra High Pressure Liquid Chromatography for Crude Plant Extract Profiling. *Journal of AOAC International*, **2011**. 94: 51-70.
- [20] M.I. Churchwell, N.C. Twaddle, L.R. Meeker, D.R. Doerge. Improving LC-MS sensitivity through increases in chromatographic performance: Comparisons of UPLC-ES/MS/MS to HPLC-ES/MS/MS. *Journal of Chromatography B*, **2005**. 825: 134-143.
- [21] J.-L. Wolfender, G. Marti, E.F. Queiroz. Advances in Techniques for Profiling Crude Extracts and for the Rapid Identification of Natural Products: Dereplication, Quality Control and Metabolomics. *Current Organic Chemistry*, **2010**. 14: 1808-1832.
- [22] K.W. Cheng, F. Cheng, M. Wang. Liquid chromatography-mass spectrometry in natural product research, in *Bioactive Natural Products: Detection, Isolation, and Structural Determination, Second Edition*, S.M. Colegate and R.J. Molyneux, Editors. **2008**, CRC press: London. p. 245-266.
- [23] C. Gardana, M. Scaglianti, P. Simonetti. Evaluation of steviol and its glycosides in *Stevia rebaudiana* leaves and commercial sweetener by ultra-high-performance liquid chromatography-mass spectrometry. *Journal of Chromatography A*, **2010**. 1217: 1463-1470.
- [24] M.I. Alarcon Flores, R. Romero-Gonzalez, A. Garrido Frenich, J.L. Martinez Vidal. QuEChERS-based extraction procedure for multifamily analysis of phytohormones in vegetables by UHPLC-MS/MS. *Journal of Separation Science*, **2011**. 34: 1517-1524.
- [25] D. Guillaume, C. Casetta, C. Bicchi, J.L. Veuthey. High throughput qualitative analysis of polyphenols in tea samples by ultra-high pressure liquid chromatography coupled to UV and mass spectrometry detectors. *Journal of Chromatography A*, **2010**. 1217: 6882-6890.
- [26] J. Schappler, R. Nicoli, D. Nguyen, S. Rudaz, J.L. Veuthey, D. Guillaume. Coupling ultra high-pressure liquid chromatography with single quadrupole mass spectrometry for the analysis of a complex drug mixture. *Talanta*, **2009**. 78: 377-387.
- [27] D. Guillaume, J. Schappler, S. Rudaz, J.-L. Veuthey. Coupling ultra-high-pressure liquid chromatography with mass spectrometry. *TrAC, Trends in Analytical Chemistry*, **2010**. 29: 15-27.
- [28] A. Pelander, P. Decker, C. Baessmann, I. Ojanpera. Evaluation of a High Resolving Power Time-of-Flight Mass Spectrometer for Drug Analysis in Terms of Resolving Power and Acquisition Rate. *Journal of the American Society for Mass Spectrometry*, **2011**. 22: 379-385.
- [29] J.J.J. van der Hooft, J. Vervoort, R.J. Bino, J. Beekwilder, R.C.H. de Vos. Polyphenol Identification Based on Systematic and Robust High-Resolution Accurate Mass Spectrometry Fragmentation. *Analytical Chemistry*, **2011**. 83: 409-416.
- [30] F. Berrue, S.T. Withers, B. Haltli, J. Withers, R.G. Kerr. Chemical Screening Method for the Rapid Identification of Microbial Sources of Marine Invertebrate-Associated Metabolites. *Marine Drugs*, **2011**. 9: 369-381.
- [31] J.V. Olsen, J.C. Schwartz, J. Griep-Raming, M.L. Nielsen, E. Damoc, E. Denisov, O. Lange, P. Remes, D. Taylor, M. Splendore, E.R. Wouters, M. Senko, A. Makarov, M. Mann, S. Horning. A Dual Pressure Linear Ion Trap Orbitrap Instrument with Very High Sequencing Speed. *Molecular & Cellular Proteomics*, **2009**. 8: 2759-2769.
- [32] T. Koecher, R. Swart, K. Mechtler. Ultra-High-Pressure RPLC Hyphenated to an LTQ-Orbitrap Velos Reveals a Linear Relation between Peak Capacity and Number of Identified Peptides. *Analytical Chemistry*, **2011**. 83: 2699-2704.
- [33] Y.O. Tsybin, L. Fornelli, A.N. Kozhinov, A. Vorobyev, S.M. Miladinovic. High-Resolution and Tandem Mass Spectrometry - the Indispensable Tools of the XXI century. *Chimia*, **2011**. 65: 641-645.
- [34] Y.-L. Lin, C.-K. Lu, Y.-J. Huang, H.-J. Chen. Antioxidative Caffeoylquinic Acids and Flavonoids from *Hemerocallis fulva* Flowers. *Journal of Agricultural and Food Chemistry*, **2011**. 59: 8789-8795.
- [35] R.P.W. Scott. *Liquid chromatography Detectors*. Chrom-Ed Book Series. **2003**: Library4Science, <http://www.library4science.com/>.

- [36] X.F. Gao, M. Dan, A.H. Zhao, G.X. Xie, W. Jia. Simultaneous determination of saponins in flower buds of *Panax notoginseng* using high performance liquid chromatography. *Biomedical Chromatography*, **2008**. 22: 244-249.
- [37] S.A. Ozkan. LC with electrochemical detection. recent application to pharmaceuticals and biological fluids. *Chromatographia*, **2007**. 66: S3-S13.
- [38] T.O. Larsen, M.A.E. Hansen. Dereplication and discovery of natural products by UV spectroscopy, in *Bioactive Natural Products: Detection, Isolation, and Structural Determination, Second Edition*, S.M. Colegate and R.J. Molyneux, Editors. **2008**, CRC press: London. p. 221-244.
- [39] N.C. Megoulas, M.A. Koupparis. Twenty years of evaporative light scattering detection. *Critical Reviews in Analytical Chemistry*, **2005**. 35: 301-316.
- [40] N. Vervoort, D. Daemen, G. Torok. Performance evaluation of evaporative light scattering detection and charged aerosol detection in reversed phase liquid chromatography. *Journal of Chromatography A*, **2008**. 1189: 92-100.
- [41] R. Russo, D. Guillarme, S. Rudaz, C. Bicchi, J.-L. Veuthey. Evaluation of the coupling between ultra performance liquid chromatography and evaporative light scattering detector for selected phytochemical applications. *Journal of Separation Science*, **2008**. 31: 2377-2387.
- [42] B. Avula, Y.H. Wang, T.J. Smillie, I.A. Khan. Quantitative Determination of Triterpenoids and Formononetin in Rhizomes of Black Cohosh (*Actaea racemosa*) and Dietary Supplements by Using UPLC-UV/ELS Detection and Identification by UPLC-MS. *Planta Medica*, **2009**. 75: 381-386.
- [43] X.M. Liang, Y. Jin, Y.P. Wang, G.W. Jin, Q. Fu, Y.S. Xiao. Qualitative and quantitative analysis in quality control of traditional Chinese medicines. *Journal of Chromatography A*, **2009**. 1216: 2033-2044.
- [44] G. Lang, N.A. Mayhudin, M.I. Mitova, L. Sun, S. van der Sar, J.W. Blunt, A.L.J. Cole, G. Ellis, H. Laatsch, M.H.G. Munro. Evolving trends in the dereplication of natural product extracts: New methodology for rapid, small-scale investigation of natural product extracts. *Journal of Natural Products*, **2008**. 71: 1595-1599.
- [45] K. Hostettmann, J.-L. Wolfender, C. Terreaux. Modern screening techniques for plant extracts. *Pharmaceutical biology*, **2001**. 39: 18-32.
- [46] L.W. Sumner, P. Mendes, R.A. Dixon. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry*, **2003**. 62: 817-836.
- [47] O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R.N. Trethewey, L. Willmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, **2000**. 18: 1157-1161.
- [48] W.B. Dunn. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, **2008**. 5: 011001.
- [49] Y.Z. Liang, P.S. Xie, K. Chan. Quality control of herbal medicines. *Journal of Chromatography B*, **2004**. 812: 53-70.
- [50] D. Guillarme, E. Grata, G. Glauser, J.L. Wolfender, J.L. Veuthey, S. Rudaz. Some solutions to obtain very efficient separations in isocratic and gradient modes using small particles size and ultra-high pressure. *Journal of Chromatography A*, **2009**. 1216: 3232-3243.
- [51] J.B. Calixto. Efficacy, safety, quality control, marketing and regulatory guidelines for herbal medicines (phytotherapeutic agents). *Brazilian Journal of Medical and Biological Research*, **2000**. 33: 179-189.
- [52] S. Sang, J.D. Lambert, C.-T. Ho, C.S. Yang. The chemistry and biotransformation of tea constituents. *Pharmacological Research*, **2011**. 64: 87-99.
- [53] L. Novakova, Z. Spacil, M. Seifrtova, L. Opletal, P. Solich. Rapid qualitative and quantitative ultra high performance liquid chromatography method for simultaneous analysis of twenty nine common phenolic compounds of various structures. *Talanta*, **2010**. 80: 1970-1979.
- [54] J.B. Wang, H.F. Li, C. Jin, Y. Qu, X.H. Xiao. Development and validation of a UPLC method for quality control of rhubarb-based medicine: Fast simultaneous determination of five anthraquinone derivatives. *Journal of Pharmaceutical and Biomedical Analysis*, **2008**. 47: 765-770.

- [55] C.X. Lu, H.X. Wang, W.P. Lv, C.Y. Ma, P. Xu, J. Zhu, J. Xie, B. Liu, Q.L. Zhou. Ionic Liquid-Based Ultrasonic/Microwave-Assisted Extraction Combined with UPLC for the Determination of Anthraquinones in Rhubarb. *Chromatographia*, **2011**. 74: 139-144.
- [56] C. Gotz, A. Fekete, I. Gebefuegi, S.T. Forczek, K. Fuksova, X. Li, M. Englmann, M. Gryndler, A. Hartmann, M. Matucha, P. Schmitt-Kopplin, P. Schroder. Uptake, degradation and chiral discrimination of N-acyl-D/L-homoserine lactones by barley (*Hordeum vulgare*) and yam bean (*Pachyrhizus erosus*) plants. *Analytical and Bioanalytical Chemistry*, **2007**. 389: 1447-1457.
- [57] B. Avula, Y.H. Wang, T.J. Smillie, I.A. Khan. Column Liquid Chromatography/Electrospray Ionization-Time of Flight-Mass Spectrometry and Ultraperformance Column Liquid Chromatography/Mass Spectrometry Methods for the Determination of Ginkgolides and Bilobalide in the Leaves of *Ginkgo biloba* and Dietary Supplements. *Journal of AOAC International*, **2009**. 92: 645-652.
- [58] J.H. Chen, F.M. Wang, J. Liu, F.S.C. Lee, X.R. Wang, H.H. Yang. Analysis of alkaloids in *Coptis chinensis* Franch by accelerated solvent extraction combined with ultra performance liquid chromatographic analysis with photodiode array and tandem mass spectrometry detections. *Analytica Chimica Acta*, **2008**. 613: 184-195.
- [59] G.W. Jin, X.Y. Xue, F.F. Zhang, Y. Jin, X.M. Liang. Computer-aided target optimization for traditional Chinese medicine by ultra-performance liquid chromatography. *Talanta*, **2009**. 78: 278-283.
- [60] P. Li, G. Xu, S.P. Li, Y.T. Wang, T.P. Fan, Q.S. Zhao, Q.W. Zhang. Optimizing ultra performance liquid chromatographic analysis of 10 diterpenoid compounds in *Salvia miltiorrhiza* using central composite design. *Journal of Agricultural and Food Chemistry*, **2008**. 56: 1164-1171.
- [61] S. Tianniam, T. Bamba, E. Fukusaki. Non-targeted metabolite fingerprinting of oriental folk medicine *Angelica acutiloba* roots by ultra performance liquid chromatography time-of-flight mass spectrometry. *Journal of Separation Science*, **2009**. 32: 2233-2244.
- [62] D.-F. Toh, L.-S. New, H.-L. Koh, E.C.-Y. Chan. Ultra-high performance liquid chromatography/time-of-flight mass spectrometry (UHPLC/TOFMS) for time-dependent profiling of raw and steamed *Panax notoginseng*. *Journal of Pharmaceutical and Biomedical Analysis*, **2010**. 52: 43-50.
- [63] Y.-j. Wu, Y.-l. Zheng, L.-j. Luan, X.-s. Liu, Z. Han, Y.-p. Ren, L.-s. Gan, C.-x. Zhou. Development of the fingerprint for the quality of Radix *Linderae* through ultra-pressure liquid chromatography-photodiode array detection/electrospray ionization mass spectrometry. *Journal of Separation Science*, **2010**. 33: 2734-2742.
- [64] G. Glauser, D. Guillarme, E. Grata, J. Boccard, A. Thiocone, P.-A. Carrupt, J.-L. Veuthey, S. Rudaz, J.-L. Wolfender. Optimized liquid chromatography-mass spectrometry approach for the isolation of minor stress biomarkers in plant extracts and their identification by capillary nuclear magnetic resonance. *Journal of Chromatography A*, **2008**. 1180: 90-98.
- [65] E. Grata, J. Boccard, D. Guillarme, G. Glauser, P.A. Carrupt, E.E. Farmer, J.L. Wolfender, S. Rudaz. UPLC-TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in *Arabidopsis thaliana*. *Journal of Chromatography B*, **2008**. 871: 261-270.
- [66] F.G. Galindo, P. Dejmek, K. Lundgren, A.G. Rasmusson, A. Vicente, T. Moritz. Metabolomic evaluation of pulsed electric field-induced stress on potato tissue. *Planta*, **2009**. 230: 469-479.
- [67] G. Glauser, L. Dubugnon, S.A.R. Mousavi, S. Rudaz, J.-L. Wolfender, E.E. Farmer. Velocity Estimates for Signal Propagation Leading to Systemic Jasmonic Acid Accumulation in Wounded *Arabidopsis*. *Journal of Biological Chemistry*, **2009**. 284: 34506-34513.
- [68] A. Bar-Akiva, R. Ovadia, I. Rogachev, C. Bar-Or, E. Bar, Z. Freiman, A. Nissim-Levi, N. Gollop, E. Lewinsohn, A. Aharoni, D. Weiss, H. Koltai, M. Oren-Shamir. Metabolic networking in *Brunfelsia calycina* petals after flower opening. *Journal of Experimental Botany*, **2010**. 61: 1393-1403.
- [69] A. Burgos, J. Szymanski, B. Seiwert, T. Degenkolbe, M.A. Hannah, P. Giavalisco, L. Willmitzer. Analysis of short-term changes in the *Arabidopsis thaliana* glycerolipidome in response to temperature and light. *Plant Journal*, **2011**. 66: 656-668.

- [70] G. Glauser, G. Marti, N. Villard, G.A. Doyen, J.-L. Wolfender, T.C.J. Turlings, M. Erb. Induction and detoxification of maize 1,4-benzoxazin-3-ones by insect herbivores. *The Plant Journal*, **2011**. 68: 901-911.
- [71] R. Ho, N. Marsousi, P. Eugster, J.P. Bianchini, P. Raharivelomanana. Detection by UPLC/ESI-TOF-MS of Alkaloids in Three Lycopodiaceae Species from French Polynesia and Their Anticholinesterase Activity. *Natural Product Communications*, **2009**. 4: 1349-1352.
- [72] S. Feng, W. Zeng, F. Luo, J. Zhao, Z. Yang, Q. Sun. Antibacterial Activity of Organic Acids in Aqueous Extracts from Pine Needles (*Pinus massoniana* Lamb.). *Food Science and Biotechnology*, **2010**. 19: 35-41.
- [73] Y. Liu, Y. Xiao, X. Xue, X. Zhang, X. Liang. Systematic screening and characterization of novel bufadienolides from toad skin using ultra-performance liquid chromatography/electrospray ionization quadrupole time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2010**. 24: 667-678.
- [74] F. Cacciola, P. Delmonte, K. Jaworska, P. Dugo, L. Mondello, J.I. Rader. Employing ultra high pressure liquid chromatography as the second dimension in a comprehensive two-dimensional system for analysis of *Stevia rebaudiana* extracts. *Journal of Chromatography A*, **2011**. 1218: 2012-2018.
- [75] L. Han, G. Pan, Y. Wang, X. Song, X. Gao, B. Ma, L. Kang. Rapid profiling and identification of triterpenoid saponins in crude extracts from *Albizia julibrissin* Durazz. by ultra high-performance liquid chromatography coupled with electrospray ionization quadrupole time-of-flight tandem mass spectrometry. *Journal of Pharmaceutical and Biomedical Analysis*, **2011**. 55: 994-1009.
- [76] A. de Villiers, D. Cabooter, F. Lynen, G. Desmet, P. Sandra. High-efficiency high performance liquid chromatographic analysis of red wine anthocyanins. *Journal of Chromatography A*, **2011**. 1218: 4660-4670.
- [77] A. Urbain, A. Marston, E. Marsden-Edwards, K. Hostettmann. Ultra-performance Liquid Chromatography/Time-of-flight Mass Spectrometry as a Chemotaxonomic Tool for the Analysis of Gentianaceae Species. *Phytochemical Analysis*, **2009**. 20: 134-138.
- [78] I.D. Wilson, J.K. Nicholson, J. Castro-Perez, J.H. Granger, K.A. Johnson, B.W. Smith, R.S. Plumb. High resolution "Ultra performance" liquid chromatography coupled to oa-TOF mass spectrometry as a tool for differential metabolic pathway profiling in functional genomic studies. *Journal of Proteome Research*, **2005**. 4: 591-598.
- [79] J. Boccard, J.L. Veuthey, S. Rudaz. Knowledge discovery in metabolomics: An overview of MS data handling. *Journal of Separation Science*, **2010**. 33: 290-304.
- [80] J. Boccard, A. Kalousis, M. Hilario, P. Lanteri, M. Hanafi, G. Mazerolles, J.-L. Wolfender, P.-A. Carrupt, S. Rudaz. Standard machine learning algorithms applied to UPLC-TOF/MS metabolic fingerprinting for the discovery of wound biomarkers in *Arabidopsis thaliana*. *Chemometrics and Intelligent Laboratory Systems*, **2010**. 104: 20-27.
- [81] J.J. Jansen, J.W. Allwood, E. Marsden-Edwards, W.H. van der Putten, R. Goodacre, N.M. van Dam. Metabolomic analysis of the interaction between plants and herbivores. *Metabolomics*, **2009**. 5: 150-161.
- [82] Y. Sawada, K. Akiyama, A. Sakata, A. Kuwahara, H. Otsuki, T. Sakurai, K. Saito, M.Y. Hirai. Widely Targeted Metabolomics Based on Large-Scale MS/MS Data for Elucidating Metabolite Accumulation Patterns in Plants. *Plant and Cell Physiology*, **2009**. 50: 37-47.
- [83] P.S. Xie, S.B. Chen, Y.Z. Liang, X.H. Wang, R.T. Tian, R. Upton. Chromatographic fingerprint analysis - a rational approach for quality assessment of traditional Chinese herbal medicine. *Journal of Chromatography A*, **2006**. 1112: 171-180.
- [84] F. van der Kooy, F. Maltese, Y. Hae Choi, H. Kyong Kim, R. Verpoorte. Quality Control of Herbal Material and Phytopharmaceuticals with MS and NMR Based Metabolic Fingerprinting. *Planta Medica*, **2009**. 75: 763-775.
- [85] P.J. Eugster, C. Funari, F. Mattioli, G. Durigan, S. Martel, P. Carrupt, D. Silva, J.-L. Wolfender. Combination of LC retention, high resolution TOF-MS information and web database search as dereplication tools in a chemotaxonomic study of *Lippia* spp. *Planta Med*, **2011**. 77: PA49.

- [86] T. Kind, O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **2007**. 8: 105-124.
- [87] J. Buckingham. *Dictionary of Natural Products on DVD*, version 19:1, **2008**, Chapman & Hall/CRC.
- [88] J.L. Wolfender, P.J. Eugster, D. Guillarme, H. Kratou, G. Glauser, S. Martel, S. Rudaz, P.A. Carrupt. Potential of UHPLC for crude plant extract analysis: profiling, dereplication and metabolomics. *Planta Medica*, **2010**. 76: 1178-1178.
- [89] P.L. Yang, G.R. Litwinski, M. Pursch, T. McCabe, K. Kuppannan. Separation of natural product using columns packed with Fused-Core particles. *Journal of Separation Science*, **2009**. 32: 1816-1822.
- [90] N. Manchon, M. D'Arrigo, A. Garcia-Lafuente, E. Guillamon, A. Villares, A. Ramos, J.A. Martinez, M.A. Rostagno. Fast analysis of isoflavones by high-performance liquid chromatography using a column packed with fused-core particles. *Talanta*, **2010**. 82: 1986-1994.

Chapter III - Optimisation of UHPLC Resolution

This chapter is based on an article published in Journal of Chromatography A, and is the result of a collaboration with Atheris Laboratories, Plan-les-Ouates, Switzerland.

Foreword

The previous chapter presented the fundamentals and applications of UHPLC technology in the field of NP analysis. This chapter specifically focuses on the optimisation of chromatographic parameters for high resolution profiling of complex natural extracts. Indeed, several chromatographic parameters are known to have a strong influence on UHPLC separation [1]. Among these parameters, the analyte size has been extensively investigated and optimal conditions were found for the separation of mixtures of small (200 – 800 Da) and large (1 – 5 kDa) molecular weight molecules. These optimal conditions represent a compromise between high resolution and reasonable analysis time. Finally, this study provided practical rules for achieving the highest

peak capacity in defined limits of gradient time, to be used in high resolution profiling of complex natural extracts.

Instead of using an artificial mixture of NPs as a test sample, two natural samples were used as models in this study, namely an extract of *Hypericum perforatum* L. containing low molecular weight secondary metabolites (200-800 Da) and the venom of *Conus consors* mainly composed of high molecular weight molecules, mainly peptides (1 to 5 kDa), to study the effect of the nature of the analytes on the separation.

Hypericum perforatum L. (Figure III.1), also



Figure III.1. *Hypericum perforatum* L. whose flower extract was used as a model mixture in the experiments mentioned below. Photo: Prof. J.-L. Wolfender.

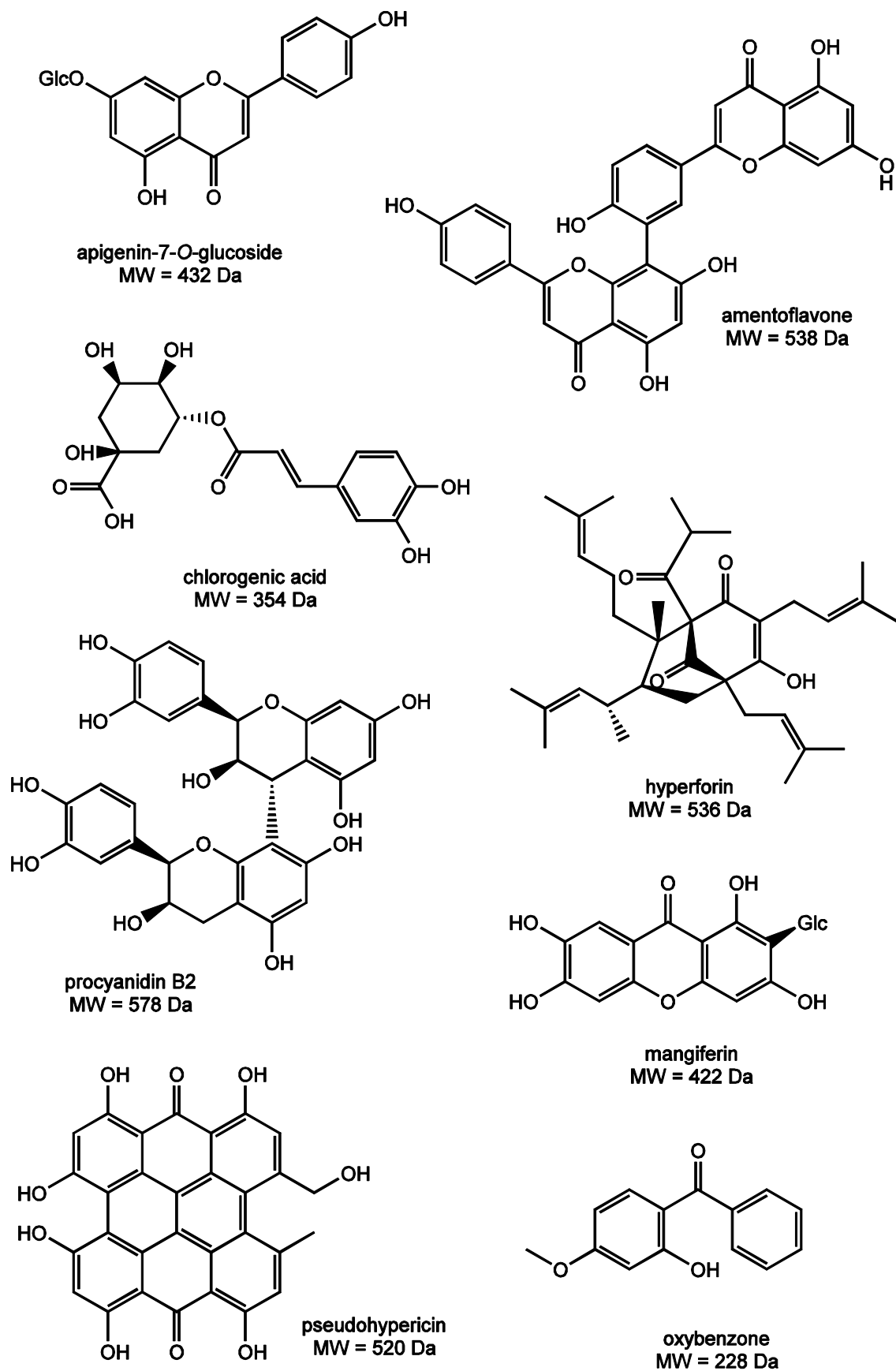


Figure III.2. Structures of some NPs of small molecular weight present in *Hypericum perforatum* extract.

known as St John's wort, is a yellow-flowering herbaceous perennial plant from the Hypericaceae family, present in Europe, Northern Africa, Asia and in the USA. Its flowering tops extract is mainly used for its antidepressant activity [2] and is widely commercialised for this purpose in Switzerland and in other countries. The main constituents of this extract are flavonoids, phenolic acids, naphthodianthrones and other polycyclic compounds (see Figure III.2 for some examples of NPs present in *Hypericum perforatum*). The compounds responsible for its mechanism of action as antidepressant are still unknown [2]. Different studies however indicate that hyperforine (Figure III.2), a phloroglucinol derivative, is active *in vitro* on the recapture of various neurotransmitters [3].

The cone snail *Conus consors* is a predatory marine gastropod from the Conidae family found in the Indo-pacific region [4]. Its shell possesses a conical shape and is typically 5 to 12 cm long (Figure III.3). The members of the *Conus* genus (700 species) developed venoms for hunting and for defence purposes. These venoms are complex mixtures of biologically active compounds that mainly consist of small disulfide-rich peptides (1-5 kDa) [5]. As an example, Figure III.4 displays the glycopeptide CcTx, a conopeptide isolated from the venom of *Conus consors* that elicits excitotoxic responses in the prey by acting on voltage-gated sodium channels. Such venom molecules have been modified and optimised during millions of years to provide specific biological activities, inducing paralysis, sleep or depression for example [6],



Figure III.3. *Conus consors* shell whose venom was used as a model mixture in the experiments mentioned below.

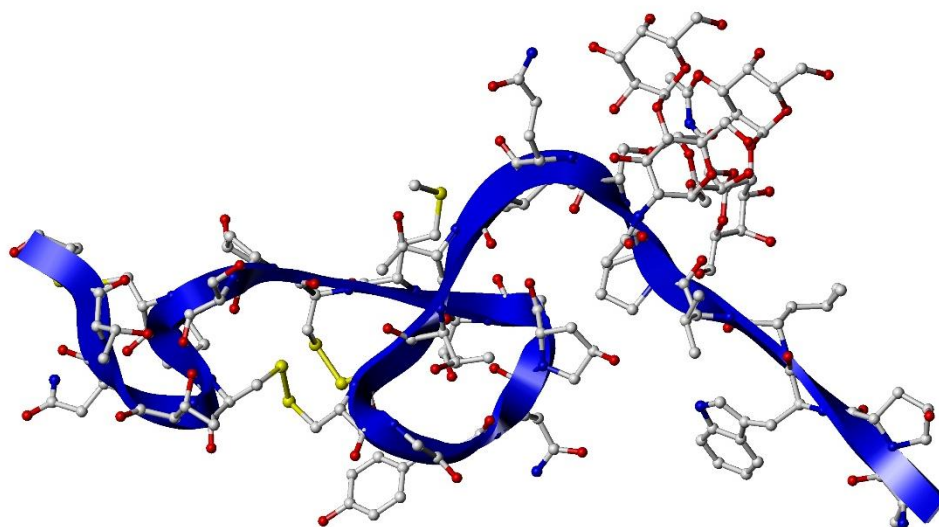


Figure III.4. The glycopeptide CcTx (4.1 kDa), isolated from the venom of *Conus consors*. This conopeptide belongs to the kA-family and elicits excitotoxic responses in the prey by acting on voltage-gated sodium channels. Adapted from [7] with the help of Dr Alessandra Nurisso.

and represent thus a sustainable source of molecules to address a wide range of targets [8]. Many venom compounds are used as research tools and even provided approved drugs, such as ziconotide (Prialt®), a neuronal calcium channel blocker that is derived from the conopeptide omega-MVIIA from *Conus magus* used to treat chronic pain [9]. The high number of molecules in a given venom and the high number of venomous animal species makes these complex mixtures highly interesting. There is therefore a great interest in the development of analytical tools for the separation and detection of the constituents of these venoms.

This article is the result of a collaboration with Atheris laboratories, a company based in Geneva (Plan-les-Ouates) that is specialised in mass spectrometry and bioinformatics with a focus on peptides and proteins. In this work, Atheris came with its expertise on peptide analysis and *Conus consors*, while our lab brought its knowledge on high resolution separations and fundamental chromatographic aspects. It resulted in a comprehensive paper that investigates theoretical and experimental aspects of the separation of complex matrices and provides practical solutions for their high resolution analyses.

Peak capacity optimisation for high resolution peptide profiling in complex mixtures by liquid chromatography coupled to time-of-flight mass spectrometry: Application to the *Conus consors* cone snail venom

Philippe J. Eugster

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Daniel Biass

Atheris Laboratories, Geneva, Switzerland

Davy Guillarme

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Philippe Favreau

Atheris Laboratories, Geneva, Switzerland

Reto Stöcklin

Atheris Laboratories, Geneva, Switzerland

Jean-Luc Wolfender

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Research article published in *Journal of Chromatography A*, 1259 (2012) 187– 199.

Available online 14 May 2012

Abstract

The high resolution profiling of complex mixtures is indispensable for obtaining online structural information on the highest possible number of the analytes present. This is particularly relevant for natural extracts, as for the venom of the predatory marine snail *Conus consors*, which contains numerous bioactive peptides with molecular masses ranging between 1000 and 5000 Da. The goal of the present work was to maximise peak capacity of peptides separations by LC–MS while maintaining a reasonable analysis time. The best gradient performance using the *C. consors* venom as a real sample was obtained with a mobile phase flow rate as high as possible to maximise performance in the gradient mode, and gradient time comprised between 75 and 350 min when using a 150 mm column length. The present study also confirmed that an elevated temperature (up to 90 °C) improves performance under ultra-high pressure liquid chromatography (UHPLC) conditions. However, the thermal stability of the analytes had to be critically evaluated. For the profiling of *C. consors*, analyte degradation was not clearly observable at 90 °C with analysis times of approximately 100 min. Finally, the MS source was found to cause significant additional band broadening in the UHPLC mode (σ_{ext}^2 was 10–24 times higher using TOF-MS vs. UV detection). Thus, if the MS contributes strongly to the peak capacity loss, classical 2.1 mm I.D. columns can be replaced by 3.0 mm I.D. to mitigate this problem. Based on these considerations, the optimal generic profiling conditions applied to the *C. consors* venom provided a peak capacity higher than

1100 for a gradient time of around 100 min, doubling the values reached by classical HPLC separation. UHPLC-QTOF-MS/MS experiments carried out in these conditions provided exploitable data that matched with peptides present in the *C. consors* venom. These optimal LC conditions are thus compatible with online peptide deconvolution and matching against transcriptomic data and, to some extent, *de novo* sequencing in such complex mixtures.

1. Introduction

The high resolution LC–MS profiling of complex mixtures is essential to rapidly generate detailed online structural information about the highest possible number of analytes present [10]. Indeed, very efficient chromatographic separation improves the quality of the MS spectra and of automated MS/MS survey analyses, while lowering the ion suppression effects that are often encountered in complex mixture analyses.

The improvement of chromatographic performance is particularly relevant for integrated analytical ‘omics’ approaches (e.g., metabolomics [11, 12], peptidomics or venomics [13, 14]). However, most of the reported studies have mainly focused on the mass spectrometric dimension, and only a few studies were dedicated to the careful optimisation of complex mixture separation.

In this respect, our group has extensively studied the influence of various chromatographic parameters on the high resolution profiling of crude plant extracts containing small molecules (molecular mass (MM) < 1000 Da) [1]. Based on previous findings and on the fundamentals of chromatography, the optimisation of profiling conditions was investigated for larger analytes (peptides with MM ranging from 1000 to 5000 Da) in complex mixtures. This work is particularly significant in light of recent findings that increased peptide separation efficiency produced an improvement in the number of peptides successfully identified by mass spectrometric methods [15].

Because our research interests focus on pharmacologically relevant natural products of

various origins, venoms containing a complex mixture of bioactive peptides were chosen as model mixtures for this study. Indeed, each venomous species (of which there are approximately 200,000) possesses a cocktail of more than 100 bioactive components, sometimes more than 1000, bringing the potential number of bioactive compounds that can be found in venoms to more than 10 million [8, 16, 17].

In this context, the venom of the marine snail *Conus consors* was studied. The venoms of cone snails have been intensively studied during the past decades and are of great interest due to the complexity of their original compounds, the conopeptides. These compounds are very often cysteine-rich peptides, which have diverse structures that provide them high specificity, potency and robustness. A conopeptide of *Conus magus* (the magician cone), omega-MVIIA, has led to the development of Ziconotide (Prialt®), a neuronal calcium channel blocker used to treat chronic pain [9].

The *C. consors* venom has been reported to contain a complex array of peptides [17]. The goal in studying this venom is to find the best possible chromatographic separation techniques in order to systematise the automated generation of high-quality MS/MS data for all bioactive peptides. Indeed, in a given venom, all peptides may have strong pharmacological effects and need to be characterised [6, 18]. In this paper, various chromatographic conditions were applied to *C. consors* venom to assess their effects on peptide separation. For comparative purposes, this set of conditions was also applied in parallel to a model

plant extract (*Hypericum perforatum*) containing small molecules spread over a large polarity range. Plant extract profiling and venom separation present certain similarities: (1) both aim at separating and detecting hundreds of constituents present in either tiny or large relative amounts and determining their different structures and physicochemical properties and (2) both require high resolution separations, usually based on a single but long generic gradient [19].

Recently, a linear relationship was reported between the peak capacity and the number of peptides identified by a MS/MS instrument coupled to a nano UHPLC (ultra-high pressure liquid chromatography) [15]. On the other hand, peak capacity optimisation was the topic of various in-depth studies on the separation of small analytes such as pharmaceuticals [20, 21] and complex plant extracts [1, 10, 22]. Because

the differences between small molecules and peptides are likely to be linked to the S parameter in the different equations used in gradient chromatography, as well as to the diffusivity (D_m) of the analytes [23], the influence of these factors are discussed in this paper.

For this purpose, different chromatographic parameters were studied in a systematic way, including the mobile phase temperature, gradient time and column geometry, both in conventional HPLC and in UHPLC conditions and using UV and time-of-flight mass spectrometry (TOF-MS) detection. To maintain generic conditions, gradient slope, mobile phase composition and stationary phase chemistry were intentionally kept constant. A significant correlation between the experimental and calculated peak capacity values was demonstrated.

The S parameter represents the slope of the relationship between the logarithm of the retention factor and the solvent composition [30].

2. Experimental

2.1. Experimental design

To determine the optimum chromatographic conditions, peak capacity values of real samples profiling were compared, varying different chromatographic parameters gathered in eight representative HPLC and UHPLC conditions (lines 1–8 of Table III.1). Based on this, the correlation between peak capacity and the number of resolved analytes was evaluated, and the coherence between experimental and calculated peak capacities assessed. The influence of the *S* parameter and diffusivity on peak capacity was experimentally evaluated based on the constituents of both venom and plant extracts. The effect of the flow rate, temperature, gradient time, column geometry and particle diameter on either the calculated or the experimental peak capacity was then investigated. Finally, because the MS detector could also strongly impact peak capacity values, the in-source dispersion was measured on the UHPLC-TOF-MS platform used in this work. Both the experimental and calculated results eventually enabled the establishment of practical generic rules for the optimal separation on LC–MS platforms.

As mentioned above, eight conditions were chosen, representing various common combinations of these parameters for standard profiling (conditions 1–8 in Table III.1). Standard particle diameters of 3.5 μm (for HPLC) and 1.7 μm (for UHPLC) were chosen. The flow rate was set to provide either 30% or 90% of the maximal backpressure recommended by the column manufacturer (*i.e.*, 400 bar for the HPLC column

and 1000 bar for the UHPLC column). Three mobile phase temperatures were considered: 30, 60 and 90 °C. The HPLC analyses using the 3.5 μm particles were not carried out at 90 °C due to the excessive flow rate for the electrospray (ESI) source, even with the fixed T-split. Detection was carried out with both an ESI-TOF-MS analyser and a UV-PDA detector to consider the additional peak dispersion that occurs in the ESI-TOF-MS device. In all cases, the re-equilibrating times and injection volumes were adjusted according to the column dead volume. Other parameters were fixed: a generic 1%/min gradient slope was used on a 150 mm C_{18} column with acetonitrile (ACN) as an organic modifier for all experiments. Because elution strength could change at 90 °C, a gradient with a final eluent composition equal to 70% of B and a slope of 0.73%/min was tested (condition 8 of Table III.1), providing the same retention time for the last peak of interest, as in condition 5 (30 °C).

2.2. Chemicals

Methylparaben, ethylparaben, propylparaben, butylparaben and uracil were provided by Sigma–Fluka (Buchs, Switzerland). Water, ACN and formic acid (FA) were of ULC/MS grade from Biosolve (Valkenswaard, The Netherlands).

2.3. Sample preparation

Two samples were prepared: a venom sample of *C. consors* (*Conus* venom) and a standardised extract of *H. perforatum* (*Hypericum* extract).

Table III.1. Investigated chromatographic conditions.

Conditions	Column geometry [mm x mm, μm]	Temp. [$^{\circ}\text{C}$]	ΔP at 30% ACN [bar]	% of max. tolerated pressure	Flow rate [$\mu\text{L}/\text{min}$]	Gradient	Reconditioning time [min]	Injected volume [μL]
1	Xbridge 150x3.0, 3.5	30	120	30	440	2-95%	17.0	2.0
2	Xbridge 150x3.0, 3.5	30	350	90	880	2-95%	9.0	2.0
3	Xbridge 150x3.0, 3.5	60	350	90	1460	2-95%	5.0	2.0
4	Acquity 150x2.1, 1.7	30	300	30	130	2-95%	28.0	1.0
5	Acquity 150x2.1, 1.7	30	900	90	380	2-95%	9.0	1.0
6	Acquity 150x2.1, 1.7	60	900	90	620	2-95%	6.0	1.0
7	Acquity 150x2.1, 1.7	90	900	90	810	2-95%	4.0	1.0
8	Acquity 150x2.1, 1.7	90	900	90	810	2-70%	4.0	1.0
9	Acquity 150x3.0, 1.7	30	900	90	776	2-95%	9.0	2.0
10	Acquity 150x3.0, 1.7	60	900	90	1110	2-95%	8.0	2.0
11	Acquity 150x3.0, 1.7	90	850	85	1450	2-95%	5.0	2.0

All specimens of *C. consors* used for this study were collected from one colony in the Chesterfield Islands (New Caledonia) as part of the CONFIELD scientific expeditions conducted in July and November 2008. The crude venoms (dissected venoms) were obtained after dissection of 11 *C. consors* specimens following a previously described method [24] and were then lyophilised. Aliquots of 0.2 mg of each of the 11 dissected venom samples were reconstituted at 1 mg/mL (protein content) in acidified water (0.1% TFA) and desalted using solid-phase extraction onto a Sep-Pak Light cartridge (130 mg C₁₈ phase) equilibrated in acidified water according to the manufacturer's instructions (Waters, Milford, MA, USA). Elution was performed with 70% ACN in acidified water. MALDI analyses were performed on the eluates to assess the venom quality prior to pooling (data not shown). Finally, the eluates were pooled and freeze-dried under vacuum in a SpeedVac concentrator (Thermo-Savant, Holbrook, NY, USA) and then stored at -80 °C. The milked venom (corresponding to the venom injected by the cone snail in its prey) was obtained from a pool of 20 milkings as previously described [13], corresponding to approximately 0.5 mg dry weight. No desalting step was undertaken for the milked venom samples. For MS/MS deconvolution purposes, samples were reduced with DTT and TCEP (following standard reduction protocols).

A standardised *H. perforatum* extract was obtained from Indena (Milan, Italy). This extract was dissolved in 85% MeOH at a final concentration of 5 mg/mL.

2.4. Instrumentation and analytical conditions

UHPLC analyses were performed on a Waters Acquity UPLC system able to withstand pressures up to 1000 bar and equipped with a column oven

able to heat samples up to 90 °C. Separations were carried out on an XBridge C₁₈ column (150 mm x 3.0 mm I.D., 3.5 µm) and on two Acquity UPLC BEH C₁₈ columns (150 mm x 2.1 mm I.D., 1.7 µm and 150 mm x 3.0 mm I.D., 1.7 µm). The mobile phase consisted of water + 0.1% FA (solvent A) and ACN + 0.1% FA (solvent B) and was used in gradient mode. The generic gradient profile consisted of an isocratic step at 2% B for 5 min, followed by a 93 min gradient from 2 to 95% B and a final isocratic step at 95% B for 2 min. Table III.1 summarises the column geometries, mobile phase temperatures, flow rates, gradient spans, reconditioning times, and injection volumes that were employed in this study.

The sample manager was thermostated at 10 °C, and the partial loop mode was used with a 10 µL injection loop. The standard Acquity PDA module was used for online UV detection in the 210–400 nm range, with a resolution of 2.4 nm, a sampling rate of 10 spectra/s and a filter response set to 0.

The UHPLC system was coupled with a Waters Micromass LCT Premier time-of-flight mass spectrometer (TOF-MS) equipped with its standard electrospray interface (ESI), using a tubing of around 75 cm length and 127 µm I.D. Because of the high mobile phase flow rates, a T-split was set with 2/5 of the eluent directed towards the MS. Analyses were operated in W positive mode with centroid data acquisition and a scan time of 0.3 s using dynamic range enhancement (DRE) and with a solution of leucin–encephalin infused through the lockspray source. Calibration of the instrument was achieved using a formate solution in the 400–1800 *m/z* range. The capillary voltage, sample cone voltage and aperture 1 voltage were set to 2800 V, 40 V, and 15 V, respectively. The source temperature, desolvation temperature, cone gas flow and desolvation gas flow were set to 120 °C, 300 °C, 800 L/h and 20 L/h, respectively. The raw data were acquired and processed with MassLynx 4.1 software from Waters.

The final experiments carried out with a UHPLC-QTOF-MS platform were performed on a Waters Acquity UPLC system with the parameters described in condition 5 of Table III.1, coupled with a Synapt G2 QTOF from Waters and equipped with an electrospray interface. Data were acquired in survey mode in the m/z range 400–2000 using a scan time of 0.5 s and a threshold of 650 cps. When ion intensity exceeded this threshold, the instrument automatically switched to MS/MS mode for 10 scans of 0.2 s each. MS/MS data were acquired using a collision energy ramp of 25–40 eV over an m/z range of 100–1500 Da. An exclusion window of 30 s was selected. For MS/MS sequencing, data were first deconvoluted with the MaxEnt 3 module of MassLynx 4.1 (Waters, Milford, MA, USA) and then matched to nucleotide or protein databases with Phenyx software (Genebio, Geneva, Switzerland).

2.5. S value determination

The S value for each type of analyte (peptide or small molecule) was calculated with HPLC optimisation software (Osiris 4.1.1.2, from Datalys, Grenoble, France). The determination was performed based on two 5–95% B gradient runs of 58 and 165 min for peptides, and of 20 and 60 min for small molecules. The average value for each type of compound, namely $S_{peptide}$ and $S_{small\ molecule}$, was used in the following equations.

2.6. System and column characterisation

2.6.1. Column dead volume (v_0) determination

The column dead volume was measured for each column based on the elution time of an unretained compound, namely uracil, after subtraction of the v_{ext} . For this purpose, 0.5 μ L of

uracil solution (0.5 mg/mL in 50% ACN) was injected in triplicate using a flow rate of 0.25 mL/min of 50% ACN, with UV detection at 265 nm. The column dead volume of the Acquity BEH C₁₈, 150 mm x 2.1 mm I.D., 1.7 μ m was found to be equal to 375 μ L, while that of the Acquity BEH C₁₈, 150 mm x 3.0 mm I.D., 1.7 μ m was measured as 763 μ L and that of the XBridge C₁₈ 150 mm x 3.0 mm I.D., 3.5 μ m was found to be equal to 588 μ L.

2.6.2. Determination of N_{col} and N_{obs}

As shown in Equation III.1, the column efficiency (N_{col}) can be calculated using the column length (l), particle size (d_p) and reduced height equivalent to a theoretical plate (h). For the BEH C₁₈ stationary phase employed in the present study, a h_{opt} value of 2.8 was previously reported [25]. However, this value is not valid for all of the conditions reported in Table III.1 because the linear velocity can be far above the optimum of the Knox curve. For this reason, N_{col} was estimated using the A, B and C parameters of the Knox equation (previously determined by our group) and the D_m values of the compounds calculated using the Wilke–Chang equation [26, 27]. The average compound molecular masses for the D_m calculation were 2400 and 500 Da for the peptides of the *Conus venom* and the small molecules of the *Hypericum* extract, respectively, while the viscosity used in this calculation was the highest value obtained during the gradient for each temperature. Thus, it is evident that there are a number of assumptions (A, B and C can vary from column to column, and the calculation of D_m with the Wilke–Chang equation can be quite imprecise) that cause the confidence interval for N_{col} to be quite large.

$$N_{col} = \frac{l}{h \times d_p} \quad (\text{Equation III.1})$$

This N_{col} value is the maximal plate number obtained with the column, considering a negligible contribution to broadening from the system. In reality, the observed plate number (N_{obs}) is lower than the N_{col} because of the additional band broadening caused by the chromatographic system. N_{obs} can thus be calculated with the following equation:

$$N_{obs} = \frac{N_{col}}{1 + \left(\frac{\sigma_{ext}^2}{\sigma_{col}^2}\right)} \quad (\text{Equation III.2})$$

where σ_{ext}^2 and σ_{col}^2 are the extra-column and column dispersions.

2.6.3. Extra-column volume (v_{ext}) and extra-column dispersion σ_{ext}^2 measurement

The extra-column volume of the UHPLC system, v_{ext} , was measured with both UV and MS detection modes. The v_{ext} was obtained by measuring the elution time of uracil at 0.5 mg/mL in 50% ACN without a column at 5 different flow rates (0.1, 0.2, 0.4, 0.8 and 1.5 mL/min). The isocratic mode with 50% ACN was selected at room temperature, with the UV detector set at 265 nm and the TOF-MS analyser operating in negative ionisation mode. Finally, v_{ext} was calculated as the slope of a plot of elution time vs. $1/F$, where F is the flow rate ($\mu\text{L}/\text{min}$). The values obtained were 18.6 and 92.2 μL for UV and MS detection, respectively.

The extra-column dispersion (σ_{ext}^2) is usually determined by injecting a 0.5 mg/mL solution of uracil without a column (replaced by a zero dead volume union) in isocratic mode with 50% ACN at a given flow rate. The σ_{ext}^2 is obtained with:

$$\sigma_{ext}^2 = \frac{(W_{50} \times F)^2}{5.54} \quad (\text{Equation III.3})$$

where W_{50} is the peak width at 50% of its height (min) and F is the flow rate ($\mu\text{L}/\text{min}$). Because the extra-column dispersion depends on both the flow rate and the system, this measurement has to be repeated for each investigated flow rate with both UV and MS detection.

In this work, this typical approach was not employed because the UV-PDA acquisition began with a 10 s delay (GPIB connection for UV-PDA instead of LAN), and thus, uracil was eluted before the acquisition started, in the absence of a column. Thus, σ_{ext}^2 was experimentally obtained by injecting a mixture of methyl-, ethyl-, propyl- and butylparaben (20 $\mu\text{g}/\text{mL}$ each in water) in isocratic mode with 40% ACN for each flow rate, using a short UHPLC column (Acquity BEH C₁₈, 30 mm x 2.1 mm I.D., 1.7 μm). According to the procedure described by Kok et al., the experimental total dispersion ($\sigma^2 = (W_{50} \times F)^2/5.54$) was plotted as a function of t_R^2 of the parabens; the intercept of this straight line represents the σ_{ext}^2 , while the slope corresponds to the column efficiency [28]. This method provides acceptable values for the present study. The σ_{ext}^2 value was measured for both UV and MS detection modes and for the eight first flow rates indicated in Table III.2.

Table III.2. Extra-column dispersion (σ^2_{ext}) values measured for chromatographic conditions 1–8 of Table III.1 using UV or TOF-MS detection, measured by the method described in Section 2.6.3.

Conditions	UV [μL^2]	MS [μL^2]
1	5.3	59.8
2	4.0	70.6
3	3.6	88.0
4	1.8	31.7
5	5.9	65.3
6	3.7	61.8
7	3.4	56.7
8	3.4	56.7

2.6.4. Column dispersion (σ^2_{col}) measurement

The dispersion due to the chromatographic column itself (σ^2_{col}) was obtained using Equation III.4:

$$\sigma_{\text{col}}^2 = \frac{(V_0 \times (1+k))^2}{N_{\text{col}}} \quad (\text{Equation III.4})$$

where k is the isocratic retention factor and is replaced by k_e in gradient experiments. The k_e represents the retention factor of the solute in the eluted mobile phase composition and is calculated using Equation III.5, which is derived from the linear solvent strength theory of Snyder and Dolan [29, 30]. Note that this equation only provides an approximation of k_e ; for example, the peak compressibility effect is not considered.

$$k_e = \frac{t_{\text{grad}}}{2.3 \times t_0 \times \Delta\Phi \times S} \quad (\text{Equation III.5})$$

where t_{grad} is the gradient time, t_0 is the column dead time, $\Delta\Phi$ is the gradient span, and S is the slope of a plot of the logarithm of the retention factor vs. the solvent composition.

2.7. Experimental peak capacity determination

The experimental peak capacity P_{exp} was calculated for each condition reported in Table III.1 using Equation III.6 [31, 32].

$$P_{\text{exp}} = 1 + \frac{t_{\text{grad}}}{W_{50} \times 1.697} \quad (\text{Equation III.6})$$

The measurement of peak width at 50% height was more accurate than at 13.4%, and a factor of

1.697 should be employed to transform $W_{50\%}$ into $W_{13.4\%}$, assuming that the chromatographic peak is Gaussian [33]. Finally, the peak width considered in Equation III.6 was the average value of the peak widths for a representative number of peaks spread over the chromatogram.

Since the retention windows (between first and last eluting peak) could be shorter than the gradient time at higher temperature, Equation III.6 may underestimate the peak capacity in these conditions. This equation was however used without adaptation of initial and final composition, because the calculation would be not generic and difficult to implement in the present study.

2.8. Theoretical peak capacity calculation

The theoretical peak capacity P_{calc} was calculated for all the conditions reported in Table III.1 using Equation III.7 [32]:

$$P_{calc} = 1 + \frac{\sqrt{N_{obs}}}{4} \times \frac{1}{b+1} \times \ln\left(\frac{b+1}{b} \times e^{S \times \Delta\Phi} - \frac{1}{b}\right) \quad (\text{Equation III.7})$$

Equation III.2 was employed for calculating N_{obs} . The S value was determined using the procedure described in Section 2.5. Finally, b , which represents the gradient steepness, was calculated using the following equation:

$$b = \frac{t_0 \times \Delta\Phi \times S}{t_{grad}} \quad (\text{Equation III.8})$$

The construction of the plots of peak capacity vs. particle diameter and column length (Figures III.9-11) required the calculation of new flow rates generating 90% of the maximal pressure of the system and the extra-column dispersions in

each case. Equation III.9 and 10 [34] were employed:

$$\frac{F_2}{F_1} = \frac{\Delta P_2}{\Delta P_1} \times \frac{I.D._2^2}{I.D._1^2} \times \frac{l_1}{l_2} \times \frac{d_{p,2}^2}{d_{p,1}^2} \quad (\text{Equation III.9})$$

where 1 and 2 refer to the 1st and 2nd analysis, ΔP is the column backpressure, $I.D.$ is the column internal diameter, l is the column length, and d_p is the particle diameter.

$$\sigma_{ext}^2 = A \times F^B \quad (\text{Equation III.10})$$

where F is the flow rate, and A and B constants determined by a least squares regression, fitting experimental values to Equation III.10 [34]. In the LC system used, A and B were equal to 6.954 and 0.3244 with MS detection and to 4.355 and 0.3244 with UV detection, respectively.

2.9. Evaluation of the number of resolved peaks

The number of peaks visually resolved on the TOF-MS chromatogram was manually determined in each chromatographic condition, in a range comprised between 2 peaks (m/z 402.23 and 1048.59). All peaks above the following limit were considered: minimum peak height = 1/8 of the average intensity of two randomly chosen peaks (m/z 602.32 and 471.51). Among the remaining peaks, only the peaks with a maximum overlap of 50% of the height of the less intense peak were considered.

2.10. Conus venom temperature stability study

The stability of samples at 90 °C was tested using following methodology: the *Conus* venom was injected in conditions 5 (30 °C) and 7 (90 °C) with the generic gradient and in condition 7 with an initial 2% isocratic hold for 93.0 min (for conditions, see Table III.1). All MS signals with

height above 10% of the relative intensity were listed by the MassLynx software, including their area, for the three separations. Areas were

normalised by the area of the peak corresponding to m/z 402.24.

3. Results and discussion

Based on the experimental design Section 2.1, the influence of different chromatographic parameters on peak capacity was investigated in a systematic way, to attain the optimum LC–MS conditions (*i.e.*, maximal resolution for the profiling of peptides with MMs ranging from 1000 to 5000 Da), in view of venomics applications. In parallel, the same analyses were performed on a complex sample containing small molecules (MMs 200–800 Da).

3.1. Peak capacity for the evaluation of profiling performance

The performance of the different chromatographic conditions was evaluated in terms of peak capacity. This parameter was experimentally determined from the chromatograms of both *Conus* venom and *Hypericum* extract and calculated based on the equations presented in Section 2. According to Equation III.7, peak capacity mainly depends on two factors: the square root of the isocratic efficiency, N , and the “gradient retention factor”, k_e , which is inversely proportional to the gradient steepness, b ($k_e = 1/2.3 b$). All of the investigated chromatographic parameters act on one or both factors, sometimes with opposite effects on the peak capacity. The column efficiency (N_{col}) mainly depends on the particle diameter and column length (fixed to 150 mm in this work) and was determined for each flow rate in the absence of extra-column dispersion. On the other hand, k_e mainly depends on the S parameter, column dead time (t_0), gradient time (t_{grad}) and gradient span ($\Delta\Phi$). Thus, all chromatographic parameters

influence N and/or k_e and consequently the performance of gradient separation [32]. Finally, it is important to note that instrumentation (σ_{ext}^2) also strongly influences the N_{col} value.

Experimental peak capacity values were obtained using Equation III.6 [31, 32], based on duplicate determinations of half height peak width for 19 representative peptides (*Conus* venom) and 20 plant secondary metabolites (*Hypericum* extract). Both the peptides and small molecules that were monitored were randomly chosen and cover a wide polarity range over the whole chromatogram; they are listed based on their molecular masses in Table III.3. The mean peak capacities experimentally obtained with this procedure for the eight selected conditions are displayed in Figure III.5 (black bars).

Because mixtures of peptides (real extracts), instead of pure products, were used to assess peak capacity, a practical estimation of the effects of changing profiling conditions was obtained by measuring the number of resolved peaks for each condition at a given threshold (see Section 2). As illustrated in Figure III.5, peak capacity is a good measure of profiling performance, because an increase in the number of resolved peaks (grey bars in Figure III.5) was well correlated with peak capacity (black bars in Figure III.5). However, because the gradient steepness (Equation III.8) and temperature were not constant for all of the tested conditions, some changes in selectivity occurred, which could explain the observed differences between grey and black bars in Figure III.5.

Table III.3. List of representative peptides from the *Conus* venom and small molecules from the *Hypericum* extract employed for peak capacity calculation, by the method described in Section 2.7. Their corresponding molecular mass (MM) and calculated S parameter (S) are indicated.

	Selected peptides of the <i>Conus</i> venom		Selected molecules of the <i>Hypericum</i> extract	
	MM	S	MM	S
1	762.455	16.2	290.079	8.9
2	776.451	16.0	464.097	7.9
3	1313.481	31.3	464.097	3.7
4	1565.927	19.6	448.100	4.5
5	1656.823	22.6	506.105	10.1
6	1949.912	24.0	302.043	13.4
7	2080.171	13.3	538.091	15.1
8	2108.185	12.9	538.092	7.7
9	2352.981	30.9	538.092	32.9
10	2352.981	29.7	517.319	7.8
11	2373.890	44.6	519.333	3.2
12	2631.520	24.6	521.349	10.3
13	2645.532	23.5	495.333	8.1
14	2909.134	32.1	517.315	3.9
15	3358.431	22.0	484.107	4.3
16	3380.378	22.7	558.125	14.8
17	3392.421	14.8	552.379	5.8
18	3570.019	22.1	568.377	8.6
19	4115.615	30.4	731.559	13.8
20			798.527	13.6
Average	2384	23.9	517	10.1

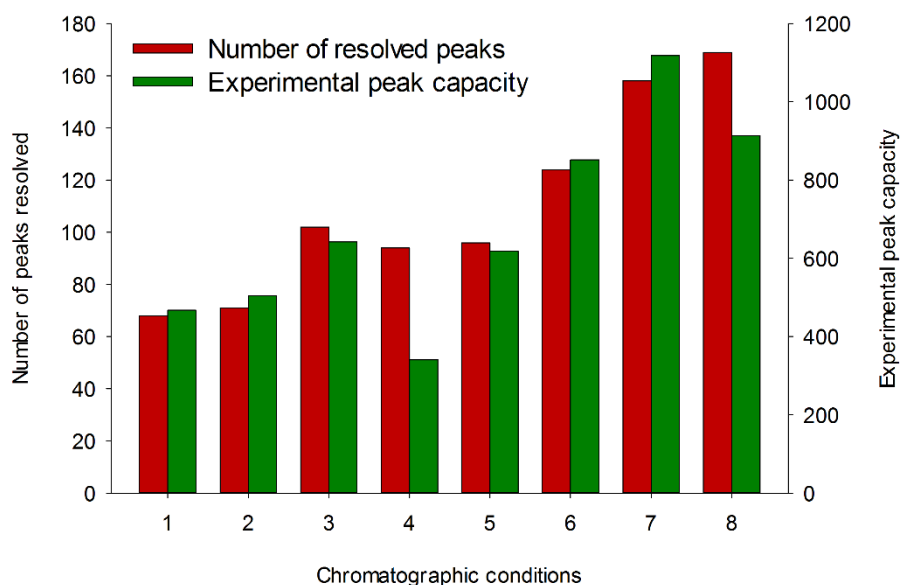


Figure III.5. Number of peaks visually resolved on the LC-TOF-MS chromatogram of *Conus* venom using a constant time window ranging from peak of m/z 402.23 and m/z 1048.59 (grey bars) compared to experimental peak capacity (black bars). All peaks with intensity above $1/8$ of the average intensity of two randomly chosen peaks (m/z 602.32 and 471.51), and with a maximum overlap of 50% of the height of the less intense peak, were considered. This comparison was made for the 8 conditions described in the first eight lines of Table III.1.

3.2. Comparison of the experimental and calculated peak capacity values

To verify the accuracy of the calculated vs. experimental peak capacities, a systematic comparison was performed for the 8 conditions described in Section 2.1 and in Table III.1 with both *Hypericum* extract and *Conus* venom.

The calculated peak capacity values for each set of conditions were obtained from Equation III.7 [32, 35]. Prior to this calculation, the column and extra-column dead volumes (v_0 and v_{ext}) and dispersions (σ_{col}^2 and σ_{ext}^2), column efficiency (N_{col}) and observed (or effective) efficiency (N_{obs}) were determined (Section 2.6). For both samples, the average S parameter was experimentally determined based on two gradient runs of different slopes (see Section 2.5 and Table III.3).

The results are summarised in Figure III.6, and representative chromatograms of *Conus* venom profiling are displayed in Figure III.7. In general, a good correlation between the experimental and calculated peak capacities was obtained. Mean differences between the experimental and calculated values of 15% and 11% were observed for the small molecules of *Hypericum* extract and the peptides of *Conus* venom, respectively. These values (Figures III.6A and B, respectively) were within the range of previous studies [20] and are acceptable for the purpose of this study. Interestingly, the calculated values of conditions 1, 4 and 5 with the *Hypericum* extract were more critical (up to 33% difference between the experimental and calculated values). These discrepancies could be related to the approximate measurement of the experimental peak width because peaks are sometimes not perfectly Gaussian. Another explanation is related to the numerous approximations that

were made for the calculation, such as (i) the inaccurate parameters, A , B and C , of the Knox equation for all of the tested columns [27]; (ii) the imprecise value of D_m ; (iii) the change in mobile phase viscosity during the gradient elution, a factor that was not accounted for in the peak capacity calculation; (iv) the imperfect peak capacity estimation of Equation III.7; and (v) the peak compressibility and frictional heating effects, which were not considered in the model.

Because the differences between the calculated and experimental determination nevertheless

remain acceptable, the results of this work confirm the validity of Equation III.7 as a reliable tool for estimation of peak capacity of both small analytes and peptides in complex mixtures.

3.3. Effect of the nature of the analytes on experimental peak capacity

Because the previous experiments produced coherent results for both small molecules and peptides, a comparison of the peak capacities

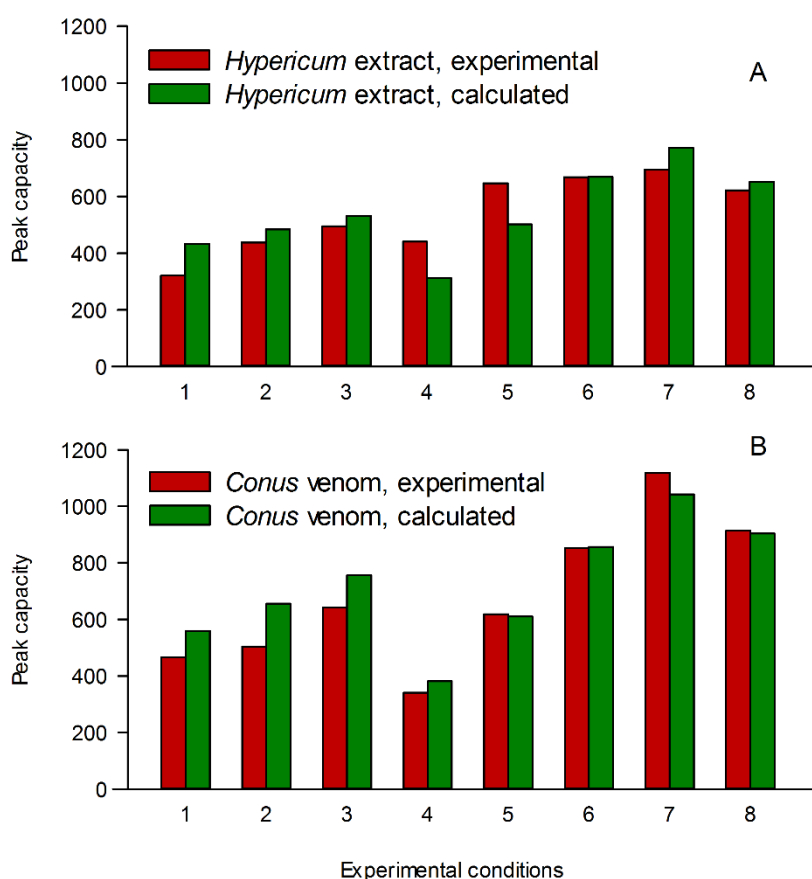


Figure III.6. Average experimental (grey bars) and calculated (black bars) peak capacity on (A) *Hypericum* extract and (B) *Conus* venom, obtained for the 8 conditions described in the first eight lines of Table III.1. Detection: TOF-MS (see conditions in Section 2).

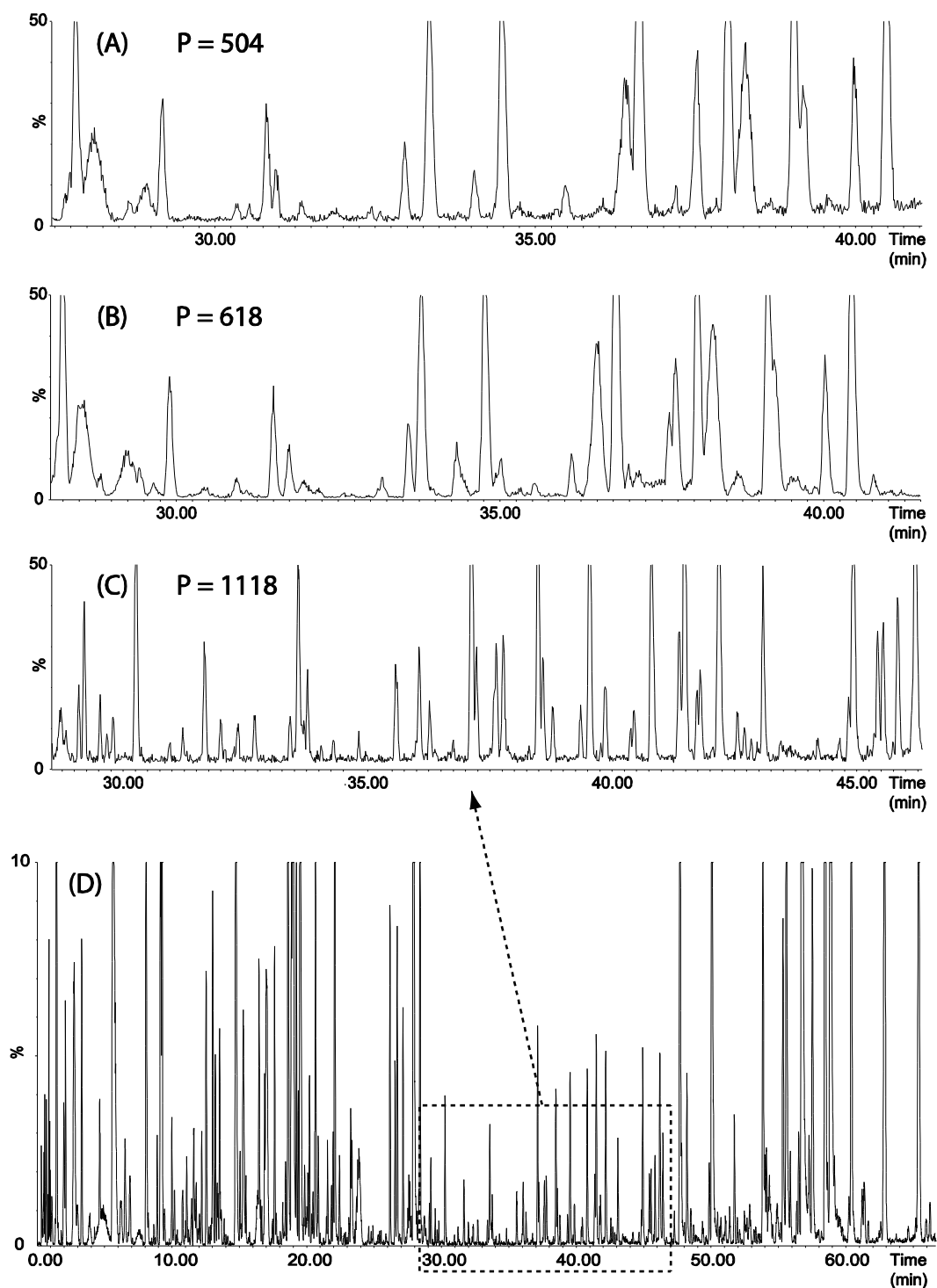


Figure III.7. Zoom-in on a representative region of the chromatograms of *Conus consors* samples. (A) Xbridge 150 x 3.0 mm I.D., 3.5 μ m column at 30 $^{\circ}$ C and 880 μ L/min (condition 2), (B) Acquity 150 x 2.1 mm I.D., 1.7 μ m column at 30 $^{\circ}$ C and 380 μ L/min (condition 5) and (C) Acquity 150 x 2.1 mm I.D., 1.7 μ m column at 90 $^{\circ}$ C and 810 μ L/min (condition 7). The lower chromatogram (D) represents the separation obtained with the Acquity 150 x 2.1 mm I.D., 1.7 μ m column at 90 $^{\circ}$ C and 810 μ L/min (condition 7). Detection: TOF-MS, experimental conditions reported in Table III.1 and in Section 2.

measured for these two types of analytes in each condition allow the assessment of the effect of the nature of the analyte. In general, the peak capacity values obtained for peptides (Figure III.6B) were higher than those recorded for small molecules (Figure III.6A). This difference was as high as 61% in condition 7 (temperature, 90 °C; flow rate, 810 $\mu\text{L}/\text{min}$; and particle diameter, 1.7 μm), but the intensity of the effect varied with the experimental conditions. According to the theory, the nature of the analyte strongly influences its chromatographic behaviour. The major differences between peptides and small analytes were the S parameter and the diffusivity (D_m). It is well established that peptides have a higher S value than small molecules [23]. The mean S value was of 23 for peptides and 10 for the small molecules, as reported in Table III.3. According to Equation III.7, an increase in peak capacity is expected for peptides when compared with small molecules. This increase corresponded well to the experimental observations reported in Figure III.6. In addition, because the S parameter also affects the gradient steepness, b , a stronger peak compression effect may occur in the case of peptides [30]. In other words, the change in elution strength may be important enough to increase the elution of the analyte between the beginning and the end of the peak and to produce thinner peaks and thus a higher peak capacity [36].

The greater peak capacities generally observed for peptides vs. small molecules were attributed to the greater S values for these analytes. This effect could be partly counterbalanced by effects related to the lower diffusivity of peptides, which resulted in a lower optimal linear velocity, u_{opt} . Indeed, most of the flow rates reported in Table III.1 were beyond u_{opt} in the case of peptides. The loss in efficiency was thus more pronounced with peptides vs. small molecules (the average reduction in efficiency was 30% for all the tested conditions with peptides), leading to a reduction of peak capacity. The impact of S on peak capacity

is higher than that of D_m , which explains the higher experimental peak capacity values obtained for the *Conus* venom samples compared with the *Hypericum* extract.

3.4. Effect of the flow rate on experimental peak capacity

Because an increase in the mobile phase flow rate decreases the column dead time, t_0 , such an increase also increases k_e and thus the peak capacity, according to Equation III.7. In the meantime, the increase in linear velocity towards the C-term dominated region of the Van Deemter curve will result in a decrease in efficiency and thus in peak capacity, depending on the particle size and the nature of the analyte.

The consequences of these opposite effects on peak capacity were investigated with the two model samples. Experimentally, comparisons of peak capacity between conditions 1 and 2 (3.5 μm particles), or between 4 and 5 (1.7 μm particles), differing only in the flow rate, show greater peak capacity at higher flow rates in all cases (+37% at 880 vs. 440 $\mu\text{L}/\text{min}$ and +47% at 380 vs. 130 $\mu\text{L}/\text{min}$ for *Hypericum* extract and +8% and +82% for *Conus* venom, respectively). The results are displayed in Figures III.6A and B, and they were confirmed by the calculated values.

The increase in peak capacity with an elevated flow rate is higher for sub-2 μm than for 3.5 μm particles for all analytes. This phenomenon is well known for small analytes, and these new results indicate that a similar trend holds for peptides. This effect is explained by the position of these analytes on the van Deemter curve relative to u_{opt} . Indeed, in the case of 3.5 μm particles, the flow rate used in conditions 1 and 2 was far above the u_{opt} , in contrast to conditions 4 and 5 [37]. In addition, the increase in peak capacity at higher flow rates is more pronounced for the *Conus*

venom than for the *Hypericum* extract. An explanation for this result is that the retention factor, k_e , is lower for peptides (0.27–2.34 for the 8 tested conditions) compared with small molecules (0.65–5.51). In fact, it is well known that peak capacity increases markedly for small k_e values, while no real improvements in peak capacity occur for k_e values beyond 10 (a plateau is observed) [21]. This increase is not linear and is much more pronounced for low k_e values.

Thus, for both peptides and small analytes, gradient separations should be performed at the highest tolerated flow rate, even when the experiments are conducted deep in the C-term region of the Van Deemter curve. This statement is particularly true in UHPLC, as the maximal flow rate is rapidly limited by the backpressure generated by sub-2 μm particles. For this reason, a safe recommendation would be to use a flow rate that generates 90% of the maximal backpressure tolerated by the UHPLC system in order to extend the column and instrument lifetimes while obtaining maximal peak capacity.

Because of the physical limitation of the ESI source, the use of a T-split is recommended for high flow rates (above 600 $\mu\text{L}/\text{min}$). The split should not alter sensitivity, since the ESI is a concentration-sensitive device, nor provide peak broadening, because the ratio of extra-column (σ_{ext}^2) to in-column dispersion (σ_{col}^2) decreases at high flow rates and with 3 mm I.D. columns.

3.5. Effect of the temperature on experimental peak capacity

As discussed, the maximal UHPLC flow rate generates the highest peak capacity. The limiting factor to further increasing the flow rate is the generated backpressure. It is well known that temperature significantly decreases mobile phase viscosity and consequently, backpressure, allowing for operation at higher flow rates [38,

39]. Furthermore, the decrease in viscosity increases the diffusion coefficient of analytes, D_m [40], thus increasing the peak capacity values of peptides [39, 41]. Last but not least, increasing the mobile phase temperature often provides better peak shape, because secondary interaction kinetics is improved at elevated temperatures [23].

The effect of temperature was thus evaluated for both peptides and small molecules. Analyses in conditions 6 and 7 were carried out in UHPLC at high temperatures (60 and 90 °C), showing a significant increase in experimental peak capacity (+38% from 30 °C to 60 °C and +31% from 60 °C to 90 °C), up to a value of 1118 for *Conus* venom samples. On the other hand, the increase was only +3% and +4% for the *Hypericum* extract, while the increases in the calculated values were higher (+33% and +15%). The difference between the calculated and experimental peak capacity value is not explained and contradicts previous works on other types of plant extracts, in which an increase in the peak capacity of approximately 30% was observed with a comparable increase in temperature [1].

Because the flow rates at 30 °C and 90 °C were set at 380 and 810 $\mu\text{L}/\text{min}$, both generating a 900 bar backpressure, the column dead time was reduced by a factor 2 at 90 °C compared with at 30 °C. Thus, an improvement in peak capacity was expected because k_e increased, as described in Equation III.7 for peptides and small molecules. However, the larger increase in peak capacity for peptides can be explained by a stronger improvement in the diffusivity of peptides as compared with small molecules, which thus increased the efficiency in Equation III.7.

Because elution strength could change at 90 °C, a 0.73%/min gradient slope was tested (condition 8 of Table III.1), providing the same retention time for the last peak of interest as in condition 5

(30 °C). Both experimental and calculated peak capacities decreased (from -10 to -20%) for both analytes (Figure III.6), which is explained by the negative influence of the gradient span ($\Delta\Phi$) on peak capacity, as shown in Equation III.7. Conversely, the number of resolved peaks increased (Figure III.5), showing the positive effect of the higher retention window. This divergence between peak capacity and number of resolved peaks has been discussed in Section 2.7.

In addition to the positive effects on peak capacity, the increase in temperature significantly affected the overall profiling pattern of a given sample, particularly because complex mixtures were considered. Indeed, depending on the nature of the analytes, significant changes in selectivity occurred. Furthermore, the high temperature also increased the eluent strength of the mobile phase. The retention factor may become too low when working with polar analytes, resulting in a poorer quality of separation [38]. Thus, unlike a flow rate change, a temperature modification cannot be considered as a geometric transfer, and its consequences on analyte profiling have to be estimated carefully, even though an increase in temperature will generally improve peak capacity.

3.5.1. Thermal stability of *Conus* venom at elevated temperatures

The use of an elevated temperature for peptide separation raises the question of thermal stability. No evidence of thermal degradation was observed during the analyses for peak capacity measurement, and two additional experiments were carried out to rapidly estimate this effect (as described in Section 2).

A comparison of all peaks detected above a 10% threshold at 30 and 90 °C (conditions 5 and 7) was performed. Because the selectivity was different, the peaks on the two chromatograms were identified according to their corresponding m/z values, and their areas were compared. Among the 44 peaks above this 10% threshold, 13 could not be considered in the 30 °C gradient because they were not eluted from the column, due to the change in elution strength between 90 and 30 °C. Additionally, 2 of the 31 remaining peaks could not be detected. All of the other peaks were found in both conditions, although their intensities were different. These differences are probably due to (i) changes in selectivity, and thus in the extent of matrix effects, and (ii) changes in the mobile phase composition during the elution of the compound, leading to a different desolvation yield in ESI.

To keep selectivity constant, an additional experiment was performed. In this case, a 2% isocratic hold was maintained for 93 min before the gradient of condition 7 was applied at 90 °C. A comparison of both 90 °C chromatograms, with and without the isocratic hold, demonstrated that among the 44 peaks monitored, 2 were missing after the 93 min isocratic hold at 90 °C, while all other peaks were present with similar areas. The results of these two experiments indicate that the possibility of degradation cannot be excluded, but the extent of such effect seems to be very limited.

In addition, injection in all the tested temperatures of a pure commercialised conopeptide (CnIIIIC) did not show the presence of any additional peak that could correspond to degraded products, as shown in Figure III.8.

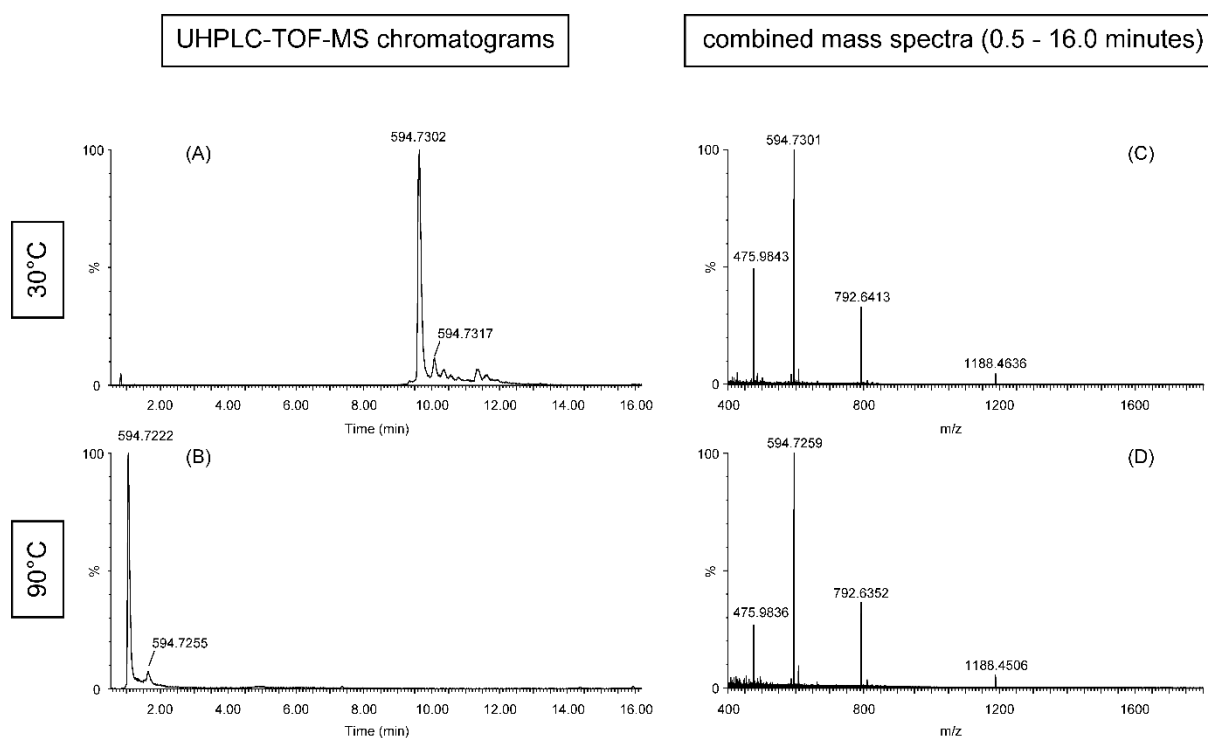


Figure III.8. UHPLC-TOF-MS chromatograms (A, B) and total mass spectra (C, D) of the CnIIIC peptide at 30 °C and 90 °C, respectively. Both chromatograms display a main peak that corresponds to the CnIIIC peptide and several smaller peaks, showing the same mass spectrum at both temperatures. The total mass spectra (TMS) obtained by combining spectra from RT 0.5 to RT 16.0 min highlighted all ions detected. Comparison of TMS at 30 °C (C) and 90 °C (D) shows the same pattern and only m/z characteristic for CnIIIC were recorded.

Moreover, the 100% stability of this conopeptide in solution at 50 °C during 10 days was previously demonstrated (personal communication of Cécile Cros, Atheris Laboratories).

In conclusion, no evidences of degradation were experimentally demonstrated. Indeed, the folded and cysteine-rich peptides that usually compose such venoms are known to be stable at high temperature [42-44]. Similarly, a previous study on plant extracts indicated that no apparent degradation occurred at high temperatures with small molecules [1].

If degradation is however suspected, separations can be carried out at lower temperature and using larger I.D. columns to provide similar peak capacity values, as discussed below (Section 3.7.3).

3.6. Effect of gradient time on the calculated peak capacity

Previously, all analyses for maximising peak capacity were performed using the same gradient time (Table III.1). However, another way to increase peak capacity is to change the gradient

time. Indeed, the gradient time influences gradient steepness and thus k_e (Equation III.5). However, gradient time also influences the column dispersion (σ_{col}^2) according to Equation III.4 and thus the observed column efficiency and the peak capacity. It is well known that for small molecules, a longer gradient time generally provides a higher peak capacity [20, 37]. In the present study, the influence of this parameter was evaluated for peptides, taking in account the effects of the instrumentation (σ_{ext}^2).

Figure III.9 presents a plot of the calculated peak capacity as a function of gradient time for both peptides and small molecules at different temperatures and with different column I.D. measurements. The white points at 93 min represent experimental values, showing the validity of the model. The column length was

fixed at 150 mm, and flow rates were set to provide a 900-bar backpressure. Peak capacity values were calculated using Equation III.7. For a given set of conditions, the curves representing the peak capacity increased with increasing gradient time, always showing a steeper slope for peptides than for small molecules. These curves flattened more rapidly for small molecules than for peptides, indicating that the optimum peak capacity was reached with higher gradient time for peptide separations. For example, at 90 °C, the peak capacity generated with the standard 1% slope gradient (93 min gradient time) increased by more than 40% for peptides when the gradient time doubled, while this increase was limited to 20% for small molecules. The reason for this behaviour is again related to the value of k_e . As shown previously [21], the gain in peak capacity becomes negligible for k_e values

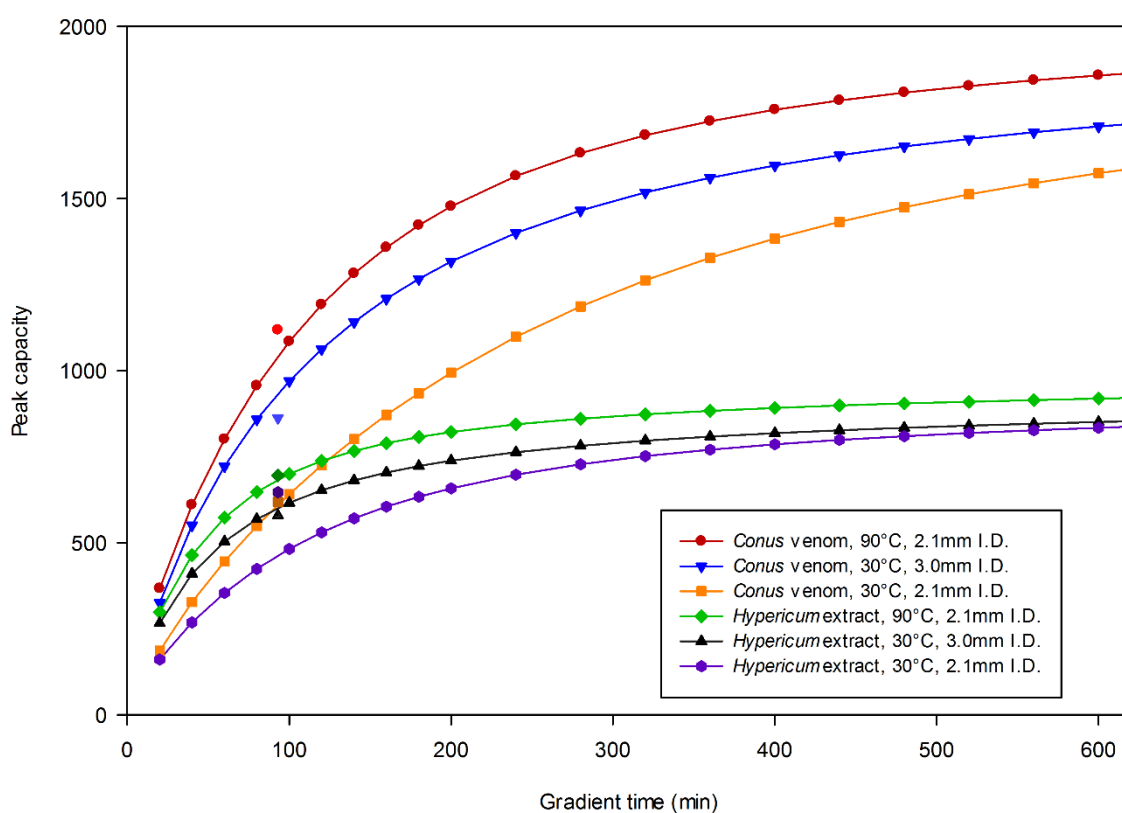


Figure III.9. Plot of the calculated peak capacity vs. gradient time for both *Conus venom* and *Hypericum extract* in 3 different conditions. White points represent experimental values, for comparison purpose.

beyond 10. Because the S value was, on average, 2-fold higher for peptides vs. small molecules, k_e values were much lower for peptides, and thus, a very high gradient time was required to attain a k_e of 10. As a general rule, a practical way to select an optimum gradient time that provides a good compromise between a high peak capacity and a reasonable analysis time is to work with a gradient k_e below 10.

In order to find a good compromise between peak capacity and analysis time, a gradient time range was selected, which limits are defined by the gradient times providing 50% and 80% of the maximum peak capacity of an infinite time. The lower limit (50% of maximum peak capacity) roughly corresponded to optimum peak capacity values previously attained in conventional HPLC profiling of venoms [17, 45]. The upper limit (80% of maximum peak capacity) provided UHPLC gradient times similar to the highest analysis times found for HPLC venom profiling studies [17, 45].

For example, in the case of peptides, a peak capacity that was 80% of the maximal achievable value was reached for a gradient of approximately 280 min, while a value that was 50% of the maximum was obtained at 90 min, at 90 °C and using a 2.1 mm I.D. column. For small molecules, these values were obtained at 190 and 50 min, respectively. Working between these values probably provides the best compromise in terms of both peak capacity and analysis time.

3.7. Effect of the particle diameter and of the column geometry on the calculated peak capacity

When optimising metabolite profiling conditions, the last variable that can be modified is the column itself. When a given chemistry has been defined, the column geometry must be chosen

with care. The effects of altering particle diameter or aspects of the column geometry, such as column length and internal diameter, were evaluated.

3.7.1. Effect of the particle diameter on the calculated peak capacity

Decreasing particle diameter (d_p) is known to increase the column efficiency; this effect has been the basis of UHPLC technology using sub-2 μm particles. However, the use of lower particle diameters increases the backpressure, limiting the flow rate.

The influence of d_p on peptides and small molecules was calculated for a fixed 93 min gradient time and for a detector with low extra-column dispersion, such as a UV detector, and the optimum peak capacity values were achieved for the same d_p with both analytes. This value was close to 1.6 m, which corresponds to that of commercially available sub-2 μm columns. As shown in Figure III.10, increasing d_p decreases the peak capacity for both peptides and small molecules in a relatively similar manner. With the TOF-MS device, the optimal d_p was slightly different (approximately 2.0–2.2 m) because of the non-negligible effect of the MS device on band broadening, but it remained close to that of the column employed

in the present study ($d_p = 1.7 \mu\text{m}$). The effect of the detector on separation performance will be discussed in more detail in Section 3.8.

3.7.2. Effect of the column length on the calculated peak capacity

Column length is a delicate parameter to optimise because the curve of peak capacity vs.

column length presents an optimum value that is related to the gradient time. Peak capacity is also strongly related to other column parameters such as particle diameter and column internal diameter.

The column efficiency N_{col} increases with the use of longer columns (Equation III.1), and the effect of instrumentation on overall performance also becomes less pronounced. However, the retention factor, k_e , decreases significantly. The lower k_e is explained by both the higher column

dead volume and the reduced flow of longer columns (Equation III.9). Because peak capacity depends on the balance between N and k_e , there is an optimum value for column length. Figure III.11 shows the peak capacity vs. gradient time and column length for both peptides and small molecules. Peak capacity values were calculated using Equation III.7. Peak capacities were higher for peptide separations than for small molecules, and all trends visible on the 3D plots were similar. As expected, the plots clearly show that maximal peak capacity values were not obtained for the

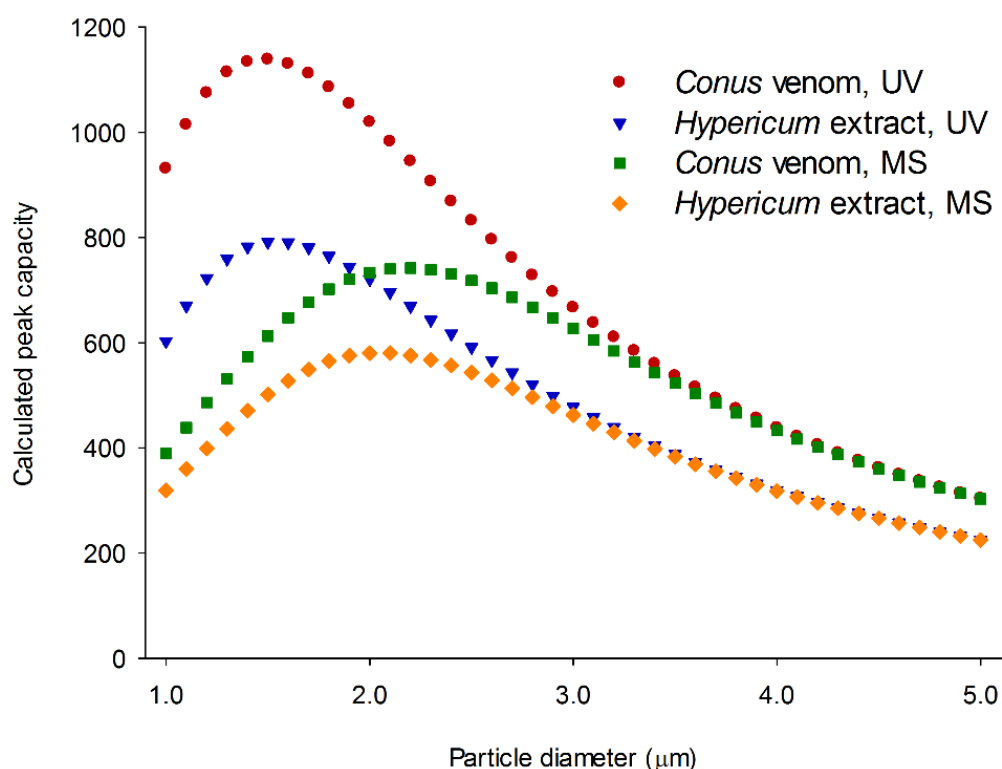


Figure III.10. Plot of the calculated peak capacity vs. particle diameter, for both *Conus venom* and *Hypericum extract*, using both UV and TOF-MS detection. Flow rates were systematically calculated using Equation III.9 to attain a maximal backpressure (90% of ΔP_{max}) for all particle diameters.

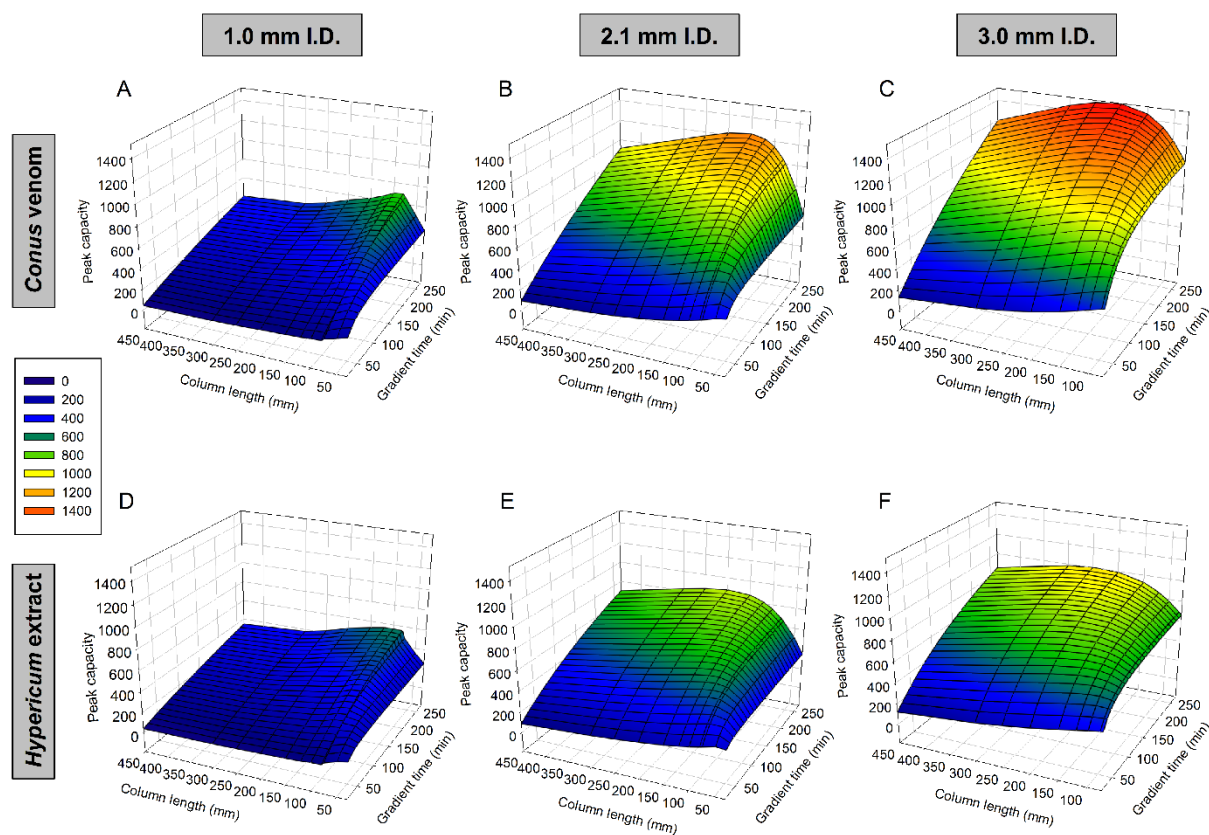


Figure III.11. 3D plots of calculated peak capacity vs. column length and gradient time at 30 °C, for both *Conus venom* (plots A–C) and *Hypericum extract* (plots D–F), using 3 different column internal diameters: 1.0 mm I.D. (plots A and D), 2.1 mm I.D. (plots B and E) and 3.0 mm I.D. (plots C and F). Flow rates were always calculated to provide the maximal backpressure (90% of ΔP_{\max}).

longest columns and that there was an optimum column length. Apart from small columns I.D. (1.0 mm), the optimum column length for gradient times higher than 30–45 min (commonly used for high resolution profiling of complex mixtures) was always found to be between 150 and 200 mm. This result also corresponds to the length of commercially available columns. In a previous study, increasing the column length up to 450 mm was found to be beneficial for separating small molecules only for gradient times exceeding 200 min [20].

3.7.3. Effect of the column internal diameter on the calculated peak capacity

As shown in Figure III.11, the column internal diameter (I.D.) strongly influenced the 3D plot representations (see, for example, the differences between 1.0 mm I.D. and 3.0 mm I.D. displayed in Figures III.11A and C). The internal diameter indeed had a strong influence on the observed efficiency, N_{col} , because a change in the flow rate affects the $\sigma_{ext}^2/\sigma_{col}^2$ ratio, as described

in Equation III.2 (see below). It must be noted, however, that column efficiency (N_{col}) should theoretically be unaffected by column I.D. changes.

For both types of analytes, Figures III.11B, C and E, F show an increase in the calculated peak capacity when using a 3.0 mm I.D. compared with a 2.1 mm I.D.; this increase was from 612 to 923 for peptides and from 465 to 599 for small molecules (93 min gradient time at 30 °C on 150

mm columns). Experimentally, the same result (30 °C, MS) can be seen in Figure III.12 for similar conditions. The use of 1.0 mm I.D. columns strongly lowered the peak capacity for both small molecules and peptides (Figures III.11A and D) because of the important contribution of σ_{ext}^2 in the UHPLC-TOF-MS configuration used. Practically, 1.0 mm I.D. columns are only useful for high-throughput screening, when high resolution is not required and when only a small amount of sample is available.

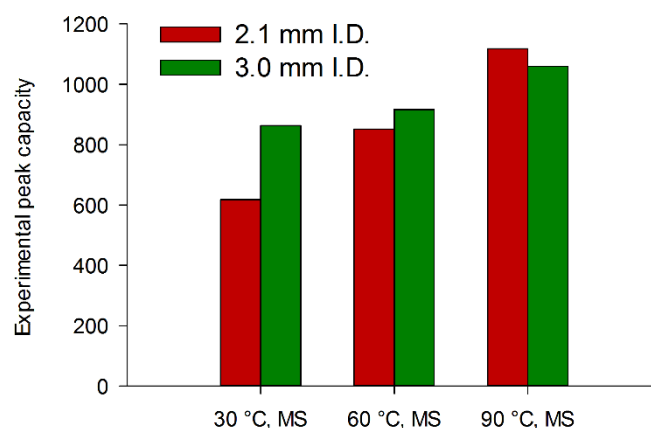


Figure III.12. Experimental peak capacity for 2.1 mm I.D. and 3.0 mm I.D. columns, at 30, 60, and 90 °C. Detection: TOF-MS (see conditions in Section 2).

The combined effects of increased temperature and increased I.D. were experimentally investigated on different columns with I.D.s of 2.1 and 3.0 mm (Figure III.12). Experiments on the 3.0 mm I.D. column were performed using conditions 9–11 (Table III.1), which were derived from conditions 5–7. Experimental peak capacities were significantly improved on the 3.0 mm column compared with the 2.1 mm I.D. column at 30 °C, while this difference was smaller at 60 °C, and the 2.1 mm I.D. column performed

slightly better at 90 °C. This slight relative decrease in peak capacity at 90 °C for the larger I.D. column is probably due to the frictional heating that occurred at high flow rates (1450 L/min) with sub-2 μm diameter particles, an effect that is known to rapidly increase the reduced plate height, h , as the flow increases above the optimum rate [46, 47].

Using 3.0 mm I.D. columns is thus an easy way to increase peak capacity without a strong increase

in temperature (Figure III.12), thus limiting the risk of putative degradation. However, the consumption of both solvent and sample is doubled because of the increase in flow rate and column dead volume. Finally, separation at 90 °C with a 3.0 mm I.D. column showed no real improvements.

3.8. Effect of the detection mode on peak capacity

It is noteworthy that the different experimental results obtained in terms of peak capacity depended on the analytical platform used, and in this respect, the extra-column dispersion, σ_{ext}^2 , which is related to the detector, played a non-negligible role [34]. Furthermore, acquisition parameters such as acquisition rate or peak smoothing may bias the real peak width measurements and thus the peak capacity.

To evaluate the extra-column dispersion (σ_{ext}^2) caused by the MS detector, analyses were carried out using both UV and TOF-MS detection. With the UHPLC-TOF-MS platform used for this study, peak capacities were shown to be 15–113% higher with UV compared with TOF-MS detection (data not shown). A correlation between these differences and the flow rate was found: the higher the flow rate, the lower the difference in peak capacity between UV and MS detection.

These differences between UV and MS detection are shown in Figure III.10, in which peak capacities are plotted against particle diameter for the maximum flow rate. As shown in this plot, for peptides, maximum peak capacities were approximately 1140 for UV and 740 for MS detection. This important difference becomes undetectable, however, when particles larger than 3.5 μm are used. Indeed, because the calculations were always made at the maximal flow rate, increasing particle size, and thus flow

rate, tends to decrease the ratio of extra-column (σ_{ext}^2) to in-column dispersion (σ_{col}^2).

To confirm that this behaviour was related to differences in extra-column dispersion between UV and MS, σ_{ext}^2 values were experimentally determined using the method described in Section 2.6.3. In general, the MS instrument itself produced a 10- to 24-fold higher σ_{ext}^2 value than UV (see Table III.2), while the contribution of tubing between UV and MS was found to be negligible. This result indicated that the difference in observed efficiency between UV and MS detection was related to differences in their respective σ_{ext}^2 . Because σ_{ext}^2 was non-negligible in TOF-MS, the use of a 3.0 mm (high σ_{col}^2) instead of 2.1 mm I.D. column was a useful approach.

It should also be mentioned that MS parameters may influence the peak capacity [19, 48]. In this respect, the acquisition rate is probably the most critical parameter. For example, the average peak width at baseline in conditions 7 (*i.e.*, 90 °C) was 6 s. For such a thin peak, a 10 points peak resolution will require an acquisition rate of 0.3 s when using the dynamic range enhancement (DRE) function needed for ensuring high TOF-MS mass accuracy. A 1 Hz acquisition rate would provide only 3 points, a scenario that will tend to significantly bias peak width measurements [48].

3.9. MS/MS deconvolution applications

Because this study aimed at maximising peak capacity for obtaining detailed venom profiles, an important aspect was to validate the possibility of acquiring MS/MS spectra for very thin UHPLC peaks obtained under optimal conditions (see the optimised chromatogram of the venom, Figure III.7D). Indeed, MS/MS is mandatory for venomics applications in order to perform peptide deconvolution and identification by *de novo*

sequencing [45, 49]. For this purpose, the UHPLC high resolution profiling techniques for *C. consors* venom optimised in this study were used for an analysis on a recent generation QTOF-MS instrument with high acquisition frequency. To obtain MS/MS data on the highest possible number of LC peaks, an automatic MS/MS survey was undertaken with a self-generated exclusion mass list to optimise the generation of non-redundant MS/MS data. The experimental MS/MS spectra were matched with recently reported peptides present in the *C. consors*

transcriptome [50]. In addition, the MS/MS data provided spectra of sufficient quality to enable complete or partial *de novo* sequencing. A complete review of the peptidomic results will be presented in a subsequent article (manuscript in preparation). Overall, these results indicate that the thin LC peaks obtained under conditions providing maximal peak capacity were compatible with online peptide deconvolution and matched with transcriptomic data and, to some extent, with *de novo* sequencing in these complex mixtures.

4. Conclusion

The present study demonstrates a generic approach for maximising peak capacity using modern LC–MS platforms for high resolution profiling of complex mixture containing peptides or small molecules. Several practical rules, which are summarised in the decision tree presented in Figure III.13, were established. These generic rules can be applied to any LC–MS platform when seeking to maximise LC resolution. Adaptation of gradient span and gradient time may be required in a second step, depending on the sample complexity, to avoid undesired coelutions.

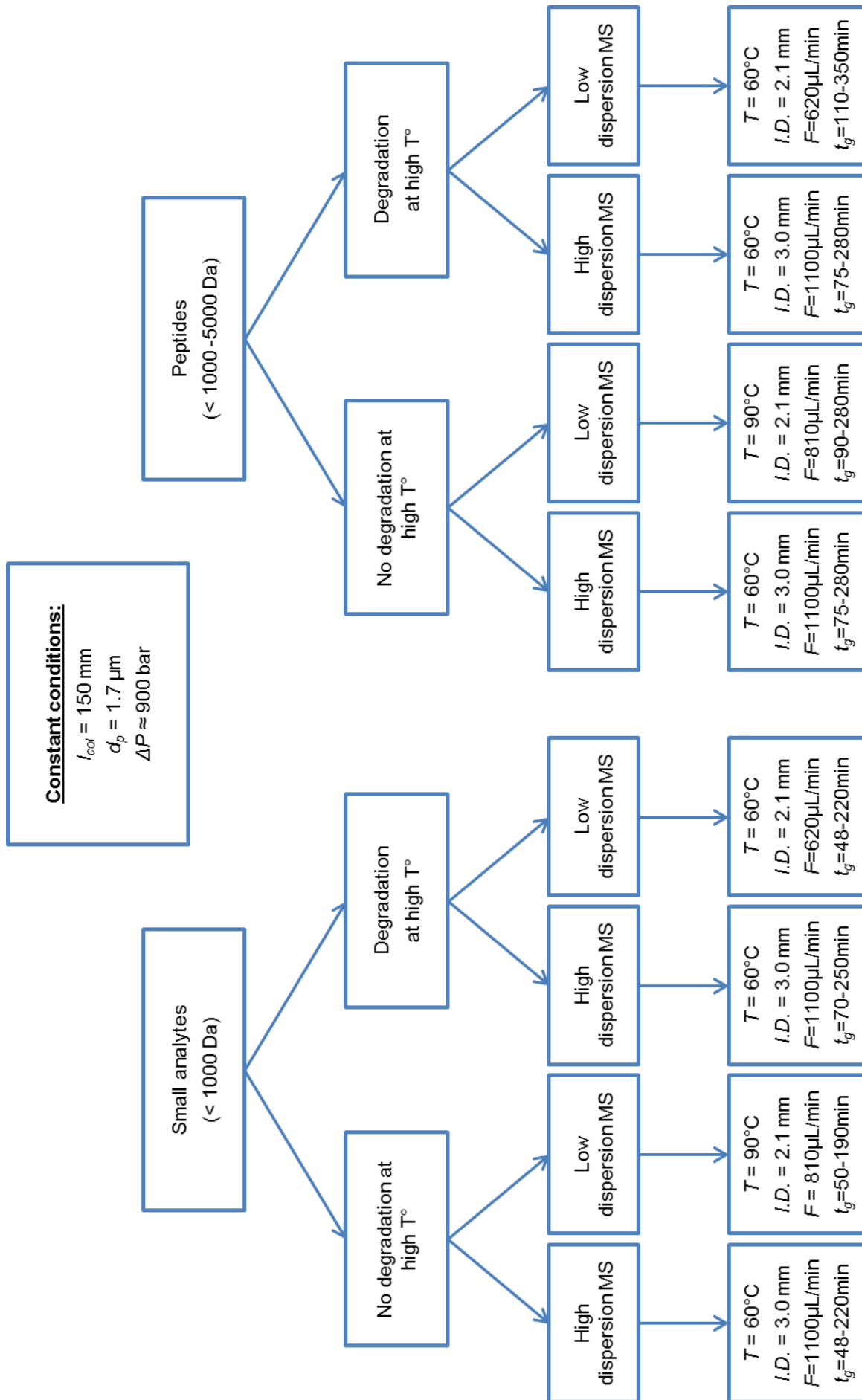
It was shown that a UHPLC strategy using 1.7 μm particles and a ΔP_{max} of 1000 bar systematically surpasses conventional HPLC methods using 3.5 μm particles and a ΔP_{max} of 400 bar, for the separation of both peptides and small molecules within a reasonable timeframe, similar to that generally used in conventional metabolite profiling studies. In contrast, the choice of mobile phase temperature, column internal diameter, mobile phase flow rate and gradient time are dictated by the nature of the analysed compounds, by the possibility of working at elevated temperatures without thermal degradation of the analytes and by the performance of the MS instrument.

As shown, the use of elevated mobile phase temperatures enhances chromatographic performance, particularly with large compounds such as peptides. In this context, the thermal stability of the analytes should be carefully evaluated. If the sample does not withstand temperatures of 90 °C, which are optimal for

maximising LC resolution on a 2.1 mm I.D. column, a mobile phase temperature of 60 °C should be selected for the first experiments.

The choice of the column internal diameter depends on the contribution of the analytical system to peak broadening and particularly to the dispersion caused by MS source. Indeed, depending on the flow rate, the UHPLC–MS dispersion of the platform used was 10- to 24-folds higher than that of UHPLC–UV. For detectors generating σ_{ext}^2 values higher than 30 μL^2 , the use of 3.0 mm I.D. columns provided better peak capacity. When σ_{ext}^2 was below 30 μL^2 , 2.1 mm I.D. columns should be preferentially used allowing a significant reduction in frictional heating effects as well as sample and mobile phase consumptions. This study also confirms that the best performance in UHPLC gradient mode was obtained with the highest mobile phase flow rate, which significantly decreased column dead time. For this reason, the flow rate corresponding to approximately 900 bar (equal to 90% of the system ΔP_{max}) was systematically selected, as reported in the decision tree presented in Figure III.13.

Finally, to achieve the best compromise between analysis time and peak capacity, the ideal duration for a 5–95% ACN gradient should be between 90 and 280 min for peptides and between 50 and 190 min for small molecules, using a 2.1 mm I.D. column at 90 °C. These gradient times provide 50% and 80% of the maximal peak capacity value in the selected conditions, respectively, that are approximately



560 and 900 for the peptides, and 350 and 560 for the small molecules. Gradient times are proposed as a range, since adaptation of this parameter, as well as gradient span, may be required according to the sample complexity. The gradient times ranges are also reported in Figure III.11.

The application of the rules deduced from this work to the profiling of *C. consors* venom

demonstrates that peak capacity values higher than 1100 can be experimentally obtained using gradient times below 100 min. UHPLC-QTOF-MS and MS/MS experiments performed in these optimised conditions enabled both the sequencing and deconvolution of peptides, despite the restricted peak width of the analytes. Compared to conventional LC-MS profiling studies usually performed in the field of venomics, these optimised conditions provided a peak capacity enhancement of 2- to 3-fold.

←

Figure III.13. Decisional tree to sum up the strategies for high resolution profiling of complex mixtures, considering the size of the compounds, compatibility of analytes with elevated mobile phase temperature and dispersion of the MS device. According to the rules established in this work, some parameters are kept constant (column length (l_{col}), particle diameter (d_p), and backpressure (ΔP)), while other parameters have to be adjusted (mobile phase temperature (T), column I.D. ($I.D.$), and flow rate (F)). Gradient times (t_g) are given as range whose minimal and maximal values provide 50% and 80% of the maximal peak capacity. This range is considered as the best compromise between high peak capacity and reasonable gradient time (see Section 3.6). The MS dispersion is considered low when $\sigma^2_{ext} < 30 \mu\text{L}^2$. The MS dispersion can be calculated using the method presented in Section 2.6.3.

Acknowledgements

The authors would like to express their deepest gratitude to the Government of New Caledonia, the French Navy, the French “Institut de Recherche pour le Développement” (IRD) and the Toxinomics Foundation for their constant support. The authors acknowledge the financial support of the European Commission. This study has been performed as part of the CONCO cone snail genome project for health (<http://www.conco.eu>) within the 6th Framework Program (LIFESCIHEALTH-6

Integrated Project LSHB-CT-2007, contract number 037592). J.L.W. is thankful to the Swiss National Science Foundation for supporting the profiling and metabolomics studies (Grant no. 205320-124667/1 and CRSII3 127187).

The authors wish also to thank Sabine Heinisch for the calculation of *S* values, and Gaétan Glauser for performing the QTOF-MS experiments.

References

- [1] E. Grata, D. Guillaume, G. Glauser, J. Boccard, P.A. Carrupt, J.L. Veuthey, S. Rudaz, J.L. Wolfender. Metabolite profiling of plant extracts by ultra-high-pressure liquid chromatography at elevated temperature coupled to time-of-flight mass spectrometry. *Journal of Chromatography A*, **2009**. 1216: 5660-5668.
- [2] J. Bruneton. *Pharmacognosie, Phytochimie, Plantes Médicinales*. 4th ed. **2009**, Paris, Tec & Doc/Lavoisier.
- [3] W.E. Müller. *St. John's Wort and Its Active Principles in Depression and Anxiety*. **2005**, Springer.
- [4] The *Conus* Biodiversity Website. [Access April 22, 2013]; Available from: <http://biology.burke.washington.edu/conus/index.php>.
- [5] R. Halaj, D.J. Craik. Conotoxins: natural product drug leads. *Natural Product Reports*, **2009**. 26: 526-536.
- [6] B.M. Olivera, J. Rivier, C. Clark, C.A. Ramilo, G.P. Corpuz, F.C. Abogadie, E.E. Mena, S.R. Woodward, D.R. Hillyard, L.J. Cruz. Diversity of *Conus* Neuropeptides. *Science*, **1990**. 249: 257-263.
- [7] H.G. Hocking, G.J. Gerwig, S. Dutertre, A. Violette, P. Favreau, R. Stöcklin, J.P. Kamerling, R. Boelens. Structure of the O-Glycosylated Conopeptide CcTx from *Conus consors* Venom. *Chemistry – A European Journal*, **2013**. 19: 870-879.
- [8] R. Stocklin, T. Vorherr. Future perspectives of venoms for drug discovery. *Pharmanufacturing: The international peptide review*, **2011**. Nov.: 21-24.
- [9] G.P. Miljanich. Ziconotide: Neuronal calcium channel blocker for treating severe chronic pain. *Current Medicinal Chemistry*, **2004**. 11: 3029-3040.
- [10] G. Glauser, E. Grata, S. Rudaz, J.L. Wolfender. High-resolution profiling of oxylipin-containing galactolipids in *Arabidopsis* extracts by ultra-performance liquid chromatography/time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry*, **2008**. 22: 3154-3160.
- [11] G. Glauser, J. Boccard, S. Rudaz, J.-L. Wolfender. Mass spectrometry-based metabolomics oriented by correlation analysis for wound-induced molecule discovery: identification of a novel jasmonate glucoside. *Phytochemical Analysis*, **2010**. 21: 95-101.
- [12] J.W. Allwood, R. Goodacre. An introduction to liquid chromatography-mass spectrometry instrumentation applied in plant metabolomic analyses. *Phytochemical Analysis*, **2010**. 21: 33-47.
- [13] S. Dutertre, D. Biass, R. Stocklin, P. Favreau. Dramatic intraspecimen variations within the injected venom of *Conus consors*: An unsuspected contribution to venom diversity. *Toxicon*, **2010**. 55: 1453-1462.
- [14] S.C. Wagstaff, L. Sanz, P. Juárez, R.A. Harrison, J.J. Calvete. Combined snake venomomics and venom gland transcriptomic analysis of the ocellated carpet viper, *Echis ocellatus*. *Journal of Proteomics*, **2009**. 71: 609-623.
- [15] T. Koecher, R. Swart, K. Mechtler. Ultra-High-Pressure RPLC Hyphenated to an LTQ-Orbitrap Velos Reveals a Linear Relation between Peak Capacity and Number of Identified Peptides. *Analytical Chemistry*, **2011**. 83: 2699-2704.
- [16] A. Menez, R. Stocklin, D. Mebs. 'Venomics' or: The venomous systems genome project. *Toxicon*, **2006**. 47: 255-259.
- [17] D. Biass, S. Dutertre, A. Gerbault, J.L. Menou, R. Offord, P. Favreau, R. Stocklin. Comparative proteomic study of the venom of the piscivorous cone snail *Conus consors*. *Journal of Proteomics*, **2009**. 72: 210-218.

- [18] P. Favreau, R. Stoecklin. Marine snail venoms: use and trends in receptor and channel neuropharmacology. *Current Opinion in Pharmacology*, **2009**. 9: 594-601.
- [19] P.J. Eugster, D. Guillarme, S. Rudaz, J.L. Veuthey, P.A. Carrupt, J.L. Wolfender. Ultra High Pressure Liquid Chromatography for Crude Plant Extract Profiling. *Journal of AOAC International*, **2011**. 94: 51-70.
- [20] D. Guillarme, E. Grata, G. Glauser, J.L. Wolfender, J.L. Veuthey, S. Rudaz. Some solutions to obtain very efficient separations in isocratic and gradient modes using small particles size and ultra-high pressure. *Journal of Chromatography A*, **2009**. 1216: 3232-3243.
- [21] P. Petersson, A. Frank, J. Heaton, M.R. Euerby. Maximizing peak capacity and separation speed in liquid chromatography. *Journal of Separation Science*, **2008**. 31: 2346-2357.
- [22] G.X. Xie, Y. Ni, M.M. Su, Y.Y. Zhang, A.H. Zhao, X.F. Gao, Z. Liu, P.G. Xiao, W. Jia. Application of ultra-performance LC-TOF MS metabolite profiling techniques to the analysis of medicinal *Panax* herbs. *Metabolomics*, **2008**. 4: 248-260.
- [23] J. Ruta, D. Guillarme, S. Rudaz, J.L. Veuthey. Comparison of columns packed with porous sub-2 μm particles and superficially porous sub-3 μm particles for peptide analysis at ambient and high temperature. *Journal of Separation Science*, **2010**. 33: 2465-2477.
- [24] P. Favreau, I. Krimm, F. Le Gall, M.J. Bobenrieth, H. Lamthanh, F. Bouet, D. Servent, J. Molgo, A. Menez, Y. Letourneux, J.M. Lancelin. Biochemical characterization and nuclear magnetic resonance structure of novel alpha-conotoxins isolated from the venom of *Conus consors*. *Biochemistry*, **1999**. 38: 6317-6326.
- [25] D.T.T. Nguyen, D. Guillarme, S. Rudaz, J.L. Veuthey. Chromatographic behaviour and comparison of column packed with sub-2 μm stationary phases in liquid chromatography. *Journal of Chromatography A*, **2006**. 1128: 105-113.
- [26] D. Guillarme. HPLC Calculator 3.0 software. [Access March 13, 2013]; Available from: <http://www.unige.ch/sciences/pharm/fanal/lcap/telechargement-en.htm>.
- [27] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *European Journal of Pharmaceutics and Biopharmaceutics*, **2008**. 68: 430-440.
- [28] W.T. Kok, U.A.T. Brinkman, R.W. Frei, H.B. Hanekamp, F. Nooitgedacht, H. Poppe. Use of conventional instrumentation with microbore column in high-performance liquid chromatography. *Journal of Chromatography*, **1982**. 237: 357-369.
- [29] L.R. Snyder, J.W. Dolan. The linear-solvent-strength model of gradient elution. *Advances in Chromatography*, **1998**. 38: 115-187.
- [30] L.R. Snyder, J.W. Dolan. *High-Performance Gradient Elution: The Practical Application of the Linear-Solvent-Strength Model*. 1st ed. Vol. 1. **2007**, New York, Wiley-Interscience.
- [31] J.M. Davis, J.C. Giddings. Statistical-Theory of Component Overlap in Multicomponent Chromatograms. *Analytical Chemistry*, **1983**. 55: 418-424.
- [32] U.D. Neue. Theory of peak capacity in gradient elution. *Journal of Chromatography A*, **2005**. 1079: 153-161.
- [33] S.A.C. Wren. Peak capacity in gradient ultra performance liquid chromatography (UPLC). *Journal of Pharmaceutical and Biomedical Analysis*, **2005**. 38: 337-343.
- [34] S. Heinisch, G. Desmet, D. Clicq, J.L. Rocca. Kinetic plot equations for evaluating the real performance of the combined use of high temperature and ultra-high pressure in liquid chromatography - Application to commercial instruments and 2.1 and 1 mm ID columns. *Journal of Chromatography A*, **2008**. 1203: 124-136.
- [35] U.D. Neue. Peak capacity in unidimensional chromatography. *Journal of Chromatography A*, **2008**. 1184: 107-130.
- [36] F. Gritti, G. Guiochon. Exact peak compression factor in linear gradient elution I. Theory. *Journal of Chromatography A*, **2008**. 1212: 35-40.

- [37] A. de Villiers, F. Lestremou, R. Szucs, S. Gélébart, F. David, P. Sandra. Evaluation of ultra performance liquid chromatography - Part I. Possibilities and limitations. *Journal of Chromatography A*, **2006**. 1127: 60-69.
- [38] D. Guillarme, S. Heinisch. Detection modes with high temperature liquid chromatography - A review. *Separation and Purification Reviews*, **2005**. 34: 181-216.
- [39] U.D. Neue, J.R. Mazzeo. A theoretical study of the optimization of gradients at elevated temperature. *Journal of Separation Science*, **2001**. 24: 921-929.
- [40] S. Heinisch, J.L. Rocca. Sense and nonsense of high-temperature liquid chromatography. *Journal of Chromatography A*, **2009**. 1216: 642-658.
- [41] X.L. Wang, W.E. Barber, P.W. Carr. A practical approach to maximizing peak capacity by using long columns packed with pellicular stationary phases for proteomic research. *Journal of Chromatography A*, **2006**. 1107: 139-151.
- [42] G.F. King. Venoms as a platform for human drugs: translating toxins into therapeutics. *Expert Opinion on Biological Therapy*, **2011**. 11: 1469-1484.
- [43] N.J. Saez, S. Senff, J.E. Jensen, S.Y. Er, V. Herzig, L.D. Rash, G.F. King. Spider-venom peptides as therapeutics. *Toxins*, **2010**. 2: 2851-71.
- [44] G.S. Shen, R.T. Layer, R.T. McCabe. Conopeptides: From deadly venoms to novel therapeutics. *Drug Discovery Today*, **2000**. 5: 98-106.
- [45] R. Stoecklin, P. Favreau, R. Thai, J. Pflugfelder, P. Bulet, D. Mebs. Structural identification by mass spectrometry of a novel antimicrobial peptide from the venom of the solitary bee *Osmia rufa* (Hymenoptera: Megachilidae). *Toxicon*, **2010**. 55: 20-27.
- [46] A. de Villiers, H. Lauer, R. Szucs, S. Goodall, P. Sandra. Influence of frictional heating on temperature gradients in ultra-high-pressure liquid chromatography on 2.1 mm I.D. columns. *Journal of Chromatography A*, **2006**. 1113: 84-91.
- [47] D.V. McCalley. Some practical comparisons of the efficiency and overloading behaviour of sub-2 μ m porous and sub-3 μ m shell particles in reversed-phase liquid chromatography. *Journal of Chromatography A*, **2011**. 1218: 2887-2897.
- [48] D. Guillarme, J. Schappler, S. Rudaz, J.-L. Veuthey. Coupling ultra-high-pressure liquid chromatography with mass spectrometry. *TrAC, Trends in Analytical Chemistry*, **2010**. 29: 15-27.
- [49] P. Favreau, O. Cheneval, L. Menin, S. Michalet, H. Gaertner, F. Principaud, R. Thai, A. Menez, P. Bulet, R. Stocklin. The venom of the snake genus *Atheris* contains a new class of peptides with clusters of histidine and glycine residues. *Rapid Communications in Mass Spectrometry*, **2007**. 21: 406-412.
- [50] Y. Terrat, D. Biass, S. Dutertre, P. Favreau, M. Remm, R. Stöcklin, D. Piquemal, F. Ducancel. High-resolution picture of a venom gland transcriptome: Case study with the marine snail *Conus consors*. *Toxicon*, **2012**. 59: 34-46.

Chapter IV - Ion Mobility Spectrometry: an Additional Separation Dimension

This chapter is based on a poster presented at the 2012 Fall Meeting of the Swiss Chemical Society in Zurich, Switzerland, and is the result of a collaboration with Dr Richard Knochenmuss from TOFWERK AG, Thun, Switzerland.

1. Introduction

In the previous chapters, it has been demonstrated that UHPLC is a well-adapted method for the separation of the constituents of raw matrices, thanks to its high resolution. LC retention is based on complex mechanisms that include partition and adsorption. In order to increase the resolution and/or to find orthogonal mechanisms for metabolite profiling, drift time ion mobility spectrometry (IMS) was evaluated

for its use as a new separation tool for the analysis of complex mixtures. Moreover, its capability of separating closely related isomers was also investigated. Indeed, IMS is based on separation mechanisms that are different from those involved in RP-LC and MS alone, and this technique is known to provide high separation efficiency.

2. Ion mobility spectrometry

Ion mobility spectrometry (IMS) is an analytical technique used to separate ionised molecules in the gas phase based on their mobility through a gas (the drift gas) [1]. Historically, IMS was first developed for military purposes, e.g. for the trace detection of explosives. It is used today in several analytical fields such as peptide identification, drug analysis and metabolomics [2]. Basically, an IMS instrument must perform the following processes: sample introduction, compound ionisation, ion separation, mass separation and ion detection. A typical instrument designed for drift time IMS is described below.

There are four types of ion mobility spectrometers available today that possess different properties: differential, travelling wave, aspiration and drift time IMS [1].

- Differential IMS (DMS) which is also called high-field asymmetric-waveform ion-mobility spectrometry (FAIMS), is able to specifically select ions with a given mobility, by applying different field strengths for different amounts of time [3]. DMS, as well as the aspiration technique, provides the highest sensitivity of all types of IMS instruments thanks to its continuous introduction of ions. This technology has been commercially available for a few years and integrated into MS instruments, for example by AB Sciex Company with the SelexION™ Technology.
- In the aspiration technique, the electric field is directed orthogonally to the gas flow. Ion mobility is measured as a function of the distance they travel through the buffer gas before impinging on an electrode.

- Drift time IMS (often simply called IMS) is the oldest IMS technology. It provides the highest resolving power, but suffers from a low sensitivity compared to other types of IMS instruments, because of its pulsed-introduction of ions. The drift time instrumentation is detailed below. In this study, the potential of this technology was investigated for the analysis of complex mixtures of natural origin in collaboration with the Swiss IMS instruments manufacturer TOFWERK.
- Travelling wave IMS is similar to the drift time technology except that a high wave form electrical field is applied to one segment of the tube. Ions are thus moved through the tube in pulses. Waters Company integrated this technology into a commercialised QTOF (Synapt G2 HDMS).

The drift time technology was used in this study because it is the most adapted of all to be used as a separative instrument thanks to its high resolving power. The drift time IMS instrumentation consists of three main parts: the ionisation source, the drift tube and the detector (Figure IV.1). Ionisation is usually performed by APPI, ESI or MALDI techniques. The latter two are actually the most frequently used for ionisation of small molecular weight (MW) molecules and polypeptides respectively. The separation occurs in tens of milliseconds in the drift tube that contains the drift gas, such as nitrogen, at ambient pressure. Separation in the drift tube depends on the chemical and physical interactions of the analytes with the drift gas. Practically, ions are separated based on their mass, charge, size and shape [4, 5]. When

needed, organic solvents may be added to the drift gas as gas phase modifiers to change the mobility of the analytes in the drift tube [3]. Another way to modify the selectivity is to add reagents to the sample, which create adducts with the analytes in the ESI source. Detectors were historically simple Faraday plates, but today IMS is often coupled to TOF-MS instruments, which are adapted to the speed of separation of IMS, thanks to their high acquisition rate.

Interestingly, drift time IMS instruments provide high resolution separations, with a number of theoretical plates higher than 100'000, similar to the values obtained in GC, but in analysis times lower than 0.1 s (Table IV.1) [6]. The effective analysis time is however often higher (e.g. minutes) in order to infuse sufficient amount of sample to allow the detection of minor peaks by summing a high number of scans.

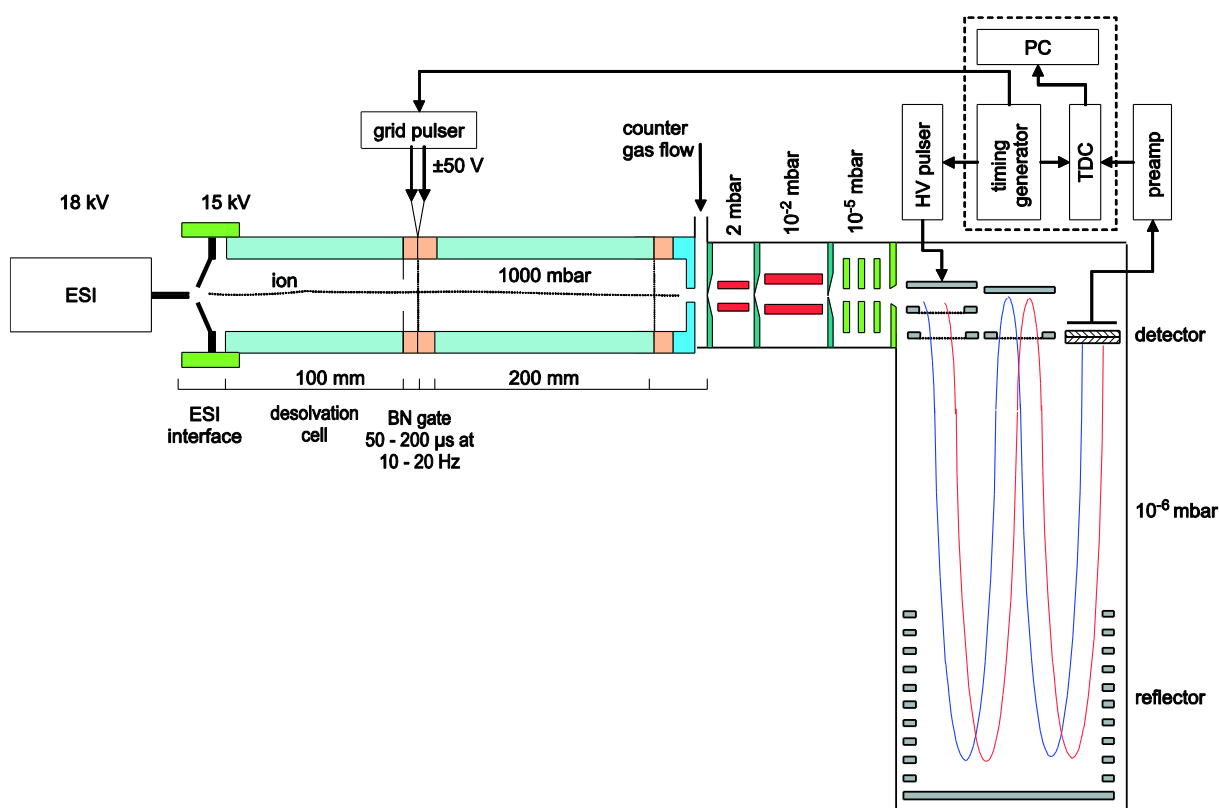


Figure IV.1. Classical scheme of a drift time IMS instrument. Analytes are ionised for example by an ESI source and enter the drift tube (blue bars) containing the drift gas and where a high electric field is applied. The detection is ensured by a TOF-MS instrument. With courtesy of Dr R. Knochenmuss, TOFWERK AG, Thun, Switzerland.

Table IV.1. Comparison of the characteristics of various separation techniques (adapted from [5-8]).

	HPLC	UHPLC	GC	CE	IMS
Number of theoretical plates	25'000	100'000	120'000	300'000	130'000
Efficiency (HETP)	10 μm	0.5 μm	400 μm	2 μm	2 μm
Resolution	65	160	145	230	150
Typical analysis time	30 min	6 min	20 min	10 min	1 - 10 min ^a or 50 ms ^b
Plates per second	14	330	100	500	216 - 2160 ^a Or 2'600'000 ^b

^a considering the total analysis time (that depends on the sensitivity and the dynamic range of the instrument)

^b considering the time per pulse

3. Evaluation of drift time IMS for the metabolite profiling of complex mixtures

The potential of IMS for separating complex mixtures was evaluated by analysing a *Ginkgo biloba* L. extract, a phytopharmaceutical containing flavonoids and terpenoids, among which are several isomers. The same extract was analysed using UHPLC using a long generic method of 30 min, frequently used for high resolution metabolite profiling (see Chapter V) to compare both techniques in terms of separation and possible applications in natural sample analysis. Both techniques allow a two-dimensional separation when coupled to MS (drift time $\times m/z$ and RT $\times m/z$, respectively). In addition, the separation of two specific flavonoids was studied using both techniques.

3.1. Metabolite profiling of a *Ginkgo biloba* extract by IMS- and UHPLC-TOF-MS

The potential of drift time IMS for metabolite profiling of complex natural samples was studied based on the investigation of the mechanisms of separation involved and on the number of features detected during the analysis of a standardised *Ginkgo biloba* extract.

3.1.1. Methods

The metabolite profiling a *Ginkgo biloba* standardised extract was performed based on two experiments. Firstly an IMS analysis was conducted using TOF-MS detection (IMS-TOF-

MS), and secondly a UHPLC-TOF-MS profiling based on the method presented in Chapter III and V. The IMS analysis was carried out in positive ESI ionisation mode over 10 min using nitrogen as drift gas. Only 1.0 μg of extract was used. Figure IV.2A and B show the two- and three-dimensional IMS-TOF-MS plot of drift time $\times m/z$ and of drift time $\times m/z \times$ intensity. The UHPLC-TOF-MS separation was performed by injecting 10.0 μg of sample on a 150 \times 2.1 mm, 1.7 μm C₁₈ column using a 30 min 5-95% ACN gradient. Figure IV.2C shows the three-dimensional UHPLC-TOF-MS plot of retention time $\times m/z \times$ intensity.

The number of detected features was determined in the IMS-TOF-MS analysis using dedicated software to the instrument that allowed automatic peak detection. The number of detected features by UHPLC-TOF-MS was evaluated using standard protocols applied for metabolomics studies (see Chapter III).

3.1.2. Results

The separation obtained by drift time IMS-TOF-MS (Figure IV.2A and B) shows a clear correlation between drift time and m/z . The ions that are not placed on a single line on the drift time $\times m/z$ plot indicated that both separation dimensions were not equivalent, and that IMS separated compounds based on a mechanism that was not only dependent on their MW. By definition, MS separates ions according to their m/z ratio,

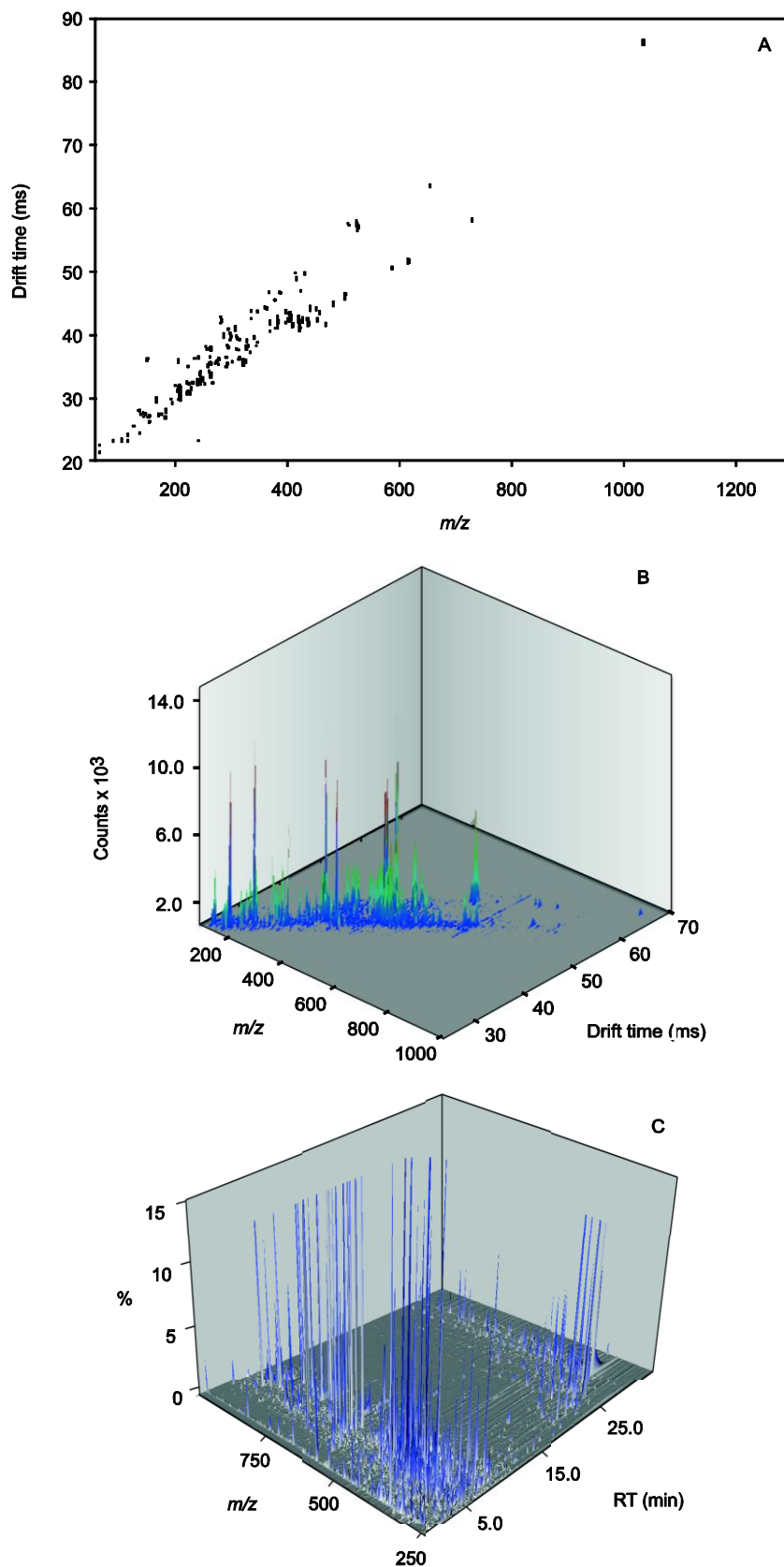


Figure IV.2. (A, B) 2D and 3D plots of a *Ginkgo biloba* standardised extract 10 min profiling using direct infusion of 1.0 μg of sample in ESI-PI-IMS-TOF-MS after blank subtraction, using nitrogen as drift gas. (C) UHPLC-ESI-PI-TOF-MS separation of 10.0 μg of the same sample, using a 150 x 2.1 mm, 1.7 μm C₁₈ column with a 5-95% ACN gradient in 30 min.

which is related to the MW in ESI ionisation. Contrarily, IMS separates compounds based on their charge, size and shape; drift time IMS specifically separates compounds based on their collisional cross-sections [1, 4, 5].

Experimentally, this difference of mechanism between IMS and MS provides a significant difference of compound separation. The combination of both separation dimensions is thus more advantageous than the use of MS only, since several ions possess the same m/z value but different drift time values. MS is however rarely used without hyphenation with LC for NP profiling studies. Still, LC is often not sufficient to provide a satisfactory separation of two closely related isomers, such as stereoisomers. The separation of stereoisomers using IMS is investigated in the next sub-chapter.

The number of detected features using LC-MS and IMS was compared and discussed. The aim was not to accurately determine the number of detected peaks, because it would require a careful optimisation of all UHPLC, IMS and TOF-MS parameters, as well as the use of the same TOF instrument, but to determine if both approaches provide similar numbers of detected features. Thus, 363 features were detected using

IMS and 1064 using UHPLC. These numbers are in the same order of magnitude. Both techniques are thus well adapted to the analysis of complex natural samples, that often contain hundreds or thousands of constituents [9].

3.2. Separation of closely related stereoisomers by drift time IMS and UHPLC

Stereoisomers are often present in natural samples and their separation is often challenging but mandatory, because they are not further separated in the MS dimension. The separation and detection of all stereoisomers in a complex sample is highly important in NP research, because they often possess different or opposite bioactivities [10]. In order to evaluate the separation efficiency obtained using IMS and LC, two stereoisomeric flavonoids present in *G. biloba* were selected, namely isoquercitrin (quercetin-3-*O*-glucoside, Figure IV.3A) and hyperoside (quercetin-3-*O*-galactoside, Figure IV.3B).

The stereoisomers were analysed separately by ESI-IMS-TOF-MS in PI mode. Solutions at 0.5 $\mu\text{g/mL}$ (hyperoside) and 1.0 $\mu\text{g/mL}$ (isoquercitrin) provided adequate intensities. Their analysis by

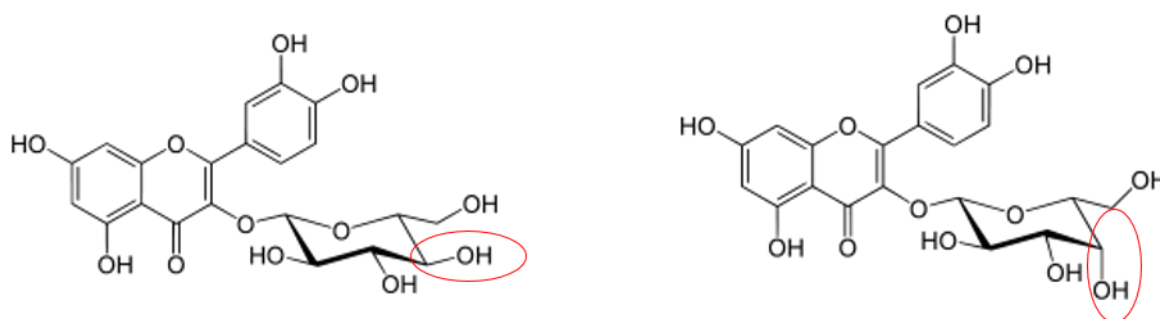


Figure IV.3. Structures of the two studied stereoisomers, isoquercitrin (quercetin-3-*O*-glucoside, A) and hyperoside (quercetin-3-*O*-galactoside, B).

direct IMS infusion provided similar drift times (see Figure IV.4A). Their separation was however strongly improved by addition of Li^+ which forms adducts with the analytes (Figure IV.4B). This reagent interacts probably differently with the sugars of the two flavonoids. The injection of a (1:3) mixture confirmed their clear separation (Figure IV.4C). The separation time in IMS was lower than 50 ms and both compounds were separated by a difference of drift time of 0.4 ms. The total analysis time was fixed to 10 min to provide a satisfactory sensitivity.

The UHPLC separation of the two stereoisomers was first performed using a fast generic method routinely used for metabolomics studies, based

on a 4.0 min 5-95% ACN gradient on a 50 x 1.0 mm, 1.7 μm C_{18} column. This method did not enable the separation of the analytes (Figure IV.4D). Another method – optimised for high resolution metabolite profiling – was then used, based on a 30.0 min 5-95% ACN gradient on a 150 x 2.1 mm, 1.7 μm C_{18} column. Only a slight separation was observed (Figure IV.4E), but the gradient had to be modified to 5-30% to provide a satisfactory separation of the two flavonoids (Figure IV.4F).

Finally, a standardised *Ginkgo biloba* (50 $\mu\text{g}/\text{mL}$) extract was analysed by ESI-IMS-TOF-MS in positive ionisation mode with the addition of Li^+ . The resulting IMS-MS two-dimensional plot is

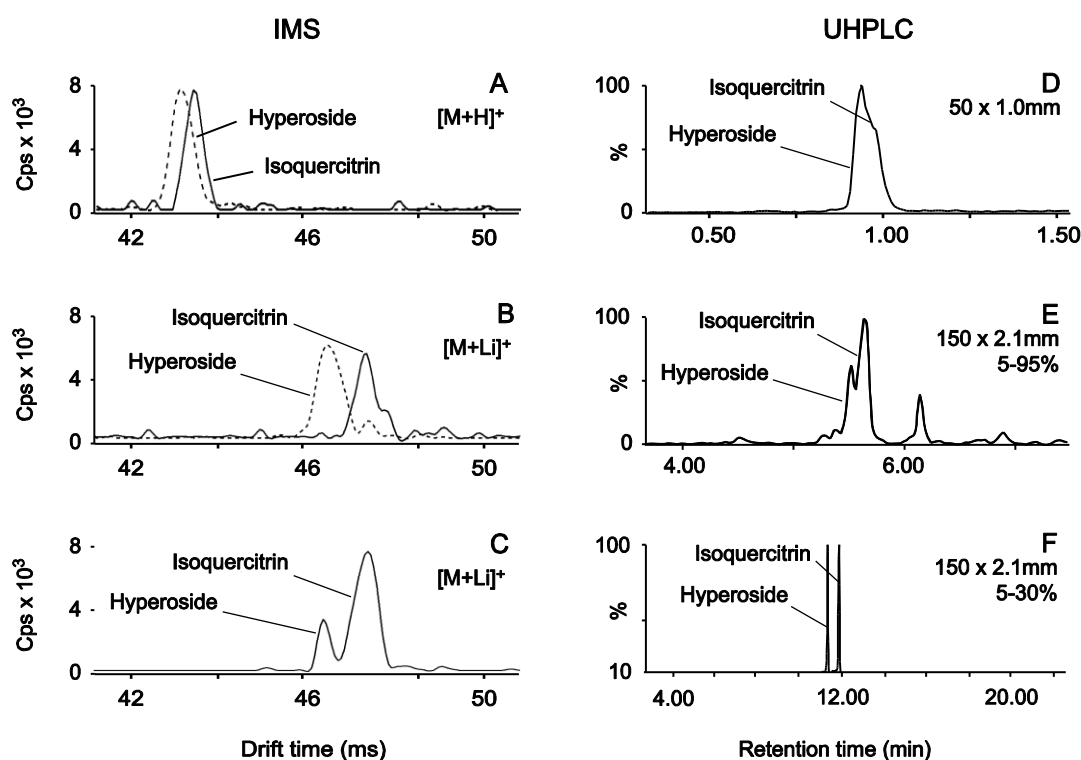


Figure IV.4. Separations of hyperoside and isoquercitrin using IMS (A-C) and UHPLC (D-F). Separate IMS injections of the flavonoids without (A) and with (B) addition of Li^+ (A and B respectively). IMS injection of the 1:3 mix of the flavonoids with addition of Li^+ (C). UHPLC-MS analyses of the flavonoids using a 4 min generic gradient on a 50 mm column (D), a long 30 min generic gradient on a 150 mm column (E), and an optimised 5-30% acetonitrile 30 min gradient on a 150 mm column (F).

displayed in Figure IV.5, where all detected features are shown by green marks. The intensity of the colour reflects the amount of detected ions. Red dots represent features highlighted by a finding algorithm. Both stereoisomers (isoquercitrin and hyperoside) were found

among the detected features and were nicely separated (see zoom in Figure IV.5). This confirmed that both stereoisomers are also well separated in a complex natural sample such as a crude plant extract.

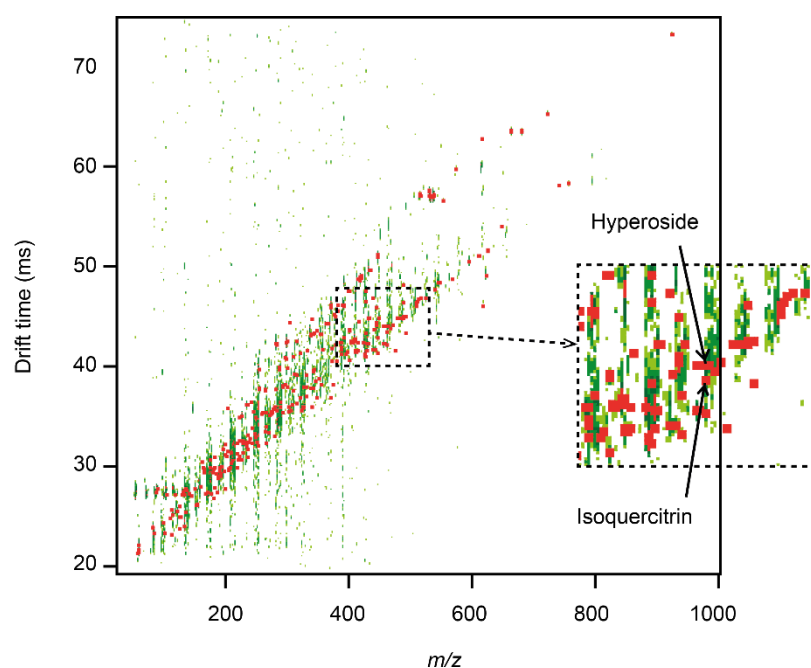


Figure IV.5. IMS-TOF-MS 2D plot of the injection of a *Ginkgo biloba* extract in positive ESI ionisation mode after blank subtraction. All detected ions are indicated by green marks. The intensity of the colour is related to the amount of detected ions. An algorithm for automated feature detection allowed metabolite finding (red dots). Hyperoside and isoquercitrin are highlighted in the inset.

4. Conclusion

In this preliminary study, the potential of drift time IMS as an additional complementary tool for high resolution profiling of complex natural samples has been investigated.

The metabolite profiling experiments on the *Ginkgo biloba* extract showed that IMS is well-adapted to this type of analysis. The IMS ion map (Figure IV.2) indicates that a significant proportion of ions is not placed on the diagonal of the plot drift time $\times m/z$ which clearly indicates the additional separation dimension provided by IMS compared to single HR-MS infusion. The ion MAP obtained is rich and the number of features is high in both experiments: 1064 using UHPLC and 363 using IMS, and could be further optimised by changing some parameters of the LC and MS systems. The IMS separation mechanism is different from that of LC and MS: the drift time depends on the collision cross-section of the molecule [4, 5]. This technique is thus able to separate closely related isomers that are slightly or not separated in LC and MS dimensions respectively.

Besides the separation power of the techniques, other parameters strongly varied between UHPLC and IMS separations, such as the analysis time. The optimised UHPLC profiling lasted 50 min (30 min of gradient plus 20 min for wash and reconditioning), while the IMS infusion time was fixed to 10 min. This value could even be lowered since it mainly depends on the sensitivity of the TOF-MS detector and on the concentration of the analytes. Moreover, the amount of sample used was very low, and was approximately 10 times lower using IMS compared to UHPLC-TOF-MS. When needed, IMS selectivity may be modified

by gas phase modifiers such as acetone, isopropanol, or methanol [3], or by adding ions such Li^+ and Na^+ to the sample to product adducts. Ion mobility may thus be considered as an orthogonal dimension for separations of complex mixtures. Finally, IMS possess two interesting features compared to LC in the frame of the profiling of complex mixtures. First, the high reproducibility of drift times is interesting for differential metabolomics applications. Second, contaminations of the drift tube should theoretically rarely occur, while this is a recurrent issue in LC columns. Both features couldn't however be practically tested in the frame of complex natural samples analysis. Based on this, IMS seems to be a promising separation method in NP research, mainly adapted for fingerprinting, with applications in different fields such as quality control and metabolomics (see Chapter II). Indeed, both applications need fast and efficient separation.

However, although its features are highly interesting, IMS is still not ready for routine use in NP research. Some of the main issues have to be solved first, such as the complexity of the data processing and the potential ionisation suppression effects because of the absence of LC prior to the ionisation. Many of these issues could be solved by LC-IMS-MS coupling. It seems to be a very promising method for tri-dimensional LC \times IM \times TOF-MS high resolution profiling of complex natural extracts. Its practical implementation is in progress [11], but the compatibility between the thin LC peaks and the IMS analysis time is still challenging and the data are difficult to exploit. The implementation of drift tubes in some commercialised MS instruments is a first

encouraging step. In my opinion, a fast-UHPLC-IM-TOF-MS setup could be interesting for fast LC analysis with high resolution.

Finally, thanks to the mechanism involved in ion mobility separations, the retention of a given compound may be easily calculated based on its structure. A correlation has been shown between the drift time of moderately flexible molecules

and their collisional cross-sectional area [12]. In another work, a model for the drift time prediction was built using several molecular descriptors such as chi path and volume parameters. The predictive ability of the model was sufficient for its use in dereplication studies [13]. This can represent an additional dimension to LC retention prediction for dereplication that we have investigated in Chapter VII.

5. References

- [1] A.B. Kanu, P. Dwivedi, M. Tam, L. Matz, H.H. Hill. Ion mobility–mass spectrometry. *Journal of Mass Spectrometry*, **2008**. 43: 1-22.
- [2] C. Laphorn, F. Pullen, B.Z. Chowdhry. Ion mobility spectrometry-mass spectrometry (IMS-MS) of small molecules: Separating and assigning structures to ions. *Mass Spectrometry Reviews*, **2013**. 32: 43-71.
- [3] E. Varesio, J.C. Le Blanc, G. Hopfgartner. Real-time 2D separation by LC × differential ion mobility hyphenated to mass spectrometry. *Analytical and Bioanalytical Chemistry*, **2012**. 402: 2555-2564.
- [4] K. Kaplan, S. Graf, C. Tanner, M. Gonin, K. Fuhrer, R. Knochenmuss, P. Dwivedi, H.H. Hill, Jr. Resistive Glass IM-TOFMS. *Analytical Chemistry*, **2010**. 82: 9336-9343.
- [5] P. Dwivedi, A.J. Schultz, H.H. Hill, Jr. Metabolic profiling of human blood by high-resolution ion mobility mass spectrometry (IM-MS). *International Journal of Mass Spectrometry*, **2010**. 298: 78-90.
- [6] G.R. Asbury, H.H. Hill. Evaluation of ultrahigh resolution ion mobility spectrometry as an analytical separation device in chromatographic terms. *Journal of Microcolumn Separations*, **2000**. 12: 172-178.
- [7] A. de Villiers, F. Lestremou, R. Szucs, S. Gélébart, F. David, P. Sandra. Evaluation of ultra performance liquid chromatography - Part I. Possibilities and limitations. *Journal of Chromatography A*, **2006**. 1127: 60-69.
- [8] D. Guilleme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *European Journal of Pharmaceutics and Biopharmaceutics*, **2008**. 68: 430-440.
- [9] J.L. Wolfender, G. Glauser, J. Boccard, S. Rudaz. MS-based Plant Metabolomic Approaches for Biomarker Discovery. *Natural Product Communications*, **2009**. 4: 1417-1430.
- [10] J.C. Leffingwell. Chirality & bioactivity I.: pharmacology. *Leffingwell Reports*, **2003**. 3: 1-27.
- [11] S.J. Valentine, X. Liu, M.D. Plasencia, A.E. Hilderbrand, R.T. Kurulugama, S.L. Koeniger, D.E. Clemmer. Developing liquid chromatography ion mobility mass spectrometry techniques. *Expert Review of Proteomics*, **2005**. 2: 553-565.
- [12] N.L. Zakharova, C.L. Crawford, B.C. Hauck, J.K. Quinton, W.F. Seims, H.H. Hill Jr, A.E. Clark. An assessment of computational methods for obtaining structural information of moderately flexible biomolecules from ion mobility spectrometry. *Journal of the American Society for Mass Spectrometry*, **2012**. 23: 792-805.
- [13] L.C. Menikarachchi, S. Cawley, D.W. Hill, L.M. Hall, L. Hall, S. Lai, J. Wilder, D.F. Grant. MolFind: A Software Package Enabling HPLC/MS-Based Identification of Unknown Chemical Structures. *Analytical Chemistry*, **2012**. 84: 9388-9394.

Chapter V – Rational Approach for LC-MS

Online Dereplication

This chapter is based on a book chapter entitled *Strategies in Biomarker Discovery. Peak annotation by MS and targeted LC-MS micro-fractionation for de novo structure identification by micro-NMR*, published in *Metabolomics Tools for Natural Products Discoveries*, of the *Methods in Molecular Biology* series.

Foreword

The next three chapters focus on the online dereplication and *de novo* structure elucidation of NPs using LC-MS. The methods and tools used for this purpose are presented in this foreword.

Since the implementation of LC-(PDA)-MS systems in laboratories dealing with NP research, this two-dimensional (RT x m/z) technique was used to get as much online information as possible on the compounds present in natural samples for identification without need of isolation. Typically, such a procedure is performed on biomarkers that are highlighted by metabolomics (see Chapter II) or for the early identification of metabolites in the frame of bioactivity guided isolation studies (see Chapter I). The first step usually aims at determining if the compound(s) of interest are already cited in the literature, to avoid unnecessary efforts on a well-known compound. This procedure is called dereplication. If the compound seems to be unknown, the second step aims at determining their structures on additional LC-MS/MS, MSⁿ, LC-NMR or microNMR data, without tedious isolation procedures. This is known as the *de novo* structure elucidation. Although it is mandatory to further isolate the pure compound to confirm its identity or to perform biological assays, the rapid

online or rapid at-line identification of biomarkers saves time and efforts.

The use of MS/MS instruments provides useful structural information, as illustrated by the routine use of database matching with GC-MS/MS experiments. However the use of MS/MS instruments in conjunction with LC is very limited because the LC-ESI-MS/MS experiments provide instrument-dependent fragmentation pattern that needs in-house databases. The development of such LC-MS/MS databases is therefore only possible for big companies, while smaller companies and academicians have to use other strategies. Therefore, modern dereplication procedures and *de novo* structure elucidation are more and more frequently based on LC-HR-MS methods that are the first of several steps providing ideally the identification of a molecule. Since the beginning of the 21st century, there were great efforts to develop efficient tools for online dereplication and *de novo* structure elucidation.

This foreword presents the common steps of a typical dereplication procedure and some recently developed tools used for this purpose. A comprehensive dereplication method is

Dereplication is the process of identifying known metabolites in a sample from online data, to avoid focusing on compounds that were already studied.

***De novo* structure elucidation (or identification)** is the complete identification procedure of a metabolite based on online data and on further experiments performed on the pure isolated compound.

described in the rest of the chapter.

Typical dereplication procedure

Almost all HR-MS-based dereplication procedures are multistep procedures based on

both the m/z and the isotopic information of the MS spectra, as well as several of the following information: UV spectra, MS/MS fragmentation, heuristic rules, NPs databases, chemotaxonomic information, LC retention, and IMS drift time [1]. Figure V.1 graphically presents the different steps

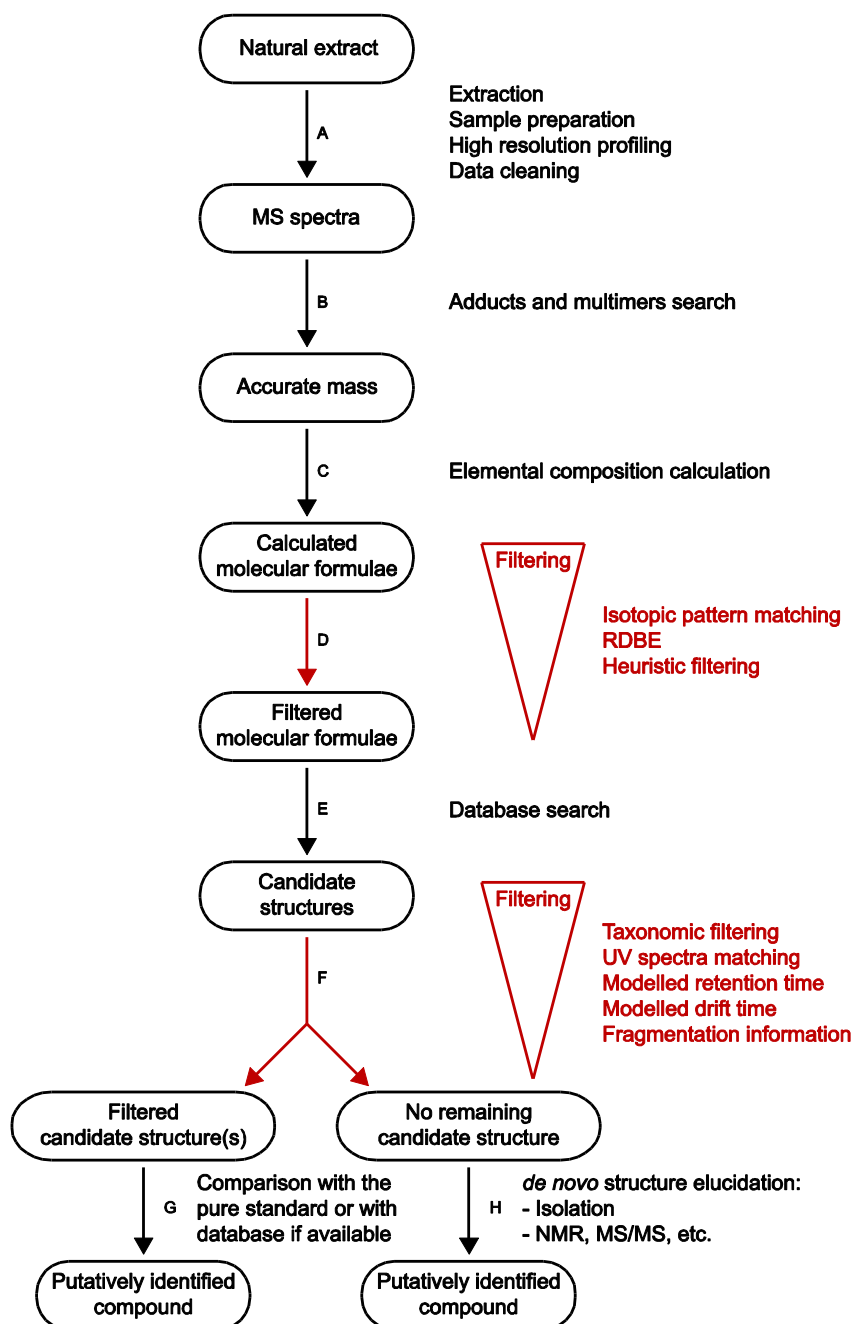


Figure V.1. Schematic representation of a typical modern dereplication procedure of NPs. The letters on the left side correspond to the paragraphs below that detail each step of the procedure.

involved in a typical natural extract dereplication procedure which is detailed below.

A – LC-MS analysis. Two main requirements have to be fulfilled to perform efficient dereplication procedures. Firstly, high mass and spectral accuracies (spectral accuracy is the accurate measurement of the isotopic pattern) are required, and secondly, high quality and clean MS spectra, *i.e.* without interfering peaks, are mandatory. High mass accuracy is obtained by the use of HR-MS instruments, such as (Q)TOF-MS, orbitrap or FT-ICR instruments, while interfering peaks maybe be avoided by using high

resolution LC separations prior to MS detection (see Chapters II and III). Figure V.2 displays a typical high resolution metabolite profiling of an *Arabidopsis thaliana* crude extract. The peaks are well resolved, but the two-dimensional map shows the presence of many adducts and/or multimers for the analytes (see below).

Before LC-MS data handling, background subtraction is usually applied by subtracting a blank chromatogram. This step is of utmost importance for profiling methods, but may be avoided for metabolomics applications, since only differences between samples are searched.

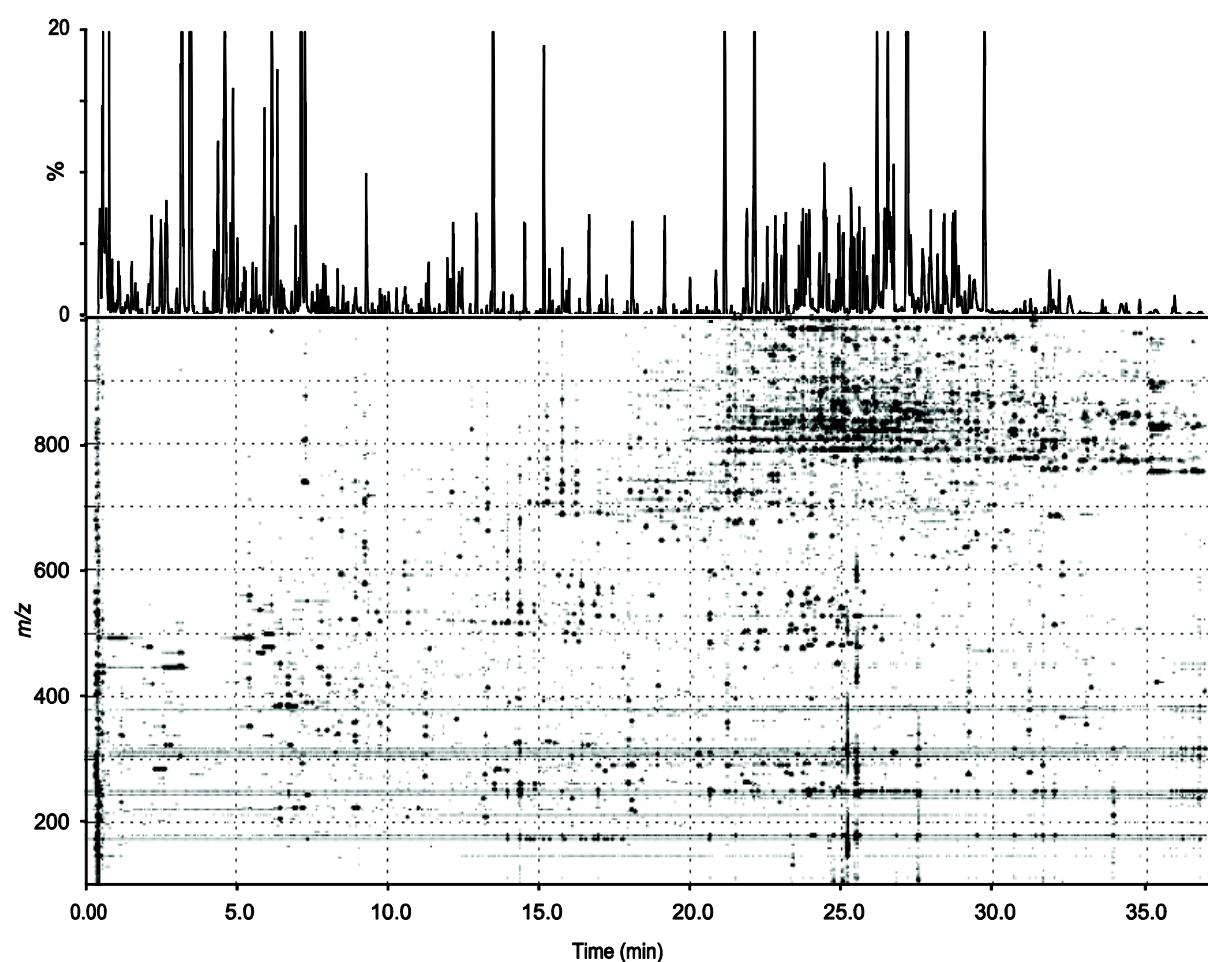


Figure V.2. High resolution UHPLC-NI-TOF-MS profiling of an *Arabidopsis thaliana* extract using a 40 min gradient separation on a 150 x 2.1 mm, 1.7 μm Acquity BEH C₁₈ column. The upper part displays the chromatogram and the lower part shows the corresponding two-dimensional map (RT x *m/z*).

Noise reduction algorithms are also often applied to clean the chromatograms and mass spectra, such as the component detection algorithm (CODA) algorithm [2]. Most of these procedure considerably reduce noise but cannot totally eliminate all interfering traces. Consequently, the baseline of LC-MS chromatograms is lower and the peaks are better highlighted, but 2D ions maps usually remain similar as before the data cleaning, *i.e.* noise traces remain visible, as illustrated in Figure V.2. This is however not an issue for the further LC-MS data processing, since these ions do not possess the typical shape of LC peaks and therefore are eliminated during the peak picking procedure [3].

B – Accurate mass determination. The next step aims at determining the accurate mass of the analyte(s) of interest. Indeed, MS spectra usually display several peaks, including the $[M+H]^+$ or $[M-H]^-$ adducts, and other adducts such as formate, ammonium, sodium, etc., as well as fragments of the ion, dimers or multimers (*i.e.* clusters of ions that are frequently created in the ESI source, often described as $[2M\pm H]^\pm$ or $[xM\pm H]^\pm$ and other peaks that are not related to the compound of interest (Figure V.3). Table V.2 from the book chapter below lists several adducts commonly found in ESI. The comparison of both PI and NI spectra, if available, enhances the efficiency of this step by comparison of the main m/z ions. The accurate mass determination procedure is illustrated in Figure V.3, for an unknown metabolite (RT = 19.20) in an UHPLC-TOF-MS profiling of *Ginkgo biloba*. The study of the peaks present in the PI spectrum (Figure V.3A) showed a dimer and a sodium adduct of an ion at m/z 583.1240. These three ions are probably related since their LC-MS traces showed a perfect coelutions (Figure V.3B). The NI spectrum similarly displayed a formate adduct and a dimer of an ion at m/z 581.1084 (Figure V.3C). The LC-MS traces of all three ions revealed also their perfect coelutions (Figure V.3D). Based on the convergent information from both spectra, the

pseudomolecular ions were determined with a high level of confidence as $[M+H]^+ = m/z$ 583.1240 and $[M-H]^- = m/z$ 581.1084. Indeed, this metabolite was further identified as amentoflavone ($C_{32}H_{22}O_{11}$, MW = 586), a known metabolite of *G. biloba*. Unfortunately, there are currently no tools to perform this task automatically [4], although there are software, such as MZmine [3, 5] and CAMERA [6], which are able to highlight typical mass differences between two peaks. This task is however not performed automatically and needs manual supervision. The CAMERA software package [6] provides an interesting tool for PI-NI spectra comparison, useful to assess the mass of the metabolite by searching for corresponding $[M+H]^+$ and $[M-H]^-$.

C – Molecular formula calculation. Molecular formulae are calculated from the selected m/z ion. For this, elemental composition calculators are integrated in almost every MS software dedicated to given HR-MS instruments. The number of potential molecular formulae depends on two factors: the mass accuracy of the MS instrument and the MW of the analytes. The mass accuracy of the HR-MS instrument is a critical feature and has to be as high as possible and reduce the number of molecular formulae for a given m/z . Today, instruments used in routine ensure mass accuracy ranging from 10 to 2 ppm, even lower than 1 ppm for the Maxis 4G from Bruker Company. However, such a high mass accuracy is not sufficient to determine unambiguously the unique chemical formula for a given analyte. For example, the $[M+H]^+$ ion 781.4374 of digoxin measured with an accuracy of 1 ppm will match with 4 different molecular formulae when the search is restricted to compounds containing C, H, O and N elements only. Moreover, the number of possible molecular formulae dramatically increases with higher MW. This is illustrated by the red plot in Figure V.4, displaying the number of possible

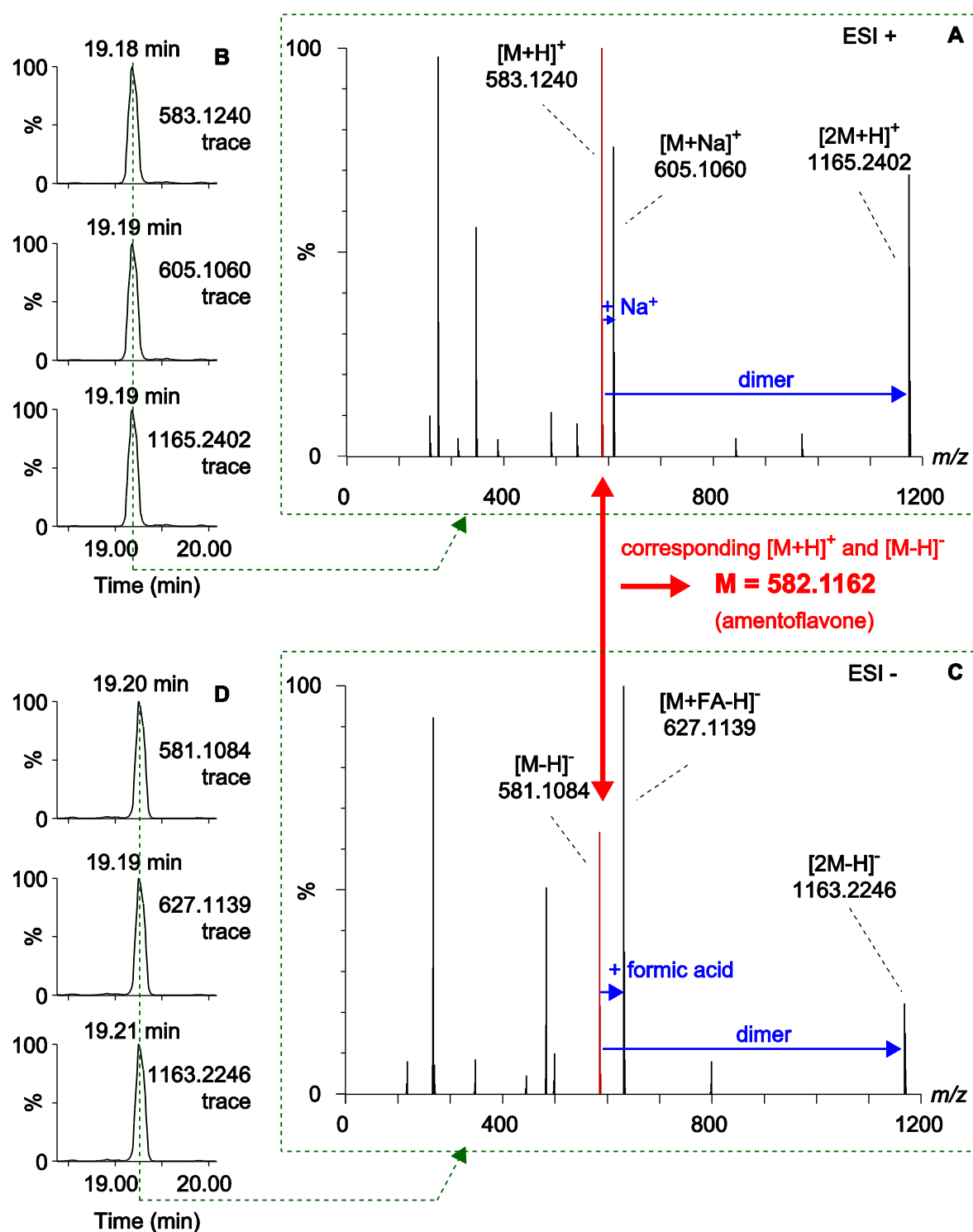


Figure V.3. Determination of the mass of an unknown analyte with RT = 19.20 min from a high resolution profiling of *Ginkgo biloba*, using UHPLC-TOF-MS in both PI and NI modes. (A) PI spectrum with a dimer and a sodium adduct of an ion at m/z 583.1240 that (B) coeluted perfectly. (C) NI spectrum with a formate adduct and a dimer of an ion at m/z 581.1084 with (D) same RT. The pseudomolecular ions provided convergent information that allowed the determination of the accurate mass of the analyte and its identification after dereplication as amentoflavone ($C_{32}H_{22}O_{11}$, MW = 586).

formulae for 48 compounds vs. the m/z value, using C, H, N, O, P, and S elements and allowing up to one of the Br, Cl, Fe, and Mg elements, for a 5 ppm error [7].

D – Filtering of putative molecular formula.

Additional orthogonal filters are required to reduce the number of potential elemental compositions. The most efficient method is based on the estimation of the isotopic pattern of the analyte. The software calculates the theoretical isotope pattern of all potential molecular formulae proposed for a given experimental measured mass. All isotopic patterns are matched and ranked against the experimental spectrum according to a matching score. For example, the search for putative molecular formulae corresponding to an ion of m/z 587.4342 using a 5 ppm range and taking into account C, H, and O atoms provides two putative molecular formulae: $C_{29}H_{63}O_{11}$ (m/z 587.4370) and $C_{36}H_{59}O_6$ (m/z 587.4312). Therefore, the correct molecular formula cannot be determined based on the mass information only (this 5 ppm range used corresponds to the mass accuracy

provided by the majority of the MS used in NP research today, although the accuracy of modern HR-MS instruments is usually below this value). The correct molecular formula can however be determined using the isotopic pattern matching, since the isotopic pattern of both species is completely different, as shown in Figure V.5, with relative intensities of 34% and 41% for the species containing one ^{13}C atom compared to that without ^{13}C atom. Kind and Fiehn showed that MS instruments with 3 ppm mass accuracy and 2% spectral accuracy outperform instruments with less than 1 ppm mass accuracy used without isotope information in the calculation of molecular formulae [8]. Many authors agree that this step is crucial for metabolite identification [7]. Moreover, isotopic information is usually more reliable than mass measurements (provided that peak intensities do not exceed the dynamic range of the instrument). The reason is that this measure is more stable over time than the mass measurement that is sensitive to external factors such as variations in temperature, even with a correction provided by a reference compound that is continuously

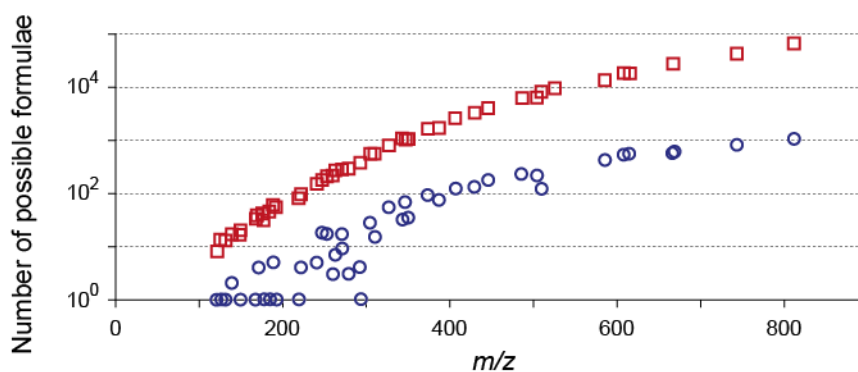


Figure V.4. Plot of the number of possible formulae vs. the m/z value for 48 compounds highlighted by a metabolomics study (red squares), using C, H, N, O, P, and S elements in the calculation, as well as up to one of each of the following elements with characteristic isotopic features, that is Br, Cl, Fe, and Mg, and a 5 ppm error. The number of possible formulae is dramatically reduced thanks to a chemical formula prediction tool based on heuristic rules and isotope pattern matching, among others (blue circle). Circles that are on the horizontal 10^0 line correspond to compounds with one possible formula only. Adapted with permission from [7]. Copyright 2012 American Chemical Society.

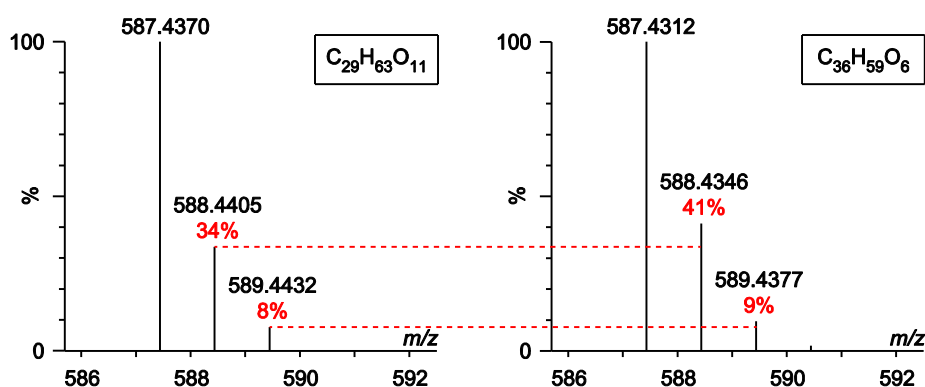


Figure V.5. MS spectra of the ions m/z 587.4370 and of m/z 587.4312 in positive ionisation mode. The mass difference is lower than 10 ppm, and a clear molecular formula determination may be impossible using the majority of the HR-MS used in NP laboratories today that provide usually a 5 ppm mass accuracy. The isotopic patterns are however completely different, and provide a clear differentiation of both ions.

injected. Another existing parameter to discriminate false candidates is the ring double bond equivalents (RDBE) value, which estimates the number of rings and unsaturated bonds in a molecule [9]. This value is however of rather limited use in the determination of molecular formulae since it produces uncertain results, particularly in the presence of heteroatoms [10]. Therefore, this tool is often not used in such procedures, and is for example replaced by the study of the valence values (see below).

The above-mentioned methods (molecular formulae calculation and isotopic pattern comparison) are routinely used and integrated in most of the modern MS spectra processing software provided by the main MS instruments vendors (e.g. Xcalibur/ Mass Frontier from Thermo Fisher Scientific, SmartFormula3D from Bruker Daltonics, MassLynx from Waters, or PeakView from AB SCIEX). Independent software, such as the Sirius tool, developed to determine the formula of unknown metabolites from isotopic patterns has been implemented in an

independent java-based program able to import LC-MS data [11].

At this stage, there are generally still several candidate molecular formulae for a given LC peak of interest, and it is thus necessary to apply additional filters to reduce the number of possibilities. For this, Fiehn et al. developed in 2007 the “Seven Golden Rules” that combine several filters to reduce the number of putative molecular formulae [10]. These rules were developed based on the statistical study of tens of thousands of compounds and are designed to annotate peaks from HR-MS spectra to provide elemental compositions for compounds containing C, H, N, S, O, P, F, Cl and Br up to 2 kDa. The seven heuristic rules are: (1) simple chemical rules (e.g. maximal number of C atoms possible for a given MW), (2) Lewis and Senior rules (e.g. « octet rule »), (3) isotopic abundance pattern matching, (4) H / C atoms ratio, (5) heteroatom / C atoms ratio, (6) sum of the heteroatoms, and (7) TMS adduct subtraction (for GC-MS analyses only). The “Seven Golden Rules” is probably the most powerful filtering tool available today for

small molecules and is able to considerably reduce the number of molecular formulae from a given list of putative hits. In many cases, however, this filtering does not generate one single candidate only. The rules were used in numerous dereplication studies [12-14] and were implemented in several tools, such as the MZmine 2.10 software [3, 5] and the original Seven Golden Rules MS Excel sheet [10, 15].

E – Database search. In dereplication studies, however, obtaining validated molecular formula only is not sufficient, and further identification steps are required to finally determine the structure of the metabolite. Indeed, even if the filtering process provides one single molecular formula, this doesn't provide the identity of the analyte. For example, a simple query in the Dictionary of Natural Products [16] for the molecular formula of colchicine ($C_{22}H_{25}NO_6$, MW = 399) provides 22 corresponding structures (or 38 hits when every possible conformation is considered, for example (+)- and (-)-colchicine).

F – Filtering of candidate structures. Thus, filters are needed to decrease the number of possible structures. Five main approaches are found in the literature, based on chemotaxonomic information, retention information and fragmentation of the compound.

Firstly, the chemotaxonomic filter is often used when natural samples are studied and if a database holding molecular formula of NPs is available. For this, a database search is performed using the genus or species of the

studied sample (for example, " $C_{22}H_{25}NO_6$ AND *Colchicum*") [1]. The Dictionary of Natural Products [16] is probably the most appropriate NP database for this purpose. This filter will only provide reliable information if the organism studied and the compound are known, and if the database is holding updated and valid taxonomic information. This last point is sometimes misleading because information on plant families are not always provided or synonyms for a given natural organism exist. Such cross search can be very efficient for dereplication if the species of interest has been well documented, but it is a very tedious process, because there is today no software to perform this task automatically.

Secondly, precious information may be extracted from the PDA-UV spectrum of the compound of interest, if available. Indeed, a candidate compound having a PDA-UV spectrum which clearly does not correspond to the experimental spectrum may thus be discarded. Exact match with reported UV spectra is however not possible for more precise assignment since UV band will shift slightly according to the mobile phase used and in gradient elution [17].

Thirdly, the retention of the studied compound may potentially be used to provide structural information. It is indeed well known that the retention time can be correlated to the $\log P$ which is the lipophilicity parameter [18]. Moreover, the retention is also linked to several physicochemical properties such as the solvatochromic parameters [19-21]. This information was used in our work to provide

Taxonomy is the science aiming at describing and classifying living species according to their presumed natural relationships.

Chemotaxonomy is the classification of living species based on similarities and differences in biochemical composition.

additional information for dereplication purposes. The log P parameter was used in Chapter VI as an additional filter for dereplication purposes in the frame of a chemotaxonomic study of several Brazilian *Lippia* species. In Chapter VII, a retention time prediction method was developed that may be used as the final step of dereplication procedure, if the standard filtering process results in more than one putative natural product.

Fourthly, the ion mobility spectrometry drift time may also be predicted and used as an additional dereplication filter [22]. This value is correlated to the collision cross section of the molecule with the correction of additional physicochemical parameters. The collision cross section and the other parameters are easily calculated from the structure of the molecule. This kind of filter will probably gain importance as the use of ion mobility is increasing in NP research. As mentioned in Chapter IV, the drift time offers an orthogonal dimension compared to LC and MS. Thus, filters based on drift and retention times are not redundant but complementary. Such a filter has been implemented in the MolFind software and successfully enabled the identification of unknown molecules [22].

Lastly, the fragmentation of the compound occurring during MS/MS experiments provides valuable structural information that may be compared to *in silico* calculated fragments, or to spectra from databases. The calculated MS/MS spectra obtained from an *in silico* fragmentation may be matched against measured fragment ions. Such an approach is very efficient, if the *in silico* fragmentation is reliable. Such a tool is integrated in the MetFrag software [23, 24] and in the MZmine 2.10 [3], among others. On the other hand, the comparison of the measured MS/MS spectrum with spectra of a database, if available, is a powerful tool to discard or confirm candidates. However, the MS instrument, the experimental parameters and the concentration

of the analyte that are used have to be similar for both measured and reference spectra that are compared. Because such a database has to contain a high number of spectra of reference compounds to be useful, this approach is possible only for big laboratories or companies.

After this step of filtering of candidate structure, two scenarios are possible (Figure V.1): (1) there is one (or some) remaining candidate structure(s), and the procedure continues with point **G** to provide a putative identification, or (2) there are no more remaining candidate structures, probably because the compound is not present in databases. A *de novo* structure elucidation is then necessary (see point **H**).

G – Identity confirmation. The use of all the filters usually provides one or very few putative structures if the compounds can be found at least in species closely related to the natural organism of interest. The easiest and fastest way to confirm this identification is to analyse a commercialised standard or previously isolated NP using the same LC-MS conditions as used for the profiling, if available. The comparison of the chromatographic behaviour, usually using two different mobile or stationary phases, as well as MS spectra and MS/MS fragments, should confirm or infirm the identity of the targeted compound.

H – *de novo* structure elucidation. If the dereplication procedure does not yield any putative structure, this may indicate a highly interesting compound that was never described before. In this case a targeted isolation of the LC peak of interest can be performed by semi-preparative HPLC. Extensive 1D and 2D NMR as well as complementary MS/MS usually enable a complete *de novo* identification of the compound of interest (see Section 3.4 and 3.5 below).

The whole procedure – from extract to data processing, including the chemotaxonomic search – is described in detail in this chapter, which is based on a methodological chapter published in the *Methods in Molecular Biology* series.

Tools available for LC peak annotation

The tools that are mentioned above are integrated in several free software or part of software of MS vendors. Some of them are described here.

The original Seven Golden Rules algorithm is implemented in a Microsoft Excel sheet [10, 15]. It efficiently filters molecular formulae obtained from the MS spectra according to the seven heuristic rules mentioned above, but in its current version it necessitates unfortunately tedious copy-paste operations of both molecular formulae and isotopic measurements from the original MS software. Interestingly, an option is implemented to perform an additional query of the resulting molecular formulae against the Chemical Structure Lookup Service [25]. This service indexes more than 100 chemical databases and aims at determining whether the submitted structure is present in any of these databases. This option is very efficient but may provide false negative results if the compound is not yet reported and/or not present in the database.

The MZMine 2.10 software [3, 5, 7] is probably today the most complete software for peak annotation purposes. It is able to perform almost all the steps mentioned in Figure V.1: data cleaning, adduct search, molecular formula calculation, isotopic pattern matching and MS/MS fragmentation prediction (see step F above). The molecular formula calculation combines several tools to predict chemical

formulae including the modified Seven Golden Rules, the RDBE value, isotopic pattern matching. The prediction capability of the software was evaluated using a real metabolomics data set of 48 compounds of the extract from the cells of the fission yeast *Schizosaccharomyces pombe*. The chemical formula of 79% of the 48 compounds analysed using an Orbitrap mass analyser were correctly determined. Figure V.4 shows the number of possible molecular formulae calculated when the filtering module is applied (blue circles) and when only the elemental composition tool is applied (red squares) [7].

The SmartFormula™ software from Bruker (and its modules) is probably the most advanced of the software proposed by an MS instrument vendor. Besides the classical tool providing molecular formulae, it allows isotopic pattern matching, a cross-search in databases such as ChemSpider, and is able to deal with MS/MS fragmentation data. For this, the software correlates fragments of the MS/MS spectrum and sums formulae to highlight specifically the fragments of the studied ion, and then export them to the MetFrag tool, to match calculated and experimental fragments from MS/MS experiments [23, 24].

The CAMERA package combines the following features: peak grouping based on retention times, isotopic peak detection, annotation of adducts and neutral losses, and comparison of data from both positive and negative ionisation modes [6]. This last feature is particularly interesting to reliably extract pseudomolecular ions from complex spectra. The algorithms are implemented in an R package.

The MolFind Java-based software [22] is designed to aid in identifying chemical structures in complex mixtures from LC-MS data. It enables compound identification by matching experimental data obtained for an unknown compound with four calculated LC-MS derived

properties. These properties are retention, energy required to fragment 50% of a selected ion, calculated drift time and CID spectra, and are calculated for all candidate compounds downloaded from a chemical database such as PubChem that possess the same mass as the targeted LC-MS peak. According to the authors, the drift time-based filter was unable to remove any candidates among the smaller compounds, but was quite useful to filter large molecular weight compounds. This procedure is highly interesting since it is the only one using the retention and drift times, but it is somewhat limited by the content of the databases used. Finally, it lacks an efficient procedure to automatically extract the mass of the compound from the LC-MS data.

The Progenesis CoMet software from Nonlinear Dynamics (Newcastle upon Tyne, UK) is a commercialised tool for the quantification of metabolites which concentration is changing in samples analysed by LC-MS in the frame of metabolomic studies. This tool performs several tasks, including adduct search, chromatographic alignment, MW and RT matching with in-house

databases, and query in the online Metlin search tool.

In summary, and as illustrated by these examples, great efforts were recently made to develop new (semi-)automated tools for peak annotation in LC-MS analyses. Although the algorithms are different, they are all based on similar rules and provide efficient filters to dramatically reduce the number of candidates in metabolite identification procedures to a few or one single molecular formula. Still, some improvements are necessary to allow a routine use of these techniques. Firstly, tools should be integrated in one single software that should be fully compatible with (or even part of) the MS acquisition software, to avoid tedious data transfer or transformation procedures. Secondly, reliable databases are required, *i.e.* databases with correct annotation and that contain all known NPs. Finally, the database search for taxonomic information has to be automated, since this step may be one of the major bottlenecks today, in particular when adapted databases are not available and when a literature search has to be performed.

Strategies in Biomarker Discovery. Peak annotation by MS and targeted LC-MS micro-fractionation for *de novo* structure identification by micro-NMR.

Philippe J. Eugster

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Gaëtan Glauser

Chemical Analytical Service of the Swiss Plant Science Web, University of Neuchâtel, Switzerland

Jean-Luc Wolfender

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Book chapter in *Methods in Molecular Biology, Metabolomics Tools for Natural Products Discoveries*, Editors U. Roessner and D.A. Dias, Humana Press, New York, USA (In press).

Abstract

In metabolomic studies the identification of biomarkers is a key step but represents a serious bottleneck since the *de novo* identification of natural products is a lengthy process. A strategy for the dereplication and peak annotation of plant biomarkers is presented based on high resolution mass spectra acquired on quadrupole-time-of-flight mass spectrometry coupled to ultra-high pressure liquid chromatography and chemotaxonomy information. A rational approach for the targeted LC-MS micro-isolation of biomarkers followed by *de novo* identification by NMR at the microgram scale is described, based on gradient transfer from the analytical scale and chromatographic modelling. The methodology is illustrated by the identification of various stress biomarkers of the plant wound response using *Arabidopsis thaliana* as a model.

1. Introduction

Metabolomics plays an increasingly important role in natural product research [26]. Within systems biology, this holistic approach actually provides the most “functional” information of the ‘omics’ technologies [27]. Metabolomics represents a new way of interrogating biological systems since it is an unbiased data-driven approach that may ultimately lead to hypotheses and new biological knowledge.

In typical metabolomic studies, complex crude natural extracts are compared by using various analytical methods that generate metabolic fingerprints [28]. For this, the main approaches are either based on nuclear magnetic resonance (NMR) [29] or mass spectrometry (MS) [30]. MS is often used in hyphenation with chromatographic techniques such as gas chromatography (GC-MS) or high performance liquid chromatography (LC-MS). Since primary and secondary metabolites have a wide chemical diversity, no single analytical method that provides a complete survey of all metabolites in a given organism exists at present [31], and combinations of methods are more and more used for metabolomics. Once data are recorded and pre-processed, the comparison of a sufficient number of fingerprints is performed with either unsupervised or supervised multivariate data analysis (MVDA) methods that may reveal features in the dataset that can be linked to biomarkers [32]. These features correspond to peak intensities of NMR chemical shifts, m/z ions and/or chromatography retention times according to the approach used.

One of the main difficulties in metabolomics resides in the correct identification of the

biomarkers highlighted in the peak lists generated by MVDA. This usually represents a bottleneck in such an approach since, unlike for proteomics, the structure determination of low molecular weight compounds does not follow generic rules, and no freely accessible comprehensive MS/MS database provides unambiguous dereplication for natural products. Despite many efforts, the identification of biomarkers still relies on the interpretation of the MS or NMR data obtained. This is especially true for secondary metabolites that are often species specific [29].

In this chapter, a practical approach for biomarker peak annotation by LC-MS is presented. The strategy proposed relies on a high resolution LC-MS profiling allowing the unambiguous determination of molecular species, the calculation of the corresponding molecular formulae and filtering for validation. Dereplication is then based on database cross search taking into account chemotaxonomic considerations. Additional information from UV photodiode array (UV-PDA) and/or MS/MS spectra is also integrated to support on-line identification. For the *de novo* identification of unknown biomarkers, a rational and rapid LC-MS micro-isolation approach is described that provides microgram amounts compatible with further at-line micro-NMR structure determination and/or bioactivity assessment. Due to the complexity of natural extracts, the purification of metabolites present in low concentrations is especially critical [33]. The micro-isolation strategy relies on the optimisation of the chromatographic analysis using ultra-high pressure liquid chromatography

(UHPLC) thanks to modelling software [34] and further transfer to semi-preparative LC conditions with MS detection [35].

The approach described is based on UHPLC coupled to quadrupole-time-of-flight mass spectrometry (QTOF-MS), but is generic and may be adapted to other types of high resolution LC-MS instruments.

2. Materials

2.1. Solvents and reagents

For extractions, sample preparations and dilutions, methanol (MeOH), acetonitrile (ACN) and isopropanol (IPA) of analytical or HPLC grade and water of milli-Q quality (18.2 M Ω ·cm at 25°C) are recommended. Solvents and additives of LC-MS quality should be used for UHPLC-QTOF-MS analyses (see Note 1). Solvents for semi-preparative LC-MS micro-isolation should be of HPLC grade (see Note 2).

1. Solvents for sample preparation prior to UHPLC-QTOF-MS analysis: IPA, MeOH:H₂O 85:15 (v/v) and MeOH 100%.
2. Final dissolution solvent for sample injection: MeOH:H₂O (70:30).
3. UHPLC and semi-preparative LC-MS mobile phases: A = water + 0.1% formic acid (FA), B = ACN + 0.1% FA.
4. Solvents for sample preparation prior to biomarker isolation: IPA, MeOH, MeOH:H₂O (5:95), MeOH:H₂O (70:30).
5. NMR solvent: methanol d-4 (CD₃OD) (see Note 3).

2.2. Equipment

2.2.1. Sample preparation and extraction for UHPLC-QTOF-MS analysis

1. Ball mill with 2 cm diameter stainless steel balls (e.g. Retsch MM200, from Schieritz & Hauenstein AG, Arlesheim, Switzerland).
2. Centrifuge.

3. Centrifugal or nitrogen evaporator.
4. Ultrasonic bath.
5. SPE cartridge 100 mg (Sep-Pak C₁₈, Vac 1 cc, 100 mg, Waters, Milford, USA) with vacuum manifold.
6. Weighed 2 mL microcentrifuge tubes.
7. Glass vials (5 mL).
8. Pipettes and pipette tips.
9. HPLC vials and caps.

2.2.2. UHPLC-QTOF-MS analysis

1. UHPLC system able to withstand a maximal pressure of 1000 bar and equipped with a binary or quaternary pump and a column oven maintaining constant temperature (e.g. Waters Acquity UPLC).
2. QTOF-MS instrument hyphenated with the UHPLC system through an electrospray (ESI) interface (e.g. Synapt G2 from Waters) (see Note 4).
3. Acquity UPLC BEH C₁₈ (150 x 2.1 mm I.D., 1.7 μ m particle size) column (Waters) with an Acquity UPLC BEH C₁₈ Van guard (5 x 2.1 mm I.D., 1.7 μ m particle size) pre-column (Waters).

2.2.3. Data processing

1. MassLynx 4.1 for LC-MS raw data processing (Waters) or any software adapted to the MS analyser used.
2. Seven Golden Rules Excel-based software, freely available [10]

(http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/Software/).

3. Databases for dereplication based on molecular formulae (The Dictionary of Natural Products (DNP) (Chapman & Hall / CRC Press) (see Note 5).
4. Osiris 4.2 for HPLC modelling (Datalys, Saint-Martin-d'Hères, France) (see Note 6).

2.2.4. Sample preparation and extraction for LC-MS isolation

1. Important amount of fresh plant tissue for separation scale-up (e.g. 500 g in the case of *Arabidopsis* leaves) (see Note 7).
2. Big mortar and pestle.
3. Erlenmeyer (2 L).
4. Agitation plate or magnetic agitator.
5. Big size filtration paper and filter funnel.
6. Rotative evaporator.
7. Ultrasonic bath.
8. LiChroprep RP-18, 40-63 μm (50 g) (Merck, Darmstadt, Germany).
9. Glass column for open chromatography with frit (porosity 4) and vacuum assembly.

2.2.5. Optional pre-fractionation and targeted LC-MS isolation

1. For pre-fractionation: XBridge BEH C_{18} (150 x 19.0 mm I.D., 5.0 μm particle size) column, or another column with the same phase chemistry than the analytical column previously used (see Note 8).
2. Fraction collector that can hold 120 x 10 mL tubes.
3. For final purification: Two Xbridge BEH C_{18} (250 x 10.0 mm I.D., 5.0 μm particle size) columns, or other columns with the same phase chemistry than the analytical column previously used (see Note 8 and Note 9).
4. Semi-preparative HPLC system, such as the Varian 9012 (Varian, Palo Alto, CA, USA), or another semi-preparative LC system able to deliver a 10 mL/min flow.
5. Column oven adapted to semi-preparative columns (40°C) (see Note 10).
6. 96-well deep plates (1 mL per well).
7. Splitter able to divide the flow and maintain a 50 $\mu\text{L}/\text{min}$ flow rate in the MS instrument (see Note 11).
8. For LC-MS isolation monitoring: MS detector, such as TSQ 7000 (Thermo Fisher Scientific Inc., Waltham, MA).

3. Method

The following protocol has been devised for biomarker identification either based on UHPLC-QTOF-MS data only (peak annotation) or complete *de novo* structure identification based on LC-MS targeted micro-isolation and subsequent micro-NMR analysis. A prerequisite for this protocol is the localisation of a biomarker of interest in the metabolite profiling chromatogram of a representative crude extract. Localisation is performed by extracting the ion trace corresponding to a specific feature (m/z x retention time (RT)) found in the loadings after the MVDA of a classical LC-MS based metabolomics study (see for example [36] and Note 12). We will refer to it as “biomarker of interest” in the following steps. The process is illustrated by the identification of biomarkers in the aerial parts of the model plant *Arabidopsis thaliana* (Biomarkers **A** and **B** for the dereplication process, and Biomarker **C** for the isolation procedure) but may be applied to other plants and metabolite types provided that slight adaptations are made.

3.1. Sample preparation for dereplication based on UHPLC-QTOF-MS

3.1.1. Extraction procedure


1. Harvest and weigh approximately 500 mg of the fresh plant tissue of interest (see Note 13).
2. Put the fresh plant tissue, 5 mL of IPA and a 2 cm diameter ball in the jar of the ball mill (see Note 14).
3. Extract for 2 minutes at 30 Hz using the ball mill.
4. Collect the content of the jar in centrifuge tubes, without the balls, centrifuge for 4 minutes.
5. Collect the supernatant; dry it under nitrogen at 40°C or using the centrifugal evaporator to obtain the crude extract.

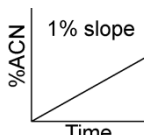
3.2. 3.1.2 Sample preparation

1. Dissolve the crude extracts (2-10 mg) in 0.7 mL MeOH:H₂O (85:15), with ultrasonic bath if needed.
2. Place a C₁₈ SPE cartridge on the manifold chamber. Prepare a glass vial under the cartridge to collect the conditioning and equilibrating solvents.
3. Condition and equilibrate the cartridge by washing with 1 mL MeOH 100% and then 1 mL MeOH:H₂O (85:15), under vacuum. Adjust the pressure to obtain an elution rate of about 1 drop per second. Discard the eluate.
4. Place a weighed 2 mL micro-centrifuge tube in the 5 mL glass vial, under the cartridge.
5. Load the dissolved extract on the cartridge (see Note 15) and elute it. Wash with 0.8 mL MeOH:H₂O (85:15).
6. Evaporate the collected eluate to dryness using a centrifugal evaporator or under a gentle nitrogen flow.
7. Weigh the 2 mL microcentrifuge tube to assess the yield of the SPE extraction.
8. Dissolve the extract in MeOH:H₂O (70:30) (see Note 16).

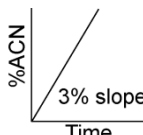
Table V.1. Gradients used for UHPLC-TOFMS analyses and pre-fractionation on semi-preparative scale.

%B	Time (min)		Time (min)
	90 min UHPLC gradient	30 min UHPLC gradient	
5.0	0.0	0.0	0.0
5.0	0.0	0.0	1.0
95.0	90.0	30.0	113.9
95.0	100.0	40.0	152.0
5.0	100.5	40.5	155.0
5.0	110.0	50.0	190.0





1% slope



3% slope

3.3. UHPLC-QTOF-MS analysis

3.3.1. UHPLC gradient conditions

For metabolite localisation and dereplication, a generic high resolution UHPLC-QTOF-MS (3%/min slope) 30 min gradient is used (see second column of Table V.1 and Note 17). Other parameters are:

1. Flow rate: 460 $\mu\text{L}/\text{min}$.
2. Column and autosampler temperatures: 40 $^{\circ}\text{C}$ and 10 $^{\circ}\text{C}$, respectively.
3. Injection volume: 2.0 μL .
4. PDA: 10 Hz at least, over 210 – 600 nm.

3.3.2. QTOF-MS conditions

Perform calibration using for example a sodium formate solution in the 100-1000 m/z range, in both positive (PI) and negative ionisation (NI) modes (see Note 18). Check the mass accuracy by subsequent injection of any selected molecule.

Run two separated analyses using both PI and NI ESI modes (see Note 19), with alternating scans at low and high collision energies (e.g. MS^E) (see Note 20). For the Synapt G2 QTOF-MS (Waters), generic source parameters are:

1. Voltages: capillary 2500 V, cone 25 V, extraction cone -4.5 V and 3.0 V in NI and PI modes, respectively.
2. Temperatures: source 120 $^{\circ}\text{C}$, desolvation gas 350 $^{\circ}\text{C}$.
3. Gas flows: desolvation gas 800 L/h, cone gas flow 20 L/h.

4. Mass range: 85-1200 Da.
5. Scan time: 0.2 s.
6. Collision energy: 4 eV and ramp of 10-30 eV, both applied on the transfer region of the collision cell.
7. Collision gas: argon, at a flow rate of 2.1 mL/min (pressure inside the collision cell 7.0×10^{-3} mbar).
8. Internal calibration (Lockspray™): infusing a 500 ng/mL solution of leucin-enkephalin at a flow-rate of 7.5 μ L/min.
9. Lockspray™ scan time and frequency: 0.5 s and 15 s, respectively, with data averaged over 5 scans for mass correction.

A typical UHPLC-NI-ESI-TOF-MS metabolite profiling chromatogram (BPI trace) of the leaf extract of *Arabidopsis thaliana* is presented in Figure V.6. All biomarkers that are identified by the procedures described below (Biomarkers A-C) have been labelled by their corresponding m/z ions.

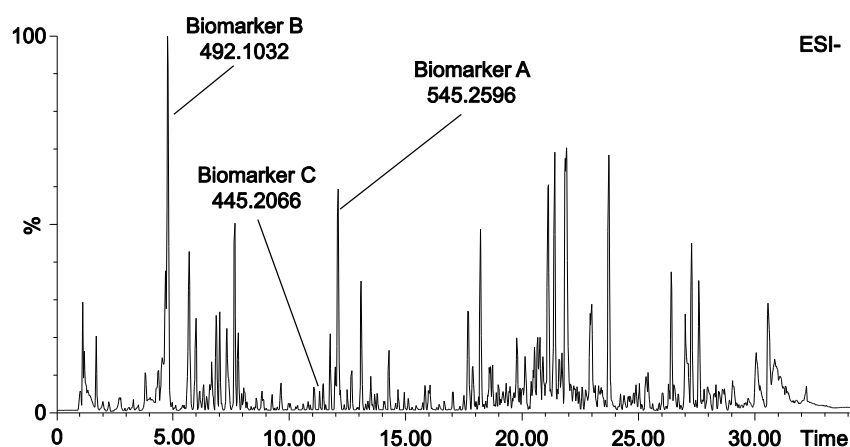


Figure V.6. High resolution UHPLC-TOF-MS metabolite profiling (BPI trace) of *Arabidopsis thaliana* (crude leaf isopropanol extract). Separation was carried out on an Acquity BEH C₁₈ (150 x 2.1 mm; 1.7 μ m particle size) column with a 5–95% ACN gradient in 30 min at 40°C, and detection was performed by TOF-MS in the NI mode. Biomarkers A-C, used to illustrate the metabolite identification process, are labelled on this chromatogram.

3.4. Data processing

Figure V.7 illustrates some steps of the dereplication of the Biomarker A (m/z 545.2596), which was detected in NI mode in the metabolite profile of *A. thaliana*, following leaf wounding (Figure V.5).

3.4.1. Determination of the molecular weight

In a given ESI spectrum, molecular species ions may be present either as protonated or deprotonated molecules ($[M+H]^+$ or $[M-H]^-$) or may form dimers or higher oligomers, or adducts with solvents, molecules present in the solvents, or LC-MS additives. Moreover, the most intense peak doesn't always correspond to $[M+H]^+$ or $[M-H]^-$. In addition, ions that are not related to the biomarker of interest may be present. The procedure below presents ways to find the MW of an unknown peak based on both PI and NI MS spectra in typical situations.

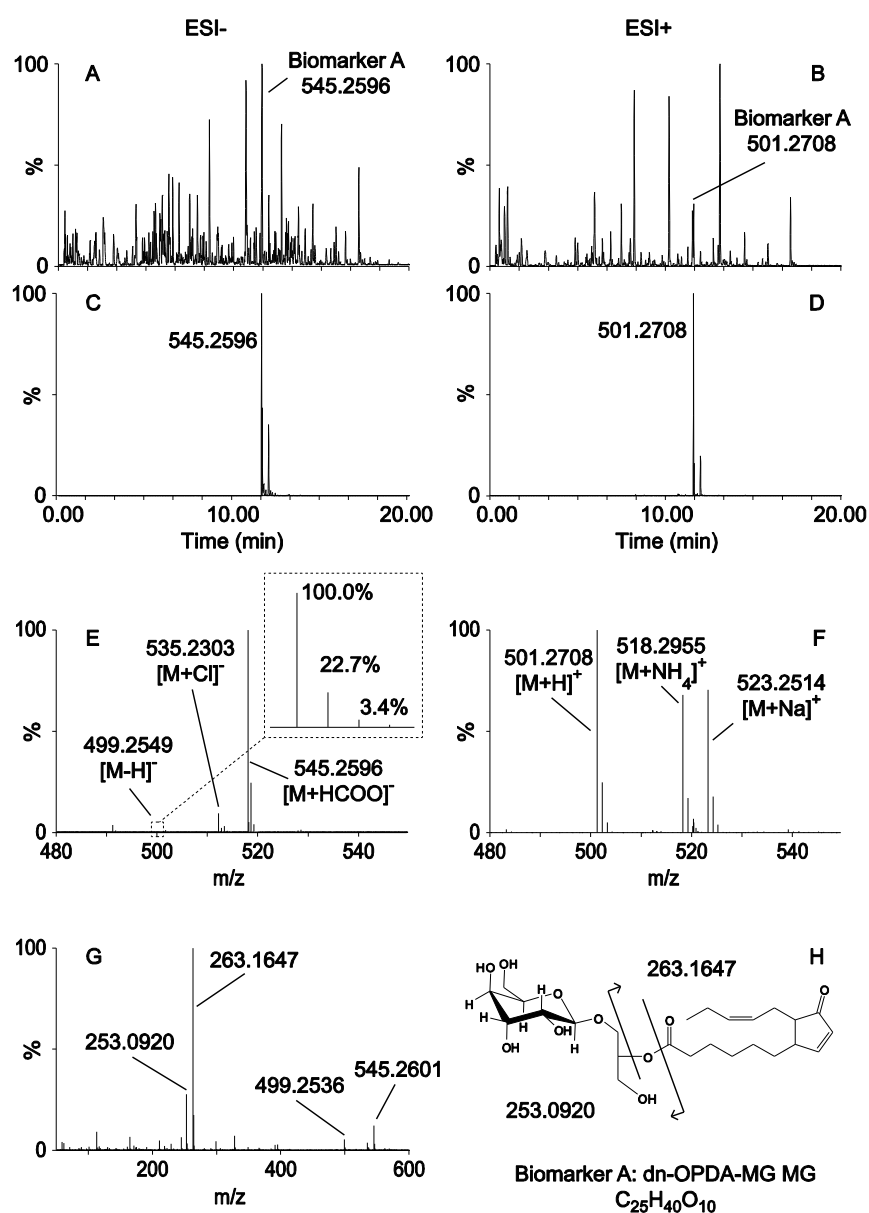


Figure V.7. Dereplication procedure for the unknown Biomarker **A** in an *Arabidopsis thaliana* extract following leaf wounding. Localisation of its main m/z ions (m/z 545.2596 and m/z 501.2708) in the corresponding NI (**A**) and PI (**B**) UHPLC-TOF-MS chromatograms and corresponding extracted ion chromatograms (m/z \pm 0.01 Da) (**C**, **D**). The isomer also detected is not discussed in more details. According to the spectra in NI mode (**E**), m/z 545.2596 represents the formate adduct $[M+HCOO]^-$, while m/z 535.2303 corresponds to $[M+Cl]^-$, and m/z 499.2549 to $[M-H]^-$. Comparison with the PI mode (**F**) confirms this, and $[M+H]^+$, $[M+NH_4]^+$ and $[M+Na]^+$ were at m/z 501.2708, m/z 518.2955 and m/z 523.2514 respectively. This combined information provides the unambiguous MW determination of Biomarker **A** (average monoisotopic MW: 500.2620). The inset in (**E**) represents an expansion of the isotopic pattern recorded for the $[M-H]^-$, the peak height ratios were used during heuristic filtering for molecular formulae determination. The hit obtained for Biomarker **A** (**H**) by the database search is confirmed by the fragments recorded in the high energy CID NI-TOF-MS^E spectrum (**G**).

Table V.2. list of commonly found adducts in ESI PI and NI modes, using acetonitrile as organic modifier and formic acid as additive. All ions are singly charged. A comprehensive list of adducts may be found in literature [10, 37, 38].

Type of ions	Ion mass
Positive ionisation mode	
$[M+H]^+$	$M + 1.007276$
$[M+NH_4]^+$	$M + 18.033823$
$[M+Na]^+$	$M + 22.989218$
$[M+K]^+$	$M + 38.963158$
$[M+ACN+H]^+$	$M + 42.033823$
$[M+ACN+Na]^+$	$M + 64.015765$
$[2M+H]^+$	$2M + 1.007276$
Negative ionisation mode	
$[M-H]^-$	$M - 1.007276$
$[M+Cl]^-$	$M + 34.969402$
$[M+HCOO-H]^-$	$M + 44.998201$
$[M+HCOO+Na-2H]^-$	$M + 67.987419$
$[2M-H]^-$	$2M - 1.007276$
$[2M+HCOO-H]^-$	$2M + 44.998201$

1. Highlight the peak of interest from the chromatogram by extracting its trace with an adapted mass range window (e.g. 0.02 Da) (see Figure V.7A-D and Note 21).
2. Combine the spectra containing the mass of interest (see Note 22).
3. Determine the presence of adducts and/or dimers, and ensure the molecular weight (MW) based on the following rules:
 - a. Look for adducts, detected by the presence of both $[M+H]^+$ and $[M+adduct]^+$ or $[M-H]^-$ and $[M+adduct]^-$, depending on the ionisation mode (see Figure V.7E-F and Notes 22-23). Table V.2 provides a list of the most frequently encountered adducts using ACN+FA mobile phases and an ESI source.
 - b. Check for the presence of dimers, characterised by $[2M+H]^+$ or $[2M-H]^-$ depending on the ionisation mode, or higher oligomers.
 - c. Compare results of PI and NI modes, if both are available: the resulting

molecular weight (MW) should be the same.

- d. If only one single ion species is present, check the corresponding MS^E spectrum to verify the possible loss of adducts.

3.4.2. Extraction of molecular formulae

1. Use the elemental composition tool provided in MassLynx (or another MS software) to obtain putative molecular formulae from the molecular ion species recorded. Work on combined spectra only (see Note 22).
 - a. Set the minimum number of elements to 0 and the maximum to 200 for C, H, O, N and S (see Note 24).
 - b. Set the mass tolerance as three times the usual mass accuracy of the instrument (e.g. $3 \times 1\text{-}2 = 5$ ppm for the Synapt G2). Table V.3 illustrates the number of molecular formulae corresponding to the [M-H]⁻ and [M+H]⁺ ions (Figure V.7E-F) (m/z 499.2549 and m/z 501.2708) of Biomarker **A** used as example in section 3.3.1, as well as for Biomarker **B**, containing CHONS. As shown, the number of calculated molecular formulae for a given exact mass around 500 Da may exceed 100 if a large tolerance window (15 ppm) is used and if CHONS elements are considered (see Note 25).
2. Export the calculated molecular formulae and correct them according to the ionisation mode used (add or remove an atom of adducts in NI or PI ESI, respectively).

3.4.3. Heuristic filtering

As shown above, high mass accuracy measurements alone do not provide

unambiguous molecular formula assignment for a given biomarker. Thus additional filtering is needed to reduce the number of possible hits and validate the molecular formula assignment. In this respect the application of heuristic filtering with the Seven Golden Rules [10] represents a rational approach.

1. Import all molecular formulae into the Seven Golden Rules software Excel sheet [10] (see Note 26).
2. On the spectrum, measure the isotopic pattern, *i.e.* the height of the ¹³C₁, ¹³C₂ and ¹³C₃ peaks of the biomarker, expressed as a percentage of its main ¹²C peak. Report it in the dedicated field of the Excel sheet.
3. Set the isotopic pattern error as the ratio of the background to the intensity of the main peak of the marker, as a percentage, but at least 3%.
4. Click the “1) Autofill”, “2) Calc” and “3) Check” buttons (see Note 27). Copy the remaining molecular formulae that are highlighted in blue in the “Pubchem”-“Found” column.

In the examples discussed in Table V.3, application of heuristic filtering reduces the number of possible formulae for Biomarker **A** (MW 500) from 26 to 2 and 27 to 2 in PI and NI modes respectively (C₂₅H₄₀O₁₀ and C₂₂H₃₂N₁₀O₄) and from 47 to 1 for Biomarker **B** (MW 493, C₁₆H₃₁NO₁₀S₃) (see Note 28).

3.4.4. Database search

1. Perform a search of the remaining molecular formulae in the Dictionary of Natural Products:

Table V.3. Number of potential molecular formulae corresponding to the dereplicated ion and using different mass windows. Biomarker **B** was detected in NI mode only. Spectral accuracy is discussed later.

Biomarker **A** (dn-OPDA-MG MG, MW 500, C₂₅H₄₀O₁₀):

Mass window (ppm)	ESI + (<i>m/z</i> 501.2700)		ESI – (<i>m/z</i> 499.2543)	
	with CHO	with CHONS	with CHO	with CHONS
15	2	74	2	79
10	1	49	1	53
5	1	26	1	27
3	1	15	1	16
1	1	6	1	4
5 + heuristic filtering	1	2 ^a	1	2 ^a

^a the two remaining molecular formulae were: C₂₅H₄₀O₁₀ and C₂₂H₃₂O₄N₁₀

Biomarker **B** (glucohirsutin, MW 493, C₁₆H₃₁NO₁₀S₃)

Mass window (ppm)	ESI – (<i>m/z</i> 492.1032)		
	with CHO	With CHON	with CHONS
15	0	36	140
10	0	27	95
5	0	14	47
3	0	7	27
1	0	3	10
5 + heuristic filtering	0	0	1 ^b

^b the remaining molecular formula was: C₁₆H₃₁NO₁₀S₃

This step may be used to discard molecular formulae not corresponding to previously reported natural products. For Biomarker **A**, this discards C₂₂H₃₂N₁₀O₄ and the only remaining formula is C₂₅H₄₀O₁₀ (see Note 29.).

For Biomarker **B**, the unique proposed formula was found to match with existing hits in the Dictionary of Natural Products.

Based on such validated molecular formulae assignment, peak annotation can be made by cross search with chemotaxonomic information of the plant studied.

2. Perform a cross search with validated formulae and chemotaxonomic keywords (e.g. species, genus, family) in the Dictionary of Natural Products or other databases.

For Biomarker **A**, a cross search based on $C_{25}H_{40}O_{10}$ and 'Arabidopsis' and 'thaliana' provided one single structure only, the galactolipid dn-OPDA-MG MG (see Figure V.7H and Note 30). For Biomarker **B**, the same cross search based on $C_{16}H_{31}NO_{10}S_3$ provided one structure only, the glucosinolate glucohirsutin.

3.4.5. Additional information from MS/MS

The use of collision induced dissociation (CID) spectra may provide additional structural information on the biomarker of interest. With the MS^E acquisition described above, this information can be retrieved as follows:

1. Verify which fragments generated in the high energy MS^E acquisition have identical retention times to the $[M+H]^+$ or $[M-H]^-$ ions and discard those which are not perfectly aligned (see Note 31). Compare CID MS spectra from the high energy MS^E with those compiled in existing databases such as Massbank or ReSpect for Phytochemicals (see Note 32).
2. Alternatively, determine whether the fragments obtained may be described by means of rules of possible fragmentation mechanisms.

The CID spectrum of Biomarker **A** displays two main diagnostic fragments (see Figure V.7G-H)

that confirm the peak annotation made based on molecular formula and chemotaxonomy cross search.

3.4.6. Additional information from UV-PDA

The use of UV-PDA spectrum as an additional filter for dereplication is advantageous provided that the biomarker is present in sufficient amount. Many UV-active natural products such as polyphenols exhibit characteristic chromophores [39] that can be exploited to strengthen the peak annotation.

1. Compare both UV-PDA spectra and wavelength(s) of maximum absorption (λ_{max}) of the biomarker of interest with values reported for the hits previously obtained (§3.3.4). Discard candidates whose values don't match with the experimental spectrum or λ_{max} (see Note 33).

No UV information could be obtained for Biomarker **A**, while Biomarker **B** displayed a UV-PDA spectrum comparable to the one of glucohirsutin.

3.4.7. Confirmatory analysis

At this stage, the number of putative structures for a given biomarker is usually relatively limited. Confirmation of the identity of the analyte may be thus obtained by injection of the standard, if it is commercially available or if synthesis is applicable. The following protocol provides a way to ascertain peak identification based on standard.

1. Prepare a 5 $\mu\text{g}/\text{mL}$ solution of the standard in an adapted solvent (preferably in the same solvent as previously used for the analysis of the extract, *i.e.* MeOH:H₂O 70:30).

2. Inject 2 μL of this solution in the previously used UHPLC-QTOF-MS conditions (§ 3.2.1 & 3.2.2). If necessary, adjust its concentration (see Note 34).
3. Verify that the retention time of the biomarker of interest and of the standard don't vary by more than 1%, and that adducts and dimers are similar in both cases, provided that the intensity of the main peak is similar (see Note 35).
4. Further confirm or infirm the identity by spiking the extract with the standard. Only one peak should be visible in the chromatogram when extracting the specific trace.

3.5. Targeted LC-MS isolation

The targeted LC-MS isolation procedure presented here is used for NMR *de novo* identification of biomarkers that cannot be annotated by the dereplication approach described above. The procedure is only based on HPLC methods and is adapted for isolation of microgram amounts of the biomarker of interest, a scale compatible with full characterisation by state of the art micro-NMR methods [40, 41]. This procedure is adapted from [33] and is illustrated by the isolation of Biomarker C, a jasmonate synthesised in *A. thaliana* in response to wounding (see Figure V.8)

3.5.1. Extraction and sample preparation for biomarker micro-isolation

The sample preparation for the micro-isolation of the biomarkers of interest is based on an upscale of the procedure used for the UHPLC-QTOF-MS metabolite profiling described under §3.1.1.

1. Harvest (see Note 13), weigh and grind using a mortar and pestle approximately 500 g of fresh plant tissue (see Note 7).
2. Extract with 1L of IPA during 3 hours at room temperature, under agitation (see Note 36).
3. Filter on paper, keep the eluate.
4. Extract the residue once again in the same conditions, filter.
5. Combine and evaporate both eluates using the rotative evaporator to obtain the crude extract.

3.5.2. Extraction and sample preparation for biomarker micro-isolation

1. Dissolve the crude extracts (about 5 g) in 40 mL MeOH, with ultrasonic bath if needed.
2. Add 10 g of LiChroprep RP-18, mix, evaporate to dryness using the rotative evaporator.
3. On the bottom of the open column, place 40 g of LiChroprep RP-18, condition and equilibrate with 250 mL of MeOH and MeOH:H₂O (5:95) respectively, and place above the mixture of extract and LiChroprep RP-18 previously prepared (see Note 37).
4. Add 250 mL MeOH:H₂O (5:95) and elute by applying an adapted vacuum to obtain an elution rate of about 10 mL/min (see Note 38).
5. Elute the compounds with 250 mL MeOH:H₂O (70:30) at the same flow rate (see Note 39).
6. Evaporate to dryness to obtain the enriched extract, weigh.
7. Dissolve the enriched extract in 500 μL MeOH:H₂O (70:30).

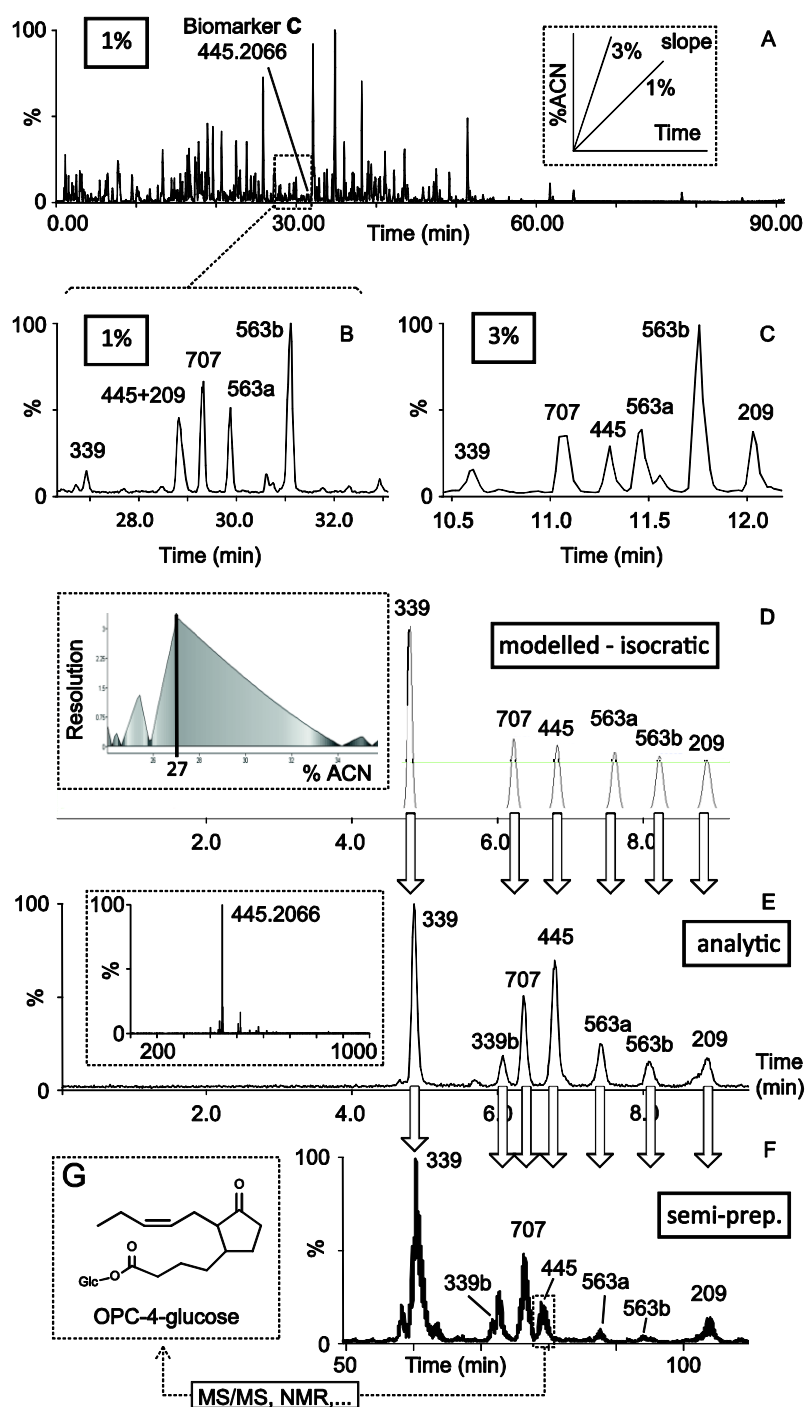


Figure V.8. LC-MS isolation procedure of the Biomarker C (m/z 445.2066 in NI mode). Two UHPLC-NI-TOF-MS gradient runs of 1%/min slope (A, B) and 3%/min slope (C) were performed using an Acquity UPLC BEH C₁₈ (150 x 2.1 mm; 1.7 μ m particle size) column with a 5–95% ACN gradient at 40°C. (D) Simulated chromatogram calculated based on the modelled isocratic condition 27% ACN (inset in D) by the Osiris software. (E) Isocratic UHPLC-TOF-MS analysis of the enriched fraction containing m/z 445. Geometrical transfer to the semi-preparative scale on two XBridge BEH C₁₈ (250 x 10 mm; 5 μ m particle size) columns (F) ensured the same selectivity (see D-F) and provided pure microfractions containing Biomarker C finally identified as OPC-4-glucose by further NMR experiments (G).

3.5.3. Optional pre-fractionation

A pre-fractionation step in gradient mode is recommended to ensure the optimal purification of minor compounds in the final isolation step. For this purpose, the original UHPLC gradient (3%/min slope gradient) is transferred to the semi-preparative scale with the following conditions (see Note 40).

1. Stationary phase: Xbridge BEH C₁₈ (150 x 19.0 mm I.D., 5.0 µm particle size).
2. Mobile phase: water and ACN, both with 0.1% FA, using the gradient described in the third column of Table V.1 (see Note 41).
3. Flow rate: 10 mL/min.
4. Column temperature: 40 °C.
5. Injection volume: 150 µL (see Note 42).
6. Divert 0.1-1% of the flow to the MS detector using a flow splitter (see Note 43)
7. Collect the fractions every minute (10 mL per fraction) in tubes, for 120 minutes.
8. Determine which fraction(s) contains the biomarker of interest based on extracted ion chromatogram obtained by MS detection (see Note 44).
9. Verify the presence of the biomarker of interest in the collected enriched fractions using the UHPLC-TOF-MS method described in §3.2.
10. Combine and evaporate fractions containing the biomarker of interest and dissolve it in MeOH:H₂O (70:30).

3.5.4. Gradient modelling and optimisation

To achieve the semi-preparative purification of biomarkers, optimal separation conditions were predicted and tested at the analytical scale using HPLC modelling software (see Note 6). Figure V.8 illustrates the method.

1. Analyse the enriched fraction by UHPLC-TOF-MS using the 3%/min slope gradient (30 min) used previously (see §3.2.1).
2. Run a second gradient in the same conditions but with 1%/min slope (see Figure V.8A-B), according to the gradient described in the first column of Table V.1 (see Note 45). MS conditions are identical to those used in §3.2.2. (see Note 46).
3. Localise and note the retention time (RT) of the biomarker of interest, and of all other compounds that elute in the same retention time window (± 1 min for the 30 min gradient, and ± 3 min for the 90 min gradient) in both chromatograms (see Figure V.8B-C).
4. In Osiris software, provide the following information:
 - a. The LC conditions as described in §3.2.1.
 - b. The dwell volume (approximately 120 µL for the Acquity UPLC system).
 - c. The column's dead volume (approximately 375 µL for the Acquity UPLC 150 x 2.1 mm, 1.7 µm column) (see Note 47).
 - d. A random value for the area of the peak (for example, 1).
 - e. RT of the biomarker of interest and the neighbour peaks in both conditions (see Figure V.8B-C).
5. Determine the optimised separation in isocratic mode and using the same column and mobile phase as described in §3.4.1 (see Figure V.8D). If an error occurs, modify the maximal retention factor value (k) up to 99.
6. Inject the pre-fractionated sample in the modelled conditions in UHPLC-TOF-MS, and verify the separation of the biomarker of interest (see Figure V.8E). If the separation is not satisfactory, model a new separation with slightly different conditions using the Osiris software.

3.5.5. LC-MS purification

The same set-up as used in section 3.4.2 is employed except that a longer column and 96 well-plates (2 ml /well) are used micro-collection.

1. Connect two Xbridge BEH C₁₈ (250 x 10.0 mm I.D., 5.0 μ m particle size) columns together with very short (30 mm) PEEK tubing (see Note 48).
2. Inject 150 μ L of the pre-purified sample using the following conditions:
 - a. Flow rate 3.5 mL/min.
 - b. Isocratic separation as modelled by the Osiris software.
 - c. Oven temperature: 40°C.
3. Collect fractions every 30 s in 96-well plates.
4. Identify the fractions containing the biomarker of interest, based on extracted ion chromatograms from the MS monitoring of the semi-preparative separation. Figure V.8F illustrates the separation obtained after injection of the enriched fraction containing the biomarker of interest.
5. Check the presence of the biomarker of interest in the fractions predicted by the model, using the UHPLC-QTOF-MS method described in §3.2.

6. Combine the fractions containing the pure biomarker of interest, dry them using the centrifugal evaporator or under gentle nitrogen flow.

3.6. Micro-flow NMR analysis

The micro-isolation procedure described above typically yields a few tenths or hundreds of μ g of biomarkers. With such amounts, weighing is often not possible, but the samples are compatible with further micro-NMR analysis. For micro-NMR measurements, samples may be dissolved in 5 μ L of deuterated methanol or another appropriate solvent. Such solution can be measured on a micro-flow probe (Protasis, Marlboro, MA, USA) [42] or in 1 mm capillary tubes. Typical hydroxyjasmonates spectra yielded by micro-flow CapNMR using such a procedure are illustrated in [33]. In the illustrated example, the structure of Biomarker **C** was elucidated as 3-oxo-2-(2Z-pentenyl)cyclopentane-1-butyric acid-1-O- β -glucose (OPC-4-Glucose, Figure V.8G), a new wound biomarker, based on the complementary MS/MS and ¹H NMR data obtained [43].

4. Notes

1. Ultra-high purity solvents are required to ensure low background noise in the LC-MS chromatograms and to maintain good instrument performances.
2. For micro-isolation it is important to verify that HPLC solvent purity is good enough to prevent signal interferences in the micro-NMR spectra of semi-preparative LC-MS blank samples.
3. CD₃OD is adapted for the dissolution of most biomarkers eluting in reversed phase C₁₈ separations. Other alternative solvents such as CDCl₃ or DMSO may be used to solve solubility issues.
4. A QTOF-MS system is necessary for selective MS/MS fragmentation experiments. Metabolite profiling may also be recorded on a TOF-MS (e.g. LCT Premier from Waters).
5. Other databases containing natural product molecular formulae and information related to their origin (family, genus, and species) can be used (e.g. SciFinder, <https://scifinder.cas.org>)
6. Other commercially available HPLC modelling software can be used.
7. It is difficult to estimate the amount of plant that is required to finally isolate a few tenth of µg of biomarkers, since MS detection is largely compound-dependent. As a rule of thumb, approximately 1000 times more plant material is needed for the micro-NMR detection of given peaks of the LC-MS metabolite profiles.
8. The same phase chemistry in both analytical (UHPLC) and semi-preparative scales is mandatory to maintain the same selectivity [35, 44].
9. The use of two columns coupled in series provides a good compromise at the semi-preparative scale between high chromatographic resolution and reasonable elution times [45]. If high resolution is not mandatory, shortest column may be used.
10. To ensure predictable chromatographic transfer, temperature has to be controlled at both analytical and semi-preparative levels.
11. A splitter with adjustable split ratio or a T-piece with tubing of adapted length may be used after careful measurement of the split ratio. For 10 mL/min flow rate (pre-fractionation step) a 1:200 ratio is advisable, while for 3.5 mL/min (high resolution isolation step) a 1:70 ratio can be used.
12. This metabolite dereplication process is illustrated for specific biomarkers but can be applied to any LC peak in a metabolite profiling chromatogram.
13. If the metabolites that need to be identified are sensitive to enzymatic reactions, the fresh plant tissues have to be frozen in liquid nitrogen immediately after harvesting.
14. Isopropanol was selected since it is miscible with water contained in the fresh leaves and since it extracts both relative apolar and polar metabolites. Other solvent such as MeOH or MeOH-water mixtures may be considered according to the physicochemical properties of the biomarkers of interest.
15. A SPE procedure is advisable for the removal of chlorophyll and other interfering apolar compounds for reversed phase LC-MS metabolite profiling. Such compounds are strongly retained in standard reversed phase conditions and may alter the

- chromatographic performances and reduce column's lifetimes after multiple injections.
16. Such injection solvent usually dissolves the extract. Its elution strength will not affect UHPLC separation since only a small volume is typically injected (2 μL). In case of solubility issues, the injection solvent should be adapted.
 17. This gradient is generic for many separations of complex mixtures, since it represents a good compromise between maximal peak capacity and reasonable gradient time [45].
 18. This calibration has to be performed once a week for instruments of new generation such as the Synapt G2 from Waters, or daily for older platforms such as the LCT Premier from Waters. Refer to manufacturer's recommendations.
 19. On the Synapt G2 QTOF-MS, PI and NI modes cannot be acquired in alternating scans during the same analysis. On MS instruments that can alternatively switch to PI and NI modes, MS accuracy and acquisition frequency are affected and, in our opinion, high quality data can only be obtained by performing separated PI and NI analyses with the current technology.
 20. This acquisition mode provides MS fragmentation of all metabolites in an untargeted manner when high collision induced dissociation (CID) energies are used. This information is useful for the dereplication process.
 21. Extraction of ion traces at ± 0.01 Da (0.02 Da window) reduces sufficiently the noise without loss of data using modern (Q-)TOF-MS analysers. For MS instruments providing a 5 ppm accuracy or higher, this window can be set to 0.05 Da.
 22. Combining spectra increases mass accuracy and provides better ion statistics. For high mass accuracy measurement, depending on the instrument used, working with very intense ions may cause saturation and m/z shifts. In such case, combine less intense spectra on the edges of the chromatographic peak only.
 23. If usually $[\text{M}+\text{H}]^+$ or $[\text{M}-\text{H}]^-$ ion are selected for the determination of elemental composition, any other ion species may be used, but the molecular formula should be corrected according to the type of adducts generated.
 24. Most natural products contain only those 5 elements. According to the adducts observed, one Na^+ or K^+ can be added to the list. Maximum number of CHONS can be overestimated (200) since molecular formulae will be filtered later. If based on chemotaxonomic information the search may be restricted to CHO only, the number of hits is considerably reduced. A link between MW and maximal number of given element for natural products is provided in [10].
 25. 15 ppm is selected on purpose here as an extreme case since a 5 ppm tolerance windows may be practically applied in routine for well-calibrated instruments of the last generation.
 26. This procedure can alternatively be performed on-line on freely accessible website:
<http://maltese.dbs.aber.ac.uk:8888/hrmet/search/gr.html>.
 27. Chemically non-viable formulae are discarded based on various filters, such as hydrogen/carbon ratio or Lewis rule [10].
 28. As shown, molecular formulae assignment may still be ambiguous even after heuristic filtering, and requires search in natural product databases.
 29. This step is valid only if the biomarker of interest was previously reported in databases. One should not completely exclude the presence of an unknown biomarker. In this case, a *de novo* identification procedure is necessary (see §3.4).
 30. The structure assignments made are only putative. However, if data previously

- reported in the database correspond to the same plant species the chances for a correct assignment are important. In the frame of a previous chemotaxonomic study, verification of the peak annotation by further isolation of standard revealed a very good prediction of such approach [12]. In the case of Biomarker **A**, another isomer is also present (Figures 2C-D) with a similar MS/MS spectrum. The dereplication made on MS data alone does not provide information on the stereochemistry of such compounds that are likely diastereoisomers.
31. On the Synapt G2, linkage between precursor and product ions may be automatically made by use of the "MS^E data viewer" software. Selective additional MS/MS experiments may be performed if precursor and fragment ions cannot easily be linked.
 32. It should be noted that these databases contain relatively few spectra and the risk of false positive matches is therefore high. Moreover, CID mass spectra acquired on different mass spectrometers may differ [46]. Unfortunately, no generic MS/MS library with free access exists for natural products and this represents a serious bottleneck for biomarker identification. In-house libraries may however provide very good match but are limited by the number of standards at hand [47].
 33. Absorption maxima may shift according to the solvent used. The UV-PDA spectra obtained in gradient mode may thus be slightly different from those reported in pure solvent in databases.
 34. The intensity in ESI strongly depends on the nature of the analyte. The mentioned 5 µg/mL concentration is an average value providing appropriate intensity for the majority of the compounds. On the Synapt G2, appropriate intensity is comprised between 10³ and 10⁶ cps.
 35. The spectrum pattern varies with the concentration of the biomarker of interest.
- For example, a high concentration may result in higher probability of dimerisation.
36. For large scale extraction, a maceration step instead of ball mill extraction is advisable and was shown to provide similar extract composition.
 37. This corresponds to a solid introduction of the extract on a large C₁₈ adapted SPE column (3 to 4 cm diameter).
 38. This step aims at eliminating highly polar compounds, in order to concentrate the sample.
 39. This procedure is similar to that made at the analytical level with SPE (§3.1.2) and is aimed at the removal chlorophyll and other interfering apolar compounds.
 40. Different software packages are available to calculate gradient transfers. Here, a freely available Excel-based program was used [35].
 41. An initial hold was introduced to take into account the differences in dwell volume between the systems.
 42. Injection volume is adapted in proportion to the volume of the column.
 43. A make-up flow may be added to dilute the mobile phase directed to the MS and avoid saturation effects. Delay between MS detection and fraction collection have to be minimised first.
 44. MS detection is important at the semi-preparative scale for the monitoring of the separation, to ensure a specific collection of the fraction containing the biomarker of interest based on its characteristic *m/z* ion. It is necessary to verify that the flow reaches the collection tubes and the MS detector at the same time, to ensure proper fraction collection.
 45. Osiris software requires at least two chromatograms with elution retention factors (*k_e*) of 3 and 10 approximately. This corresponds to gradient slopes of 1%/min and 3%/min in the present case.
 46. This comparison of the 1%/min and 3%/min gradient can be made on a specific enriched

fraction, but may also be performed on the crude extract directly, to calculate optimised separation conditions for several biomarkers at the same time.

47. To obtain this value, a porosity of 0.7 was assumed.

48. The selected column has the same phase chemistry as the analytical UHPLC-QTOF-MS used at the analytical level. This is important to ensure similar selectivity as shown by comparing Figures 3E and 3F. A very long column (500 mm) was selected to ensure a high chromatographic efficiency (*see Note 9*).

Acknowledgments

This work was supported by Swiss National Science Foundation grants no. 205320_135190 and CRSII3_127187. G.G. and J.L.W. are also grateful to the National Center of Competence in Research (NCCR) Plant Survival and to the Swiss Plant Science Web (SPSW).

References

- [1] W.F. Smyth, T.J.P. Smyth, V.N. Ramachandran, F. O'Donnell, P. Brooks. Dereplication of phytochemicals in plants by LC-ESI-MS and ESI-MSn. *TrAC Trends in Analytical Chemistry*, **2012**. 33: 46-54.
- [2] W. Windig, J.M. Phalp, A.W. Payne. A Noise and Background Reduction Method for Component Detection in Liquid Chromatography/Mass Spectrometry. *Analytical Chemistry*, **1996**. 68: 3602-3606.
- [3] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *Bmc Bioinformatics*, **2010**. 11: 395-405.
- [4] T. Kind, O. Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews*, **2010**. 2: 23-60.
- [5] T. Pluskal. MZMine. [Access November 29, 2011]; 2.10:[Available from: <http://mzmine.sourceforge.net/>].
- [6] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann. CAMERA: An Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid Chromatography/Mass Spectrometry Data Sets. *Analytical Chemistry*, **2011**. 84: 283-289.
- [7] T. Pluskal, T. Uehara, M. Yanagida. Highly Accurate Chemical Formula Prediction Tool Utilizing High-Resolution Mass Spectra, MS/MS Fragmentation, Heuristic Rules, and Isotope Pattern Matching. *Analytical Chemistry*, **2012**. 84: 4396-4403.
- [8] T. Kind, O. Fiehn. Metabolomic database annotations via query of elemental compositions: Mass accuracy is insufficient even at less than 1 ppm. *BMC Bioinformatics*, **2006**. 7: 234-243.
- [9] A. Pontet. Method for calculation of the number of rings in the structure of organic compounds. *Chimia*, **1951**. 5: 39-40.
- [10] T. Kind, O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **2007**. 8: 105-124.
- [11] S. Böcker, M.C. Letzel, Z. Lipták, A. Pervukhin. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics*, **2009**. 25: 218-224.
- [12] C.S. Funari, P.J. Eugster, S. Martel, P.-A. Carrupt, J.-L. Wolfender, D.H.S. Silva. High resolution ultra high pressure liquid chromatography–time-of-flight mass spectrometry dereplication strategy for the metabolite profiling of Brazilian *Lippia* species. *Journal of Chromatography A*, **2012**. 1259: 167–178.
- [13] S.M. Al-Massarani, S. Bertrand, A. Nievergelt, A.M. El-Shafae, T.A. Al-Howiriny, N.M. Al-Musayeib, M. Cuendet, J.-L. Wolfender. Acylated pregnane glycosides from *Caralluma sinaica*. *Phytochemistry*, **2012**. 79: 129-140.
- [14] K.A. Aliferis, S. Jabaji. Metabolite Composition and Bioactivity of *Rhizoctonia solani* Sclerotial Exudates. *Journal of Agricultural and Food Chemistry*, **2010**. 58: 7604-7615.
- [15] T. Kind, O. Fiehn. Seven Golden Rules Software. [Access February 16, 2011]; Available from: http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/.
- [16] J. Buckingham. *Dictionary of Natural Products on DVD*, version 21:2, **2012**, CRC Press.
- [17] C. Reichardt, T. Welton. *Solvents and solvent effects in organic chemistry*. **2011**, John Wiley & Sons.

- [18] S. Martel, D. Guillarme, Y. Henchoz, A. Galland, J.L. Veuthey, S. Rudaz, P.-A. Carrupt. Chromatographic Approaches for Measuring LogP, in *Molecular Drug Properties. Measurement and Prediction.*, R. Mannhold, Editor. **2008**. p. 331-355.
- [19] A. Nasal, R. Kaliszan. Progress in the Use of HPLC for Evaluation of Lipophilicity. *Current Computer - Aided Drug Design*, **2006**. 2: 327-340.
- [20] M.H. Abraham. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. *Chemical Society Reviews*, **1993**. 22: 73-83.
- [21] C. Stella, A. Galland, X.L. Liu, B. Testa, S. Rudaz, J.L. Veuthey, P.A. Carrupt. Novel RPLC stationary phases for lipophilicity measurement: Solvatochromic analysis of retention mechanisms for neutral and basic compounds. *Journal of Separation Science*, **2005**. 28: 2350-2362.
- [22] L.C. Menikarachchi, S. Cawley, D.W. Hill, L.M. Hall, L. Hall, S. Lai, J. Wilder, D.F. Grant. MolFind: A Software Package Enabling HPLC/MS-Based Identification of Unknown Chemical Structures. *Analytical Chemistry*, **2012**. 84: 9388-9394.
- [23] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann. *In silico* fragmentation for computer assisted identification of metabolite mass spectra. *BMC bioinformatics*, **2010**. 11: 148-159.
- [24] S. Wolf, S. Schmidt, M. Müller-Hannemann, S. Neumann. MetFrag. [Access April 25, 2013]; Available from: <http://msbi.ipb-halle.de/MetFrag/>.
- [25] Chemical Structure Lookup Service. [Access June 20, 2013]; Available from: cactus.nci.nih.gov/cgi-bin/lookup/search.
- [26] C. Guy, J. Kopka, T. Moritz. Plant metabolomics coming of age. *Physiologia Plantarum*, **2008**. 132: 113-116.
- [27] O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R.N. Trethewey, L. Willmitzer. Metabolite profiling for plant functional genomics. *Nature Biotechnology*, **2000**. 18: 1157-1161.
- [28] B. Zhou, J.F. Xiao, L. Tuli, H.W. Ressom. LC-MS-based metabolomics. *Molecular BioSystems*, **2012**. 8: 470-481.
- [29] K.A. Leiss, Y.H. Choi, R. Verpoorte, P.G. Klinkhamer. An overview of NMR-based metabolomics to identify secondary plant compounds involved in host plant resistance. *Phytochemistry Reviews*, **2011**. 10: 205-216.
- [30] J.L. Wolfender, G. Glauser, J. Bocard, S. Rudaz. MS-based Plant Metabolomic Approaches for Biomarker Discovery. *Natural Product Communications*, **2009**. 4: 1417-1430.
- [31] W.B. Dunn. Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes. *Physical Biology*, **2008**. 5: 011001.
- [32] J. Bocard, J.L. Veuthey, S. Rudaz. Knowledge discovery in metabolomics: An overview of MS data handling. *Journal of Separation Science*, **2010**. 33: 290-304.
- [33] G. Glauser, D. Guillarme, E. Grata, J. Bocard, A. Thiocone, P.-A. Carrupt, J.-L. Veuthey, S. Rudaz, J.-L. Wolfender. Optimized liquid chromatography-mass spectrometry approach for the isolation of minor stress biomarkers in plant extracts and their identification by capillary nuclear magnetic resonance. *Journal of Chromatography A*, **2008**. 1180: 90-98.
- [34] S. Heinisch, E. Lesellier, C. Podevin, J.L. Rocca, A. Tchaplá. Computerized optimization of RP-HPLC separation with nonaqueous or partially aqueous mobile phases. *Chromatographia*, **1997**. 44: 529-537.
- [35] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *European Journal of Pharmaceutics and Biopharmaceutics*, **2008**. 68: 430-440.
- [36] E. Grata, J. Bocard, D. Guillarme, G. Glauser, P.A. Carrupt, E.E. Farmer, J.L. Wolfender, S. Rudaz. UPLC-TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in *Arabidopsis thaliana*. *Journal of Chromatography B*, **2008**. 871: 261-270.
- [37] N. Huang, M.M. Siegel, G.H. Kruppa, F.H. Laukien. Automation of a Fourier transform ion cyclotron resonance mass spectrometer for acquisition, analysis, and e-mailing of high-resolution exact-mass electrospray ionization mass spectral data. *Journal of the American Society for Mass Spectrometry*, **1999**. 10: 1166-1173.

- [38] K.F. Nielsen, M. Månsson, C. Rank, J.C. Frisvad, T.O. Larsen. Dereplication of Microbial Natural Products by LC-DAD-TOFMS. *Journal of Natural Products*, **2011**. 74: 2338-2348.
- [39] K.R. Markham. *Techniques of flavonoid identification*. **1982**, Academic press.
- [40] T.F. Molinski. NMR of natural products at the 'nanomole-scale'. *Natural product reports*, **2010**. 27: 321-329.
- [41] J.L. Wolfender, E.F. Queiroz, K. Hostettmann. Phytochemistry in the microgram domain - a LC-NMR perspective. *Magnetic Resonance in Chemistry*, **2005**. 43: 697-709.
- [42] D.L. Olson, J.A. Norcross, M. O'Neil-Johnson, P.F. Molitor, D.J. Detlefsen, A.G. Wilson, T.L. Peck. Microflow NMR: concepts and capabilities. *Analytical chemistry*, **2004**. 76: 2966-2974.
- [43] G. Glauser, J. Boccard, S. Rudaz, J.-L. Wolfender. Mass spectrometry-based metabolomics oriented by correlation analysis for wound-induced molecule discovery: identification of a novel jasmonate glucoside. *Phytochemical Analysis*, **2010**. 21: 95-101.
- [44] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part I: Isocratic separation. *European Journal of Pharmaceutics and Biopharmaceutics*, **2007**. 66: 475-482.
- [45] D. Guillarme, E. Grata, G. Glauser, J.L. Wolfender, J.L. Veuthey, S. Rudaz. Some solutions to obtain very efficient separations in isocratic and gradient modes using small particles size and ultra-high pressure. *Journal of Chromatography A*, **2009**. 1216: 3232-3243.
- [46] P. Waridel, J.-L. Wolfender, K. Ndjoko, K.R. Hobby, H.J. Major, K. Hostettmann. Evaluation of quadrupole time-of-flight tandem mass spectrometry and ion-trap multiple-stage mass spectrometry for the differentiation of C-glycosidic flavonoid isomers. *Journal of Chromatography A*, **2001**. 926: 29-41.
- [47] J.J.J. van der Hooft, J. Vervoort, R.J. Bino, J. Beekwilder, R.C.H. de Vos. Polyphenol Identification Based on Systematic and Robust High-Resolution Accurate Mass Spectrometry Fragmentation. *Analytical Chemistry*, **2011**. 83: 409-416.

Chapter VI - LC-MS Online Dereplication - Practical Application to a Crude Plant Extract

This chapter is based on an article published in Journal of Chromatography A, and is the result of a collaboration with a team of University of São Paulo State, Brazil.

Foreword

This work resulted from a close collaboration of our laboratory with a Brazilian team from the University of Araquara - São Paulo involved in phytochemical investigations of plants from the Brazilian flora. The aim of this collaboration was to study the chemotaxonomic relationships between Brazilian *Lippia* species. Indeed, various taxonomic problems involving some species of this genus have been highlighted. In this collaboration, the Brazilian researchers brought their phytochemical knowledge on the *Lippia* genus and provided several isolated NPs, while

we used the tools presented in the previous chapter for profiling the extracts and comparing the different species.

A comprehensive dereplication strategy using the techniques presented in the previous chapters was carried out to get online peak annotations of numerous metabolites. Because this annotation was not unambiguous for several compounds, another filter was added to the procedure based on the standard lipophilicity parameter, $\log P$.



Figure VI.1. *Lippia salviaefolia*. Photo CS Funari, 2010.

more than 40 NPs that were further used for the chemotaxonomic study.

The species of the *Lippia* genus (Verbenaceae) that are studied in this work are distributed in South and Central Americas and in tropical Africa

with ca. 70% of all known species found in Brazil [1]. They are flowering plants widely used in folk medicine for gastrointestinal and respiratory disorders and hypertension [2], as well as food, sweetener and beverage flavouring. They are also known to contain essential oils. *Lippia salviaefolia* is shown in Figure VI.1.

High resolution ultra-high pressure liquid chromatography–time-of-flight mass spectrometry dereplication strategy for the metabolite profiling of Brazilian *Lippia* species

Cristiano S. Funari *

NuBBE – Nucleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais, São Paulo State University, Araraquara, Brazil

Philippe J. Eugster *

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Sophie Martel

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Pierre-Alain Carrupt

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Jean-Luc Wolfender

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Dulce Helena S. Silva

NuBBE – Nucleo de Bioensaios, Biossíntese e Ecofisiologia de Produtos Naturais, São Paulo State University, Araraquara, Brazil

* PJ Eugster and CS Funari contributed equally to this work. The whole isolation procedure was performed in Brazil by CSF, the analytical part was done in Geneva by PJE and CSF, while the data processing was performed by PJE.

Research article published in Journal of Chromatography A, 1259 (2012) 167– 178.

Abstract

Plants belonging to the *Lippia* genus have been widely used in ethnobotany throughout South and Central America and in tropical Africa as foods, medicines, sweeteners and in beverage flavouring. Various taxonomic problems involving some genera from Verbenaceae, including *Lippia*, have been reported. In this study, the metabolite profiling of fifteen extracts of various organs of six *Lippia* species was performed and compared using UHPLC-PDA-TOF-MS. Fourteen phenolic compounds that were previously isolated from *L. salviaefolia* Cham. and *L. lupulina* Cham. were used as references. The annotation of the remaining LC peaks was based on concomitant online high mass accuracy measurements and subsequent molecular formula assignments following these different steps: (i) elimination of non-coherent putative molecular formulae by heuristic filtering, (ii) verification of the occurrence of remaining molecular formulae in databases, (iii) cross search with reported compounds in the *Lippia* genus, (iv) match with reported UV spectra, (v) estimation of the chromatographic retention behaviour based on the log *P* parameter of reference compounds. This strategy is generic and time-saving, avoids isolation/purification procedures, enables an efficient LC peak annotation of most of the studied compounds and is well adapted for plant chemotaxonomic studies. Within this study, the interconversion of four flavanone glucoside isomers was additionally highlighted by analytical HPLC isolation and immediate analysis using fast UHPLC gradients. Dereplication results and hierarchical data analysis

demonstrated that *L. salviaefolia*, *L. balansae*, *L. velutina* and *L. sidoides* displayed significant chemical similarities, while the compositions of *L. lasiocalicina* and *L. lupulina* differed substantially.

1. Introduction

The family Verbenaceae comprises approximately 1035 species and 36 genera, including the genus *Lippia*, which is distributed throughout South and Central America and in tropical Africa, with ca. 70-75 % of the known *Lippia* species occurring in Brazil [1]. Because most *Lippia* species are aromatic, chemical and pharmacological studies of this genus have mainly focused on their essential oils. In contrast, relatively few have been focused on the non-volatile constituents [3], although infusions, decoctions and hydroethanolic extracts are often used in ethnomedicine [2, 4, 5]. Flavonoids, phenylpropanoids, naphthoquinoids and iridoid glucosides are the non-volatile compounds that are commonly reported in *Lippia* [3]. Various taxonomic problems involving some Verbenaceae genera have been reported. As a result, these plants have often been incorrectly classified, and difficulties involving the determination of geographical distributions and the number of species in certain genera have been reported [6]. Such discrepancies indicate that the taxonomic relationships within the Verbenaceae family remain unresolved, and botanical classification continues to be difficult. Efforts to solve this problem have predominantly focused on cytogenetic studies [7-9]. Plants of the *Lippia* genus have been widely used in ethnobotany around the world in foods, medicines, sweeteners and beverage flavouring [2]. In Brazil, infusions of the aerial parts of *Lippia sidoides* Cham. and of *Lippia lupulina* Cham., both investigated in the present work, are commonly employed in ethnomedicine. The former is widely used throughout the north-eastern region as a general-use antiseptic [4], whereas the latter is a medicinal plant from Cerrado bioma (south-

eastern Brazil) that is frequently used to treat mouth and throat infections [10]. Recently, *L. sidoides* was included in the Brazilian Health Ministry's priority list of 71 species for phytotherapeutic product development [11].

Therefore, with the goal of improving the chemical knowledge on species belonging to this genus, a rational dereplication strategy based on ultra-high pressure liquid chromatography (UHPLC) coupled to both photodiode array detection and time-of-flight mass spectrometry (UHPLC-PDA-TOF-MS) was developed. This platform was selected because UHPLC, which uses sub-2 μm packed columns and operates at pressures of up to 1200 bar, provides high peak capacity, sensitivity and reproducibility, while TOF-MS analysers are able to acquire broad m/z range data with high mass accuracy at an acquisition rate that is adapted to the thinner peaks of UHPLC. UHPLC-TOF-MS is now a well-established and powerful platform for both high-resolution metabolite profiling and rapid fingerprinting of crude plant extracts [12-15]. In most phytochemical profiling studies, the dereplication process, which aims at identifying known compounds in a mixture, is performed either by with low resolution LC-PDA-MS methods [16] or by extensive MS/MS or MS^n measurements [17] and recently by high resolution orbitrap LC-MS/MS measurements [18]. While comparisons of MS/MS spectra have been demonstrated to be very efficient for dereplication [18] they require however the creation of in-house databases and thus access to a high number of pure natural products. On the other hand, molecular formula has been recognised to be very useful for dereplication

[19] and this information is easily accessible in open access natural products databases and is searchable from literature data.

Based on these considerations, a dereplication approach kept as generic as possible has been developed. It is based on an original combination of high resolution profiling of the crude extracts by UHPLC, and extraction of molecular formulae based on TOF-MS data together with a rational use of various filters (heuristic filtering, retention prediction based on the $\log P$ parameter, UV

spectra matching, and chemotaxonomic cross search). In addition, a method using hierarchical cluster analysis that globally compares the chemical composition of the extracts dereplicated in this way was developed.

This generic UHPLC-PDA-TOF-MS metabolite profiling method has been applied to rapidly determine the chemical composition of different organs of selected *Lippia* species and to evaluate the potential of the method to assess chemotaxonomic relationships.

2. Experimental

2.1. Chemicals

Ethanol 95% (analytical grade) from Labsynth (Brazil) was used for plant extractions. Acetonitrile, methanol, water and formic acid used for UHPLC-PDA-TOF-MS analyses were ULC/MS grade from Biosolve (Valkenswaard, The Netherlands). Phenolics (2*R*)- and (2*S*)-3',4',5,6-tetrahydroxyflavanone 7-*O*- β -D-glucopyranoside (**1a/1b**), 6-hydroxyluteolin-7-*O*- β -D-glucopyranoside (**2**), (2*R*)- and (2*S*)-3',4',5,8-tetrahydroxyflavanone 7-*O*- β -D-glucopyranoside (**3a/3b**), (2*R*)- and (2*S*)-eriodictyol 7-*O*- β -D-glucopyranoside (**4a/4b**), lariciresinol 4'-*O*- β -D-glucopyranoside (**5**), aromadendrin (**6**), forsythoside B (**7**), verbascoside (**8**), naringenin (**10**), phloretin (**12**), asebogenin (**13**) and sakuranetin (**14**) were previously isolated from *L. salviaefolia* Cham. whereas piceid (**9**) and biochanin A 7-*O*- β -D-apiofuranosyl-(1 \rightarrow 5)- β -D-apiofuranosyl-(1 \rightarrow 6)- β -D-glucopyranoside (**11**) were isolated previously from *L. lupulina* Cham. (Figure VI.2). Their identities and purities were confirmed by 1D and 2D NMR, MS and circular dichroism analyses [20, 21].

2.2. Plant material

The aerial parts of *L. salviaefolia* Cham. and *L. velutina* were collected in Mogi-Guaçu (São Paulo State, Brazil) in 2006 (voucher specimens no Lima 90 and no. Brumati T173, respectively) and identified by Dr Inês Cordeiro ("Herbarium Maria Eneida P. Kaufmann" - Instituto Botânico de São Paulo, São Paulo, Brazil). Aerial parts of *L. balansae* Briq. and *L. lasiocalycina* Cham. were collected in Santa Cruz do Rio Pardo and Pratânia

(São Paulo State), respectively, in 2008 (voucher specimens no. FEA 402 and no. FEA 3556, respectively) and identified by Dr Giselda Durigan (Herbarium Coleção Botânica da Floresta Estadual de Assis, São Paulo, Brazil). Aerial parts and roots of *L. lupulina* Cham. and *L. sidoides* Cham. were collected in Iaras (São Paulo State, Brazil) in 2009 (voucher specimens no. FEA 3638 and no. FEA 3639, respectively) and identified by Dr Giselda Durigan (Herbarium Coleção Botânica da Floresta Estadual de Assis, São Paulo, Brazil).

2.3. Extraction and concentration

The plants were dried in an oven with forced air circulation (Fanem 320 SE, Brazil) at 45 °C and were ground in a knife mill. Each extraction was performed by maceration with three aliquots of fresh ethanol at approximately 25-30 °C. The extractions were performed in closed glass flasks covered with aluminium foil. The ratio of solvent and plant material was 7:2 (v/w). The ethanolic (EtOH) solutions were gathered, concentrated at 40 °C, dried under nitrogen, and kept at -5 °C until use. Extraction from leaves (2982.0 g) and stems (2625.2 g) of *L. salviaefolia* Cham. yielded the crude extracts *salviaefolia_L* (405.5 g; 14% yield) and *salviaefolia_S* (93.8 g; 4% yield), respectively; the extraction from flowers (83.7 g), leaves (78.9) and stems (61.3 g) of *L. balansae* Briq. gave samples *balansae_FL* (17.9 g; 21% yield), *balansae_L* (14.3 g; 18% yield) and *balansae_S* (2.8 g; 5% yield), respectively; the extraction from combined leaves and stems (380.7 g) of *L. lasiocalycina* Cham. led to sample

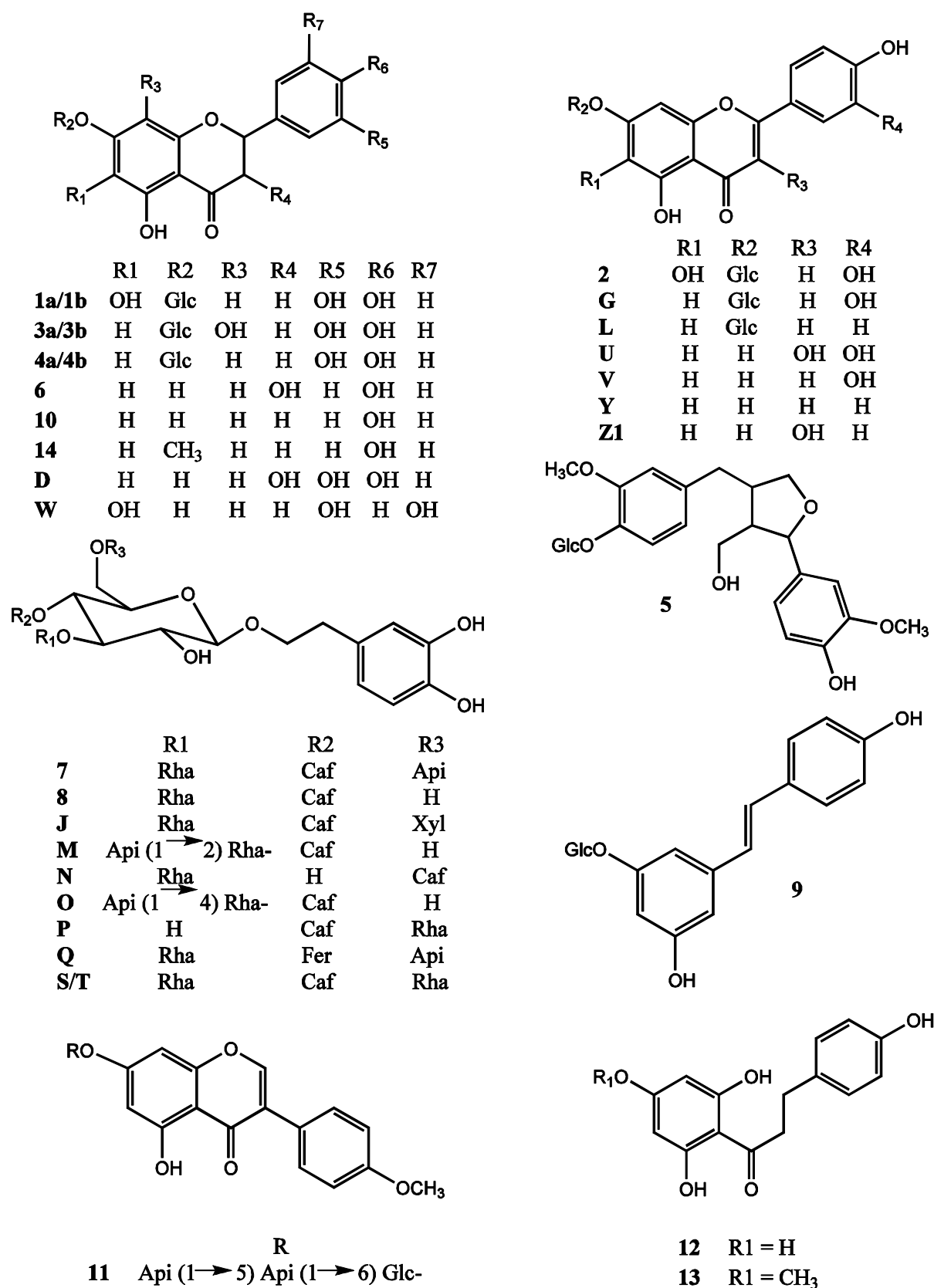


Figure VI.2. Compounds identified in the fifteen investigated extracts after the dereplication procedure corresponding to categories (I) to (III) in Table VI.1. Glucopyranosyl, rhamnopyranosyl, apiofuranosyl, xylanopyranosyl, caffeoyl and feruloyl are indicated as Glc, Rha, Api, Xyl, Caf and Fer, respectively.

lasiocalycina_LS (18.6 g; 5% yield); the extraction from leaves (106.1 g), stems (233.9 g) and roots (129.9 g) of *L. sidoides* Cham. yielded samples *sidoides_L* (33.6 g; 32% yield), *sidoides_S* (3.6 g; 2% yield) and *sidoides_R* (4.9 g; 4% yield), respectively; the extraction from leaves (15.5 g), stems (42.6 g), flowers (2.2 g) and roots (158.5 g) of *L. lupulina* Cham. gave samples *lupulina_L* (2.5 g; 16% yield), *lupulina_S* (2.6 g; 6% yield), *lupulina_F* (0.3 g; 14% yield) and *lupulina_R* (21.9 g; 14% yield), respectively; and the extraction from leaves (27.5 g) and stems (52.5 g) of *L. velutina* led to samples *velutina_L* (5.9 g; 21% yield) and *velutina_S* (2.2 g; 4% yield), respectively. A previous comparison of extracts from fresh leaves and stems of *L. salviaefolia* immediately after their collection by HPLC-UV did not show any possible degradation related to the drying and extraction process [21].

2.4. Sample preparation

EtOH extracts were treated prior to analysis using solid phase extraction (Waters Sep-Pak C₁₈, Vac 1 cm³, 100 mg). The stationary phase was activated with 1 mL of methanol (MeOH) and was equilibrated with 1 mL of 85:15 MeOH:H₂O (v/v). The cartridge was loaded with 5.0 mg of each extract, which was solubilised in 500 µL of 85:15 MeOH:H₂O (v/v). The elution was performed with 1 mL 85:15 MeOH:H₂O (v/v) to eliminate the chlorophylls and other low-polarity compounds [13]. The eluate was dried under N₂ and was solubilised in 85:15 MeOH:H₂O (v/v) to afford 1 mg/mL solutions. The isolated standards (**1** - **14**) were dissolved in the same solvents to a concentration of 1.5 µg/mL.

2.5. HPLC fractionation of isomeric flavanone glucosides

A mixture of **1a**, **1b**, **3a** and **3b** (0.15 mg) was separated using HPLC-PDA on an Agilent 1100 HPLC system (Agilent, Waldbronn, Germany)

using a C₁₈ column (Phenomenex Synergi Hydro-RP, 250 x 4.6 mm, 4 µm) with either 17:83 MeOH:H₂O (v/v) (pH ≈ 6) or 17:83 MeOH:H₂O (v/v) containing 0.1% trifluoroacetic acid (TFA) (v/v) (pH ≈ 2.5) at 1 mL/min, and the separation was monitored at 287 nm [20]. Immediately after its collection, each collected isomer was analysed by UHPLC-PDA-TOF-MS. Subsequent rapid isocratic UHPLC analyses (see below) were performed over 21 h.

2.6. UHPLC-PDA-ESI-TOF-MS experiments

Analyses were performed on a Waters Acquity UPLC system that was coupled to a Waters Micromass LCT Premier Time-of-Flight mass spectrometer (Milford, MA, USA), which was equipped with an electrospray interface (ESI). Separations were performed on a C₁₈ column (Waters Acquity UPLC BEH C₁₈, 150 x 2.1 mm, 1.7 µm). For the high-resolution metabolite profiling of the extracts, the mobile phases were H₂O (A) and acetonitrile (B), each containing 0.1% formic acid (v/v), with the following gradient: 5-60% B (0.0-52.0 min), 60-100% B (52.0-52.1 min), and 100% B (52.1-60.0 min). The flow rate was set to 600 µL/min. The temperatures in the auto sampler and in the column oven were fixed at 10 and 60°C, respectively. The UV traces were recorded from 210 to 450 nm. Analyses of each extract (1.0 µL injected, 1.5 µg on column) and chemical markers (2 ng of each) were performed in both positive (PI) and negative ionisation (NI) modes in the 100-1000 Da range with acquisition times of 0.3 s in centroid mode. The ESI conditions were set as follows: capillary voltage 2800 V, cone voltage 40 V, source temperature 120°C, desolvation temperature 330°C, cone gas flow 20 L/h, desolvation gas flow 600 L/h, and MCP (micro-channel plate) detector voltage 2400 V. The same MS conditions were used to monitor the interconversions of isomers **1a**, **1b**, **3a** and **3b** after HPLC fractionation. In this case, however, UHPLC isocratic conditions were used (17:83

MeOH:H₂O + 0.1% of formic acid (v/v) at 420 μ L/min for 12 min), the temperatures in the auto sampler and in the column oven were fixed at 25 and 40°C, respectively, and the ESI desolvation temperature was set to 250°C. The flow rate was optimised in each case to generate 90% of the maximum operating pressure (900 bar) to ensure the highest possible peak capacity.

2.7. UHPLC-PDA-ESI-TOF-MS data processing and analysis

Data were processed using MassLynx software, version 4.1 SCN#639 (Waters Corporation, Milford, MA, USA). The comparison of all of the LC peaks was performed based on a retention time shift tolerance of ± 0.05 min and an exact mass tolerance of ± 0.05 Da. For unidentified peaks, all possible molecular formulae were extracted (elements C, H, N, O, tolerance of 15 ppm, at least 2 carbons) with the Elemental Composition tool of MassLynx. The extracted

formulae were corrected by adding a hydrogen (in negative mode) and were sorted using the Seven Golden Rules of Kind and Fiehn in a Microsoft Excel file [22] with an isotopic pattern error set to 5% (10% for compounds **V**, **W**, **Y** whose peaks were less intense). The log P_{calc} values were calculated with ADME Suite 5.0 (ACD/Labs, Toronto, Canada).

2.8. Hierarchical clustering analyses

Hierarchical clustering analyses (HCAs), dendrograms and heat maps were built under the MATLAB® 7 environment (The MathWorks, Natick, MA, USA) with the clustergram routine that was implemented in the Bioinformatics Toolbox™ (version 3.3). Data were analysed without standardisation or default parameters, *i.e.*, the Euclidean distance and average linkage were used

3. Results and discussion

3.1. Optimisation of the UHPLC-PDA-TOF-MS conditions

The high-resolution metabolite profiling UHPLC-PDA-ESI-TOF-MS method was optimised using *salviaefolia_L* (Figure VI.3) and *salviaefolia_S* extracts. Based on our earlier studies [12] and in order to keep the maximum peak capacity (calculated peak capacity of 330) [23, 24] within a reasonable gradient time [12], a 150 mm column was selected, and a maximum flow rate was determined based on the backpressure that was generated to obtain the optimum peak capacity.

For an extensive survey of the metabolite compositions of the extracts, both MS positive (PI) and negative ionisation (NI) ESI modes were used. The NI mode permitted the ionisation of a greater number of compounds in the reference extracts and the available standards. Most of the profiles presented in this study were thus based on NI detection. However, for online MW assignments of the unknowns, both PI and NI data were used to ascertain the molecular formula determinations.

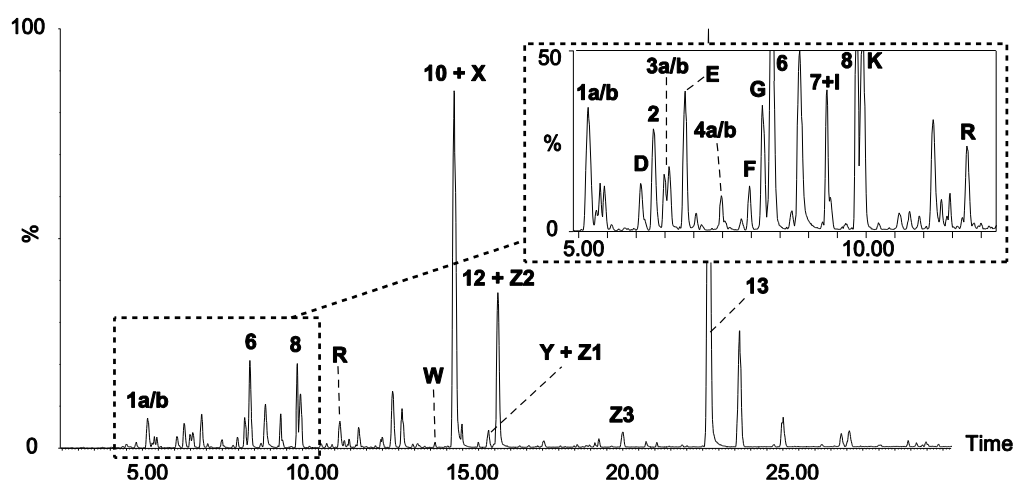


Figure VI.3. NI BPI trace of the UHPLC-ESI-TOF-MS profiling of the EtOH leaf extract of *L. salviaefolia* (*salviaefolia_L*) after blank subtraction. The compounds that were identified with standards are labelled with numbers, and the peaks that were annotated online are labelled with capital letters (for conditions, see Section 2).

3.2. Study of the interconversions of some flavanones

During the phytochemical investigation of *L. salviaefolia* [20], the isomeric flavanone glucosides **1a**, **1b**, **3a** and **3b** could not be successfully purified. Indeed, each one generated the other three compounds following isolation and storage. All of these flavanones were detected in the metabolite profile of *L. salviaefolia* (Figure VI.3), suggesting a possible interconversion. To investigate their instabilities, **1a**, **1b**, **3a** and **3b** were isolated on an analytical scale using a 120 min HPLC-PDA isocratic separation. The same separation was geometrically transferred on UHPLC-ESI-TOF-MS by a modelling software, without change in selectivity [23, 25]. This provided the separation of all isomers in less than 12 minutes for further studies of the interconversion kinetics. The collected samples were protected from light, and the stabilities were monitored at 25°C. Because high pH was previously shown to induce the non-enzymatic isomerisation of flavanones [26], the investigated compounds were maintained in acidified and non-acidified solutions (pH ≈ 2.5

and pH ≈ 6, respectively) for further comparison. Repeated injections of **1b** just after LC peak collection and at different storage times up to 21 h were performed using the fast UHPLC method. The same procedure was followed after the isolation of **1a**, **3a** and **3b**. Immediately after isolation (t = 0), at both pH ≈ 2.5 and pH ≈ 6, it was possible to observe pure **1a**, **1b**, **3a** and **3b** with very small amounts of their respective epimers. This could have been due to contamination during the LC peak collection from partially coeluting epimers (see **1a/1b** in Figure VI.4). Further analyses 1.7 h after collection showed evidence of small amounts of the constitutional isomers of each specific isolated compound in addition to its epimer. The partial interconversion of each isolated compound with its three isomers increased as a function of time. Thus, the acidified solution (pH ~ 2.5) was not efficient at stabilising the flavanone glucosides **1a/1b** and **3a/3b**, and it resulted in similar observed interconversion behaviour as the non-acidified solution (pH ~ 6). This interconversion reaction is likely due to an acid-promoted Wessely-Moser rearrangement, which often occurs to flavones or flavonoids having an unprotected hydroxyl group at position 5 [20, 27].

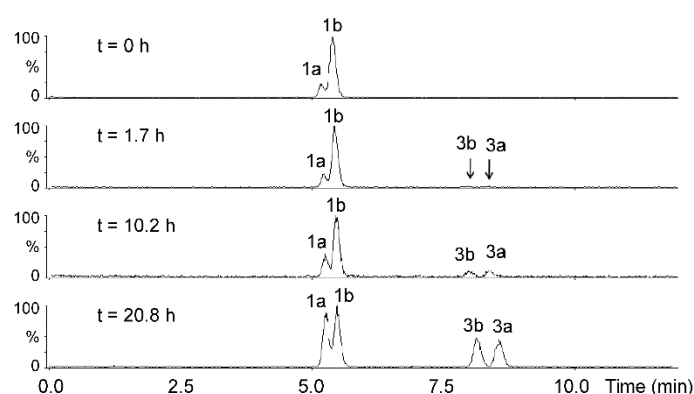


Figure VI.4. Interconversion of the isomer (*2R*)-3',4',5,6-tetrahydroxyflavanone-7-*O*- β -glucopyranoside after isolation from HPLC (**1b**). Isocratic UHPLC-ESI-TOF-MS chromatograms at selected times after collection of **1b**. UHPLC conditions: Acquity BEH C₁₈ column (150 x 2.1 mm, 1.7 μ m), 20:80 MeOH:H₂O containing 0.1% FA at 420 μ L/min.

Figure VI.4 presents selected analyses of compound (2*R*)-3',4',5,6-tetrahydroxyflavanone-7-*O*- β -glucopyranoside (**1b**) in acidified solution (pH \sim 2.5), which shows the gradual interconversion to its epimer (2*S*)-3',4',5,6-tetrahydroxyflavanone-7-*O*- β -glucopyranoside (**1a**) and its constitutional isomers (2*S*)-3',4',5,8-tetrahydroxyflavanone-7-*O*- β -glucopyranoside (**3a**) and (2*R*)-3',4',5,8-tetrahydroxyflavanone-7-*O*- β -glucopyranoside (**3b**). The order of elution and retention time of each isomer was previously established by circular dichroism (CD) from HPLC-CD-PDA analyses and nuclear magnetic resonance (NMR) experiments [20]. After 20.8 h, the chromatographic profiles were similar to those acquired from the original mixture of isomers that were obtained from the crude extract, suggesting an equilibrium state (Figure VI.4).

3.3. Comparison of the phenolic profiles of all extracts

Based on this information, all previously isolated compounds (chemical markers **1-14**) were unambiguously localised in the metabolite profile of *L. salviaefolia*, and this provided valuable information for the further assignments of non-isolated compounds. Compounds **1-14** and all *Lippia* extracts were analysed using the same UHPLC-PDA-TOF-MS conditions. The occurrences of **1-14** were examined in each extract (Table VI.1). Assessments of the presence of these compounds were based on the signal to noise ratio (S/N) of their LC peaks in their corresponding extracted single ion traces ($m/z \pm 0.05$ Da) [S/N from 3 to 10 (+); S/N from 10 to 40 (++) and S/N > 40 (+++)]. This semi-quantitative estimation could not be related to compound concentration because MS detection is compound-dependent. However, because the amounts of extract that were injected were always identical, this provided a satisfactory estimation of the relative amounts of each

metabolite that were present and facilitated the interpretation of the results in view of the statistical treatment that was applied for chemotaxonomic comparisons (see below). All extracts were compared by similarity of their chromatographic profiles and by the occurrences of the selected biomarkers [28]. Figure VI.5 shows the UHPLC-ESI-TOF-MS NI BPI (base peak intensity) traces of the leaf extracts of the six *Lippia* species that were investigated.

The comparison showed similar chromatographic profiles of the EtOH extracts from the leaves of four *Lippia* species, which included *L. salviaefolia*, *L. balansae*, *L. velutina* and *L. sidoides*, and striking differences in *L. lasiocalycina* and *L. lupulina*, in which no flavonoids were detected (Table VI.1). In addition, several LC peaks not related to the isolated compounds were detected. For a more comprehensive chemotaxonomic comparison, the assignment of some of these unknown LC peaks was made putatively based on the high-resolution MS information that was acquired online.

3.4. Dereplication procedure

Twenty-eight minor peaks (**A-Z3**) were tentatively identified (peak annotation). The TOF-MS detection provided accurate molecular weights (< 5 ppm) and retention time information for all of these compounds. Different successive filters were applied to extract and ascertain molecular formulae to reduce the number of structural possibilities. Chemotaxonomic information was then also added for the final selection of putative structures. For the most abundant unknown compounds, UV spectra were recorded and were used during the dereplication process. Based on these structural hypotheses, correlations between retention time, lipophilicity, and elution behaviour within a series of related compounds were performed.

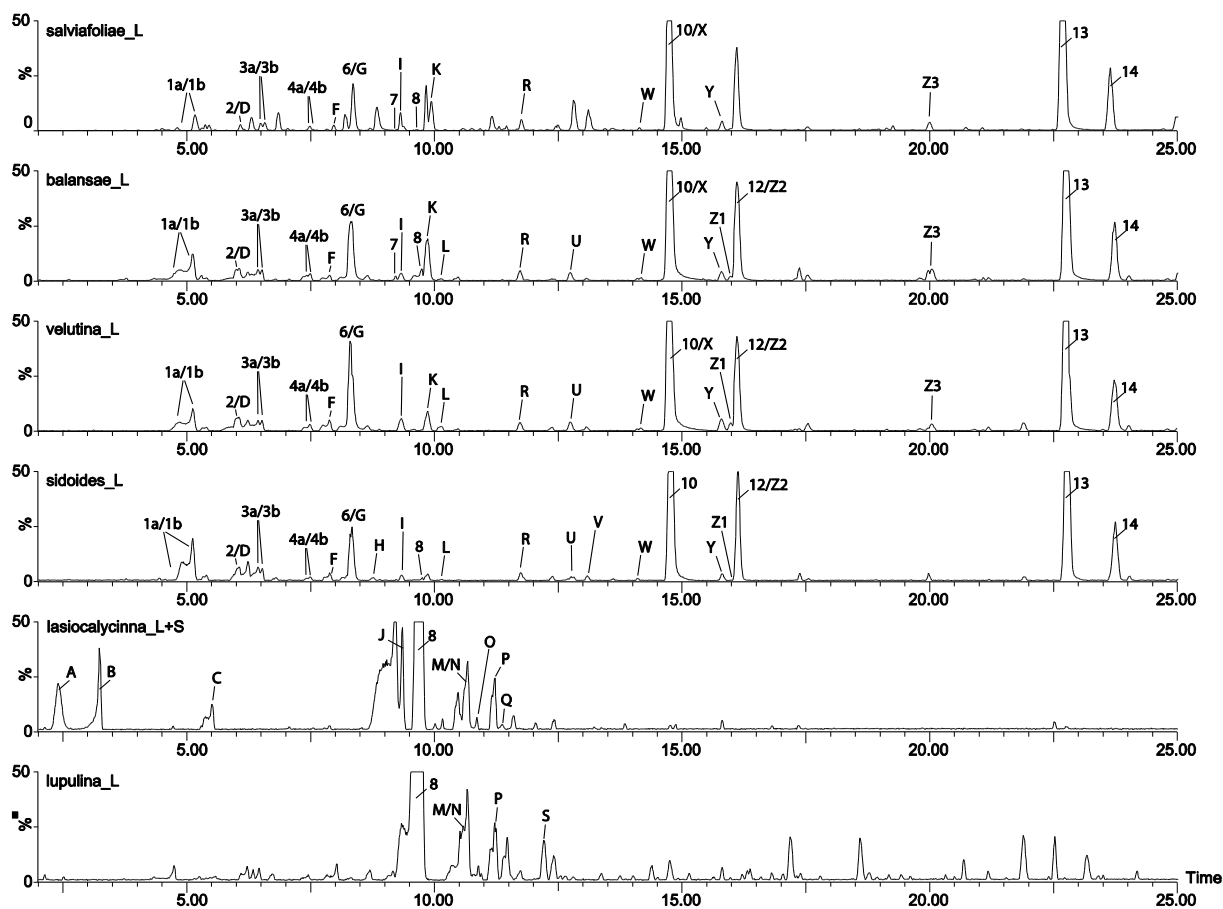


Figure VI.5. UHPLC-TOF-MS NI BPI chromatograms of EtOH leaf extracts: *salviaefolia_L*, *sidoides_L*., *velutina_L*, *balansae_L*, *lupulina_L* and *lasiocalycina_LS* (for conditions, see Section 2).

The dereplication procedure is illustrated by the dereplication of peak **V**:

(i) The first filter involved the application of a 15 ppm mass tolerance for the calculation of possible formulae that were compatible with the accurate mass obtained from the TOF-MS, with no restrictions on the numbers of C, H, O and N. The 15 ppm tolerance was applied to avoid the possible exclusion of the correct molecular formula due to possible mass accuracy shifts during the analysis despite the 5 ppm mass accuracy granted by the analyser. The use of such a large precision tolerance window provided in average 10 to 15 formulae per LC peak. This can be considered as a pitfall of the applied strategy,

however as described below, the use of heuristic filtering unambiguously reduced the number of possibilities to a maximum of 2. In the case of peak **V** ($[M-H]^-$ 285.0399) this first step generated 9 molecular formulae.

(ii) The Seven Golden Rules from Kind and Fiehn [29], which were compiled in a Microsoft Excel file, were used as the second filter. Those rules, which include the octet rule, the isotopic abundance pattern, the carbon/hydrogen ratio, and the maximum number of heteroatoms, reduced the potential formulae for peak **V** from 9 to 2. To minimise the noise contributions to the isotopic pattern, the ratio was measured at the apex of the chromatographic peak.

(iii) The third filter assessed the presence of the selected molecular formulae in databases such as the PubChem database and Chapman & Hall/CRC Dictionary of Natural Products Database [30]. This filter looked for the reported products and provided information on their nature (natural products vs. synthetic compounds). The putative molecular formulae for peak **V** were reduced to 1 single possibility, $C_{15}H_{10}O_6$, which was a formula that matched products from natural origin. However, this filter must be applied with caution because truly unknown compounds are obviously not reported in databases. Regardless, it provides significant hit reductions in dereplication studies and is well fitted for this type of study.

(iv) The fourth filter consisted of a cross search of the retained molecular formulae with phytochemical data that have previously been reported in plants of the *Lippia* genus in SciFinder and the Dictionary of Natural Products Database [30] (chemotaxonomy). In this respect, peak **V** could correspond to luteolin, which was previously isolated from *L. sidoides* [31] and *L. triphylla* [32], or to its isomer scutellarein, which was identified in *L. graveolens* [33].

(v) The orthogonal spectroscopic information provided by the UV spectra that were recorded online was used to assist in the final peak annotation, which works especially well for natural products containing characteristic chromophores such as flavonoids. In the case of peak **V**, the UV spectra of both luteolin and scutellarein could have corresponded to the UV spectrum of peak **V**; therefore, this information was not conclusive.

(vi) The last filter was a simplified retention prediction based on the standard lipophilicity parameter, $\log P$. It is well known that there is a relationship between chromatographic retention and the $\log P$ of neutral compounds [34, 35] under certain chromatographic conditions. In

UHPLC in particular, this relation has been demonstrated for a series of homologous compounds that have been analysed in both isocratic and gradient modes using Acquity BEH C_{18} and RP18 Shield columns [36]. A general estimation of the retention behaviour of natural products with very diverse structures is not currently applicable from our own experience using simple correlation models [37]. This relationship can, however, be used in homologous series of compounds under well-defined conditions (e.g., analytes in neutral form). This was the case in this study, where this filter helped to differentiate between isomeric flavonoid aglycones that possessed different $\log P$ values. The estimation of the retention behaviour was based on the use of selected internal standards (identified flavonoids) for which the $\log P$ values were calculated using the same software and conditions ($\log P_{calc}$). The compounds were aromadendrin (**6**) ($\log P_{calc} = 1.75$, RT = 8.3 min), naringenin (**10**) ($\log P_{calc} = 2.57$, RT = 14.8 min) and sakuranetin (**14**) ($\log P_{calc} = 2.74$, RT = 23.8 min). Similarly, the $\log P_{calc}$ values were calculated for all putative structures that corresponded to a given molecular formula, and these data were used to discriminate between several possibilities. In the case of peak **V**, luteolin (RT = 13.1 min) was the best candidate according to the $\log P$ filter ($\log P_{calc} = 2.27$ for luteolin and 2.71 for scutellarein). This filter could be used in another way when more than one compound possessed the same molecular formula, but they eluted with different retention times. For example, the MS filters indicated the same molecular formula ($C_{15}H_{12}O_7$) for peaks at RT 6.0 (**D**), 6.8 (**E**) and 14.2 min (**W**). Based on a chemotaxonomy cross search at the *Lippia* genus level, the hits that matched this formula were taxifolin, which was previously isolated from *L. sidoides* [31] and *L. graveolens* [33], or (2S)-5,6,7,3',5'-pentahydroxyflavanone, which was previously identified in *L. graveolens* [38].

Table VI.1. List of fully or partially identified compounds with retention times, molecular formulae and their occurrence in the *Lippia* extracts.

Cat ^a	Name (number or letter of identification)	RT (min)	Measured mass (negative mode) (Δ ppm)	Theoretical exact mass (negative mode)	Molecular formula (neutral form)	<i>salviaefolia_L</i>	<i>salviaefolia_S</i>	<i>balansae_L</i>	<i>balansae_S</i>	<i>balansae_FL</i>	<i>velutina_L</i>	<i>velutina_S</i>	<i>sidaoides_L</i>	<i>sidaoides_S</i>	<i>sidaoides_R</i>	<i>lasiocalycina_LS</i>	<i>lupulina_L</i>	<i>lupulina_S</i>	<i>lupulina_FL</i>	<i>lupulina_R</i>
(IV)	(A)	2.4	451.1444 (11.3)	451.1393	C ₂₅ H ₂₄ O ₈											++				
(IV)	(B)	3.2	451.1444 (11.3)	451.1393	C ₂₅ H ₂₄ O ₈											++				
(I)	(2R)- and (2S)-3',4',5,6-tetrahydroxyflavanone 7-O- β -glucopyranoside (1a/1b)	4.9/5.1	465.1029 (0.9)	465.1033	C ₂₁ H ₂₂ O ₁₂	++ ^b	+++	++	+++	+++	++	++	++							
(IV)	(C)	5.5	435.1473 (6.6/6.7)	435.1444 435.1502	C ₂₅ H ₂₄ O ₇ C ₁₈ H ₂₈ O ₁₂											+				
(II)	Taxifolin (D)	6.0	303.0499 (2.0)	303.0505	C ₁₅ H ₁₂ O ₇	+	++	+	++	+	+	++	+							
(I)	6-hydroxyluteolin-7-O- β -glucopyranoside (2)	6.2	463.0875 (0.4)	463.0877	C ₂₁ H ₂₀ O ₁₂	++	+	++	++	++	++		++							
(I)	(2R)- and (2S)- 3',4',5,8-tetrahydroxyflavanone 7-O- β -glucopyranoside (3a/3b)	6.4/6.5	465.1028 (1.1)	465.1033	C ₂₁ H ₂₂ O ₁₂	+	++	+	++	++	++	+	++							
(IV)	(E)	6.8	303.0523 (1.6)	303.0505	C ₁₅ H ₁₂ O ₇					+			+							

(II)	Apigenin 7-O-glucoside (L)	10.1	431.0969 (2.1)	431.0978	C ₂₁ H ₂₀ O ₁₀			++		+++		+							
(III)	Betonyoside F (M) ^c	10.6	755.2399 (0)	755.2399	C ₃₄ H ₄₄ O ₁₉		++		+			+	++	++	++	++	+		
(II)	Isoverbascoside (N)	10.7	623.1954 (3.5)	623.1976	C ₂₉ H ₃₆ O ₁₅		+		+				+	+	++	+	++	+	
(III)	Samioside (O) ^c	10.9	755.2399 (0)	755.2399	C ₃₄ H ₄₄ O ₁₉								+	++	+		++	+	
(III)	Forsythoside A (P)	11.2	623.1929 (7.5)	623.1976	C ₂₉ H ₃₆ O ₁₅		+		+			+	+	+	++	++	+	++	+
(III)	Alyssonoside (Q) ^d	11.4	769.2592 (4.8)	769.2555	C ₃₅ H ₄₆ O ₁₉		+		+			+	+	+			+		
(IV)	(R)	11.7	287.0490 (8.7)	287.0515	C ₁₀ H ₁₂ N ₂ O ₈		+	+	+	+	+	+							
(III)	Poliumoside (S) ^d	12.1	769.2592 (4.8)	769.2555	C ₃₅ H ₄₆ O ₁₉												+	+	
(III)	Poliumoside (T) ^d	12.2	769.2592 (4.8)	769.2555	C ₃₅ H ₄₆ O ₁₉				+				+						
(II)	Quercetin (U)	12.7	301.0372 (8.0)	301.0348	C ₁₅ H ₁₀ O ₇		+		+++		++	+++	++	++	++		+		
(II)	Luteolin (V)	13.1	285.0400 (0.4)	285.0399	C ₁₅ H ₁₀ O ₆							+					+		
(III)	(2S)-5,6,7,3',5'- pentahydroxyflavanone (W)	14.2	303.0531 (8.6)	303.0505	C ₁₅ H ₁₂ O ₇		+		+			+					+		
(I)	Naringenin (10)	14.8	271.0607 (0.4)	271.0606	C ₁₅ H ₁₂ O ₅		+++	++	+++	+++	+++	+++	+++	++	+++	++	+++	++	+

(IV)	(X)	14.8	317.0544 (5.7)	317.0562	C ₁₈ H ₁₀ N ₂ O ₄	+	+	+	+	+	+				
(I)	Biochanin A 7-O-β-D-apiofuranosyl-(1→5)-β-D-apiofuranosyl-(1→6)-β-D-glucopyranoside (11)	15.6	709.1994 (2.0)	709.1980	C ₃₂ O ₁₈ H ₃₇										++
(II)	Apigenin (Y)	15.8	269.0421 (10.8)	269.0450	C ₁₅ H ₁₀ O ₅	+		+		+	+				+
(II)	Kaempferol (Z1)	15.9	285.0407 (2.8)	285.0399	C ₁₅ H ₁₀ O ₆			+	++	++	++	+++	+		+
(I)	Phloretin (12)	16.1	273.0772 (3.3)	273.0763	C ₁₅ H ₁₄ O ₅	+++	+	+++		++	+++				++
(IV)	(Z2)	16.1	301.0705 (2.3)	301.0712	C ₁₆ H ₁₂ O ₆	++		+	+	+	+	+	+	+	+
(IV)	(Z3)	20.0	285.0758 (1.8)	285.0763	C ₁₆ H ₁₄ O ₅	+		+		+	+				
(I)	Asebogenin (13)	22.7	287.1922 (1.0)	287.1919	C ₁₆ H ₁₆ O ₅	+		+			+				+
(I)	Sakuranetin (14)	23.8	285.0768 (1.7)	285.0763	C ₁₆ H ₁₄ O ₅	+++	++	+++	++	+++	+++	+	+++	+	+

^a Categories:

- (I) Identified after dereplication and validated by injection of the previously isolated compounds
- (II) Identified after dereplication and validated by injection of available standards
- (III) Putatively identified after dereplication: Forsythoside F (**J**) was previously isolated from *L. canescens* [39]; Betonyoside F (**M**) and Samioside (**O**) were isolated from *Lantana trifolia* [40] and *Aloysia virgata* [41]; Forsythoside A (**P**) was isolated from *L. triphylla* [42]; Alyssonoside (**Q**), Poliumoside (**S** and **T**) were isolated from *L. salviaefolia* [20] and *Callicarpa* spp [43, 44]; (2S)-5,6,7,3',5'-pentahydroxyflavanone (**W**) was identified in *L. graveolens* [38].
- (IV) Unidentified peaks with validated molecular formulae

^b Intensities ($m/z \pm 0.05$): S/N from 3 to 10 (+); S/N from 10 to 40 (++) and S/N > 40 (+++).

^c **J**, **M** and **O** could be exchanged.

^d **Q**, **S** and **T** could be exchanged.

The calculated $\log P_{calc}$ values were 1.12 for taxifolin and 2.04 for (2*S*)-5,6,7,3',5'-pentahydroxyflavanone. This indicated that taxifolin could have corresponded either to the LC peak at RT 6.0 (**D**) or 6.8 (**E**) min but not to the peak at 14.2 min (**W**), which should have a $\log P > 1.75$ based on the internal standards (see Figure VI.3). However, the peak at 14.2 min (**W**) could have corresponded to (2*S*)-5,6,7,3',5'-pentahydroxyflavanone.

(vii) After application of these filters, one or two candidate compounds remained in many cases. Thus, the last step of the peak annotation procedure relied on the comparison with the corresponding pure compound if it was available or if it was previously isolated. MS/MS experiments could provide additional complementary information to support the peak annotation process; however they were not performed in the frame of this study.

The whole process could not be fully automated since especially the molecular weight assignment in each TOF-MS spectra recorded required manual processing and comparison of both PI and NI spectra. This step may probably be improved in future by the development of dedicated deconvolution algorithms. The semi-automated processing used required about 5 to 15 min for the dereplication of a given LC peak. The whole process was however much more efficient when working on different plants from the same genus since several compounds are shared by the species, and series of analogues are often detected.

In the case of this *Lippia* study, a total of 42 LC peaks among the various extracts compared led to interpretable spectra. Indeed, 14 peaks (**1-14**) corresponded to the phenolics previously isolated (category I in Table 1), 28 peaks **A-Z3** were unknown at this stage (Table VI.1, Figure VI.2). After application of all the filters described

above, only a single molecular formula was obtained for each peak in the large majority of all 42 detected compounds. In the other cases, two possibilities were left, and in only one case (**N**), 3 molecular formulae corresponding to natural products were possible. At the end of the dereplication process, the unknowns were finally sorted into 3 additional categories: categories II and III comprised compounds reported in the *Lippia* genus and/or more generally in the Verbenaceae family; for category II, standards were obtained and confirmed the dereplication, while no standards were available for category III; in category IV, molecular formulae were ascertained but no match in related plant species could be found and therefore no definitive peak assignment was made (see Table VI.1).

In order to illustrate representative cases that were submitted to the dereplication protocol, and besides the above discussion on compound **V**, two examples are presented below.

For example, peak **N** (RT = 10.7 min, $[M-H]^- = 623.1954$) provided 70 molecular formulae in the 15 ppm precision limit. After application of the heuristic filtering, 5 molecular formulae remained valid and only 3 among them were reported in natural products databases. According to a literature search for these formulae, two compounds were reported in the *Lippia* genus, corresponding to $C_{29}H_{36}O_{15}$: isoverbascoside which was previously isolated or identified in six *Lippia* spp., including *L. javanica* [45], *L. alba* [46], *L. dulcis* [39], *L. triphylla* [42], *L. citriodora* [47], and *L. multiflora* [48], or forsythoside A, which was isolated from *L. triphylla* [42]. The online UV spectrum observed for **N** was compatible with both compounds, with λ_{max} at 246, 294 and 331 nm [49, 50]. The $\log P$ values calculated for both compounds were 2.29 and 2.12, respectively. Those values are compatible with the retention of both compounds but are not different enough to allow a reliable discrimination. Finally, the injection of

pure isoverbascoside provided exactly the same RT, UV spectrum and MS pattern as obtained for peak **N**, and allowed its unambiguous identification.

Compounds **H** (RT = 8.8 min, $[M-H]^- = 301.0367$), and **U** (RT = 12.7 min, $[M-H]^-$, 301.0372) both provided the same molecular formula, $C_{15}H_{10}O_7$, that corresponded to the flavonoid quercetin or its isomer 6-hydroxyluteolin, which were previously isolated from *L. sidoides* [31] and *L. dulcis* [39], respectively. The UV spectrum of **U** was compatible with both flavonoids, while **H** displayed no spectrum, probably due to its low concentration. The RT of **U** was compatible with $\log P_{calc}$ of both compounds (1.82 and 1.91, for quercetin and hydroxyluteolin respectively), while the RT of **H** was too low to correspond to one of those compounds. Based on the injection of the pure standard, compound **U** was identified as quercetin, and whereas compound **H** remained unidentified.

The other peaks putatively identified after dereplication following the same strategy are indicated in Table VI.1. The putative structures are shown in Figure VI.2.

The reliability of the peak annotation procedure was verified by comparison with pure standards from our natural product library. Compounds **D**, **G**, **L**, **U**, **V**, **Y** and **Z1** (category II of Table VI.1) were thus confirmed to be the flavonoids taxifolin, luteolin 7-*O*-glucoside, apigenin 7-*O*-glucoside, quercetin, luteolin, apigenin and kaempferol, respectively, whereas **N** was the phenylpropanoid isoverbascoside. In this manner, 8 additional compounds were rapidly identified, and all matched with the available standards. This indicated that the putative structures that were determined for the other additional peaks had a high probability of being correct if they had already been reported in related plants. For compounds having no

previously reported chemotaxonomic relationships with the plants that were investigated, the acquired molecular formulae were not sufficient to formulate valuable putative structural assignments (**A**, **B**, **C**, **E**, **F**, **H**, **I**, **K**, **R**, **X**, **Z2** and **Z3**, category IV of Table VI.1). However, this enabled to unambiguously label a given LC peak, and all similar peaks in different extracts could be compared for chemotaxonomic purposes. This dereplication strategy enabled the identification of 30 phenolic compounds (categories I, II and III), among which 22 were confirmed by injection of the corresponding pure compound for confirmation and validation of the strategy. Moreover, the molecular formula of the 12 remaining compounds could be unambiguously established, while they are still unidentified. Such an approach represented thus a good compromise for a rapid and rational dereplication process, without the need to isolate all minor constituents in metabolite profiling studies, when the composition of botanically related plant species are compared.

3.5. Phytochemical and chemotaxonomic considerations

All the 42 LC peaks (categories I to IV Table VI.1) were used to evaluate the chemical composition of the fifteen extracts from 6 *Lippia* species studied. Based on their S/N ratio, their relative abundances were categorised as described in Section 3.3 for further comparison through multivariate data analysis. The comparison highlighted similar chromatographic profiles of the EtOH extracts from the leaves of four *Lippia* species, which included *L. salviaefolia* (*salviaefolia_L*), *L. balansae* (*balansae_L*), *L. velutina* (*velutina_L*) and *L. sidoides* (*sidoides_L*) (Figure VI.5). These extracts contained the flavanone glucosides **1a/1b**, **3a/3b** and **4a/4b**, flavone glucoside **2**, flavanones **6**, **10** and **14** and dihydrochalcones **12** and **13**. Furthermore, compounds **D**, **F**, **G**, **I**, **R**, **W**, **Y** and **Z2** were detected in all of the leaf extracts from these four

Lippia species (Table VI.1). A global overview of all of these observations is displayed in the hierarchical clustering analysis (HCA) of the data that are found in Table VI.1 (Figure VI.6). HCA highlighted the underlying structures from the acquired information. For example, when comparing the composition similarities among the leaf extracts from the six species, it was possible to observe a cluster composed of *balanseae_L*, *velutina_L*, *salviaefolia_L* and *sidoides_L* (see cluster in the bottom left of Figure VI.6). Also in accordance with Figure VI.6, the HCA

among the species indicated that *balanseae_L* and *velutina_L* were more closely related than *salviaefolia_L* and *sidoides_L*. The cluster on the right side of Figure VI.6 also indicated the occurrence of phenylpropanoids **7** and **8** in most of the species. A close relationship between these compounds is also highlighted by the HCA among constituents. Compounds **5**, **9**, **11**, **A**, **B**, **C**, and **S** were often detected in one species only and did not clearly assist in the differentiation of the species.

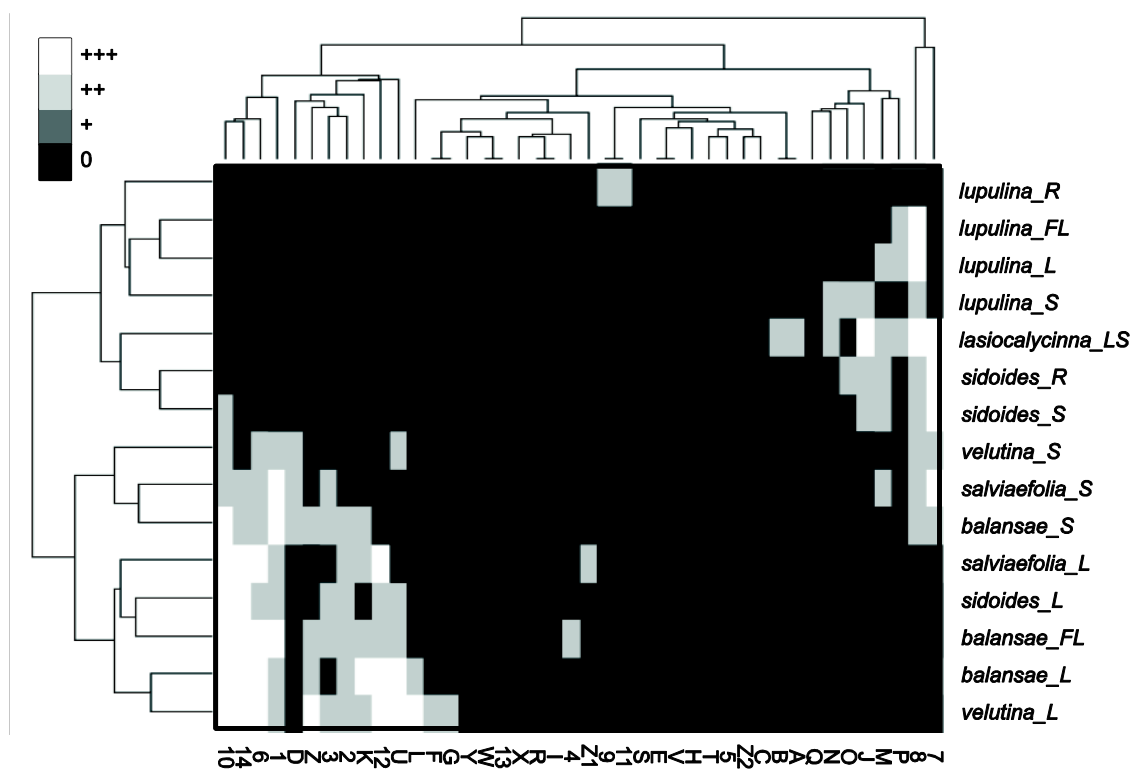


Figure VI.6. Hierarchical clustering and heat map based on the occurrences of the compounds in all extracts (ranked intensities). The grouping of the extracts is displayed horizontally. The clusters of compounds showing co-occurrence are visible in the top dendrogram. In zones of high intensities (+++), the heat map revealed clusters of similar extracts and the co-occurrence of related compounds, which explained this similarity. Extract abbreviations correspond to the species followed by the initial letter of the organ extracted: leaves (L), stems (S), flowers (F), roots (R) and leaves plus stems (L+S). The compounds that were identified with standards are labelled with numbers, and the peaks that were annotated online are labelled with capital letters.

The significant differences among the previously mentioned extracts and those from the leaves and stems of *L. lupulina* (*lupulina_L* and *lupulina_S*, respectively) and *L. lasiocalycina* (*lasiocalycina_LS*) were confirmed by Figure VI.6. This figure also provides information about the relationships between the occurrence of compounds in the extracts under investigation. For example, compounds **W** and **13** were closely related because both were present in *salviaefolia_L*, *velutina_L*, *balansae_L* and *sidoides_L*, and they were absent in the other investigated extracts. Other closely related pairs were **F** and **G**, **R** and **I**, **9** and **11**, **E** and **V**, and **B** and **A**.

It should be noted that the phenylpropanoid verbascoside (8) was present in all investigated species. Its derivative forsythoside B (7) was also widespread except in the L. lupulina Cham. extracts. The putative phenylpropanoids M, P and Q were detected in all six Lippia species that were investigated. This set of compounds thus represented a common set of markers, even if their abundances varied largely among the investigated species (Table VI.1). Phenylpropanoids are characteristic components found in the Lamiales order [51] and have been reported as verbascoside derivatives in the Verbenaceae family [52].

Regarding the chemical composition of *L. sidoides*, which has been included in the Brazilian Health Ministry's priority list of 71 species for phytotherapeutic product development [11], the following compounds were identified in this study for the first time: (2*R*)- and (2*S*)-3',4',5,6-tetrahydroxyflavanone-7-*O*- β -glucopyranoside (**1a/1b**), 6-hydroxyluteolin-7-*O*- β -glucoside (**2**), (2*R*)- and (2*S*)-3',4',5,8-tetrahydroxyflavanone-7-*O*- β -glucopyranoside (**3a/3b**), (2*R*)- and (2*S*)-eriodictyol 7-*O*- β -*D*-glucopyranoside (**4a/4b**),

aromadendrin (**6**), forsythoside B (**7**), verbascoside (**8**), naringenin (**10**), phloretin (**12**), asebogenin (**13**), sakuranetin (**14**), apigenin 7-*O*-glucoside (**L**), isoverbascoside (**N**), apigenin (**Y**) and kaempferol (**Z1**) (putative peaks later confirmed by comparison with standard) (Table VI.1). Flavonoids taxifolin (**D**), luteolin 7-*O*-glucoside (**G**), quercetin (**U**) and luteolin (**V**) detected in low amounts in the ethanol extracts have previously been reported in the polar extracts of this species together with β -sitosterol, thymol, carvacrol, isolariciresinol, lapachenol I, 6-oxo-3,4,4a,5-tetrahydro-3-hydroxy-2,2-dimethylnaphtho-1,2-pyran, tectoquinone, tecomaquinone I, tectol, acetylated tectol, and lippsidoquinone as well as palmitic, estearic, behenic, arachidic, lignoceric and 3-*O*-acetyl-oleanonic acids [31, 53-55]. Furthermore, based on the peak annotation procedure described above, there is strong evidence for the occurrence of at least seven additional compounds that have not yet been reported in *L. sidoides*, including forsythoside F, betonyoside F and samioside (**J**, **M** and **O**); forsythoside A (**P**); alyssonoside and poliumoside (**Q** or **T**); and (2*S*)-5,6,7,3',5'-pentahydroxyflavanone (**W**).

Since, to our knowledge, this is the first report on the non-volatile chemical composition of *L. balansae*, *L. velutina*, *L. lasiocalycina* and *L. lupulina*, the following compounds are reported for the first time in some of these species: compounds **1-14**, taxifolin (**D**), luteolin 7-*O*-glucoside (**G**), apigenin 7-*O*-glucoside (**L**), isoverbascoside (**N**), quercetin (**U**), luteolin (**V**), apigenin (**Y**) and kaempferol (**Z1**). Furthermore, these species could potentially contain forsythoside F, betonyoside F and samioside (**J**, **M** and **O**); forsythoside A (**P**); alyssonoside and poliumoside (**Q**, **S** or **T**); and (2*S*)-5,6,7,3',5'-pentahydroxyflavanone (**W**), as shown in Table VI.1.

4. Conclusion

High-resolution metabolite profiling by UHPLC-PDA-TOF-MS provided a rational approach to obtain a detailed analysis of the secondary metabolite composition of six *Lippia* species. In addition, the fast separation capacity of UHPLC allowed the monitoring of kinetic measurements of the interconversions of four flavanone glucosides. The high quality of the obtained profiling data and the applied multivariate data analysis provided a precise picture of the chemical relationships that exist between the various investigated species. For example, *L. salviaefolia*, *L. balansae*, *L. velutina* and *L. sidoides* displayed significant chemical similarities that differed substantially from *L. lasiocalicyna* and *L. lupulina*. Such data could contribute to the current reclassifications of this genus. The combined use of heuristic filters, chemotaxonomic information and retention time estimations based on calculated log *P* efficiently reduced this number and provided a good mean to obtain a precise picture of the chemical compositions of these various *Lippia* species without the need for the tedious isolation of all of their minor constituents. The generated data thus represent a satisfactory compromise between complete *de novo* structure determination and online putative assignments of LC peaks. The generated data contribute to the current chemical knowledge of non-volatile compounds in the *Lippia* genus, and the hierarchical clustering analyses of the results provides an efficient approach to discover cluster relationships between chemically related species. In order to enable a more significant comparison between species, a larger collection of independent specimens and the extension to a larger number of species would have been

favourable. The results obtained here already demonstrate a high degree of similarity between some of the *Lippia* species studied, and the approach is fully compatible with the study of a much larger set of samples. Although the methodology presented could not be fully automated, it is generic, practically applicable and provides a rational and robust dereplication protocol that takes advantage of the high resolution provided by conventional bench-top UHPLC-TOF-MS platform for plant profiling on both LC and MS dimensions. As this has been demonstrated, the application of different filters provides a good confidence for molecular formulae determination and subsequent LC peak annotation without the need of ultra-high resolution MS detector. The other advantage of the approach is that, since dereplication is based on a cross search between chemotaxonomy information and molecular formulae, it does not require the constitution of a dedicated database of LC-MS/MS spectra, that is known to be very efficient but unfortunately instrument dependant. In the profiling process, the simultaneous acquisition of MS/MS spectra from an untargeted manner and using generic collision-induced dissociation (CID) conditions may be considered if a quadrupole time-of-flight (QTOF) instrument is available. In the absence of MS/MS database for direct matching of the spectra, such data (assignment of characteristic fragments) can be used as an additional filter at the end of the dereplication procedure for definitive structure assignment. With the advent of mass spectrometers providing 1 ppm precision in a routine basis at a high acquisition rate, and with the development of robust algorithms for deconvolution and automated molecular weight

assignment, the efficiency of such an approach will even be improved. This procedure will be applied to other much extended plant chemotaxonomic studies.

Acknowledgements

The authors would like to thank São Paulo State Research Foundation (FAPESP), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and National Research Council (CNPq) for students (to CSF) and researcher fellowships (to DHSS); Dr. Maria Inês Cordeiro, from Instituto de Botânica do Estado de São Paulo, Dr. Giselda

Durigan, from Instituto Florestal, Floresta Estadual de Assis, and Dr. Fátima Salimena, from Universidade Federal de Juiz de Fora, for the botanical identifications. Julien Boccard from the School of Pharmaceutical Sciences in Geneva is thanked for the multivariate data analysis.

References

- [1] W.S. Judd, C.S. Campbell, E.A. Kellogg, P.F. Stevens, M.J. Donoghue. *Plant Systematics: A Phylogenetic Approach*. Systematic Biology. **1999**, Sinauer Associates.
- [2] M.E. Pascual, K. Slowing, E. Carretero, D. Sánchez Mata, A. Villar. *Lippia*: traditional uses, chemistry and pharmacology: a review. *Journal of Ethnopharmacology*, **2001**. 76: 201-214.
- [3] C.A.N. Catalan, M.E.P. Lampasona. *Oregano: the Genera Origanum and Lippia*. **2002**, London, Taylor and Francis.
- [4] H. Lorenzi, A.F.J. Matos. *Plantas medicinais no Brasil: nativas e exóticas cultivadas*. **2002**, Nova Odessa, Instituto Plantarum.
- [5] F.J.A. Matos. *Farmácias Vivas - sistema de utilização de plantas medicinais projetado para pequenas comunidades*. 4th ed. **2002**, Fortaleza, UFC.
- [6] F.R.G. Salimena. Novos Sinônimos e Tipificações em *Lippia* Sect. *Rhodolippia* (Verbenaceae). *Darwinia*, **2002**. 40: 121-125.
- [7] A.D. Brandao, L.F. Viccini, F.R.G. Salimena, A.L.L. Vanzela, S.M. Recco-Pimentel. Cytogenetic characterization of *Lippia alba* and *Lantana camara* (Verbenaceae) from Brazil. *Journal of Plant Research*, **2007**. 120: 317-321.
- [8] L.F. Viccini, D.C. Souza da Costa, M.A. Machado, A.L. Campos. Genetic diversity among nine species of *Lippia* (Verbenaceae) based on RAPD Markers. *Plant Systematics and Evolution*, **2004**. 246: 1-8.
- [9] L.F. Viccini, P.M.O. Pierre, M.M. Praca, D.C.S. da Costa, E.D. Romanel, S.M. de Sousa, P.H.P. Peixoto, F.R.G. Salimena. Chromosome numbers in the genus *Lippia* (Verbenaceae). *Plant Systematics and Evolution*, **2005**. 256: 171-178.
- [10] V.E.G. Rodrigues, D.A. Carvalho. *Plantas Medicinais no Domínio dos Cerrados*. 1st ed. **2001**, UFLA, Lavras.
- [11] R. Ministério da Saúde. *Relação Nacional de Plantas Medicinais de Interesse ao SUS*. **2009**, Brasília, DAF/SCTIE/MS.
- [12] D. Guillaume, E. Grata, G. Glauser, J.L. Wolfender, J.L. Veuthey, S. Rudaz. Some solutions to obtain very efficient separations in isocratic and gradient modes using small particles size and ultra-high pressure. *Journal of Chromatography A*, **2009**. 1216: 3232-3243.
- [13] E. Grata, J. Boccard, D. Guillaume, G. Glauser, P.A. Carrupt, E.E. Farmer, J.L. Wolfender, S. Rudaz. UPLC-TOF-MS for plant metabolomics: A sequential approach for wound marker analysis in *Arabidopsis thaliana*. *Journal of Chromatography B*, **2008**. 871: 261-270.
- [14] T.A. van Beek, K.K.R. Tetala, I.I. Koleva, A. Dapkevicius, V. Exarchou, S.M.F. Jeurissen, F.W. Claassen, E.J.C. van der Klift. Recent developments in the rapid analysis of plants and tracking their bioactive constituents. *Phytochemistry Reviews*, **2009**. 8: 387-399.
- [15] P.J. Eugster, D. Guillaume, S. Rudaz, J.L. Veuthey, P.A. Carrupt, J.L. Wolfender. Ultra High Pressure Liquid Chromatography for Crude Plant Extract Profiling. *Journal of AOAC International*, **2011**. 94: 51-70.
- [16] M. Politi, R. Sanogo, K. Ndjoko, D. Guilet, J.L. Wolfender, K. Hostettmann, I. Morelli. HPLC-UV/PAD and HPLC-MSn analyses of leaf and root extracts of *Vismia guineensis* and isolation and identification of two new bianthrone. *Phytochemical Analysis*, **2004**. 15: 355-364.
- [17] Y. Konishi, T. Kiyota, C. Draghici, J.-M. Gao, F. Yeboah, S. Acoca, S. Jarussophon, E. Purisima. Molecular Formula Analysis by an MS/MS/MS Technique To Expedite Dereplication of Natural Products. *Analytical Chemistry*, **2006**. 79: 1187-1197.

- [18] J.J.J. van der Hooft, J. Vervoort, R.J. Bino, J. Beekwilder, R.C.H. de Vos. Polyphenol Identification Based on Systematic and Robust High-Resolution Accurate Mass Spectrometry Fragmentation. *Analytical Chemistry*, **2011**. 83: 409-416.
- [19] K.F. Nielsen, M. Månsson, C. Rank, J.C. Frisvad, T.O. Larsen. Dereplication of Microbial Natural Products by LC-DAD-TOFMS. *Journal of Natural Products*, **2011**. 74: 2338-2348.
- [20] C.S. Funari, T.G. Passalacqua, D. Rinaldo, A. Napolitano, M. Festa, A. Capasso, S. Piacente, C. Pizza, M.C.M. Young, G. Durigan, D.H.S. Silva. Interconverting flavanone glucosides and other phenolic compounds in *Lippia salviaefolia* Cham. ethanol extracts. *Phytochemistry*, **2011**. 72: 2052-2061.
- [21] C.S. Funari. *Estudos químicos e biológicos de espécies do gênero Lippia (Verbenaceae) nativas no Cerrado paulista*. PhD Thesis, Universidade Estadual Paulista, **2010**. Araraquara, Brazil.
- [22] T. Kind, O. Fiehn. Seven Golden Rules Software. [Access February 16, 2011]; Available from: http://fiehnlab.ucdavis.edu/projects/Seven_Golden_Rules/.
- [23] D. Guillarme. HPLC Calculator 3.0 software. [Access March 13, 2013]; Available from: <http://www.unige.ch/sciences/pharm/fanal/lcap/telechargement-en.htm>.
- [24] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *European Journal of Pharmaceutics and Biopharmaceutics*, **2008**. 68: 430-440.
- [25] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part I: Isocratic separation. *European Journal of Pharmaceutics and Biopharmaceutics*, **2007**. 66: 475-482.
- [26] D. Wistuba, O. Trapp, N. Gel-Moreto, R. Galensa, V. Schurig. Stereoisomeric Separation of Flavanones and Flavanone-7-O-glycosides by Capillary Electrophoresis and Determination of Interconversion Barriers. *Analytical Chemistry*, **2006**. 78: 3424-3433.
- [27] Z. Wang. Wessely-Moser Rearrangement, in *Comprehensive organic name reactions and reagents*. **2009**, John Wiley & Sons, Inc.: Hoboken, NJ, USA. p. 2983.
- [28] A. Urbain, A. Marston, E. Marsden-Edwards, K. Hostettmann. Ultra-performance Liquid Chromatography/Time-of-flight Mass Spectrometry as a Chemotaxonomic Tool for the Analysis of Gentianaceae Species. *Phytochemical Analysis*, **2009**. 20: 134-138.
- [29] T. Kind, O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **2007**. 8: 105-124.
- [30] J. Buckingham. *Dictionary of Natural Products on DVD*, version 18:2, **2009**, Chapman & Hall/CRC.
- [31] S.M.O. Costa, T.L.G. Lemos, O.D.L. Pessoa, C. Pessoa, R.C. Montenegro, R. Braz-Filho. Chemical Constituents from *Lippia sidoides* and Cytotoxic Activity. *Journal of Natural Products*, **2001**. 64: 792-795.
- [32] H. Skaltsa, G. Shammis. Flavonoids from *Lippia citriodora*. *Planta Medica*, **1988**. 54: 465.
- [33] L.-Z. Lin, S. Mukhopadhyay, R.J. Robbins, J.M. Harnly. Identification and quantification of flavonoids of Mexican oregano (*Lippia graveolens*) by LC-DAD-ESI/MS analysis. *Journal of Food Composition and Analysis*, **2007**. 20: 361-369.
- [34] K. Valko. Application of high-performance liquid chromatography based measurements of lipophilicity to model biological distribution. *Journal of Chromatography A*, **2004**. 1037: 299-310.
- [35] A. Nasal, R. Kaliszan. Progress in the Use of HPLC for Evaluation of Lipophilicity. *Current Computer - Aided Drug Design*, **2006**. 2: 327-340.
- [36] Y. Henchoz, D. Guillarme, S. Rudaz, J.L. Veuthey, P.A. Carrupt. High-throughput log P determination by ultraperformance liquid chromatography: A convenient tool for medicinal chemists. *Journal of Medicinal Chemistry*, **2008**. 51: 396-399.
- [37] P. Eugster, S. Martel, D. Guillarme, P.A. Carrupt, J.L. Wolfender. Rapid log P determination of natural products in crude plant extracts from UHPLC-TOF-MS profiling data - an additional parameter for dereplication and bioavailability. *Planta Medica*, **2009**. 75: 913-914.

- [38] M.C. González-Guërecia, M. Soto-Hernández, M. Martínez-Vázquez. Isolation of (-)(2S)-5,6,7,3',5'-pentahydroxyflavanone-7-O- β -D-glucopyranoside, from *Lippia graveolens* H.B.K. var. *berlandieri* Schauer, a new anti-inflammatory and cytotoxic flavanone. *Natural Product Research*, **2010**. 24: 1528-1536.
- [39] F. Abe, T. Nagao, H. Okabe. Antiproliferative Constituents in Plants 9. Aerial Parts of *Lippia dulcis* and *Lippia canescens*. *Biological and Pharmaceutical Bulletin*, **2002**. 25: 920-922.
- [40] L.d.S. Julião, S.G. Leitão, C. Lotti, A.L. Picinelli, L. Rastrelli, P.D. Fernandes, F. Noël, J.-P.B. Thibaut, G.G. Leitão. Flavones and phenylpropanoids from a sedative extract of *Lantana trifolia* L. *Phytochemistry*, **2010**. 71: 294-300.
- [41] C.M.A. de Oliveira, C.C. da Silva, H.D. Ferreira, G.D. Lemes, E. Schmitt. Kauranes, phenylethanoids and flavone from *Aloysia virgata*. *Biochemical Systematics and Ecology*, **2005**. 33: 1191-1193.
- [42] T. Nakamura, E. Okuyama, A. Tsukada, M. Yamazaki, M. Satake, S. Nishibe, T. Deyama, A. Moriya, M. Maruno, H. Nishimura. Acteoside as the analgesic principle of cedron (*Lippia triphylla*), a Peruvian medicinal plant. *Chemical & Pharmaceutical Bulletin*, **1997**. 45: 499-504.
- [43] T. Yamasaki, C. Masuoka, T. Nohara, M. Ono. A new phenylethanoid glycoside from the fruits of *Callicarpa japonica* Thunb. var. *luxurians* Rehd. *Journal of Natural Medicines*, **2007**. 61: 318-322.
- [44] K.A. Koo, S.H. Sung, O.H. Park, S.H. Kim, K.Y. Lee, Y.C. Kim. *In vitro* neuroprotective activities of phenylethanoid glycosides from *Callicarpa dichotoma*. *Planta Medica*, **2005**. 71: 778-780.
- [45] D.K. Olivier, E.A. Shikanga, S. Combrinck, R.W.M. Krause, T. Regnier, T.P. Dlamini. Phenylethanoid glycosides from *Lippia javanica*. *South African Journal of Botany*, **2010**. 76: 58-63.
- [46] P. Timoteo, A. Karioti, S.G. Leitão, F.F. Vincieri, A.R. Bilia. HPLC/DAD/ESI-MS Analysis of Non-volatile Constituents of Three Brazilian Chemotypes of *Lippia alba* (Mill.) N. E. Brown. *Nat. Prod. Commun.*, **2008**. 3: 2017-2020.
- [47] A.R. Bilia, M. Giomi, M. Innocenti, S. Gallori, F.F. Vincieri. HPLC-DAD-ESI-MS analysis of the constituents of aqueous preparations of verbena and lemon verbena and evaluation of the antioxidant activity. *Journal of Pharmaceutical and Biomedical Analysis*, **2008**. 46: 463-470.
- [48] K. Taoubi, M.T. Fauvel, J. Gleye, C. Moulis, I. Fouraste. Phenylpropanoid glycosides from *Lantana camara* and *Lippia multiflora*. *Planta Medica*, **1997**. 63: 192-193.
- [49] R.W. Owen, R. Haubner, W. Mier, A. Giacosa, W.E. Hull, B. Spiegelhalder, H. Bartsch. Isolation, structure elucidation and antioxidant potential of the major phenolic and flavonoid compounds in brined olive drupes. *Food and Chemical Toxicology*, **2003**. 41: 703-717.
- [50] J. Petreska, M. Stefova, F. Ferreres, D.A. Moreno, F.A. Tomas-Barberan, G. Stefkov, S. Kulevanova, A. Gil-Izquierdo. Potential bioactive phenolics of Macedonian *Sideritis* species used for medicinal "Mountain Tea". *Food Chemistry*, **2011**. 125: 13-20.
- [51] A. Delazar, S. Gibbons, Y. Kumarasamy, L. Nahar, M. Shoeb, S.D. Sarker. Antioxidant phenylethanoid glycosides from the rhizomes of *Eremostachys glabra* (Lamiaceae). *Biochemical Systematics and Ecology*, **2005**. 33: 87-90.
- [52] F.G. Barbosa, M.A.S. Lima, R. Braz-Filho, E.R. Silveira. Iridoid and phenylethanoid glycosides from *Lippia alba*. *Biochemical Systematics and Ecology*, **2006**. 34: 819-821.
- [53] T.L.G. Lemos, S.M.O. Costa, O.D.L. Pessoa, R. Braz-Filho. Total assignment of ¹H and ¹³C NMR spectra of tectol and tecomaquinone I. *Magnetic Resonance in Chemistry*, **1999**. 37: 908-911.
- [54] L.M.A. Macambira, C.H.S. Andrade, F.J.A. Matos, A.A. Craveiro, R.B. Filho. Naphthoquinoids from *Lippia sidoides*. *Journal of Natural Products*, **1986**. 49: 310-312.
- [55] A.K.L. Santos, J.C. Assuncao, A.M. Fonseca, O.D.L. Pessoa, F.J.Q. Monte, T.L.G. Lemos, R. Braz-Filho. Total assignments of ¹H and ¹³C NMR spectra of isocatalpanol and a derivative of tecomaquinone. *Magnetic Resonance in Chemistry*, **2005**. 43: 582-584.

Chapter VII – Retention Prediction: an Additional Tool for Dereplication

This chapter is based on an article submitted to the journal Analytical Chemistry in July 2013.

Foreword

The use of LC-MS is more and more frequent in NP research, in particular for dereplication and identification purposes. Metabolite identification however mainly relies on the MS information alone, while the chromatographic information is almost unexploited. The LC dimension, however, holds valuable structural information on the analytes, since many models used in medicinal chemistry aim at relating the retention to several physicochemical parameters. As an example, correlation between solvatochromic parameters [1, 2] and the $\log P$ parameter (see below), were often highlighted, while on the contrary several models aim at predicting the retention time (RT) from selected physicochemical parameters [3].

Based on these considerations, the study of the chromatographic behaviour of the analytes can be of great use in NP research, e.g. to evaluate their (a) *drugability*, or (b) pK_a , or (c) to be used as an additional filter in dereplication procedures.

(a) The *drugability* of the compounds (see Chapter I) may be partly evaluated online, avoiding the tedious isolation of compounds that cannot become drugs. For example, the $\log P$, which is one of the features of the famous Lipinski's Rule of Five, may be determined from the LC retention in specific conditions [4, 5]. The $\log P$ is the octanol–water partition coefficient used as a lipophilicity parameter, and is a key parameter involved in pharmacokinetic (ADME) and pharmacodynamic processes (ligand-target interactions).

(b) The ionisation of the analytes may be experimentally determined by the study of their

chromatographic behaviour at different pH values [6]. Figure VII.1 represents schematically the retention of codeine and escin at pH 2.5, 5.5 and 10.5. The behaviour of codeine (Figure VII.1A) provides much information on its ionisable site(s). Firstly, because the retention is similar at pH 2.5 and 5.5, the number of charges is the same in both conditions. Moreover, given that the ionisation is incomplete in a range of 2 pH units around the pK_a , and that the ionisation is similar at both pH, one can consider that there is no pK_a between 0.5 and 7.5. Secondly, there is one (or more) supplementary charge at pH 10.5, making the molecule more polar compared to its form at pH 2.5/5.5. This is probably due to an alkaline function in which the pK_a is comprised between 7.5 and 12.5. Indeed, codeine is a tertiary amine with a pK_a of 8.22. Escin, on the contrary, is an acidic compound with a pK_a lower than 3.5 since the retention of the molecule is similar at pH 10.5 and 5.5, but higher at pH 2.5 (Figure VII.1B). In summary, some interesting ionisation information of analytes can be directly extracted from the retention behaviour measured at different pH. Provided that NPs can be ionised in MS under these various pH conditions, a systematic tracking of these changes in retention time can be performed in crude extract metabolite profiling. The change in retention provides valuable information on functional groups of NPs that need to be identified.

(c) In dereplication procedures, the chromatographic information complements ideally the one extracted from MS. Indeed, the

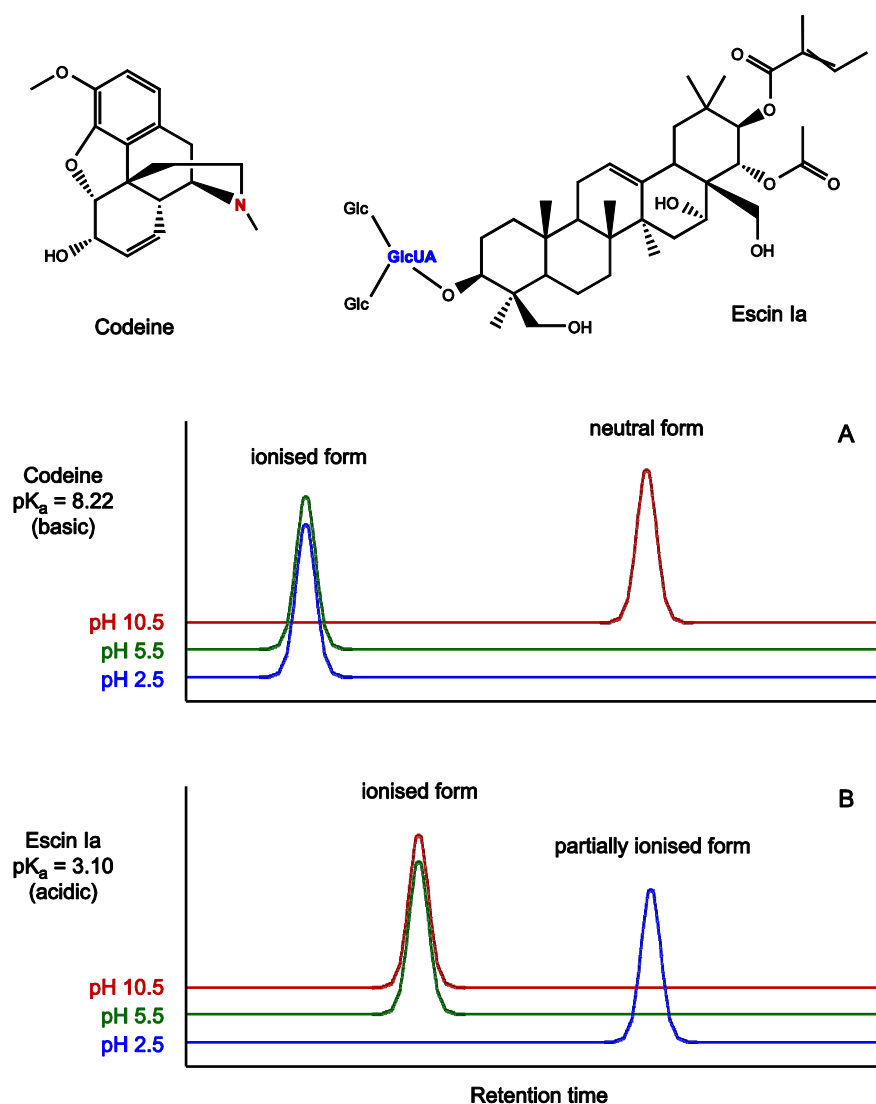


Figure VII.1. Schematic representation of the reversed-phase LC retention of codeine (basic compound, pK_a 8.22, **A**) and escin (acidic compound, pK_a 3.10, **B**) in three different chromatographic conditions (pH 2.5, 5.5 and 10.5).

previous chapters described efficient dereplication tools based on MS data designed to assign the correct molecular formula to a given LC peak. However, these tools do not provide an unambiguous identification for a single LC peak, and this often leads to more than one putative structure. In this respect, the retention information may be used as an additional dereplication tool thanks to a new RT prediction method based on the structure of the analyte. The predicted RT of the structures previously selected by the dereplication procedure can be compared with the experimental RT value of the

LC peak to assess the putative identification of the compound.

The chromatographic dimension is thus not only used to separate analytes in NP analysis, but may also bring valuable structural information on these compounds. This information is however difficult to obtain because of inter-instrument variability (in terms of column geometry and chemistry, analytical parameters and even LC system characteristics). Therefore, it is mandatory to standardise the chromatographic conditions at least in a given laboratory.

This chapter presents a method for the prediction of the retention of NPs in UHPLC based on physicochemical parameters calculated from their structures. This provides predicted RTs that can be matched with that of unknown metabolites analysed with the generic method presented in Chapter V that is now routinely used in our laboratory for high resolution profiling of complex natural samples.

Retention time prediction for dereplication of natural products ($C_xH_yO_z$) in LC-MS metabolite profiling

Philippe J. Eugster

Julien Boccard

Benjamin Debrus

Lise Bréant

Jean-Luc Wolfender

Sophie Martel

Pierre-Alain Carrupt

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Research article submitted to the journal Analytical Chemistry in July 2013.

Abstract

The detection and identification of natural products (NPs) require efficient, high-resolution methods for the profiling of crude natural extracts. This task is difficult because of the high number of NPs in these complex biological matrices and because of their very high chemical diversity. Metabolite profiling using ultra-high pressure liquid chromatography coupled to high-resolution mass spectrometry (UHPLC-HR-MS) is very efficient for the separation of complex mixtures and provides molecular formula information as a first step in dereplication. This structural information alone or even combined with chemotaxonomic information is often not sufficient for unambiguous metabolite identification. In this study, a representative set of 260 NPs containing C, H, and O atoms only was analysed in generic UHPLC-HR-MS profiling conditions. Two quantitative structure retention relationship (QSRR) models were built based on the measured retention time and on eight simple physicochemical parameters calculated from the structures using partial least square regressions and an artificial neural network. Prediction by PLS required that several models based on phytochemical classes be built to obtain satisfactory results, while a unique ANN model was able to provide similar results on the whole set of NPs. The retention prediction methods described in this study were found to improve the level of confidence of the identification of a given analyte among putative isomeric structures. Its applicability was verified for the dereplication of NPs in model plant extracts.

1. Introduction

Natural products (NPs) play a central role in the pharmaceutical, food, flavour, and fragrance industries and are important in chemical-biology studies [7, 8]. They occur in all living organisms as parts of very complex mixtures that may contain up to a few thousand metabolites (the metabolome) [9]. In all studies dealing with NP research (e.g., bioactivity-guided isolation [10], metabolomics [9], phytochemical investigations [11], and quality control [12]), the unambiguous identification of the NPs is a key procedure. This procedure, however, still represents a major bottleneck when profiling natural extracts because of the lack of generic databases (especially LC-MS and LC-MS/MS databases) that could provide efficient early metabolite identification of previously reported NPs. This putative or partial online identification is either known as LC peak annotation (mainly in metabolomics [13]) or as dereplication (mainly in NP research [14]). The latter term will be mainly used in this paper.

Because pure standards are often not available, dereplication of NPs has to rely on published data and chemotaxonomy [14]. This process has already been significantly improved by the recent increase in the use of high resolution mass spectrometers (HR-MS), such as time-of-flight instruments (TOF-MS) in hyphenation with LC [15]. Indeed, the high mass and spectral accuracies provided by LC-HR-MS instruments offer unambiguous molecular formula assignment when used with an adapted heuristic filtering procedure [16]. In NP research, molecular formulae can be used in combination with chemotaxonomic information to generate

putative structure attribution to a given LC peak [14, 17].

The preliminary study of the composition of natural extracts (metabolite profiling) is thus mainly performed using LC-HR-MS [15]. In this process, the efficient separation of the metabolites in a mixture prior to MS detection is important to generate high quality data for dereplication. In the majority of applications, this profiling is performed using HPLC on reversed phase C18 columns using generic gradients. The recent introduction of ultra-high pressure liquid chromatography (UHPLC) has further improved the resolution and throughput of the separation as well as the reproducibility of the retention times (RTs) [15, 18, 19].

In addition to the molecular formula, which can be used as a generic parameter for database searches, the RT of the analytes is the other accessible information in metabolite profiling. RT is correlated to the physicochemical parameters of the analyte, and efforts to predict this parameter based on its structure are needed, especially for NPs that have extremely diverse properties and are known to occupy a large chemical space [8, 20].

In this respect, quantitative structure-retention relationship (QSRR) approaches attempt to model the relationship between chromatographic retention and molecular parameters derived from the analyte. Several QSRR models have already been developed and are summarised in three comprehensive reviews [3, 21, 22]. Most of the models provided satisfying RT predictions for series of homologues

or closely congeneric compounds, such as small neutral compounds [23] or small peptides [24], or, rarely, for homologous series of NPs, such as phenolic compounds [25]. However, to our knowledge no model able to predict the RT of a large array of NPs has been published.

Most QSRR models are built using multiple linear regression (MLR) [24, 26], partial least squares (PLS) regression [27], and artificial neural networks (ANN) [25, 28]. The models usually predict retention factors (k), their logarithm ($\log k$) or the RT from a representative calibration set of analytes. The predicted RT (RT_{pred}) of new compounds is then valid only for the chromatographic conditions used when analysing the calibration set. Several different physicochemical parameters are used as predictor variables in the models. Solvatochromic parameters, such as Abraham's descriptors [2], are often used as explanatory variables to provide models based on linear solvation energy relationships (LSER) [1]. $\log P$, the octanol–water

partition coefficient used as the lipophilicity parameter, is frequently determined using reversed-phase HPLC [4, 5] and has also been used for retention prediction [3]. Finally, other types of informative molecular descriptors may be used, such as the total dipole moment, water accessible molecular surface area or molecular volumes [21].

The aim of this work was to develop a QSRR method for the RT prediction of NPs based on physicochemical parameters easily calculable from their structure. The model had to fit to generic LC-MS conditions for the metabolite profiling of complex matrices, such as crude extracts [15]. The prediction of RTs should ideally be accurate enough to be used as an additional filter in dereplication studies. To accomplish this goal, PLS and ANN models relating RTs and physicochemical parameters were developed using a selected dataset of 260 NPs that are not ionised at pH 2-4 and that represent a large chemical diversity.

2. Experimental Section

2.1. Chemicals and Sample Preparation

The NPs of the calibration and validation sets (Table VII.S1 in Supporting information) were obtained from Sigma Aldrich-Fluka (Buchs, Switzerland) or Roth Chemicals (Karlsruhe, Germany) in the highest commercially available purity. Samples were prepared in solvents of analytical grade that were as similar to the mobile phase as possible, *i.e.*, water, acetonitrile (ACN), or water-ACN mixtures or, rarely, acetone, ensuring their total dissolution. Concentrations were adjusted to provide intensities ranging from 10^3 to 10^4 in MS detection and were usually between 0.1 and 10 $\mu\text{g}/\text{mL}$. The water, ACN and formic acid used for UHPLC-TOF-MS analyses were ULC/MS grade from Biosolve (Valkenswaard, The Netherlands).

Panax ginseng and *Ginkgo biloba* extracts used in the examples are standardised extracts obtained from Indena (Milan, Italy). The extracts were dissolved in 85% MeOH at a final concentration of 5 mg/mL.

2.2. UHPLC-TOF-MS Experiments

Analyses were performed on a Waters Acquity UPLC system coupled to a Waters Micromass LCT Premier Time-of-Flight mass spectrometer (Waters, Milford, MA, USA), which was equipped with an electrospray interface (ESI). Separations were performed on a C18 column (Waters Acquity UPLC BEH C18, 150 mm \times 2.1 mm, 1.7 μm). The mobile phase was composed of water

(A) and ACN (B), each containing 0.1% formic acid (v/v). A gradient (5–95% B) was carried out in 30.0 min, followed by a 10.0 min isocratic step at 95% B, a decrease of B from 95 to 5% in 0.2 min and a second 9.8 min isocratic step at 5% B for column reconditioning. The flow rate was set to 460 $\mu\text{L}/\text{min}$. The temperatures in the auto sampler and in the column oven were fixed at 10 and 40 $^{\circ}\text{C}$, respectively. Analyses of each sample (2.0 μL injected in the partial loop with needle overflow mode) were separately performed in both positive (PI) and negative (NI) ionisation modes in the 100–1000 Da range with acquisition times of 0.3 s in the centroid mode. The ESI conditions were set as follows: capillary voltage 2800 V, cone voltage 40 V, source temperature 120 $^{\circ}\text{C}$, desolvation temperature 300 $^{\circ}\text{C}$, cone gas flow 20 L/h, desolvation gas flow 800 L/h, and MCP (microchannel plate) detector voltage 2450 V. The MassLynx software 4.1, SCN 639 (Waters, Milford, USA) was used to extract the RTs from the chromatograms. A solution containing both rutin (20 $\mu\text{g}/\text{mL}$, RT = 5.37 min) and glycyrrhetic acid (10 $\mu\text{g}/\text{mL}$, RT = 20.39 min) was injected before and after each series of analyses to check the reliability of the measured RTs.

2.3. Calculation of the Physicochemical Parameters

The structure of each NP was downloaded from the Pubchem Library [29, 30] as SDF files and converted into a molecular formula and a SMILES code using TSAR 3D software (version 3.3, Oxford Molecular Ltd, Oxford, UK). The physicochemical parameters used in the models as independent variables were calculated for each compound using the SMILES codes as the input in the

predictive tools. The online ACD/I-lab [31] provided the solvatochromic parameters, *i.e.*, hydrogen-bond acidity (α^H) and basicity (β^H), and the dipolarity-polarisability parameter (π^*), as well as the McGowan volume (V). The calculated $\log P$ ($S+\log P$), the molecular weight (MW), and the topological polar surface area in square angstroms ($TPSA$) were provided by the free MedChem Designer software from Simulations Plus, Inc. [32]. The online calculator from Molinspiration Cheminformatics [33] provided the number of rotatable bonds (rot_bond). Note that $TPSA$, MW , and rot_bond may also be obtained from the Pubchem Library.

2.4. Clustering and PLS Regressions

The clustering of the NPs was based on the classification from the Dictionary of Natural Products (DNP) [34] coded as $Vxyyyy$, where Vx corresponds to the main NPs classes and $yyyy$ to the subclass codes. The codes are described in the table inset in Figure VII.2. PLS models were built on calibration sets (75% of the compounds contained in the database or in the cluster) using the TSAR software with leave-one-out (LOO) cross validation. The number of latent variables

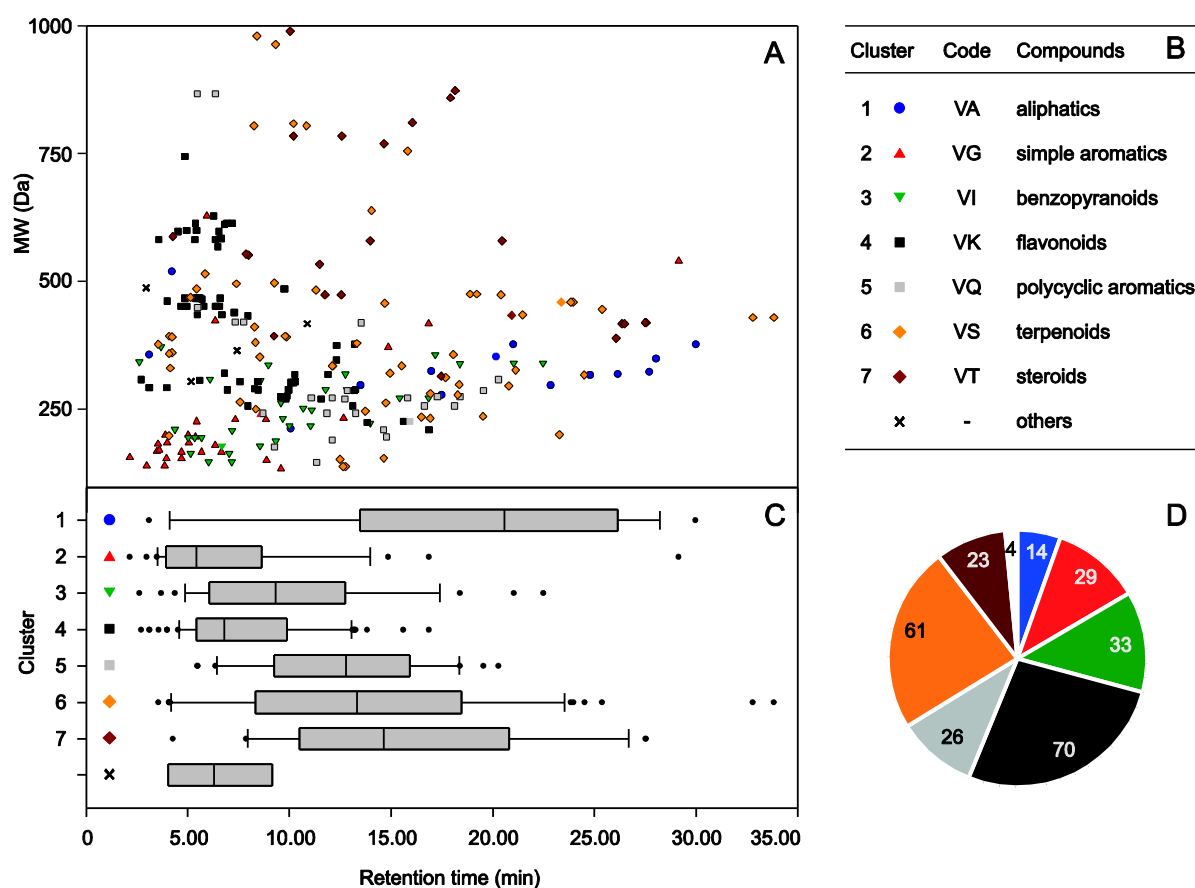


Figure VII.2. Description of the selected NPs for the construction of the model and its validation. **(A)** MW vs. RT plot showing the diversity of the NPs of the database. **(B)** Clustering of the NPs based on the classification from the DNP [34] ('code'). **(C)** Box plots representing the RT ranges of the NPs contained in each cluster. The central black vertical line represents the median, the extremities of the grey boxes indicate the 25th and 75th percentiles, the vertical lines placed at both ends of the box plots indicate the 10th and 90th percentiles, and compounds above and below the 90th and 10th percentiles are marked by single dots. **(D)** Numbers of NPs in the clusters.

was chosen to maximise the predictive ability index (Q^2_{LOO}) but was limited to 1/5 of the number of NPs included in the calibration set as an upper value. The RTs of the 25% remaining compounds were predicted using the new model for validation purposes. An Excel sheet containing the equations of the final PLS models is available as Supporting Information.

2.5. ANN Models

ANNs models were computed with the Neural Network Toolbox™ version 8.0 in the MATLAB® environment (The MathWorks, Natick, USA). A fully interconnected feedforward multilayer perceptron architecture was built with an input layer of eight units related to computed physicochemical parameters, a hidden layer optimised to five units and an output layer of one unit corresponding to the RT values of the compounds. The default tan-sigmoid and linear

transfer functions were used in the hidden and output layers, respectively.

A nested LOO cross-validation was carried out with an outer loop implementing the LOO procedure and with an inner loop devoted to network training using the Levenberg–Marquardt backpropagation algorithm [35, 36]. Optimisation was performed by randomly dividing the dataset into calibration (75%) and validation sets (25%). ANN weights and bias values were then iteratively adjusted to fit the training set, and the generalisation ability of the network was assessed with the test set using the mean square error (MSE) of prediction. Weight optimisation ended when generalisation stopped improving. After training the network, the remaining observation of the LOO outer loop was predicted. This strategy allowed proper network optimisation and avoided overfitting. A Matlab file containing the ANN model is available as Supporting Information.

3. Results and Discussion

3.1. NP Database of RT and Physicochemical Parameters

NPs available as standards and compatible with LC-ESI-MS analysis were selected. The compounds were chosen to ideally represent the main class of NPs encountered in crude plant extracts in particular. In this selection, only compounds that are not ionised in the acidic mobile phase widely used for profiling were considered. Compounds containing nitrogen atom(s) were discarded, and only NPs containing C, H, and O atoms were kept. Indeed, the retention mechanisms of neutral and ionised compounds are different, and therefore, the retention of ionised compounds cannot be predicted using the models built in this study. In the database, all structures of the selected NPs were gathered and associated with their corresponding experimental RT (RT_{exp}) and with their calculated physicochemical parameters (see below). After LC-MS measurement, some NPs had to be removed from the database for analytical reasons: (1) the compounds not compatible with the 30 min gradient separation on the C18 column ($RT < 2$ min or > 35 min) and (2) the compounds not detected using ESI-TOF-MS. The 260 NPs of the final dataset used to build the QSRR models are listed in Table VII.S1. The molecular structures were introduced into this database as SDF files and coded in SMILES to enable full compatibility with the predictive tools used in the following steps. Figure VII.2A displays the diversity of the compounds of the database in both RT and MW dimensions.

All standards were analysed using LC-MS in small sets of non-isomeric compounds to reduce the

analysis time. The generic UHPLC 30 min gradient chosen (5–95% ACN + 0.1% formic acid on a 150 mm UHPLC column) is considered an optimal compromise between a reasonable analysis time and a high peak capacity for high resolution metabolite profiling [37] and has already been successfully used elsewhere [38]. Such types of LC-MS conditions are commonly used in NP research and in metabolomic studies. The ESI-TOF-MS detection performed in both PI and NI modes ensures sensitive detection for the majority of the NPs analysed and the determination of their MW with high mass accuracy (< 5 ppm).

Although many QSRR models from the literature link the logarithm of the retention factor ($\log k$) to physicochemical parameters [21], the measured RTs were used as dependent variables in this study because the back-transformation of $\log k$ into RT increases the error of the predicted value, while the relative residues obtained from the PLS models were similar when using the RT or $\log k$ as dependent variables.

All compounds were also described by eight independent variables used to predict the RTs. These molecular descriptors were chosen according to previous knowledge of RT prediction [3]. To facilitate the usage of the method, only descriptors calculated using free software or online predictive tools and using SMILES codes as entries were selected. The solvatochromic parameters (α^H , β^H , and π^*) widely used in LSER studies for the characterisation of stationary phases and for lipophilicity determination are derived from Abraham's parameters [2]. The McGowan volume (V) is obtained using a

fragment contribution method. The $\log P$ calculated by the Simulation Plus algorithm ($S+\log P$) is a property-based model for $\log P$ prediction and was chosen because it provides a reliable variable for RT prediction [4, 5, 39]. $\log P$ calculation was preferred to experimental $\log P$ value retrieval because of the lack of reliable experimental values for many NPs. The other descriptors used in the models are simple molecular parameters: topological polar surface area ($TPSA$), number of rotatable bonds (rot_bond), and molecular weight (MW).

3.2. Development of Preliminary Models

Several regression models based on the eight descriptors previously calculated were built for RT prediction. The use of PLS regression was preferred to MLR because PLS takes advantage of the collinearity of independent variables [40], while MLR cannot handle this collinearity properly, leading to meaningless results. Similar to MLR, a PLS model can be summarised in a single linear equation, such as Equation 1:

$$RT_{pred} = a \times A + b \times B + \dots + constant$$

(Equation 1)

where a, b, \dots are the PLS coefficients and A, B, \dots are the variables (physicochemical parameters) of the NPs.

The following parameters were used to compare the models: the coefficient of determination between the RT_{pred} and RT_{exp} values (R^2), the predictive ability index estimated with a leave-one-out procedure (Q^2_{LOO}), the external predictive ability coefficient (Q^2_{ext}), the number of compounds included in the model (n), the time range including the RT_{pred} errors of 75% and 90% of the compounds, and the standard deviation.

A first PLS model (*Method 1*) was built based on the four solvatochromic parameters known to be correlated with the reversed-phase retention variables ($\alpha^H, \theta^H, \pi^*$, and V) [1]. The model was built on a calibration set made from 75% of all of the NPs in the database that were randomly chosen. The remaining 25% NPs of the database were gathered in a validation set used to evaluate the generalisation ability of the models, *i.e.*, the ability to predict the RT of NPs other than those used to build the model, as described by Q^2_{ext} . This model provided encouraging results ($R^2_{cal} = 0.77, Q^2_{LOO} = 0.76, Q^2_{ext} = 0.74, n_{cal} = 198$, see Table VII.1 and Figure VII.3), although the number of outliers and the average error on the RT_{pred} were considered too high for reliable RT prediction.

A new model (*Method 2*) was obtained by including four more variables ($MW, TPSA, S+\log P$, and rot_bond) chosen according to previous works on retention prediction [3]. The calibration and validation sets of *Method 1* were used without change. Although the prediction performance was higher than that of *Method 1* ($R^2_{cal} = 0.89, Q^2_{LOO} = 0.88, Q^2_{ext} = 0.88, n_{cal} = 198$, see Table VII.1 and Figure VII.3), the errors on both calibration and validation RT_{pred} were still too high. Unfortunately, attempts to increase the prediction ability of *Method 2* by adding other physicochemical parameters were not successful.

3.3. Development of Refined Models based on NP Clusters

To improve the prediction ability of *Method 2*, a new strategy based on compound clustering was applied. According to previous studies, the most reliable QSRR models and $\log P$ prediction tools are usually built on homologous series of compounds [3, 5, 41], whereas the NPs used in this work possess a wide range of physicochemical parameters. Therefore, the clustering aimed to divide the database into

Table VII.1. Description of the intermediate *Methods* built during the final model development, as discussed in Sections 3.2 and 3.3.

		<i>Method 1</i>	<i>Method 2</i>	<i>Method 3</i>
Calibration	Number of latent variables of the PLS model	4	6	.. ^a
	Compounds in calibration set	198	198	194
	R ²	0.77	0.89	0.95
	Cross-validation (LOO) Q ²	0.76	0.88	.. ^a
	Avg. error on RT _{pred} for the calib. set	2.53	1.76	1.09
	RT _{pred} error for 75% of the compounds (min)	3.61	2.64	1.44
	RT _{pred} error for 90% of the compounds (min)	5.00	3.62	2.16
Validation	Compounds in validation set	62	62	62
	External validation Q ²	0.74	0.88	0.92
	Avg. error on RT _{pred} for the valid. set	2.80	1.85	1.37
	RT _{pred} error for 75% of the compounds (min)	4.17	2.55	2.55
	RT _{pred} error for 90% of the compounds (min)	5.10	3.81	3.81

^a This value is not available because several models are built according to the clustering.

clusters according to their physicochemical parameters to provide several PLS models for the prediction of homologous series of NPs.

In the first attempt, the NPs of the database were clustered by HCA based on 35 molecular

parameters, such as topological indices. This clustering clearly gathered compounds of similar phytochemical classes, such as flavonoids or terpenoids, and the large majority of the results could be simply extrapolated from the structure-based classification used by the DNP [34]. Thus, instead of using HCA, the clustering was based on

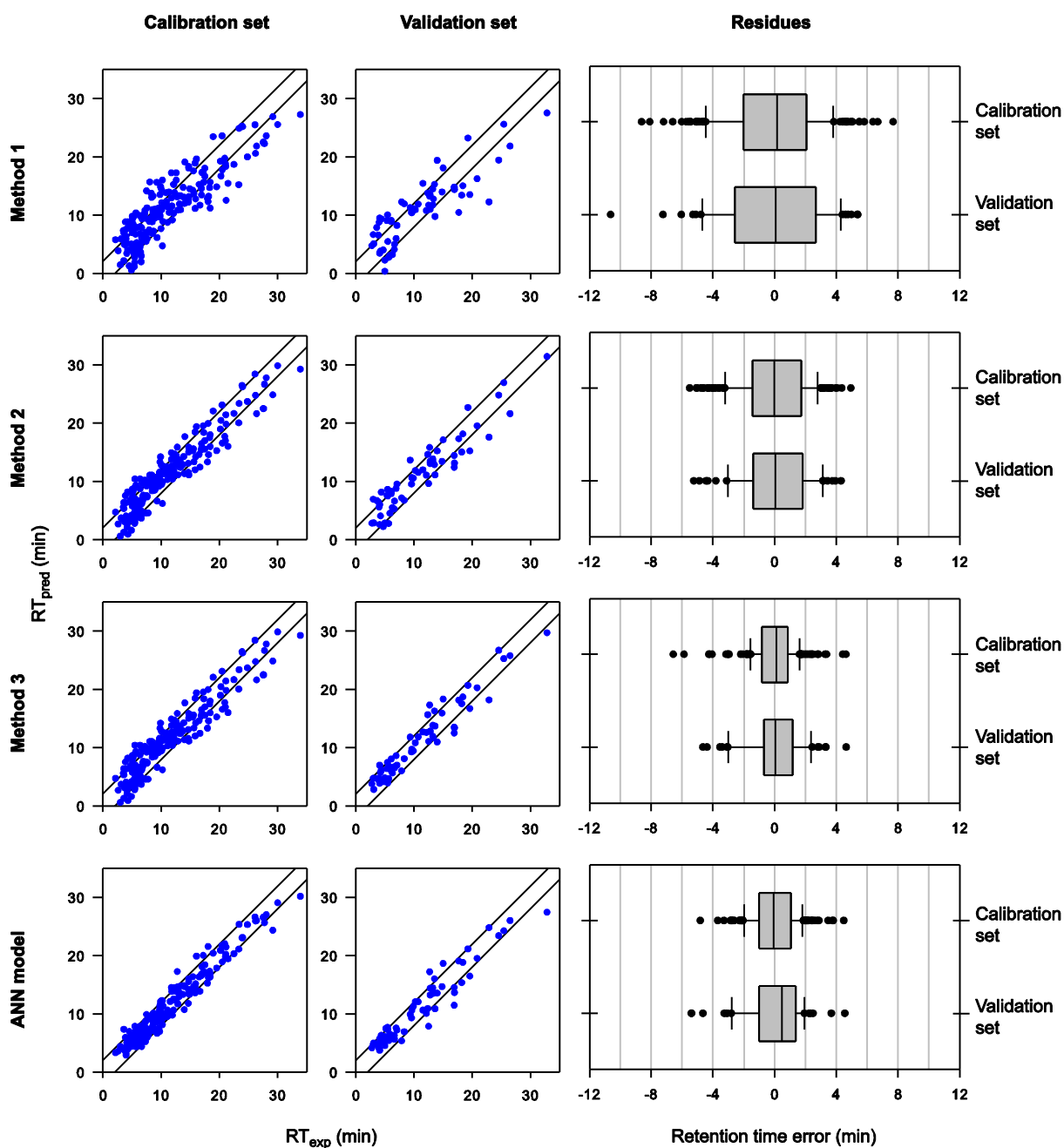


Figure VII.3. Plots of the RT_{pred} vs. RT_{exp} values of the calibration and validation sets (first and second columns, respectively) of the three PLS methods (*Method 1* – *Method 3*) and the ANN model. The diagonal black bars represent a ± 2 min range. The third column shows box plots representing the prediction error on the RT of the developed methods. The central black vertical line represents the median, the extremities of the grey boxes indicate the 25th and 75th percentiles, the vertical lines placed at both ends of the box plots indicate the 10th and 90th percentiles and compounds above and below the 90th and 10th percentiles are marked by single dots.

the classes proposed by the DNP. This approach was advantageous because it is easy to implement, it requires no calculation, and chemical class information is readily available.

All of the 260 compounds in the database were easily classified into one of the seven classes of

the DNP, and only four needed to be discarded. The seven new clusters are described in Figure VII.2B. The wide distribution of NPs' RT in each cluster is shown in the box plots of Figure VII.2C. The number of compounds in each cluster is illustrated in Figure VII.2D, and the structure of a representative NP of each cluster is displayed in Figure VII.4. A specific PLS model was built for

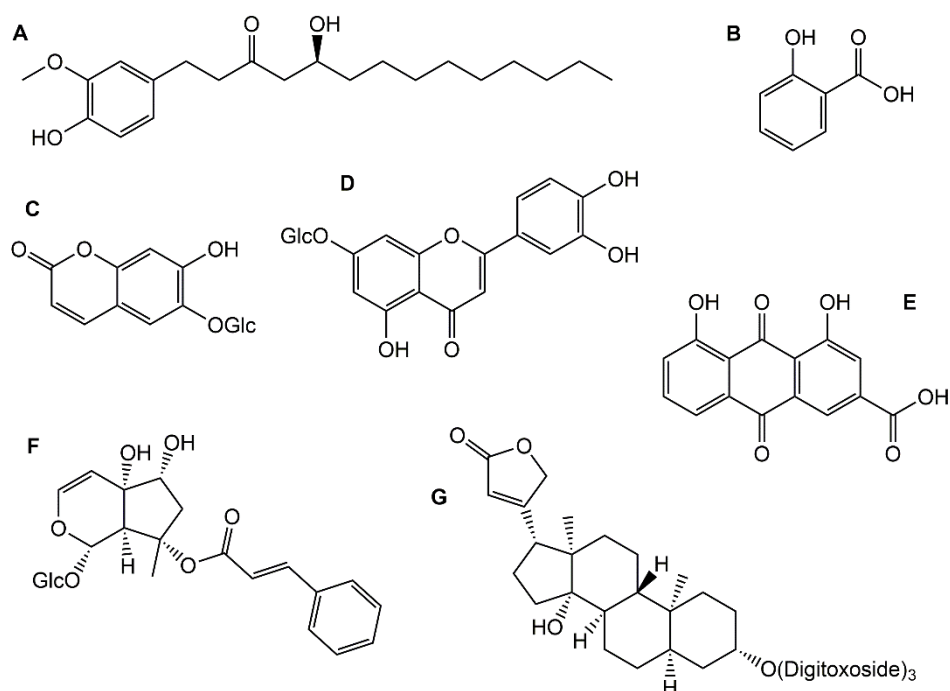


Figure VII.4. Structures of seven representative NPs. (A) 10-gingerol, an aliphatic NP. (B) Salicylic acid, a simple aromatic. (C) Esculin, a benzopyranoid. (D) Luteolin-7-O-glucoside, a flavonoid. (E) Rheic acid, a polycyclic aromatic. (F) Harpagoside, a terpenoid. (G) Digitoxin, a steroid.

→

Table VII.2. Parameters used to evaluate the prediction quality of the developed methods.

* *Method 3* (clustering) represents a combination of the seven models applied to individual clusters. In this case, the statistical parameters of the method correspond to an overall estimation of the prediction results obtained for each NP according to the model that corresponds to its class.

** *Method 2* (global model) is built on all compounds, without previous clustering, and may be used for any compound that does not fit into one of the seven clusters.

*** The final methods are built without a validation set to provide models that taking into account as many compounds as possible.

		Aliphatics	Simple aromatics	Benzopyranoids	Flavonoids	Polycyclic aromatics	Terpenoids	Steroids	Method 3 * (clustering)	Method 2 ** (global PLS model)	ANN model
	Cluster's DNP code	VA	VG	VI	VK	VQ	VS	VT	-	-	-
Calibration	Latent variables of the PLS / ANN hidden layer size	2	3	3	5	4	6	3	-	6	5
	Number of compounds in the set	11	22	27	52	20	44	18	194	198	197
	Cross-validation (LOO) Q ²	0.87	0.96	0.92	0.82	0.83	0.85	0.94	-	0.88	0.91
	R ²	0.90	0.97	0.93	0.87	0.93	0.90	0.96	0.95	0.89	0.95
	Avg. error on RT _{pred}	1.67	0.73	1.02	0.85	0.93	1.50	1.15	1.09	1.76	1.51
	RT _{pred} error for 75% of the compounds (min)	1.95	0.94	1.22	1.17	1.27	1.85	1.37	1.44	2.64	1.53
	RT _{pred} error for 90% of the compounds (min)	2.34	1.60	1.77	1.61	1.85	3.97	2.23	2.16	3.62	2.27
Validation	Number of compounds in the set	3	7	6	18	6	17	5	62	62	63
	External validation Q ²	0.86	0.94	0.91	0.87	0.90	0.93	0.96	0.92	0.88	0.92
	Avg. error on RT _{pred}	2.58	1.64	0.83	0.99	1.03	1.73	1.49	1.37	1.85	1.49
	RT _{pred} error for 75% of the compounds (min)	3.73	1.87	1.62	1.23	1.44	2.89	1.97	1.89	2.55	1.93
	RT _{pred} error for 90% of the compounds (min)	4.29	2.95	1.84	2.11	2.22	3.42	2.66	3.27	3.81	3.03
Final***	R ²	0.90	0.95	0.94	0.87	0.93	0.91	0.96	0.94	0.89	0.95
	RT _{pred} error for 75% of the compounds (min)	2.67	1.25	1.29	1.27	1.23	1.97	1.57	1.57	2.66	1.56
	RT _{pred} error for 90% of the compounds (min)	4.00	1.94	1.60	2.02	1.94	3.49	2.36	2.50	3.60	2.65
	Standard deviation of cross-validation on RT _{pred} (min)	2.69	1.19	1.37	1.10	1.20	2.12	1.45	1.57	2.21	2.00

each DNP class based on separate calibration sets containing 75% of the NPs of each cluster. The predicted values obtained from all seven regressions were gathered into *Method 3*. Compared to *Method 2*, the prediction ability of *Method 3* was greatly improved, showing the relevance of the clustering ($R^2_{\text{cal}} = 0.95$, $Q^2_{\text{ext}} = 0.92$, $n_{\text{total,cal}} = 198$, see Table VII.2). See Table VII.1 and Figure VII.3 for a comparison of all three methods.

The prediction ability of the seven PLS regressions of *Method 3* can be evaluated by the statistical parameters displayed in Table VII.2.

The generalisation ability of this method ($Q^2_{\text{ext}} = 0.92$) is suitable for the prediction of previously unseen NPs. Moreover, the error on the RT_{pred} was less than 1.89 min or 3.27 min for 75% or 90% of the compounds of the validation set, respectively, which is satisfactory given that the total UHPLC analysis time was 30 min. The plot of the RT_{pred} vs. RT_{exp} values (Figure VII.5A) obtained by *Method 3* for the validation set shows the satisfactory prediction ability for low and high RT values. The RT of the compounds that are not part of one of the seven clusters can be predicted using *Method 2*, although its prediction ability is not as good as that of *Method 3* (see Table VII.2).

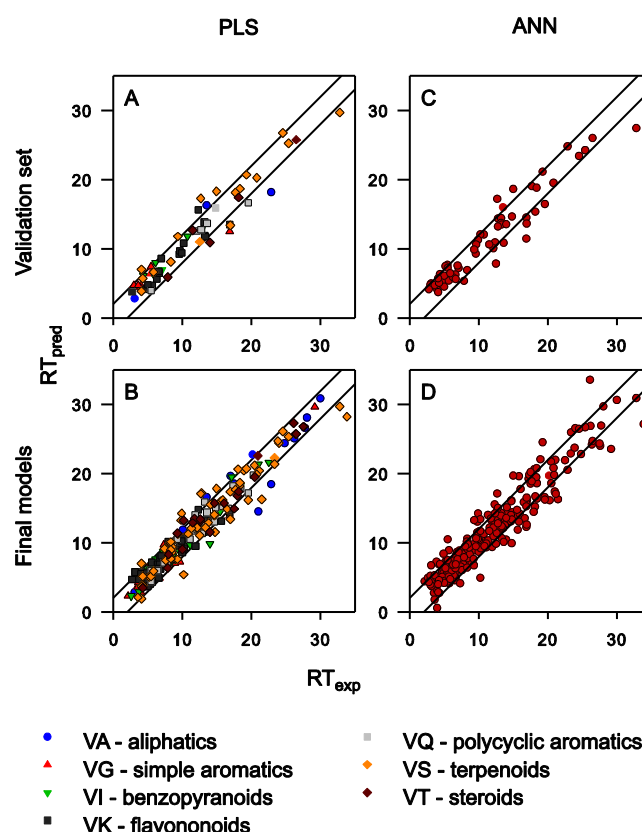


Figure VII.5. Plots of the RT_{pred} vs. RT_{exp} values of the validation set (plots A and C) and of the 260 NPs of the database (plots B and D) using the PLS with clustering (*Method 3*) and ANN methods, respectively. For the PLS models, the types of the compounds are indicated by different dots according to the cluster they belong to. The diagonal black bars represent a ± 2 min range.

To maximise the number of compounds used to build the final models, the seven PLS regressions of *Method 3*, as well as the global model (*Method 2*), were calculated on the 260 compounds available. The coefficients of these final methods are displayed in Table VII.3, the residues of all compounds are represented in the box plot in Figure VII.6, and the correlation of RT_{pred} vs. RT_{exp} is shown in Figure VII.5B. These figures clearly show that the error on the RT_{pred} calculated using *Final Method 3* is less than 2 min for 90% of the NPs, which is satisfactory considering that the total gradient time is 30 min. However, the prediction ability of the seven PLS regressions of *Final Method 3* is not identical (Table VII.2). For example, the RT of 90% of the simple aromatic NPs was predicted with an error less than 2 min, while this value was attained for only 75% of the terpenoids.

In summary, the RT of any uncharged NP containing C, H, and O atoms can be predicted based on its structure for the chromatographic conditions described in Section 2.2 using the following procedure. (1) The eight variables (α^H , β^H , π^* , V , MW , $TPSA$, $S+\log P$, and rot_bond) are calculated from the structure as described in Section 2.3. (2) The phytochemical class of the NP of interest is determined, if needed, using the DNP (from among *aliphatic*, *simple aromatic*, *benzopyranoid*, *flavonoid*, *polycyclic aromatic*, *terpenoid*, and *steroid*). If the compound does not fit into any class, it is considered "other". (3) The RT is calculated based on a linear equation (Equation 1), and using the coefficients of the corresponding cluster of *Final Method 3* or *Final Method 2* for "other" compounds (Table VII.3).

Table VII.3. Coefficients from the PLS regressions used for RT prediction using Equation 1 for the seven clusters (*Final Method 3*) and for the global model (*Final Method 2*). The compounds of both calibration and validation sets are merged to provide final models that are based on the largest possible number of NPs.

Descriptor	Aliphatics	Simple aromatics	Benzopyranoids	Flavonoids	Polycyclic aromatics	Terpenoids	Steroids	Global model
α^H	-1.62	-0.655	-2.91	-2.18	-2.44	-2.82	1.98	-3.12
β^H	-2.36	-0.477	-0.896	-0.256	-2.48	1.93	0.625	1.02
π^*	-1.75	0.00216	3.20	0.284	-5.27	0.194	-1.49	0.121
V	3.30	1.65	2.88	0.452	5.45	-1.31	2.44	-0.223
$MW \times 10^{-3}$	6.87	8.91	21.0	2.37	26.0	-1.20	11.7	1.62
$TPSA \times 10^{-2}$	-1.96	-1.75	-0.324	-0.286	6.50	1.67	-0.940	3.39
$S+\log P$	0.891	2.74	2.88	1.26	2.22	3.18	2.94	3.14
rot_bond	0.542	0.309	-0.773	0.286	-2.16	0.376	-1.28	0.0125
constant	6.27	-1.21	-8.14	7.38	3.19	4.55	-2.23	1.67

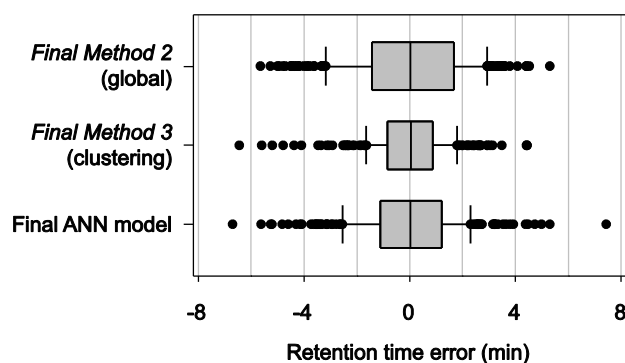


Figure VII.6. Box plot representing the error on the RT_{pred} for the 260 compounds studied, calculated using *Final Method 3* (using clustering) and *Final Method 2* (global model), as well as using the final ANN model for the 30 min UHPLC gradient. The central black vertical line represents the median. The extremities of the grey boxes indicate the 25th and 75th percentiles, and the vertical lines placed at both ends of the box plots indicate the 10th and 90th percentiles. Compounds above and below the 90th and 10th percentiles are marked by single dots.

3.4. Development of the ANN Model

As an alternative approach to RT prediction, a non-linear data modelling strategy based on ANN was tested. ANNs are very efficient function approximation tools that rely on layers of interconnected units, called neurons, which form a network [42]. The connections possess adjustable weights, and each neuron receives an input computed as a weighted sum of the outputs from upstream units. As a starting point, data are provided to the input neurons, and the signal is transmitted to downstream units through the connections until it generates activation values when it reaches the output layer. Connection weights are adjusted iteratively from a set of random values to produce the correct output. ANNs allow complex global fitting problems to be decomposed into simple sub-problems solved locally and have a great ability to model non-linear systems without specifying a mathematical model prior to fitting. However, their interpretability often remains limited.

For the purpose of comparison, the data used in the PLS models with similar calibration and validation sets were used, and a nested LOO cross-validation procedure was used to assess the model fit and its generalisation ability. The same quality indices were evaluated (R^2 , Q^2_{LOO} and Q^2_{ext}). The ANN was trained to predict RT values for the studied compounds, while the eight physicochemical parameters were used as input variables. A multi-layer perceptron topology was chosen in this study as it provides general-purpose and flexible non-linear models. The number of units in the hidden layer was optimised by varying its value from one to ten according to MSE (data not shown). If the addition of supplementary hidden neurons does not provide better prediction outputs, the model size should not be inflated to ensure the best generalisation ability and a fast training time. The network that included five hidden neurons was chosen as the optimum MSE value, leading to an 8-5-1 topology.

The trained network approximated the calibration set adequately, as revealed by the high R^2 value ($R^2_{\text{cal}} = 0.95$). A satisfactory prediction ability was highlighted by the good overall agreement between the measured RT and the prediction obtained by cross-validation ($Q^2_{\text{LOO}} = 0.91$, see Table VII.2). Additionally, the generalisation ability of the ANN model evaluated with the same validation set as the one used for validating PLS models was found to be suitable for reliable prediction of the RT of previously unseen NPs ($Q^2_{\text{ext}} = 0.92$, see Figure VII.5C). The error on the ANN prediction was less than 1.93 and 3.03 min for 75% and 90% of the molecules of the validation set, respectively.

The high values of all performance indices suggest an adequate number of units in the hidden layer and indicate that the risk of overfitting during the training phase is unlikely. The final ANN model was then built with the 260 compounds to maximise the dataset. The resulting correlation between RT_{pred} and RT_{exp} is shown in Figure VII.5D.

Finally, to add information similar to the clustering for the PLS models, a second ANN model was built using the DNP class of the compounds as a ninth input variable. However, this model was discarded because its prediction ability was not increased by the addition of this information (data not shown).

3.5. Applicability of the Prediction Models

Both the PLS and ANN approaches provide a retention prediction of NPs based on simple physicochemical parameter calculations made on the structure of the compound.

The two strategies possess different advantages and drawbacks. On the one hand, the PLS strategy is based on seven different models

corresponding to the seven main types of NPs containing C, H, and O atoms only. The models are well adapted to the specificities of each class of NPs. The RT_{pred} calculation is very easy using the PLS approach because it is only based on single linear equations that can be calculated manually or implemented in a spreadsheet. Finally, the PLS models provide a slightly better prediction of the retention compared to the ANN models, according to the results displayed in Tables VII.2. However, the number of compounds used to build some of the models was low and therefore not optimal from a statistical viewpoint. Moreover, the RT of NPs that are not part of these seven types of compounds can be predicted using the global model (*Final Model 2*).

On the other hand, the ANN strategy based on a unique model is advantageous because it does not require class information. Moreover, this model is built on the full set of calibration compounds ($n = 260$) and thus reflects the chemical diversity of NPs and is statistically robust.

Based on these results, the model has to be chosen according to the needs of the user. If the studied NP fits into one of the seven clusters, the PLS approach is best and easily provides slightly better results. However, when the studied NP is not part of one of the seven classes of NPs, or if several NPs of different clusters are studied or need to be compared, the ANN model is preferred. It is also possible to use both approaches in parallel to confirm the RT_{pred} range of the first method using the second one.

However, the ANN and PLS models possess two methodological limitations according to the choices made. First, because retention mechanisms are very different between charged and uncharged molecules, models were only built with compounds that were neutral in the acidic conditions of the mobile phase (a constraint from

the LC-ESI-MS analysis). This requirement is why nitrogen-containing compounds were removed from the database. Therefore, the models described here are not designed to predict the RT of charged and nitrogen-containing molecules. Second, because the physicochemical parameters used to predict the RT do not take into account the three-dimensional structure, the values of these calculated parameters are identical for stereoisomers. For example, the RT_{pred} of isoquercitrin (quercetin-3-*O*-glucoside) and hyperoside (quercetin-3-*O*-galactoside) are identical (4.31 min), although the RT_{exp} values are different (5.59 and 5.44 min, respectively). This issue is frequently encountered in retention prediction models [26]. However, although the use of three-dimensional parameters would provide different RT_{pred} values for two stereoisomers, the accuracy of the prediction

(typically ± 1 min) in our case would not provide an unambiguous discrimination. In the example discussed, the RT variation between isoquercitrin and hyperoside is less than 0.15 min. The use of three-dimensional parameters, however, could still slightly increase the prediction ability of the models, in particular for flexible and complex molecules.

The QSRR models presented herein provide retention information that is valid only for the conditions used to calibrate the models, described in Section 2.2. However, it is possible to transfer this value to other chromatographic conditions by changing the kinetic parameters (e.g., column geometry, flow rate, gradient span and slope) using Equation 2:

$$RT_2 = \frac{\Phi_{\text{init},1} - \Phi_{\text{init},2} + \text{slope}_1 \left(RT_1 - t_{\text{col},1} - t_{\text{isoc},1} - \frac{v_{d,1}}{F_1} \right)}{\text{slope}_2} + t_{\text{col},2} + t_{\text{isoc},2} + \frac{v_{d,2}}{F_2} \quad (\text{Equation 2})$$

where 1 and 2 represent the initial and final conditions, respectively, Φ_{init} is the fraction of organic modifier in the mobile phase at the beginning of the gradient (no unit), slope is the slope of the gradient (%/min), t_{col} is the column dead time (min), t_{isoc} is the initial isocratic hold (min), v_d is the dwell volume (μL), and F is the flow rate ($\mu\text{L}/\text{min}$). Equation 2 is presented in more detail in the Supporting Information. Conversely, the thermodynamic parameters have to be kept constant (e.g., stationary phase chemistry, mobile phase composition, and temperature).

The models proposed in this study represent an efficient tool for the retention prediction of NPs and can be used as an additional filter in LC-MS dereplication procedures. This filtering can be particularly useful for the differentiation of

isomers with relatively different physicochemical parameters.

For example, the models were used to correctly annotate the peak of one of the ginseng saponin, ginsenoside B2 (a NP that was not included in the calibration set) in the UHPLC-TOF-MS profiling of a crude *Panax ginseng* extract (Figure VII.7A). Ginsenoside B2 was searched by extracting the ion trace corresponding to its $[M-H]^-$ m/z 945.54 (Figure VII.7B). The trace revealed two LC peaks at RTs of 8.21 and 12.84 min. Analysis of both HR-MS spectra revealed the same molecular formula ($\text{C}_{48}\text{H}_{82}\text{O}_{18}$) for both compounds, which also matched that of ginsenoside B2. To discriminate which LC peak corresponded to the target compound, its RT_{pred} was calculated. The PLS model for the terpenoid class (VS) predicted a RT

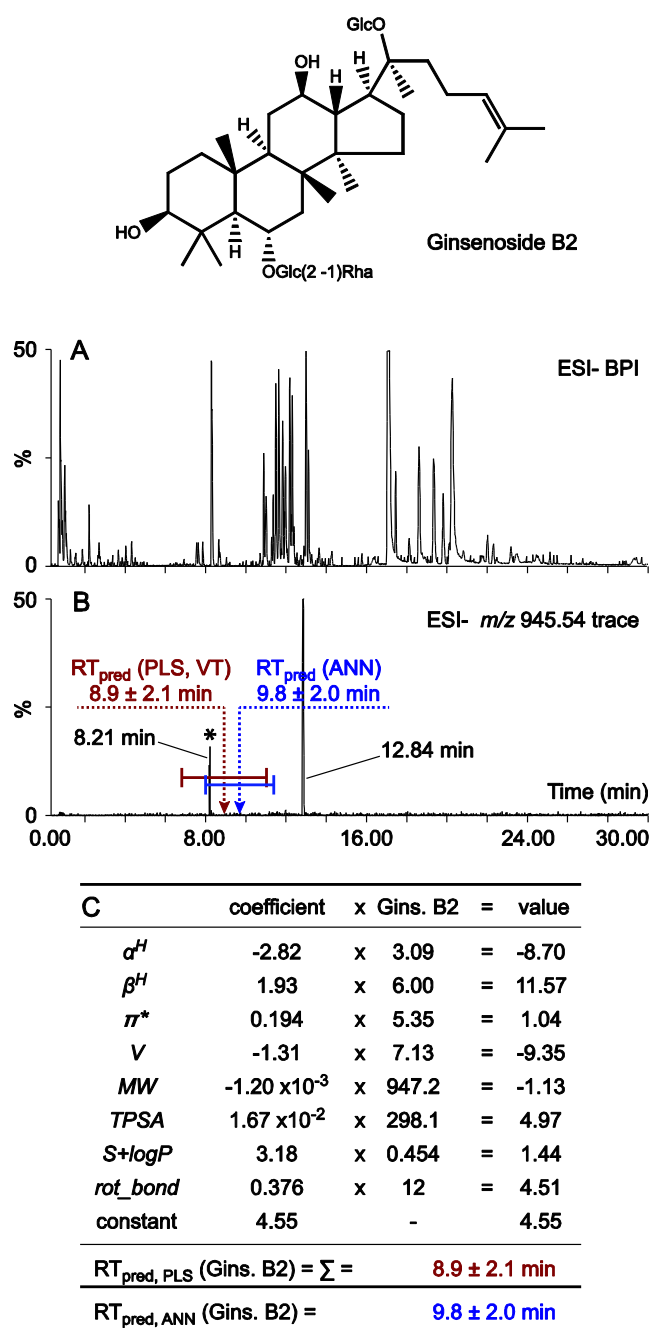


Figure VII.7. Example of the application of the retention prediction tool for the peak annotation of ginsenoside B2 in the UHPLC-TOF-MS profiling of *Panax ginseng*. **(A)** NI BPI trace of the crude extract using a 5–95% ACN 30 min gradient on an Acquity BEH C18 150 x 2.1 mm 1.7 μ m column at 460 μ L/min and 40 $^{\circ}$ C. **(B)** Extracted ion trace of the m/z 945.54. Red arrow highlights the RT_{pred} values calculated by *Final Method 3* for terpenoids using PLS regression ($RT_{pred} = 8.9$ min \pm 2.1). Blue arrow shows the RT_{pred} values calculated by the ANN model ($RT_{pred} = 9.9$ min \pm 2.0). Error bars represent the standard deviation. Final assignment of the LC peak corresponding to ginsenoside B2 is indicated with a * ($RT_{exp} = 8.21$ min) **(C)** Table describing the calculation of the RT_{pred} of ginsenoside B2 using *Final Method 3*.

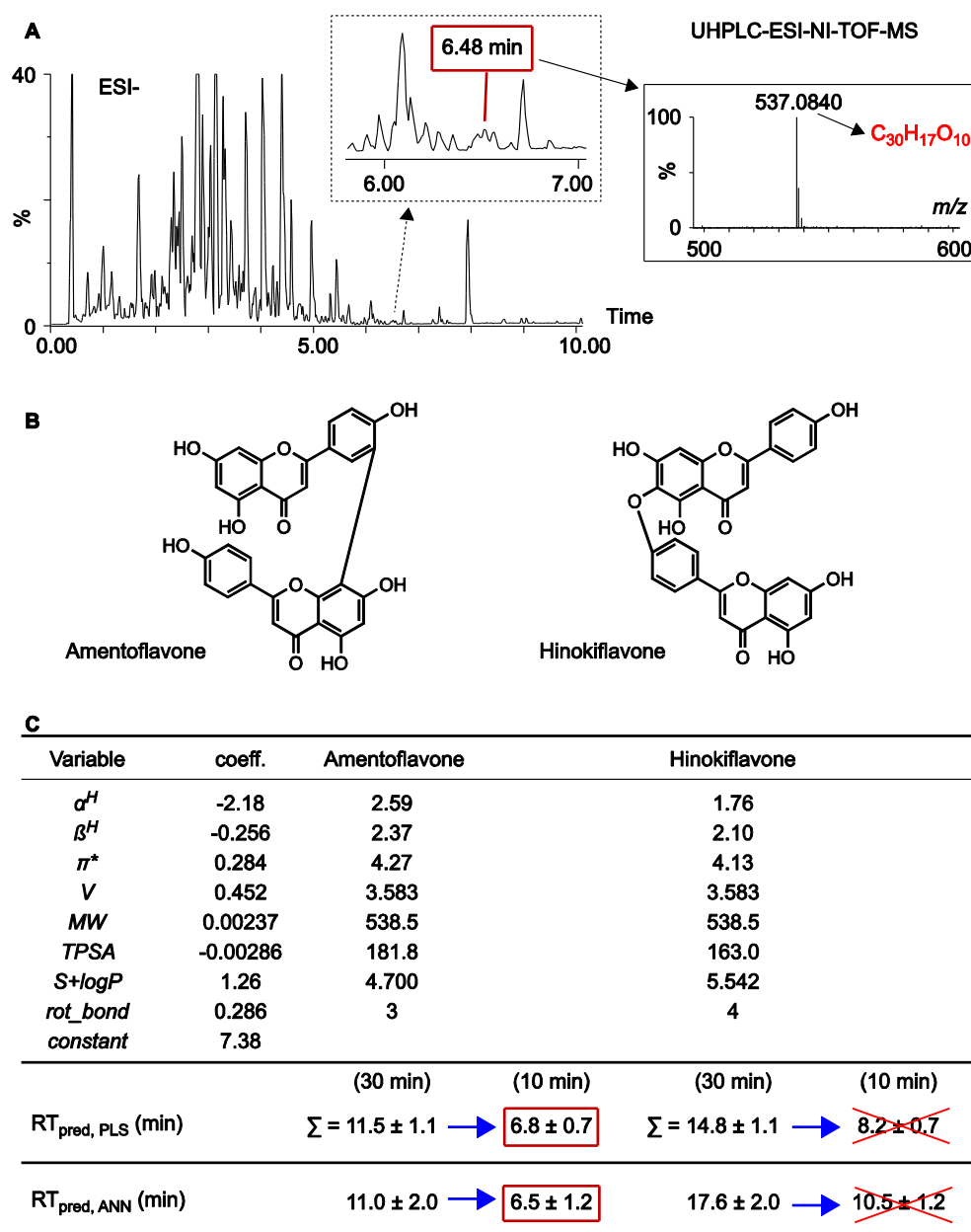


Figure VII.8. Example of the application of the retention time prediction tool for the dereplication of the LC-MS peak at 6.48 min in UHPLC-TOF-MS profiling of a *Ginkgo biloba* extract. (A) NI BPI chromatogram analysed using a 10 min 5-55% ACN gradient at 650 $\mu\text{L}/\text{min}$ and at 40 $^\circ\text{C}$, on a 100 x 2.1 mm column. The $[M-H]^-$ ion 537.0840 provided the $C_{30}H_{17}O_{10}$ molecular formula (Δppm 3.2). (B) A subsequent search in the DNP using “*Ginkgo* OR plants” and “ $C_{30}H_{17}O_{10}$ ” as filters provided two hits (amentoflavone and hinokiflavone). (C) Their RT_{pred} in the 30 min generic gradient described in Section 2.2 (11.5 ± 1.1 and 14.8 ± 1.1 min, as calculated by the *Final Method 3* (using the PLS model for flavonoids) and 11.0 ± 2.0 and 17.6 ± 2.0 , as calculated by the ANN model, respectively) were transferred to the experimental conditions used for this analysis and provided 6.8 and 6.5 min for amentoflavone and 8.2 and 10.5 min for hinokiflavone, using PLS and ANN models, respectively. The unknown analyte was therefore putatively identified as amentoflavone.

of 8.9 ± 2.1 min for ginsenoside B2 (Figure VII.7C), while the ANN model gave an RT_{pred} of 9.9 ± 2.0 min (the indicated range corresponds to one standard deviation). In both cases, the models indicated that the peak at 8.21 min in the extract corresponded to ginsenoside B2. For verification purposes, this identity was further confirmed by injecting the standard.

To further validate the applicability of the tool, the models were also used to predict the retention time of NPs analysed using other kinetic chromatographic parameters (e.g. column length, gradient time, and flow rate). For example, in the case of the metabolite profiling of a standardised extract of *Ginkgo biloba* used in phytopreparation, an LC peak at 6.48 min with an $[M-H]^-$ at m/z 537.0840 (relative isotope abundance: 37% for M+1 and 9% for M+2) was dereplicated (Figure VII.8). The UHPLC-TOF-MS analysis was carried out using a 10 min 5-55% ACN gradient at 650 $\mu\text{L}/\text{min}$ and at 40 $^\circ\text{C}$, on a 100 x 2.1 mm column. The obtained molecular formula ($\text{C}_{30}\text{H}_{17}\text{O}_{10}$) was validated by mass and spectral accuracies and heuristic filtering [16].

The UV trace did not provide further information because of the low concentration of the metabolite. Chemotaxonomic information indicated that the molecular formula matched with two previously reported biflavones from this plant, namely amentoflavone and hinokiflavone. Their RT_{pred} in the 30 min generic gradient described in Section 2.2 were 11.5 ± 1.1 and 14.8 ± 1.1 min, as calculated by the *Final Method 3* (using the PLS model for flavonoids) and 11.0 ± 2.0 and 17.6 ± 2.0 , as calculated by the ANN model, respectively. The corresponding value of these RTs in the 10 min gradient were transferred from the 30 min generic gradient using Equation 2 and provided 6.8 and 6.5 min for amentoflavone and 8.2 and 10.5 min for hinokiflavone, using the same models. Based on these RT_{pred} , amentoflavone was assigned to the LC-MS peak at 6.48 min.

These two examples show the ability of the models to provide orthogonal information to HR-MS to further improve the accuracy of the dereplication.

4. Conclusion

The NPs containing C, H, and O atoms that are neutral at acidic pH can be predicted based on their structure with a satisfactory accuracy by the models presented in this work. The RT_{pred} is valid for specific chromatographic conditions and can be further transposed to any other conditions if the thermodynamic parameters are kept constant.

Two different methods, namely, PLS and ANN, provided similar prediction ability. The PLS-based method provides RT_{pred} based on a simple equation. The prediction ability of PLS was found to be satisfactory when working on subsets of NP homologues, while direct application of PLS to a wide range of compounds with very diverse physicochemical parameters failed to give reliable predictions. In this study, we propose a simple way of clustering homologous NPs using DNP classification and verify the quality of the models for each class. However, one unique ANN model was found to be sufficient to provide a satisfactory prediction for the whole range of NPs tested.

The prediction ability of the models (the RT of 90% of the NPs was predicted with an error less than 2.50 and 2.65 min using the PLS and ANN models, respectively) reflects what can be practically achieved based on simple structure calculation and high resolution LC-MS of complex molecules such as NPs. The accuracy of prediction is not very high (at best 7% with the PLS models based on clusters) but it has been demonstrated that the tools developed are applicable for dereplication purposes in natural extracts and that they provide useful orthogonal filters for MS-based metabolite identification.

Future improvement of the models should include the use of three-dimensional parameters to better predict the retention of stereoisomers and of flexible or complex NPs and the possibility of predicting the retention of charged compounds. The development of similar models to orthogonal chromatographic LC-MS methods, such as HILIC, could further enhance the quality of dereplication if both HILIC and C18 predictions are convergent.

5. Supporting information

For sake of readability, two objects were displaced at the end of the paper (see next pages):

- Table VII.S1
- The development of Equation 2.

Note that several tables and figures of the chapter were provided as Supporting Information in the original paper in the Analytical Chemistry journal, and were placed at the appropriate place in the text of the thesis.

Table S1. List of the 260 NPs of the database used in the calibration and validation sets for model building, with the RT_{exp} and the eight physicochemical parameters used as variables in the PLS and ANN models.

Compound	CAS	Molecular formula	Cluster ^a	RT	α^H	β^H	π^*	V	MW	TPSA	S+logP	rot_bond
chlorogenic acid	327-97-9	C ₁₆ H ₁₈ O ₉	1	3.08	2.020	2.250	2.530	2.416	354.31	164.750	-0.293	5
cynarin	30964-13-7	C ₂₅ H ₂₄ O ₁₂	1	4.21	2.620	2.880	3.840	3.537	516.45	211.280	1.304	9
jasmonic acid	6894-38-8	C ₁₂ H ₁₈ O ₃	1	10.04	0.570	0.790	1.300	1.738	210.27	54.370	2.098	5
6-gingerol	23513-14-6	C ₁₇ H ₂₆ O ₄	1	13.47	0.590	1.110	1.700	2.458	294.39	66.760	2.981	10
8-gingerol	23513-08-8	C ₁₉ H ₃₀ O ₄	1	16.97	0.590	1.110	1.700	2.740	322.44	66.760	3.950	12
6-shogaol	555-66-8	C ₁₇ H ₂₄ O ₃	1	17.46	0.270	0.870	1.540	2.356	276.37	46.530	4.293	9
10-gingerol	23513-15-7	C ₂₁ H ₃₄ O ₄	1	20.15	0.590	1.120	1.710	3.022	350.49	66.760	4.957	14
embelin	550-24-3	C ₁₇ H ₂₆ O ₄	1	22.83	0.630	1.390	0.850	2.458	294.39	74.600	4.386	10
urushiol (15:3)	83543-37-7	C ₂₁ H ₃₀ O ₂	1	24.77	0.770	0.750	1.340	2.818	314.46	40.460	7.088	12
urushiol (15:2)	83258-37-1	C ₂₁ H ₃₂ O ₂	1	26.14	0.770	0.700	1.280	2.861	316.48	40.460	7.437	12
ginkgolic acid (C13:0)	20261-38-5	C ₂₀ H ₃₂ O ₃	1	27.69	0.700	0.430	1.100	2.822	320.47	57.530	7.859	13
ginkgolic acid (C15:1)	22910-60-7	C ₂₂ H ₃₄ O ₃	1	28.03	0.700	0.500	1.200	3.061	346.50	57.530	8.198	14
ginkgolic acid (C17:1)	111047-30-4	C ₂₄ H ₃₈ O ₃	1	29.97	0.700	0.510	1.210	3.343	374.56	57.530	8.873	16
protocatechuic acid	99-50-3	C ₇ H ₆ O ₄	2	2.13	1.270	0.860	1.460	1.049	154.12	77.760	0.980	1
<i>p</i> -salicylic acid	99-96-7	C ₇ H ₆ O ₃	2	2.96	1.000	0.720	1.290	0.990	138.12	57.530	1.704	1

phthalic acid	88-99-3	C ₈ H ₆ O ₄	2	3.48	1.140	0.770	1.460	1.147	166.13	74.600	1.191	2
caffeic acid	331-39-5	C ₉ H ₈ O ₄	2	3.54	1.350	0.930	1.570	1.288	180.16	77.760	1.225	2
vanillic acid	121-34-6	C ₈ H ₈ O ₄	2	3.58	0.780	0.800	1.350	1.190	168.15	66.760	1.262	2
<i>m</i> -salicylic acid	99-06-9	C ₇ H ₆ O ₃	2	3.86	1.060	0.720	1.290	0.990	138.12	57.530	1.614	1
syringic acid	530-57-4	C ₉ H ₁₀ O ₅	2	3.87	0.670	0.890	1.410	1.390	198.17	75.990	1.289	3
3-hydroxyphenylacetic acid	621-37-4	C ₈ H ₈ O ₃	2	3.95	1.070	0.720	1.290	1.131	152.15	57.530	0.841	2
homovanillic acid	306-08-1	C ₉ H ₁₀ O ₄	2	3.98	0.850	0.800	1.350	1.331	182.17	66.760	0.983	3
<i>p</i> -coumaric acid	501-98-4	C ₉ H ₈ O ₃	2	4.67	1.070	0.790	1.390	1.229	164.16	57.530	1.564	2
vanillin	121-33-5	C ₈ H ₈ O ₃	2	4.68	0.440	0.760	1.460	1.131	152.15	46.530	1.205	2
ethyl gallate	831-61-8	C ₉ H ₁₀ O ₅	2	5.01	1.410	1.020	1.600	1.390	198.17	86.990	1.302	3
syringaldehyde	134-96-3	C ₉ H ₁₀ O ₄	2	5.05	0.330	0.850	1.520	1.331	182.17	55.760	1.173	3
ferulic acid	1135-24-6	C ₁₀ H ₁₀ O ₄	2	5.36	0.850	0.870	1.460	1.429	194.18	66.760	1.609	3
sinapic acid	530-59-6	C ₁₁ H ₁₂ O ₅	2	5.42	0.730	0.960	1.520	1.628	224.21	75.990	1.676	4
<i>m</i> -coumaric acid	588-30-7	C ₉ H ₈ O ₃	2	5.67	1.070	0.790	1.390	1.229	164.16	57.530	1.489	2
verbascoside	61276-17-3	C ₂₉ H ₃₆ O ₁₅	2	5.94	2.710	4.250	4.230	4.297	624.59	245.290	-0.278	11
ferulaldehyde	458-36-6	C ₁₀ H ₁₀ O ₃	2	6.35	0.270	0.830	1.560	1.370	178.18	46.530	1.640	3
rhapontin	155-58-8	C ₂₁ H ₂₄ O ₉	2	6.35	1.770	2.650	2.940	2.969	420.41	149.070	0.617	6
<i>o</i> -coumaric acid	614-60-8	C ₉ H ₈ O ₃	2	6.63	1.070	0.790	1.390	1.229	164.16	57.530	1.496	2

<i>TRANS</i> -resveratrol	501-36-0	C ₁₄ H ₁₂ O ₃	2	7.31	1.500	1.040	1.820	1.739	228.24	60.690	2.907	2
3,4,5-trimethoxycinnamic acid	90-50-6	C ₁₂ H ₁₄ O ₅	2	8.56	0.570	1.150	2.080	1.769	238.24	64.990	1.954	5
<i>CIS</i> -resveratrol	61434-67-1	C ₁₄ H ₁₂ O ₃	2	8.84	1.500	1.040	1.820	1.739	228.24	60.690	2.907	2
<i>TRANS</i> -cinnamic acid	140-10-3	C ₉ H ₈ O ₂	2	8.88	0.570	0.510	1.180	1.171	148.16	37.300	2.060	2
cinnamaldehyde	104-55-2	C ₉ H ₈ O ₁	2	9.59	0.000	0.470	1.290	1.112	132.16	17.070	1.958	2
curcumin	458-37-7	C ₂₁ H ₂₀ O ₆	2	14.84	0.550	1.670	2.850	2.773	368.38	93.060	3.368	8
garcinone C	76996-27-5	C ₂₃ H ₂₆ O ₇	2	16.85	2.000	1.610	2.500	3.090	414.45	131.360	3.349	5
atranorin	479-20-9	C ₁₉ H ₁₈ O ₈	2	20.99	0.560	0.900	2.210	2.651	374.34	130.360	3.569	6
hyperforin	11079-53-1	C ₃₅ H ₅₂ O ₄	2	29.14	0.310	1.840	2.160	4.714	536.78	71.440	6.294	11
kawain	3155-48-4	C ₁₄ H ₁₄ O ₃	2	12.66	0.000	0.820	1.430	1.782	230.26	35.530	2.980	3
esculin	531-75-9	C ₁₅ H ₁₆ O ₉	3	2.61	1.270	2.370	2.630	2.210	340.28	149.820	-0.985	3
fraxetin-8- <i>O</i> -glucoside	524-30-1	C ₁₆ H ₁₈ O ₁₀	3	3.67	1.160	2.450	2.700	2.409	370.31	159.050	-1.090	4
fraxetin	574-84-5	C ₁₀ H ₈ O ₅	3	4.35	0.550	0.860	1.610	1.379	208.17	79.900	1.320	1
isoscopletin	776-86-3	C ₁₀ H ₈ O ₄	3	5.01	0.270	0.780	1.450	1.320	192.17	59.670	1.446	1
7-hydroxycoumarin	93-35-6	C ₉ H ₆ O ₃	3	5.14	0.500	0.700	1.360	1.121	162.14	50.440	1.742	0
7-hydroxy-6-methoxycoumarin	92-61-5	C ₁₀ H ₈ O ₄	3	5.33	0.270	0.780	1.450	1.320	192.17	59.670	1.527	1
8-hydroxy-7-methoxycoumarin	19492-03-6	C ₁₀ H ₈ O ₄	3	5.66	0.160	0.720	1.400	1.320	192.17	59.670	1.599	1
chromone	491-38-3	C ₉ H ₆ O ₂	3	6.03	0.000	0.680	1.410	1.062	146.14	30.210	1.453	0

cimifugin	37921-38-3	C ₁₆ H ₁₈ O ₆	3	6.08	0.540	1.730	1.980	2.174	306.31	89.130	0.873	3
4-methylumbelliferone	90-33-5	C ₁₀ H ₈ O ₃	3	6.67	0.500	0.720	1.310	1.262	176.17	50.440	2.187	0
4-hydroxycoumarin	1076-38-6	C ₉ H ₆ O ₃	3	7.04	0.310	0.780	1.380	1.121	162.14	50.440	1.915	0
coumarin	91-64-5	C ₉ H ₆ O ₂	3	7.13	0.000	0.490	1.160	1.062	146.14	30.210	1.926	0
6,7-dimethoxycoumarin	120-08-1	C ₁₁ H ₁₀ O ₄	3	7.18	0.000	0.880	1.780	1.461	206.19	48.670	1.804	2
7-methoxycoumarin	531-59-9	C ₁₀ H ₈ O ₃	3	8.55	0.000	0.640	1.250	1.262	176.17	39.440	2.016	1
oxypeucedanin hydrate	2643-85-8	C ₁₆ H ₁₆ O ₆	3	8.56	0.500	1.410	1.890	2.131	304.29	93.040	1.326	4
byakangelicin	482-25-7	C ₁₇ H ₁₈ O ₇	3	8.94	0.500	1.610	2.020	2.331	334.32	102.270	0.932	5
psoralen	66-97-7	C ₁₁ H ₆ O ₃	3	9.31	0.000	0.580	1.360	1.251	186.16	43.350	2.108	0
khellin	82-02-0	C ₁₄ H ₁₂ O ₅	3	9.56	0.000	1.140	1.780	1.791	260.24	61.810	1.761	2
visnagin	82-57-5	C ₁₃ H ₁₀ O ₄	3	9.67	0.000	0.940	1.650	1.591	230.22	52.580	2.131	1
8-methoxypsoralen	298-81-7	C ₁₂ H ₈ O ₄	3	9.98	0.000	0.780	1.460	1.450	216.19	52.580	2.186	1
Trolox	53188-07-1	C ₁₄ H ₁₈ O ₄	3	10.68	0.880	0.940	1.270	1.927	250.29	66.760	2.880	1
5-methoxypsoralen	484-20-8	C ₁₂ H ₈ O ₄	3	11.04	0.000	0.740	1.450	1.450	216.19	52.580	2.238	1
isopimpinellin	482-27-9	C ₁₃ H ₁₀ O ₅	3	11.10	0.000	0.930	1.580	1.650	246.22	61.810	1.896	2
heraclenin	2880-49-1	C ₁₆ H ₁₄ O ₅	3	11.77	0.000	1.050	1.640	1.964	286.28	65.110	2.696	3
byakangelicol	26091-79-2	C ₁₇ H ₁₆ O ₆	3	12.74	0.000	1.190	1.750	2.164	316.31	74.340	2.435	4
oxypeucedanin	737-52-0	C ₁₆ H ₁₄ O ₅	3	12.80	0.000	1.000	1.620	1.964	286.28	65.110	2.748	3

eugenitin	480-12-6	C ₁₂ H ₁₂ O ₄	3	13.96	0.130	0.780	1.410	1.602	220.22	59.670	2.186	1
imperatorin	482-44-0	C ₁₆ H ₁₄ O ₄	3	15.42	0.000	0.870	1.520	1.971	270.28	52.580	3.372	3
isoimperatorin	482-45-1	C ₁₆ H ₁₄ O ₄	3	16.85	0.000	0.830	1.510	1.971	270.28	52.580	3.457	3
epoxybergamottin	206978-14-5	C ₂₁ H ₂₂ O ₅	3	17.16	0.000	1.100	1.690	2.626	354.40	65.110	4.551	6
dicoumarol	66-76-2	C ₁₉ H ₁₂ O ₆	3	18.37	0.630	1.570	2.480	2.274	336.29	100.880	3.485	2
8-geranyloxypsoralen	7437-55-0	C ₂₁ H ₂₂ O ₄	3	21.03	0.000	0.970	1.590	2.633	338.40	52.580	5.344	6
bergamottin	7380-40-7	C ₂₁ H ₂₂ O ₄	3	22.48	0.000	0.920	1.570	2.633	338.40	52.580	5.443	6
(-)-epigallocatechin	970-74-1	C ₁₅ H ₁₄ O ₇	4	2.70	2.390	1.860	2.430	2.049	306.27	130.610	0.589	1
(+)-catechin	154-23-4	C ₁₅ H ₁₄ O ₆	4	3.10	2.000	1.700	2.260	1.991	290.27	110.380	0.759	1
(-)-catechin	18829-70-4	C ₁₅ H ₁₄ O ₆	4	3.11	2.000	1.700	2.260	1.991	290.27	110.380	0.759	1
procyanidin B2	29106-49-8	C ₃₀ H ₂₆ O ₁₂	4	3.57	4.000	3.410	4.270	3.872	578.52	220.760	1.843	3
(-)-epicatechin	490-46-0	C ₁₅ H ₁₄ O ₆	4	3.96	2.000	1.700	2.260	1.991	290.27	110.380	0.759	1
(-)-epigallocatechingallate	989-51-5	C ₂₂ H ₁₈ O ₁₁	4	3.97	3.560	2.470	3.650	2.990	458.37	197.370	1.939	4
apigenin-6-C-glucoside-7-O-glucoside	20310-89-8	C ₂₇ H ₃₀ O ₁₅	4	4.51	2.640	4.300	4.440	3.906	594.52	260.200	-1.255	6
luteolin-6-C-glucoside	4261-42-1	C ₂₁ H ₂₀ O ₁₁	4	4.65	2.590	2.910	3.560	2.935	448.38	201.280	-0.549	3
6-hydroxyluteolin-7-O-glucoside	54300-65-1	C ₂₁ H ₂₀ O ₁₂	4	4.85	2.120	2.940	3.680	2.994	464.38	210.510	-0.531	4
robinin	301-19-9	C ₃₃ H ₄₀ O ₁₉	4	4.85	2.810	5.390	5.190	4.878	740.66	308.120	-0.871	8
luteolin-8-C-glucoside	28608-75-5	C ₂₁ H ₂₀ O ₁₁	4	4.86	2.590	2.910	3.560	2.935	448.38	201.280	-0.750	3

quercetin-3- <i>O</i> -arabinoglucoside	23284-18-6	C ₂₆ H ₂₈ O ₁₆	4	4.94	3.020	4.290	4.560	3.824	596.49	269.430	-1.116	6
luteolin-5- <i>O</i> -glucoside	20344-46-1	C ₂₁ H ₂₀ O ₁₁	4	4.96	2.430	3.190	3.670	2.935	448.38	190.280	-0.471	4
vitexin-4'-rhamnoside	32426-34-9	C ₂₇ H ₃₀ O ₁₄	4	5.33	2.540	3.970	4.200	3.848	578.52	239.970	-0.614	5
myricetin-3- <i>O</i> -rhamnoside	17912-87-7	C ₂₁ H ₂₀ O ₁₂	4	5.38	2.680	2.990	3.640	2.994	464.38	210.510	-0.141	3
eriodictyol-7- <i>O</i> -rutinoside	13463-28-0	C ₂₇ H ₃₂ O ₁₅	4	5.41	2.360	3.980	4.280	3.949	596.53	245.290	-0.601	6
quercetin-3- <i>O</i> -galactoside	482-36-0	C ₂₁ H ₂₀ O ₁₂	4	5.44	2.570	3.160	3.700	2.994	464.38	210.510	-0.567	4
apigenin-8- <i>C</i> -glucoside	3681-93-4	C ₂₁ H ₂₀ O ₁₀	4	5.46	2.310	2.760	3.380	2.876	432.38	181.050	-0.256	3
quercetin-7- <i>O</i> -glucoside	491-50-9	C ₂₁ H ₂₀ O ₁₂	4	5.54	2.210	3.160	3.700	2.994	464.38	210.510	-0.471	4
taxifolin	480-18-2	C ₁₅ H ₁₂ O ₇	4	5.58	1.660	1.690	2.340	2.006	304.25	127.450	1.275	1
quercetin-3- <i>O</i> -glucoside	482-35-9	C ₂₁ H ₂₀ O ₁₂	4	5.59	2.570	3.160	3.700	2.994	464.38	210.510	-0.567	4
luteolin-7- <i>O</i> -glucuronide	29741-10-4	C ₂₁ H ₁₈ O ₁₂	4	5.70	2.200	2.870	3.680	2.951	462.36	207.350	0.005	4
luteolin-7- <i>O</i> -glucoside	5373-11-5	C ₂₁ H ₂₀ O ₁₁	4	5.78	1.900	2.870	3.490	2.935	448.38	190.280	-0.420	4
isorhamnetin-3- <i>O</i> -rutinoside	604-80-8	C ₂₈ H ₃₂ O ₁₆	4	6.25	2.520	4.260	4.440	4.106	624.54	258.430	-0.634	7
apigenin-7- <i>O</i> -rutinoside	552-57-8	C ₂₇ H ₃₀ O ₁₄	4	6.34	2.080	3.890	4.160	3.848	578.52	228.970	-0.240	6
kaempferol-3- <i>O</i> -glucoside	480-10-4	C ₂₁ H ₂₀ O ₁₁	4	6.34	2.290	3.020	3.530	2.935	448.38	190.280	-0.246	4
quercetin-3- <i>O</i> -rhamnoside	522-12-3	C ₂₁ H ₂₀ O ₁₁	4	6.41	2.300	2.840	3.460	2.935	448.38	190.280	0.165	3
apigenin-7- <i>O</i> -apioglucoside	26544-34-3	C ₂₆ H ₂₈ O ₁₄	4	6.47	2.100	3.860	4.120	3.707	564.49	228.970	-0.730	7
kaempferol-3- <i>O</i> -rutinoside	17650-84-9	C ₂₇ H ₃₀ O ₁₅	4	6.53	2.750	4.180	4.370	3.906	594.52	249.200	-0.624	6

kaempferol-7- <i>O</i> -glucoside	16290-07-6	C ₂₁ H ₂₀ O ₁₁	4	6.56	1.940	3.020	3.530	2.935	448.38	190.280	-0.118	4
quercetin-4'- <i>O</i> -glucoside	20229-56-5	C ₂₁ H ₂₀ O ₁₂	4	6.58	2.380	3.160	3.700	2.994	464.38	210.510	-0.454	4
naringenin-7- <i>O</i> -rhamnoside	10236-47-2	C ₂₇ H ₃₂ O ₁₄	4	6.63	2.100	3.860	4.090	3.891	580.53	225.060	-0.187	6
apigenin-7- <i>O</i> -glucoside	578-74-5	C ₂₁ H ₂₀ O ₁₀	4	6.65	1.630	2.730	3.310	2.876	432.38	170.050	0.182	4
diosmin	520-27-4	C ₂₈ H ₃₂ O ₁₅	4	6.78	1.860	3.970	4.220	4.047	608.54	238.200	-0.365	7
myricetin	529-44-2	C ₁₅ H ₁₀ O ₈	4	6.79	2.270	1.780	2.820	2.022	318.24	151.590	1.481	1
hesperidin	520-26-3	C ₂₈ H ₃₄ O ₁₅	4	6.90	1.860	3.920	4.160	4.090	610.56	234.290	-0.233	7
fisetin	528-48-3	C ₁₅ H ₁₀ O ₆	4	6.94	1.750	1.710	2.620	1.905	286.24	111.130	2.139	1
neohesperidin	13241-33-3	C ₂₈ H ₃₄ O ₁₅	4	7.17	1.880	3.940	4.150	4.090	610.56	234.290	-0.321	7
phloretin-2'- <i>O</i> -glucoside	60-81-1	C ₂₁ H ₂₄ O ₁₀	4	7.28	2.290	2.740	3.300	3.028	436.41	177.140	0.251	7
morin	480-16-0	C ₁₅ H ₁₀ O ₇	4	7.56	2.110	1.710	2.660	1.963	302.24	131.360	1.785	1
formononetin-7- <i>O</i> -glucoside	486-62-4	C ₂₂ H ₂₂ O ₉	4	7.93	1.000	2.740	3.180	2.958	430.40	138.820	1.042	5
daidzein	486-66-8	C ₁₅ H ₁₀ O ₄	4	7.94	1.160	1.270	2.230	1.787	254.24	70.670	2.484	1
eriodictyolchalcone	73692-51-0	C ₁₅ H ₁₂ O ₆	4	8.26	1.920	1.280	2.440	2.013	288.25	118.220	2.467	3
quercetin	117-39-5	C ₁₅ H ₁₀ O ₇	4	8.43	1.880	1.630	2.640	1.963	302.24	131.360	1.919	1
luteolin	491-70-3	C ₁₅ H ₁₀ O ₆	4	8.44	1.570	1.340	2.420	1.905	286.24	111.130	2.346	1
butein	21849-70-7	C ₁₅ H ₁₂ O ₅	4	9.54	1.570	1.160	2.310	1.954	272.25	97.990	2.513	3
naringenin	480-41-1	C ₁₅ H ₁₂ O ₅	4	9.59	1.300	1.140	2.190	1.889	272.25	86.990	2.741	1

genistein	446-72-0	C ₁₅ H ₁₀ O ₅	4	9.64	1.300	1.200	2.250	1.846	270.24	90.900	2.695	1
silybin	22888-70-6	C ₂₅ H ₂₂ O ₁₀	4	9.74	1.390	2.580	3.570	3.245	482.44	155.140	1.905	4
apigenin	520-36-5	C ₁₅ H ₁₀ O ₅	4	9.75	1.300	1.200	2.250	1.846	270.24	90.900	2.805	1
coumestrol	479-13-0	C ₁₅ H ₈ O ₅	4	9.80	1.000	1.100	2.130	1.737	268.22	83.810	2.715	0
phloretin	60-82-2	C ₁₅ H ₁₄ O ₅	4	9.88	1.650	1.080	2.160	1.997	274.27	97.990	2.943	4
kaempferol	520-18-3	C ₁₅ H ₁₀ O ₆	4	9.94	1.610	1.490	2.460	1.905	286.24	111.130	2.277	1
chrysoeriol	491-71-4	C ₁₆ H ₁₂ O ₆	4	10.06	1.070	1.280	2.310	2.045	300.26	100.130	2.555	2
diosmetin	520-34-3	C ₁₆ H ₁₂ O ₆	4	10.18	1.070	1.280	2.310	2.045	300.26	100.130	2.685	2
isorhamnetin	480-19-3	C ₁₆ H ₁₂ O ₇	4	10.26	1.380	1.570	2.530	2.104	316.26	120.360	2.067	2
hesperetin	520-33-2	C ₁₆ H ₁₄ O ₆	4	10.28	1.070	1.230	2.250	2.088	302.28	96.220	2.553	2
formononetin	485-72-3	C ₁₆ H ₁₂ O ₄	4	11.54	0.660	1.210	2.120	1.928	268.26	59.670	2.947	2
rhamnetin	90-19-7	C ₁₆ H ₁₂ O ₇	4	11.88	1.220	1.570	2.530	2.104	316.26	120.360	2.044	2
eupatilin	22368-21-4	C ₁₈ H ₁₆ O ₇	4	12.30	0.460	1.390	2.720	2.386	344.32	98.360	2.986	4
sinensetin	2306-27-6	C ₂₀ H ₂₀ O ₇	4	12.32	0.000	1.840	3.340	2.668	372.37	76.360	2.625	6
wogonin	632-85-9	C ₁₆ H ₁₂ O ₅	4	12.99	0.570	0.990	2.130	1.987	284.26	79.900	2.710	2
chrysin	480-40-0	C ₁₅ H ₁₀ O ₄	4	13.11	0.800	0.920	2.040	1.787	254.24	70.670	3.238	1
vitexicarpin	479-91-4	C ₁₉ H ₁₈ O ₈	4	13.19	0.180	1.560	2.800	2.586	374.34	107.590	2.149	5
acacetin	480-44-4	C ₁₆ H ₁₂ O ₅	4	13.20	0.800	1.140	2.130	1.987	284.26	79.900	3.341	2

sakuranetin	520-29-6	C ₁₆ H ₁₄ O ₅	4	13.25	0.630	1.080	2.080	2.030	286.28	75.990	2.967	2
phenylchromone	525-82-6	C ₁₅ H ₁₀ O ₂	4	13.82	0.000	0.780	1.820	1.670	222.24	30.210	3.362	1
flavanone	487-26-3	C ₁₅ H ₁₂ O ₂	4	15.60	0.000	0.730	1.760	1.713	224.25	26.300	3.133	1
chalcone	94-41-7	C ₁₅ H ₁₂ O ₁	4	16.85	0.000	0.580	1.700	1.720	208.26	17.070	3.633	3
beta-naphthoflavone	6051-87-2	C ₁₉ H ₁₂ O ₂	4	17.27	0.000	0.840	2.150	2.039	272.30	30.210	4.675	1
gamma-naphthoflavone	6051-88-3	C ₁₉ H ₁₂ O ₂	4	18.37	0.000	0.840	2.150	2.039	272.30	30.210	4.543	1
quercetin-3-O-rutinoside	153-18-4	C ₂₇ H ₃₀ O ₁₆	4	5.37	3.020	4.320	4.550	3.965	610.52	269.430	-0.873	6
hamamelitannin	469-32-9	C ₂₀ H ₂₀ O ₁₄	VM	2.95	3.540	3.250	3.930	3.079	484.36	243.900	-0.709	8
ellagic acid	476-66-4	C ₁₄ H ₆ O ₈	VM	5.15	1.770	1.410	2.600	1.772	302.19	141.340	1.704	0
secoisolariciresinol	29388-59-8	C ₂₀ H ₂₆ O ₆	VO	7.42	1.170	1.540	2.180	2.804	362.42	99.380	1.857	9
podophyllotoxin	518-28-5	C ₂₂ H ₂₂ O ₈	VO	10.88	0.310	2.170	3.670	2.834	414.41	92.680	1.971	4
sennoside B	128-57-4	C ₄₂ H ₃₈ O ₂₀	5	5.46	3.380	5.170	6.280	5.644	862.74	347.960	0.508	9
rhein-8-glucoside	34298-86-7	C ₂₁ H ₁₈ O ₁₁	5	5.47	1.730	2.870	3.730	2.892	446.36	191.050	0.416	4
sennoside A	81-27-6	C ₄₂ H ₃₈ O ₂₀	5	6.35	3.380	5.170	6.280	5.644	862.74	347.960	0.508	9
aloin B	28371-16-6	C ₂₁ H ₂₂ O ₉	5	7.33	1.740	2.530	3.010	2.860	418.39	167.910	-0.148	3
aloin A	5133-19-7	C ₂₁ H ₂₂ O ₉	5	7.74	1.740	2.530	3.010	2.860	418.39	167.910	-0.148	3
2,6-dihydroxyanthraquinone	84-60-6	C ₁₄ H ₈ O ₄	5	8.68	1.450	1.330	2.470	1.646	240.21	74.600	2.666	0
juglone	481-39-0	C ₁₀ H ₆ O ₃	5	9.24	0.190	0.720	1.780	1.219	174.15	54.370	1.936	0

lucidin	478-08-0	C ₁₅ H ₁₀ O ₅	5	11.07	1.190	1.370	2.420	1.846	270.24	94.830	2.359	1
alpha-naphthol	90-15-3	C ₁₀ H ₈ O ₁	5	11.32	0.500	0.450	1.230	1.144	144.17	20.230	2.909	0
1,2-dihydroxyanthraquinone	72-48-0	C ₁₄ H ₈ O ₄	5	11.82	0.690	0.880	2.250	1.646	240.21	74.600	2.997	0
plumbagin	481-42-5	C ₁₁ H ₈ O ₃	5	12.09	0.190	0.740	1.730	1.359	188.18	54.370	2.244	0
aloe-emodin	481-72-1	C ₁₅ H ₁₀ O ₅	5	12.10	0.810	1.230	2.320	1.846	270.24	94.830	2.238	1
rubiadin 1-methylether	7460-43-7	C ₁₆ H ₁₂ O ₄	5	12.72	0.720	1.220	2.290	1.928	268.26	63.600	3.179	1
rheic acid	478-43-3	C ₁₅ H ₈ O ₆	5	12.82	1.040	1.150	2.560	1.862	284.22	111.900	2.319	1
1,5-dihydroxyanthraquinone	117-12-4	C ₁₄ H ₈ O ₄	5	13.24	0.390	0.690	2.100	1.646	240.21	74.600	3.061	0
frangulin	69686-05-1	C ₂₁ H ₂₀ O ₉	5	13.53	1.220	2.250	3.140	2.817	416.38	153.750	1.682	2
anthraquinone	84-65-1	C ₁₄ H ₈ O ₂	5	14.62	0.000	0.790	2.050	1.529	208.21	34.140	2.934	0
anthrone	90-44-8	C ₁₄ H ₁₀ O ₁	5	14.77	0.000	0.480	1.560	1.513	194.23	17.070	3.331	0
emodin	518-82-1	C ₁₅ H ₁₀ O ₅	5	15.80	1.180	1.040	2.320	1.846	270.24	94.830	2.990	0
1-hydroxyanthraquinone	129-43-1	C ₁₄ H ₈ O ₃	5	15.91	0.190	0.740	2.080	1.588	224.21	54.370	3.269	0
rubiadin	117-02-2	C ₁₅ H ₁₀ O ₄	5	16.64	0.880	0.960	2.220	1.787	254.24	74.600	3.084	0
chrysophanol	481-74-3	C ₁₅ H ₁₀ O ₄	5	18.11	0.490	0.820	2.120	1.787	254.24	74.600	3.421	0
physcione	521-61-9	C ₁₆ H ₁₂ O ₅	5	19.53	0.490	0.980	2.200	1.987	284.26	83.830	3.473	1
anthanthron	641-13-4	C ₂₂ H ₁₀ O ₂	5	20.27	0.000	0.890	2.730	2.158	306.31	34.140	5.288	0
swertiamarine	17388-39-5	C ₁₆ H ₂₂ O ₁₀	6	3.54	1.310	2.850	2.540	2.495	374.34	155.140	-1.592	4

cantharidin	56-25-7	C ₁₀ H ₁₂ O ₄	6	4.07	0.000	1.030	2.460	1.341	196.20	52.600	0.349	0
gentiopicroside	20831-76-9	C ₁₆ H ₂₀ O ₉	6	4.08	1.000	2.620	2.380	2.394	356.32	134.910	-1.181	4
loganin	18524-94-2	C ₁₇ H ₂₆ O ₁₀	6	4.08	1.310	2.850	2.510	2.679	390.38	155.140	-1.111	5
anisatin	5230-87-5	C ₁₅ H ₂₀ O ₈	6	4.11	0.940	2.320	3.310	2.171	328.31	133.520	-1.275	0
cornin	548-37-8	C ₁₇ H ₂₄ O ₁₀	6	4.20	1.000	2.850	2.740	2.636	388.37	151.980	-0.970	5
sweroside	14215-86-2	C ₁₆ H ₂₂ O ₉	6	4.20	1.000	2.570	2.320	2.437	358.34	134.910	-1.079	4
agnuside	11027-63-7	C ₂₂ H ₂₆ O ₁₁	6	5.12	1.970	3.250	3.290	3.162	466.44	175.370	-0.523	7
catalposide	6736-85-2	C ₂₂ H ₂₆ O ₁₂	6	5.44	1.890	3.510	3.350	3.155	482.43	187.900	-0.791	7
picroside II	39012-20-9	C ₂₃ H ₂₈ O ₁₃	6	5.84	1.660	3.590	3.410	3.354	512.46	197.130	-0.903	8
picroside I	27409-30-9	C ₂₄ H ₂₈ O ₁₁	6	7.38	1.200	3.330	3.230	3.335	492.47	167.670	0.028	8
helenalin	6754-13-8	C ₁₅ H ₁₈ O ₄	6	7.57	0.310	1.210	1.600	1.959	262.30	63.600	0.941	0
ginsenoside RG1	22427-39-0	C ₄₂ H ₇₂ O ₁₄	6	8.24	2.620	4.810	4.520	6.154	801.01	239.220	1.243	10
ginkgolide A	15291-75-5	C ₂₀ H ₂₄ O ₉	6	8.28	0.360	2.520	4.120	2.674	408.40	128.590	0.399	1
cnicin	24394-09-0	C ₂₀ H ₂₆ O ₇	6	8.32	0.850	1.930	2.050	2.862	378.42	113.290	0.472	6
santamarin	4290-13-5	C ₁₅ H ₂₀ O ₃	6	8.34	0.310	0.890	1.070	1.943	248.32	46.530	2.074	0
madecassoside	34540-22-2	C ₄₈ H ₇₈ O ₂₀	6	8.38	3.320	6.210	4.870	7.091	975.12	335.440	0.494	10
andrographolide	5508-58-7	C ₂₀ H ₃₀ O ₅	6	8.55	0.850	1.590	1.510	2.765	350.45	86.990	1.699	3
harpagoside	19210-12-9	C ₂₄ H ₃₀ O ₁₁	6	9.25	1.540	3.280	3.300	3.444	494.49	175.370	0.001	7

asiaticoside	16830-15-2	C ₄₈ H ₇₈ O ₁₉	6	9.31	3.010	5.910	4.620	7.032	959.12	315.210	1.347	10
neoquassin	76-77-7	C ₂₂ H ₃₀ O ₆	6	9.78	0.310	1.930	2.130	2.954	390.47	82.060	2.416	2
quassin	76-78-8	C ₂₂ H ₂₈ O ₆	6	9.85	0.000	1.860	2.840	2.911	388.45	78.900	2.452	2
stevioside	57817-89-7	C ₃₈ H ₆₀ O ₁₈	6	10.18	2.740	5.490	4.290	5.673	804.87	294.980	-0.984	10
ginsenoside RF	52286-58-5	C ₄₂ H ₇₂ O ₁₄	6	10.83	2.680	4.790	4.530	6.154	801.01	239.220	1.727	10
neoandrographolide	27215-14-1	C ₂₆ H ₄₀ O ₈	6	11.29	1.000	2.420	2.120	3.678	480.59	125.680	1.885	7
dehydroandrographolide	134418-28-3	C ₂₀ H ₂₈ O ₄	6	12.10	0.630	1.260	1.410	2.664	332.43	66.760	2.637	3
14-deoxy-11,14-didehydroandrographolide	42895-58-9	C ₂₀ H ₂₈ O ₄	6	12.12	0.630	1.260	1.410	2.664	332.43	66.760	2.597	3
carvone	99-49-0	C ₁₀ H ₁₄ O ₁	6	12.47	0.000	0.510	0.810	1.339	150.22	17.070	2.094	1
beta-pinene	127-91-3	C ₁₀ H ₁₆ O ₀	6	12.62	0.000	0.160	0.280	1.257	136.23	0.000	4.392	0
alpha-pinene	80-56-8	C ₁₀ H ₁₆ O ₀	6	12.75	0.000	0.180	0.330	1.257	136.23	0.000	4.415	0
7alpha-O-methyl-conacytone	-	C ₂₁ H ₂₈ O ₆	6	13.30	0.630	1.950	1.190	2.813	376.44	93.060	2.296	2
linderalactone	728-61-0	C ₁₅ H ₁₆ O ₃	6	13.73	0.000	0.640	1.070	1.857	244.29	39.440	2.822	0
echinocystic acid-3-glucoside	78285-90-2	C ₃₆ H ₅₈ O ₉	6	14.03	1.880	2.830	2.340	4.972	634.84	156.910	4.262	4
geranial	141-27-5	C ₁₀ H ₁₆ O ₁	6	14.63	0.000	0.500	0.840	1.447	152.23	17.070	2.721	4
obacunone	751-03-1	C ₂₆ H ₃₀ O ₇	6	14.66	0.000	1.850	3.640	3.273	454.51	95.340	2.507	1
linderane	13476-25-0	C ₁₅ H ₁₆ O ₄	6	14.73	0.000	1.050	2.650	1.851	260.29	51.970	2.002	0
andrograpanin	82209-74-3	C ₂₀ H ₃₀ O ₃	6	14.94	0.310	0.890	1.060	2.648	318.45	46.530	4.362	4

marrubiin	465-92-9	C ₂₀ H ₂₈ O ₄	6	15.52	0.310	1.030	1.990	2.598	332.43	59.670	3.672	3
alpha-hederin	27013-91-8	C ₄₁ H ₆₆ O ₁₂	6	15.81	2.090	3.630	2.960	5.744	750.96	195.600	4.039	6
(+)-costunolide	553-21-9	C ₁₅ H ₂₀ O ₂	6	16.49	0.000	0.630	0.850	1.950	232.32	26.300	3.563	0
dehydrocostus lactone	477-43-0	C ₁₅ H ₁₈ O ₂	6	16.91	0.000	0.640	0.830	1.842	230.30	26.300	3.067	0
dihydrotanshinone I	87205-99-0	C ₁₈ H ₁₄ O ₃	6	16.92	0.000	1.270	3.210	2.043	278.30	43.370	3.791	0
demethylfruticulic acid	106664-42-0	C ₁₉ H ₁₈ O ₄	6	17.66	0.810	1.540	1.230	2.351	310.34	74.600	4.428	1
fruticulic acid	-	C ₂₁ H ₂₂ O ₅	6	18.04	0.310	1.710	1.220	2.691	354.40	72.830	4.732	3
tanshinone I	568-73-0	C ₁₈ H ₁₂ O ₃	6	18.27	0.000	1.070	3.270	2.000	276.29	47.280	4.125	0
cryptotanshinone	35825-57-1	C ₁₉ H ₂₀ O ₃	6	18.34	0.000	1.230	2.860	2.269	296.36	43.370	4.339	0
hederagenin	465-99-6	C ₃₀ H ₄₈ O ₄	6	18.89	1.200	1.270	1.300	3.941	472.70	77.760	6.384	2
echinocystic acid	510-30-5	C ₃₀ H ₄₈ O ₄	6	19.20	1.200	1.300	1.280	3.941	472.70	77.760	6.568	1
valerenic acid	3569-10-6	C ₁₅ H ₂₂ O ₂	6	19.52	0.570	0.610	0.820	1.993	234.33	37.300	4.218	2
glycyrrhetic acid	471-53-4	C ₃₀ H ₄₆ O ₄	6	20.39	0.880	1.300	1.520	3.898	470.68	74.600	6.417	1
tanshinone IIA	568-72-9	C ₁₉ H ₁₈ O ₃	6	20.77	0.000	1.040	2.920	2.226	294.34	47.280	4.791	0
fruticulic acid	106664-40-8	C ₂₀ H ₂₀ O ₄	6	21.12	0.310	1.480	1.110	2.492	324.37	63.600	4.882	2
7alpha-19-diacetoxy-royleanone	-	C ₂₄ H ₃₂ O ₇	6	21.46	0.310	1.910	1.340	3.317	432.51	106.970	2.979	6
guaiazulene	489-84-9	C ₁₅ H ₁₈ O ₀	6	23.28	0.000	0.450	0.600	1.790	198.30	0.000	5.685	1
betulinic acid	472-15-1	C ₃₀ H ₄₈ O ₃	6	23.35	0.880	1.190	2.080	3.883	456.70	57.530	6.742	2

oleanolic acid	508-02-1	C ₃₀ H ₄₈ O ₃	6	23.82	0.880	1.000	1.020	3.883	456.70	57.530	7.808	1
ursolic acid	77-52-1	C ₃₀ H ₄₈ O ₃	6	23.96	0.880	1.020	1.060	3.883	456.70	57.530	7.718	1
tetrahydrocannabinol	1972-08-3	C ₂₁ H ₃₀ O ₂	6	24.50	0.500	0.710	1.040	2.687	314.46	29.460	7.317	4
uvaol	545-46-0	C ₃₀ H ₅₀ O ₂	6	25.37	0.630	0.970	0.890	3.867	442.72	40.460	7.896	1
alpha-amyrin	638-95-9	C ₃₀ H ₅₀ O ₁	6	32.78	0.310	0.690	0.620	3.808	426.72	20.230	9.367	0
lupeol	545-47-1	C ₃₀ H ₅₀ O ₁	6	33.82	0.310	0.860	1.640	3.808	426.72	20.230	8.606	1
ouabain	630-60-4	C ₂₉ H ₄₄ O ₁₂	7	4.25	2.290	3.730	4.290	4.162	584.65	206.600	-1.587	4
convallatoxin	508-75-8	C ₂₉ H ₄₂ O ₁₀	7	7.85	1.350	3.140	4.050	4.001	550.64	162.980	0.165	4
strophanthin K	508-77-0	C ₃₀ H ₄₄ O ₉	7	7.98	0.810	2.820	3.660	4.083	548.66	131.750	1.335	5
gitoxigenin	545-26-6	C ₂₃ H ₃₄ O ₅	7	9.23	0.940	1.660	2.720	3.014	390.51	86.990	1.701	1
lanatoside C	17575-22-3	C ₄₉ H ₇₆ O ₂₀	7	10.02	2.080	5.860	5.690	7.080	985.12	288.280	0.436	12
digoxin	20830-75-5	C ₄₁ H ₆₄ O ₁₄	7	10.18	1.580	4.320	4.460	5.753	780.94	203.060	1.362	7
proscillaridin	466-06-8	C ₃₀ H ₄₂ O ₈	7	11.48	1.040	2.680	3.440	3.982	530.65	129.590	1.933	3
withaferin A	5119-48-2	C ₂₈ H ₃₈ O ₆	7	11.74	0.540	2.120	3.150	3.582	470.60	96.360	2.922	3
12-deoxywithastramonolide	60124-17-6	C ₂₈ H ₃₈ O ₆	7	12.53	0.540	2.120	3.150	3.582	470.60	96.360	2.783	3
gitoxin	4562-36-1	C ₄₁ H ₆₄ O ₁₄	7	12.57	1.580	4.320	4.460	5.753	780.94	203.060	1.413	7
oleandrin	465-16-7	C ₃₂ H ₄₈ O ₉	7	13.96	0.500	2.610	3.280	4.365	576.72	120.750	2.551	6
digitoxin	71-63-6	C ₄₁ H ₆₄ O ₁₃	7	14.65	1.270	4.020	4.200	5.694	764.94	182.830	2.403	7

acetyldigitoxin	1111-39-3	C ₄₃ H ₆₆ O ₁₄	7	16.04	0.950	4.170	4.280	5.991	806.98	188.900	3.094	9
guggulesterone	39025-24-6	C ₂₁ H ₂₈ O ₂	7	17.45	0.000	1.090	2.550	2.579	312.45	34.140	3.634	0
ophiopogonin D	945619-74-9	C ₄₄ H ₇₀ O ₁₆	7	17.92	1.970	5.170	4.880	6.227	855.02	235.680	1.325	6
dioscin	19057-60-4	C ₄₅ H ₇₂ O ₁₆	7	18.14	1.910	5.240	4.840	6.368	869.04	235.680	1.814	7
trillin	14144-06-0	C ₃₃ H ₅₂ O ₈	7	20.45	1.000	2.780	3.220	4.425	576.76	117.840	3.487	3
ruscogenin	472-11-7	C ₂₇ H ₄₂ O ₄	7	20.91	0.630	1.550	2.420	3.453	430.62	58.920	4.538	0
cholesterol	57-88-5	C ₂₇ H ₄₆ O ₁	7	26.06	0.310	0.810	1.760	3.494	386.65	20.230	8.372	5
yamogenin	512-06-1	C ₂₇ H ₄₂ O ₃	7	26.36	0.310	1.250	2.160	3.394	414.62	38.690	5.802	0
diosgenin	512-04-9	C ₂₇ H ₄₂ O ₃	7	26.46	0.310	1.250	2.160	3.394	414.62	38.690	5.802	0
smilagenin	126-18-1	C ₂₇ H ₄₄ O ₃	7	27.49	0.310	1.200	2.100	3.437	416.64	38.690	6.102	0
sarsasapogenin	126-19-2	C ₂₇ H ₄₄ O ₃	7	27.54	0.310	1.200	2.100	3.437	416.64	38.690	6.102	0

^a clusters :

1: aliphatics (VA)

2: simple aromatics (VG)

3: benzopyranoids (VI)

4: flavonoids (VK)

5: polycyclic aromatics (VQ)

6: terpenoids (VS)

7: steroids (VT)

Compounds that are not in clusters: VM = tannins; VO = lignans.

Geometrical Transfer Equation

The RT_{pred} can be calculated for new chromatographic conditions. All kinetic parameters can be changed (e.g., column geometry or gradient time), but the thermodynamic ones have to be kept constant (e.g., mobile and stationary phases and column temperature).[43] The parameter that is constant when transferring a gradient is the composition of the organic modifier at the elution of the analytes, which is defined by Equation S1:

$$CE = \Phi_{init} + slope(RT - t_{col} - t_{isoc} - \frac{v_d}{F}) \quad (\text{Equation S1})$$

where CE is the composition of the organic modifier at elution (no unit), Φ_{init} is the fraction of organic modifier in the mobile phase at the beginning of the gradient (no unit), $slope$ is the gradient slope (no unit) defined by Equation S2, RT is the retention time (min), t_{col} is the column dead time (min) calculated using Equation S3, t_{isoc} is the initial isocratic hold (min), v_d is the system dwell volume (μL) and F is the mobile phase flow rate ($\mu\text{L}/\text{min}$).

$$slope = \frac{\Phi_{fin} - \Phi_{init}}{t_{grad}} \quad (\text{Equation S2})$$

where Φ_{fin} is the fraction of organic modifier in the mobile phase at the end of the gradient (no unit) and t_{grad} is the gradient time (min).

$$t_{col} = \frac{\pi \times p_{col} \times \frac{1}{4} d_{col}^2 \times l_{col}}{F} \quad (\text{Equation S3})$$

where p_{col} is the column porosity (no unit), d_{col} is the column internal diameter (mm) and l_{col} is the column length (mm).

If the CE is unchanged after the transfer from condition 1 to condition 2, then the RT_2 is calculated using Equation S4.

$$\Phi_{init,2} + slope_2(RT_2 - t_{col,2} - \frac{v_{d,2}}{F_2}) = \Phi_{init,1} + slope_1(RT_1 - t_{col,1} - \frac{v_{d,1}}{F_1}) \quad (\text{Equation S4a})$$

or

$$RT_2 = \frac{\Phi_{init,1} - \Phi_{init,2} + slope_1(RT_1 - t_{col,1} - t_{isoc,1} - \frac{v_{d,1}}{F_1})}{slope_2} + t_{col,2} + t_{isoc,2} + \frac{v_{d,2}}{F_2} \quad (\text{Equation S4b} = \text{Equation 2}).$$

Note that the dead volume between the column and the detector is not considered in Equation S4b, which could change with a modification of the flow rate. However, this change has no real effects on the final RT calculation.

Acknowledgements

The authors are thankful to Dr. Davy Guillarme, Dr. Guillaume Marti and Dr. Samuel Bertrand from the University of Geneva for the fruitful discussions. They want to acknowledge also Prof. Annelise Lobstein from the University of Strasbourg and the French National Chemical

Library for the gift of many NP standards, and Marie Ades from the University of Strasbourg for their preparation. J.L.W is thankful to the Swiss National Science Foundation for the financial support for the development of metabolomics methods (Grant 200020_146200).

References

- [1] P.C. Sadek, P.W. Carr, R.M. Doherty, M.J. Kamlet, R.W. Taft, M.H. Abraham. Study of Retention Processes in Reversed-Phase High-Performance Liquid-Chromatography by the Use of the Solvatochromic Comparison Method. *Analytical Chemistry*, **1985**. 57: 2971-2978.
- [2] M.H. Abraham. Scales of solute hydrogen-bonding: their construction and application to physicochemical and biochemical processes. *Chemical Society Reviews*, **1993**. 22: 73-83.
- [3] K. Héberger. Quantitative structure–(chromatographic) retention relationships. *Journal of Chromatography A*, **2007**. 1158: 273-305.
- [4] A. Nasal, R. Kaliszan. Progress in the Use of HPLC for Evaluation of Lipophilicity. *Current Computer - Aided Drug Design*, **2006**. 2: 327-340.
- [5] S. Martel, D. Guillarme, Y. Henchoz, A. Galland, J.L. Veuthey, S. Rudaz, P.-A. Carrupt. Chromatographic Approaches for Measuring LogP, in *Molecular Drug Properties. Measurement and Prediction.*, R. Mannhold, Editor. **2008**. p. 331-355.
- [6] P. Eugster, S. Martel, D. Guillarme, P.A. Carrupt, J.L. Wolfender. Rapid log P determination of natural products in crude plant extracts from UHPLC-TOF-MS profiling data - an additional parameter for dereplication and bioavailability. *Planta Medica*, **2009**. 75: 913-914.
- [7] F.E. Koehn, G.T. Carter. The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery*, **2005**. 4: 206-220.
- [8] M. Feher, J.M. Schmidt. Property Distributions: Differences between Drugs, Natural Products, and Molecules from Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences*, **2002**. 43: 218-227.
- [9] J.L. Wolfender, G. Glauser, J. Boccard, S. Rudaz. MS-based Plant Metabolomic Approaches for Biomarker Discovery. *Natural Product Communications*, **2009**. 4: 1417-1430.
- [10] K. Hostettmann, A. Marston, M. Hostettmann. *Preparative Chromatography Techniques: Applications in Natural Product Isolation*. 2nd ed. **1997**, Berlin, Springer.
- [11] J.-L. Wolfender, G. Marti, E.F. Queiroz. Advances in Techniques for Profiling Crude Extracts and for the Rapid Identification of Natural Products: Dereplication, Quality Control and Metabolomics. *Current Organic Chemistry*, **2010**. 14: 1808-1832.
- [12] F. van der Kooy, F. Maltese, Y. Hae Choi, H. Kyong Kim, R. Verpoorte. Quality Control of Herbal Material and Phytopharmaceuticals with MS and NMR Based Metabolic Fingerprinting. *Planta Medica*, **2009**. 75: 763-775.
- [13] G. Glauser, N. Veyrat, B. Rochat, J.-L. Wolfender, T.C.J. Turlings. Ultra-high pressure liquid chromatography-mass spectrometry for plant metabolomics: a systematic comparison of high-resolution quadrupole-time-of-flight and single stage Orbitrap mass spectrometers. *Journal of Chromatography A*, **2013**. 1292: 151-159.
- [14] W.F. Smyth, T.J.P. Smyth, V.N. Ramachandran, F. O'Donnell, P. Brooks. Dereplication of phytochemicals in plants by LC-ESI-MS and ESI-MSn. *TrAC Trends in Analytical Chemistry*, **2012**. 33: 46-54.
- [15] P.J. Eugster, D. Guillarme, S. Rudaz, J.L. Veuthey, P.A. Carrupt, J.L. Wolfender. Ultra High Pressure Liquid Chromatography for Crude Plant Extract Profiling. *Journal of AOAC International*, **2011**. 94: 51-70.
- [16] T. Kind, O. Fiehn. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics*, **2007**. 8: 105-124.
- [17] C.S. Funari, P.J. Eugster, S. Martel, P.-A. Carrupt, J.-L. Wolfender, D.H.S. Silva. High resolution ultra high pressure liquid chromatography–time-of-flight mass spectrometry dereplication

- strategy for the metabolite profiling of Brazilian *Lippia* species. *Journal of Chromatography A*, **2012**. 1259: 167–178.
- [18] J.R. Mazzeo, U.D. Neue, M. Kele, R.S. Plumb. A new separation technique takes advantage of sub-2- μm porous particles. *Analytical Chemistry*, **2005**. 77: 460A-467A.
- [19] P.J. Eugster, J.L. Wolfender. UHPLC in Natural Products Analysis, in *UHPLC in Life Sciences*, D. Guillarme and J.-L. Veuthey, Editors. **2012**, RSC Publishing. p. 354.
- [20] J. Larsson, J. Gottfries, S. Muresan, A. Backlund. ChemGPS-NP: Tuned for navigation in biologically relevant chemical space. *Journal of Natural Products*, **2007**. 70: 789-794.
- [21] R. Kaliszan. QSRR: Quantitative Structure-(Chromatographic) retention relationships. *Chemical Reviews*, **2007**. 107: 3212-3246.
- [22] R. Put, Y.V. Heyden. Review on modelling aspects in reversed-phase liquid chromatographic quantitative structure-retention relationships. *Analytica Chimica Acta*, **2007**. 602: 164-172.
- [23] A. Tellez, M. Roses, E. Bosch. Modeling the Retention of Neutral Compounds in Gradient Elution RP-HPLC by Means of Polarity Parameter Models. *Analytical Chemistry*, **2009**. 81: 9135-9145.
- [24] R. Kaliszan, T. Bączek, A. Cimochovska, P. Juszczuk, K. Wiśniewska, Z. Grzonka. Prediction of high-performance liquid chromatography retention of peptides with the use of quantitative structure-retention relationships. *PROTEOMICS*, **2005**. 5: 409-415.
- [25] J. Akbar, S. Iqbal, F. Batool, A. Karim, K. Chan. Predicting Retention Times of Naturally Occurring Phenolic Compounds in Reversed-Phase Liquid Chromatography: A Quantitative Structure-Retention Relationship (QSRR) Approach. *International Journal of Molecular Sciences*, **2012**. 13: 15387-15400.
- [26] D.J. Creek, A. Jankevics, R. Breitling, D.G. Watson, M.P. Barrett, K.E.V. Burgess. Toward Global Metabolomics Analysis with Hydrophilic Interaction Liquid Chromatography–Mass Spectrometry: Improved Metabolite Identification by Retention Time Prediction. *Analytical Chemistry*, **2011**. 83: 8703-8710.
- [27] H. Kempe, M. Kempe. QSRR analysis of β -lactam antibiotics on a penicillin G targeted MIP stationary phase. *Analytical and Bioanalytical Chemistry*, **2010**. 398: 3087-3096.
- [28] A.A. D'Archivio, M.A. Maggi, F. Ruggieri. Multiple-column RP-HPLC retention modelling based on solvatochromic or theoretical solute descriptors. *Journal of Separation Science*, **2010**. 33: 155-166.
- [29] E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant. PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry*, **2008**. 4: 217-241.
- [30] Pubchem. [Access May 5, 2013]; Available from: <http://pubchem.ncbi.nlm.nih.gov/>.
- [31] ACD/I-lab tool. [Access May 4, 2013]; Available from: <https://ilab.acdlabs.com/iLab2/>.
- [32] *MedChem Designer*, version 2.0.034, **2012**, Simulations Plus Inc.
- [33] M. Cheminformatics. Molinspiration Cheminformatics calculator. [Access May 4, 2013]; Available from: <http://www.molinspiration.com/cgi-bin/properties>.
- [34] J. Buckingham. *Dictionary of Natural Products on DVD*, version 21:2, **2012**, CRC Press.
- [35] K. Levenberg. A method for the solution of certain problems in least squares. *Quarterly of applied mathematics*, **1944**. 2: 164-168.
- [36] D.W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial & Applied Mathematics*, **1963**. 11: 431-441.
- [37] P.J. Eugster, D. Biass, D. Guillarme, P. Favreau, R. Stöcklin, J.-L. Wolfender. Peak capacity optimisation for high resolution peptide profiling in complex mixtures by liquid chromatography coupled to time-of-flight mass spectrometry: Application to the *Conus consors* cone snail venom. *Journal of Chromatography A*, **2012**. 1259: 187-199.
- [38] S.M. Al-Massarani, S. Bertrand, A. Nievergelt, A.M. El-Shafae, T.A. Al-Howiriny, N.M. Al-Musayeib, M. Cuendet, J.-L. Wolfender. Acylated pregnane glycosides from *Caralluma sinaica*. *Phytochemistry*, **2012**. 79: 129-140.

- [39] R. Mannhold, G.I. Poda, C. Ostermann, I.V. Tetko. Calculation of molecular lipophilicity: State-of-the-art and comparison of logP methods on more than 96,000 compounds. *Journal of Pharmaceutical Sciences*, **2009**. 98: 861-893.
- [40] S. Wold, A. Ruhe, H. Wold, I. Dunn, W. The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM Journal on Scientific and Statistical Computing*, **1984**. 5: 735-743.
- [41] Y. Henchoz, D. Guillarme, S. Martel, S. Rudaz, J.L. Veuthey, P.A. Carrupt. Fast log P determination by ultra-high-pressure liquid chromatography coupled with UV and mass spectrometry detections. *Analytical and Bioanalytical Chemistry*, **2009**. 394: 1919-1930.
- [42] J.J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, **1982**. 79: 2554-2558.
- [43] D. Guillarme, D.T.T. Nguyen, S. Rudaz, J.L. Veuthey. Method transfer for fast liquid chromatography in pharmaceutical analysis: Application to short columns packed with small particle. Part II: Gradient experiments. *European Journal of Pharmaceutics and Biopharmaceutics*, **2008**. 68: 430-440.

Chapter VIII – Concluding Remarks

1. High resolution metabolite profiling and online dereplication

This thesis work mainly focused on the high resolution metabolite profiling and online dereplication of NPs. It aimed first at optimising the separation of the constituents of a complex natural matrix to provide high quality spectroscopic data, and secondly to test and develop efficient tools for the online dereplication and identification of compounds in natural samples.

The first chapters provided some solutions to increase chromatographic resolution. In this respect, the UHPC technology based on the use of sub-2 μm particles is well adapted to provide high resolution profiling of complex biological matrices, and its efficiency either in terms of throughput, resolution, and reproducibility is very advantageous as compared to classical HPLC. There is also a significant reduction in solvent and sample consumption because of the smaller diameter of the columns. The growing number of semi-preparative columns available with similar phase chemistries as those developed for UHPLC allows the easy transfer of UHPLC to semi-preparative scale for the isolation of NPs which is important for *de novo* structure assignment of unknown NPs or for assessing their bioactivities.

In order to provide efficient UHPLC separations, the chromatographic conditions, e.g. mobile phase temperature, column internal diameter, mobile phase flow rate and gradient time have to be carefully chosen. Figure III.13 provided some solutions to get the highest peak capacity in LC separations, based on the results of a systematic

study of the influence of several chromatographic parameters on the peak capacity (see Chapter III). The use of elevated mobile phase temperatures is particularly interesting and should be considered after evaluation of the thermal stability of the analytes and of the selectivity change. The use of optimised chromatographic conditions for the metabolite profiling of natural samples provided a high chromatographic resolution and a baseline separation of most constituents on complex matrices, very useful for obtaining MS data with the least possible interferences from coeluting metabolites that greatly helps dereplication.

Even a UHPLC separation providing the highest peak capacity may however not be sufficient to separate all the constituents of natural samples. In this respect, the development of ion mobility spectrometry is promising. Our first trials showed that this technology is adapted to the analysis of complex natural samples since it provides a high resolution separation which is orthogonal to LC and MS dimensions. Interestingly, this technique is capable of separating closely related isomers such as stereoisomers that are frequently present in natural samples, and that are not separated by the HR-MS alone and only with difficulty in the LC dimension. However the coupling of LC, IMS and MS is today not yet applicable for technical reasons.

The second part of the thesis work investigated some aspects of the high resolution separation and the online dereplication and identification of metabolites in natural samples.

Since MS spectra with the least possible interferences were obtained by the high resolution separation methods developed, we could use the high quality HR-MS information for estimating how far dereplication can rely on this information. In the HR-MS spectra obtained, the exact mass of the metabolite of interest was determined taking into account all peaks of the spectrum to detect the presence of adducts and/or dimers. This is a crucial step that still lacks efficient automated tools. The subsequent application of the dereplication procedure allowed the dramatic reduction of the number of candidate molecular formulae, by filtering based on heuristic rules (see Chapters V and VI) and on database searches. In a second step, the identification of one or a few corresponding NPs was provided by chemotaxonomic information

(see Chapters V and VI) and by retention information ($\log P$ or retention time, see Chapters VI-VII). The identity of the metabolite of interest is then confirmed by the analysis of the pure standard spiked in the natural sample, when it is available. The application of the whole procedure in the frame of the chemotaxonomic study of the genus *Lippia* (see Chapter VI) showed that this is an efficient method for dereplication purposes. It is also useful in the frame of *de novo* structural elucidation, since it provides a validated molecular formula of the unknown compound. Still, additional online analyses have to be performed to ascertain the structure of the analyte, such as MS/MS, as well as NMR and IR, after up-scaling of the separation method to semi-preparative LC to obtain the pure NP.

2. Maturity of LC-MS instrumentation

The studies performed during this thesis work showed that the high resolution profiling based on UHPLC-(Q)TOF-MS is an efficient approach to provide a detailed metabolite profiling. LC-MS systems offer indeed the required features to deal with the complexity of natural samples: a high resolution in both LC and MS dimensions to separate hundreds of analytes, the possibility to develop generic methods to analyse compounds with a high physicochemical diversity, and a high dynamic range to detect analytes over a wide range of concentrations.

2.1. The LC dimension

Thanks to the development of UHPLC technology, the efficiency of chromatographic separations has been greatly improved and high resolution separations are obtained in a reasonable analysis time. This technique has thus become today the gold standard for the detailed profiling of complex natural samples. The wide choice of pH and high temperature-resistant phase chemistries allows the separation of many types of natural matrices, although C18 columns that are well-adapted to mid-polar metabolites are the most used (see subchapter 4).

On the other hand, the recent development of fused-core columns packed with sub-3 μm superficially porous particles offers equivalent or even higher resolution compared to UHPLC without requiring important changes on the conventional LC systems. Although these columns have not yet really been adopted by NP researchers (see Figure II.1), one can reasonably think that this tendency will be inverted in the next years.

Moreover, classical one-dimensional LC is not the only separative technique available. New approaches such as ion mobility spectrometry (see Chapter IV) and 2D-LC [1] open new perspectives, for example for the separation of stereoisomers and for the profiling of very complex mixtures.

2.2. The MS dimension

On the MS side, different key advances have been recorded over the last years that were implemented in the profiling of complex natural samples [2, 3]. Today's QTOF-MS instruments attain resolutions higher than 40'000 and provide mass accuracies of < 1 ppm. Such performances are of utmost importance in dereplication procedures, since a higher mass accuracy provides a lower number of molecular formulae for a given LC peak. Furthermore, acquisition rates were greatly improved, and allow the efficient hyphenation of QTOF-MS instruments with UHPLC systems that provide very thin LC peaks. This parameter should however be improved to provide a higher number of spectra per chromatographic peak, and to allow the parallel acquisition of several MS/MS spectra in the same analysis in hyphenation with UHPLC, as this becomes possible with the latest commercialised instruments. The acquisition of alternate positive and negative ion mode spectra without compromising spectral quality at such high frequencies would also be an important advantage for metabolite profiling.

Until now, (Q)TOF-MS instruments were almost the only high resolution MS instruments used in hyphenation with LC in NP research, although

Orbitrap-MS are also adapted to these applications [4]. However, since the acquisition rate of Orbitrap instruments was greatly improved, their hyphenation with UHPLC becomes possible. Such a UHPLC-Orbitrap-MS platform is very promising for the high resolution profiling of complex natural samples. Indeed, Orbitrap spectrometers provide very high resolution (up to 70'000 at m/z 400 for a 0.76 s scan [5]).

Finally, one can expect the future MS instruments to be mainly bench-top sized and their prices to decrease, thanks to their growing use in many analytical domains. This will probably be beneficial to NP research, where MS instruments have become mandatory in many aspects of the studies.

In summary, modern LC-MS instruments are today able to provide high quality LC-MS data for online metabolite dereplication and identification. Moreover, interesting improvements are probably coming up in the next few years that will be beneficial for NP research. One can indeed think that the (probable) future increase in resolution in both LC and MS dimensions will provide not only a higher number of features detected in a single analysis and more efficient dereplication of NPs, but also some sort of "high throughput high resolution profiling", *i.e.* an analysis providing the same peak capacity as obtained by long high resolution profiling methods but in a few minutes only. Such developments would be beneficial for classical bioactivity-guided isolation studies in NP research as well as for metabolomics.

3. Constraints in post-LC-MS analysis steps

Online dereplication is a useful tool to avoid unnecessary isolation of known compounds and thus saves time in classical phytochemical investigations. The duration of an entire metabolite profiling study with a dereplication procedure such as that presented in Chapters V and VI, including the preparation of 30 samples and the dereplication of the same 30 peaks in each sample, was estimated to be 52 hours of work (Figure VIII.1), which is much lower than the

time needed for the isolation of 30 compounds that may last for months.

Interestingly, the LC-MS analysis itself is not the limiting step anymore today, since high throughput and high resolution LC-MS instruments provide a huge amount of valuable data in a very short time. This actually took only 4 hours of acquisition in the example selected.

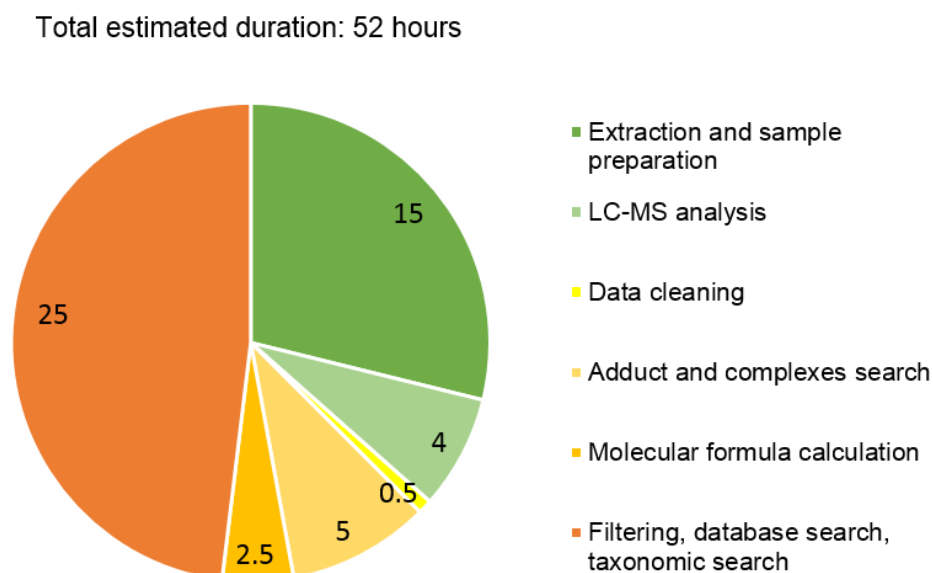


Figure VIII.1. Estimation of the time needed (in hours) for each step in a typical natural sample analysis with dereplication, including the preparation of 30 samples and the peak annotation of 30 analytes, as performed in the study presented in Chapter VI. The LC-MS analysis comprises only the preparation of the instruments, not the effective analysis time. The duration of the entire process was estimated to be 52 hours.

When considering the time needed to perform each following step separately, it is clear that today's real bottleneck in the online dereplication and metabolite identification processes is comprised in the steps that follow the LC-MS

These post-analytical steps may be divided into two main parts: molecular formula assignment and metabolite identification (yellow and orange parts of the Figure VIII.1 respectively). The limiting factors of both steps are described below.

3.1. Automated molecular formulae annotation

As described in Chapter V, section 3.3, there are several steps that have to be performed to get the molecular formula corresponding to a given LC peak, that are, in summary, determination of the exact MW of a given analyte in the spectrum, confirmation of this mass by comparison of PI and NI spectra, determination of the possible molecular formula based on these MW, and finally reduction of the number of possible molecular formula by a series of adapted filters.

Two of these steps are long and tedious and represent a serious bottleneck in dereplication, namely the determination of the MW of the metabolite in the MS spectrum and the application of filters to reduce the number of possible molecular formulae. Although many software exist that propose tools able to perform these tasks, none of them is able to provide a comprehensive, automated and integrated solution to provide a molecular formula directly from an MS spectrum. However, the recent advances in terms of software for molecular formulae annotation (such as the CAMERA package or the Progenesis CoMet software) are promising. Ideal software should automatically

analysis. This comprises molecular weight determination, molecular formulae assignment and metabolite identification, as shown in Figure VIII.1.

provide molecular formulae on all LC peaks with indices of confidence for the assignment.

3.2. Online metabolite identification

The second part of the procedure aims at providing a structure starting from a molecular formula. Several steps are necessary, including the determination of the possible structures for each molecular formula based on database searches such as the DNP, the reduction of the number of possible structures based on various filters (chemotaxonomic, retention-based and drift-time-based filters etc.), the confirmation by the UV spectrum, and the calculation of *in silico* MS/MS fragments to match with the experimental data, if available.

The main hurdle of this procedure is related to the type of the databases used and how the information is organised. In NP research, indeed, the chemotaxonomic cross-search is efficient only if the database used is up-to-date, complete and well-organised. Unfortunately, no such database exists today:

- NP-related databases are often updated years after the publication of the original article containing new information. This is often an issue since organisms often become interesting in a short period, as illustrated by the sudden interest of many scientists in *Hoodia gordonii* (Apocynaceae) in the beginning of the 2000's (this plant is traditionally used in South Africa for its appetite suppressant properties [6] and has known an intense media coverage that raised questions in the scientific community). This

issue could be partly solved by the building of a participative (and free) NP-dedicated database where scientists could easily upload new information.

- DB are often not complete, in particular because widely spread NPs such as flavonoids are not linked to all the organisms that are known to contain these molecules. For example, if quercetin is not linked to *G. biloba* in the database, the quercetin LC peak of the *G. biloba* profiling won't provide any result or maybe a false positive, even if quercetin is present in the studied sample and is very well-known. This is also related to the policy of scientific journals that encourage only new NPs to be published. All results of dereplication are however worthy to be reported, but probably in the form of an open database.
- Finally, the keywords and the organisation of the data have to be carefully chosen to provide reliable data for cross-searches with information obtained from metabolite profiling studies. For example, numerous natural organisms possess more than one

systematic and common name; these should be linked to avoid false negative results. Indeed, the database search for $C_{35}H_{52}O_4$ (molecular formula of hyperforin) with the keyword "St John's wort" could provide no result if this metabolite is linked with other names of the plant only, such as *Hypericum perforatum*, Tiptons weed, or rosin rose.

It is therefore necessary to develop comprehensive chemotaxonomic databases that should be integrated or fully compatible with the dereplication software, to get structures starting from the molecular formula.

In summary, online dereplication of NPs from complex samples using LC-MS becomes possible today, but efficient and automated tools as well as integrated and comprehensive databases are required for its routine use in an automated way with high confidence. Although it seems to be a 'black box', the ideal tool is a software that is integrated to the MS program and that provides molecular formulae directly in the chromatographic peaks, and then putative identification based on chemotaxonomic and other filters.

4. Non-universality of analytical techniques used in NP research

The lack of universal analysis techniques is another actual issue in NP research. Indeed, the main techniques of extraction, separation, and detection used in NP discovery are far from being generic.

During the extraction procedure, the range of extracted compounds is limited by the solvent used for the extraction. It is still possible to repeat the extraction procedure using another solvent to overcome this problem, but this costs time and money.

The separation step suffers the same limitations. Most of the UHPLC separations of NPs are performed on C₁₈ columns (Figure II.4). Because of this, the NPs that are not C₁₈-compatible, *i.e.* very polar, such as sugars, or apolar compounds, such as steroids, are not or are totally retained by the column. Moreover, the sample preparation usually aims at removing such compounds to increase column lifetime – and thus reduces the number of compounds in the sample and might be an issue in metabolomic studies. To overcome this problem, the LC separation may be repeated using another phase chemistry (C₄, cyano, HILIC, amide), but this is tedious and needs an adapted sample preparation.

The MS detection itself is not universal, and compounds are ionised depending on their

physicochemical properties and the ionisation source used. As an example, more than 1600 components were detected in a recent proteomic study of the venom of the marine snail *Conus consors* that was studied in Chapter III [7]. Because of the physicochemical diversity of the metabolites, two different techniques were used. MALDI-MS revealed a total of 889 components and ESI-MS 1078 components and only 339 components were detected by both techniques. Both approaches were found to be complementary and necessary to get the broadest possible proteome mapping of this natural extract.

In summary, there is nowadays no extraction procedure nor analysis technique providing a comprehensive overview of the composition of a given natural extract. This issue is partly solved by performing several analyses using orthogonal analytical techniques to combine their results, such as RP-LC, NP-LC, or ultra-high performance supercritical fluid chromatography (UHPSFC) [8], but increasing the number of analyses is tedious and time-consuming. The solution that is proposed to overcome this problem in human sample analysis is to perform analyses using RP and HILIC columns, and thus to combine the advantages of both orthogonal separations methods [9].

5. The future of the natural product research

It is difficult today to foresee if the pharmaceutical industry will increase their efforts in NP research again. Several experts think that this will happen since NPs represent a valuable source of diversified lead compounds that could compensate the decrease in new scaffolds (see Chapter I). I do agree but in my opinion, there are some conditions to this.

One of the main issues in NP research today is its low throughput. Indeed, NP discovery programs are not interesting for the pharmaceutical industry because of the low efficiency, the slowness and the high costs of the methods traditionally used. The future of NP research depends on the development of efficient methods for a rapid and rational evaluation of natural extract composition. In this respect, metabolomic approaches that are very efficient to highlight specific biomarkers and microfractionation techniques that allow fast isolation of low amounts of compounds are promising tools. The dereplication methods such as those described in this thesis, which improve the throughput of NP research by avoiding the isolation of known compounds, can be efficient. Still, there are some bottlenecks in NP analysis that were described in subchapter 4, in particular data processing. But the continuous development of new tools will undoubtedly allow the efficient online identification of metabolites in the near future.

If bio-guided activity studies have to be considerably accelerated by this means bioassays should also be adapted to HPLC profiling. Thus,

bioassays are of utmost importance in NP research, and have to be applicable to complex crude extracts as well as to few micrograms of pure NP [10]. Bioassays have thus to be sensitive, selective, fast, and, if possible, *in vivo*. The development of the *in vivo* assays based on zebrafish (see Introduction) represent a satisfying solution since they are adapted to crude extracts, microfractions and pure NPs [11, 12]. Furthermore such assays are generic and generate high content information that can then direct further target mechanistic studies on specific enzymes. Still, the low number of available tests based on zebrafish is still an important issue. The increasing number of online bioassays for HPLC profiling is also very promising [13].

Another condition is to increase the probability of finding new lead compounds in NPs. There are three main approaches to this. Firstly, the investigated organisms should be carefully selected. In this respect, the secular knowledge of traditional medicine has to be considered with much interest. Indeed, almost 75% of the currently used drugs originating from plants were previously used in traditional medicine [14]. Ethnopharmacology thus has an important role to play, providing some sort of holistic approach that may lead to empirical discovery of new bioactive compounds [15]. Secondly, some families of organisms that are rich in bioactive compounds should be investigated in the first line. Indeed, a recent study revealed that nature-related drugs have been extracted mostly from drug-productive families that tend to be

clustered rather than scattered in the phylogenetic tree. [16]. The search for bioactive compounds in genera or species close to bioactivity-rich ones should thus be more efficient. Finally, it is highly interesting to specifically target elicited instead of constitutive metabolites. Indeed, these molecules are often unknown and have specific bioactivities in relation to their role in defence. For this, metabolomics approaches that are based on stress biomarkers are well-adapted, and such approaches were already successfully used in our laboratory to highlight new antifungal compounds [17].

In my opinion, another (subjective) factor that explains the lack of interest in NPs is the old-fashioned image of NPs and of NP research, as opposed to brand new high-tech techniques. But on the contrary NPs should be seen as great opportunities to provide novelty thanks to their

physicochemical diversity. In addition, the number of NPs yet to be discovered is huge and represents an almost infinite reservoir of potential bioactive compounds. For example, the microbial and marine worlds remain largely unknown (see Table I.1) [18]. Moreover, the development of more sensitive techniques allows the detection of new bioactive metabolites present in very low concentrations in previously studied organisms. These compounds are potentially very interesting because this low concentration may be naturally compensated by a high bioactivity.

Based on these elements, one can reasonably think that NP research has a key role in modern drug discovery. Integration of state of the art analytical methods in this field will considerably increase the pace at which new bioactive NPs will be discovered and this promises a bright future for modern pharmacognosy and metabolomics.

References

- [1] Y.L. Wei, T. Lan, T. Tang, L.Y. Zhang, F.Y. Wang, T. Li, Y.P. Du, W.B. Zhang. A comprehensive two-dimensional normal-phase x reversed-phase liquid chromatography based on the modification of mobile phases. *Journal of Chromatography A*, **2009**. 1216: 7466-7471.
- [2] T. Kind, O. Fiehn. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews*, **2010**. 2: 23-60.
- [3] G.A. Theodoridis, H.G. Gika, E.J. Want, I.D. Wilson. Liquid chromatography–mass spectrometry based global metabolite profiling: A review. *Analytica Chimica Acta*, **2012**. 711: 7-16.
- [4] G. Glauser, N. Veyrat, B. Rochat, J.-L. Wolfender, T.C.J. Turlings. Ultra-high pressure liquid chromatography-mass spectrometry for plant metabolomics: a systematic comparison of high-resolution quadrupole-time-of-flight and single stage Orbitrap mass spectrometers. *Journal of Chromatography A*, **2013**. 1292: 151-159.
- [5] R.A. Zubarev, A. Makarov. Orbitrap Mass Spectrometry. *Analytical Chemistry*, **2013**. 85: 5288-5296.
- [6] P. Russell, C. Swindells. Chemical characterisation of *Hoodia gordonii* extract. *Food and Chemical Toxicology*, **2012**. 50: S6-S13.
- [7] D. Biass, S. Dutertre, A. Gerbault, J.L. Menou, R. Offord, P. Favreau, R. Stocklin. Comparative proteomic study of the venom of the piscivorous cone snail *Conus consors*. *Journal of Proteomics*, **2009**. 72: 210-218.
- [8] A. Grand-Guillaume Perrenoud, J.-L. Veuthey, D. Guillarme. Comparison of ultra-high performance supercritical fluid chromatography and ultra-high performance liquid chromatography for the analysis of pharmaceutical compounds. *Journal of Chromatography A*, **2012**. 1266: 158-167.
- [9] H.G. Gika, G.A. Theodoridis, I.D. Wilson. Hydrophilic interaction and reversed-phase ultra-performance liquid chromatography TOF-MS for metabolomic analysis of Zucker rat urine. *Journal of Separation Science*, **2008**. 31: 1598-1608.
- [10] F.E. Koehn, G.T. Carter. The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery*, **2005**. 4: 206-220.
- [11] L.I. Zon, R.T. Peterson. *In vivo* drug discovery in the zebrafish. *Nature Reviews Drug Discovery*, **2005**. 4: 35-44.
- [12] S. Challal, N. Bohni, O.E. Buenafe, C.V. Esguerra, P.A.M. de Witte, J.-L. Wolfender, A.D. Crawford. Zebrafish Bioassay-guided Microfractionation for the Rapid *in vivo* Identification of Pharmacologically Active Natural Products. *Chimia*, **2012**. 66: 229-232.
- [13] O. Potterat, M. Hamburger. Concepts and technologies for tracking bioactive compounds in natural product extracts: generation of libraries, and hyphenation of analytical processes with bioassays. *Natural Product Reports*, **2013**. 30: 546-564.
- [14] D.J. Newman, G.M. Cragg, K.M. Snader. The influence of natural products upon drug discovery. *Natural Product Reports*, **2000**. 17: 215-234.
- [15] R. Verpoorte, Y. Choi, H. Kim. Ethnopharmacology and systems biology: a perfect holistic match. *Journal of ethnopharmacology*, **2005**. 100: 53-56.
- [16] F. Zhu, C. Qin, L. Tao, X. Liu, Z. Shi, X. Ma, J. Jia, Y. Tan, C. Cui, J. Lin, C. Tan, Y. Jiang, Y. Chen. Clustered patterns of species origins of nature-derived drugs and clues for future bioprospecting. *Proceedings of the National Academy of Sciences*, **2011**. 108: 12943-12948.
- [17] S. Bertrand, O. Schumpp, N. Bohni, A. Bujard, A. Azzollini, M. Monod, K. Gindro, J.-L. Wolfender. Detection of metabolite induction in fungal co-cultures on solid media by high-throughput

- differential ultra-high pressure liquid chromatography–time-of-flight mass spectrometry fingerprinting. *Journal of Chromatography A*, **2013**. 1292: 219-228.
- [18] J. Berdy. Bioactive microbial metabolites. *The Journal of antibiotics*, **2005**. 58: 1-26.

Acknowledgements

Bien qu'une thèse de doctorat soit un projet individuel, il est clairement le fruit de collaborations internes et externes et nécessite l'aide et le soutien de nombreuses personnes. Je tiens donc à remercier ici toutes les personnes qui m'ont enseigné, aidé, guidé, accompagné ou soutenu dans l'aboutissement de ce projet.

Parce qu'un travail de doctorat ne peut se faire sans directeur de thèse, je tiens à remercier à ce titre, les Professeurs Jean-Luc Wolfender et Pierre-Alain Carrupt qui m'ont énormément apporté en m'acceptant comme candidat dans leurs équipes. Pierre-Alain, je te suis extrêmement reconnaissant pour m'avoir accordé ta confiance dès le début de ce projet. Travailler dans ton laboratoire est une chance, car tu as su y réunir un équipement hors-normes. De plus, l'aspect pluridisciplinaire y est très développé, ce qui est extrêmement positif et enrichissant dans la recherche du XXI^e siècle. Tu laisses à tes doctorants une grande liberté et leur accordes entière confiance. Je suis très heureux que notre projet de prédiction de rétention aboutisse sur un beau résultat, ce qui permet de clôturer de la plus belle des manières notre travail commun. Travail qui a donné lieu à d'intéressantes discussions, pour lesquelles je te remercie. Jean-Luc, je te suis reconnaissant de m'avoir aidé à grandir. Les quatre ans passés dans ton laboratoire m'ont donné l'occasion de travailler sur des sujets très divers, avec de grandes responsabilités. Cela m'a permis de m'épanouir en tant que scientifique et d'apprendre beaucoup de nouvelles choses. De plus, ton laboratoire offre une ambiance studieuse et socialement riche dans laquelle il est agréable de travailler et d'avancer. Je n'oublie pas non plus les nombreuses opportunités d'échanges, de collaborations externes et, cerise sur le gâteau, le *internship* au Brésil. Enfin, le grand nombre de publications, posters, communications orales et de chapitres de livres qui ont été produits durant ma thèse montrent bien la confiance que tu m'as accordée et la motivation que tu as su m'insuffler – tu es une vraie locomotive pour ton groupe.

I also would like to sincerely acknowledge Prof. Carlo Bicchi, Prof. Eric Alléman, and Dr Michael Affolter for accepting to join my Ph.D. committee. You contributed to valorise this work by your careful reading and the relevant questions that provided an interesting discussion during the defence.

Un travail de thèse interdisciplinaire et qui touche à plusieurs domaines tel que celui-ci nécessite des compétences très diverses. C'est donc très sincèrement que je voudrais remercier mes trois « meneurs d'allure » qui m'ont coaché au cours des différents projets et avec qui j'ai beaucoup appris : Sophie, Davy, et Julien B. Sophie, tu m'as aidé sans failles des premiers aux derniers jours de ma thèse. Ta grande curiosité scientifique et tes larges connaissances en chimie médicinale m'ont permis d'élargir mon domaine de compétence et ton soutien dans les moments difficiles m'a été d'une aide cruciale. Je n'oublierai pas non plus les innombrables cafés que nous avons « fumés » et tous les autres bons moments. Davy, ton expertise en LC m'a été d'une grande aide et j'ai énormément appris durant nos projets communs. Je m'estime sincèrement chanceux d'avoir pu travailler avec toi. Tu es de plus pour moi un exemple d'efficacité que je souhaite suivre. Julien, tu es une référence dans toutes ces sciences obscures au commun des mortels que sont les statistiques et analyses de données. Tu possèdes

également une grande capacité à expliquer efficacement ces concepts. Félicitations pour le calme olympien dont tu fais preuve lorsque tu partages tes connaissances. Sophie, Davy, Julien, vous avez en commun une grande qualité qui fait de vous d'excellents scientifiques académiciens : vous avez une grande motivation à enseigner et vous êtes d'une disponibilité et serviabilité incomparables. Merci.

Pour réussir un travail de thèse, il est aussi primordial de travailler dans les meilleures conditions possibles. Cela signifie d'abord que le laboratoire soit géré efficacement. Dans ce domaine, Martine, Fabrice (frit !), Christophe ; je tiens à vous remercier pour votre aide au quotidien. Soyons clair : vous portez vos laboratoires sur vos épaules. Il est également important de pouvoir compter sur des personnes compétentes qui nous aident à survivre aux nombreux tracasseries administratives qui jalonnent notre quotidien. Pour cela, je tiens à remercier chaleureusement Sylvia et Natalie pour leur soutien et efficacité sans faille dans ce domaine.

Le travail en laboratoire est une affaire d'équipe. J'ai ainsi eu l'occasion de travailler avec les membres de plusieurs laboratoires : au quotidien, le laboratoire de phytochimie, pharmacognosie, pharmacochimie ; ainsi que les laboratoires de chimie analytique pharmaceutique et spectrométrie de masse du vivant dans le cadre des TPs ou d'autres projets. Quelle richesse ! Merci à tous mes collègues de phyto, doctorants, post-doc, MER, professeurs, PAT, qui m'ont accompagné : Adeline, Adlin, Amina, Andreas, Antonio, Aziza, Caroline, Chantal, Charlotte, Claudia, Elia, Elisabeth, Emerson, Florence, Gaétan, Guillaume, Hakim, Jean-Luc, Karine, Laurence, Lise, Marcos, Mark, Martine, Mélanie, Muriel, Nadine, Natalie, Olivier, Philippe C, Quentin, Raimana, Samuel, Soura, Sylvain, Sylvian, Trixie, Vera, Vincent, Yildiz... ainsi que tous les stagiaires et étudiants ! Je tiens à décerner une mention spéciale pour quelques personnalités qui m'ont particulièrement touché. Guillaume, champion ! Nos discussions, courses à pied, et le travail avec toi étaient un réel plaisir. Je n'oublierai pas ton juron préféré bien appuyé ! Samuel, merci pour tous les débats scientifiques ou non, et tes bons conseils ... personne n'oubliera tes longs et fameux monologues avec ton ordinateur ! Nadine, petite sœur de thèse, merci pour tous les bons moments scientifiques, sportifs, et ton amitié... viel Glück und Erfolg ! Caro, mon poisson rouge préféré, tu as toujours le sourire et je voudrais te remercier pour les discussions scientifiques et autres que nous avons eues ! Flo, tu as réussi l'exploit d'être autant à l'aise aux fourneaux que devant ton ordinateur, bravo ! J'ai adoré ton amitié et ta spontanéité et je t'en remercie. Lise, ta gentillesse et ton écoute m'ont été très précieux. Je me souviendrai longtemps de nos belles discussions et bons moments passés ensemble. Chantal, ich möchte dich herzlichen danken für unsere freundlichen Diskussion auf Deutsch... Soura, nous avons dû batailler ferme ensemble pour maintenir notre Toffy. Je sais que je le laisse entre de bonnes mains ! Merci et bon courage... Vincent, bon à rien, j'ai beaucoup aimé nos entraînements de course à pieds et longues discussions en tout genre. Gaétan, tu m'as appris au début de ma thèse à dompter notre monstre UPLC-TOF... un tout grand merci. Philippe C, j'ai vraiment apprécié ta franchise et tout comme les discussions du café du matin ! Trixie, you took so much time to help me improving my bad English writing, I am very grateful and I want to thank you also for your nice friendship! Tiffany, Dr Porta dois-je dire, nous avons fini à 24h d'intervalle ... et je te suis très reconnaissant d'avoir pu partager avec toi nos doutes et expériences de fin de thèse, ainsi que nos exercices de défense, tu sais combien c'était précieux... Bonne chance pour la suite de ta carrière ! Il y a beaucoup de monde à remercier mais je ne peux pas citer tout le monde ici car il n'y a pas assez de place et je pourrais écrire une thèse à ce sujet ! Ma tête est pleine de beaux souvenirs et je voudrais souligner toute l'aide, le soutien et les bons moments passés en votre compagnie, que ce soit lors de discussions scientifiques, débats, réparations de machines, cafés, surveillances de TPs, afterwork drinks, congrès, séminaires à Zermatt, repas, réunions, préparations d'articles, entraînement à la soutenance

de thèse, course à pied, group meeting, discussions de couloir, etc. Vous m'avez tous beaucoup apporté et je ne l'oublierai jamais. Merci !

Bien que j'aie passé la majorité de mon temps dans le laboratoire de phytochimie, j'ai eu beaucoup de contacts avec les gens des laboratoires de pharmacochimie (par la codirection) et du LCAP et SMV (par le biais des TPs). Ces contacts enrichissants m'ont été très précieux et ont débouché autant sur des publications et des collaborations que sur de belles amitiés.

L'assistantat, qui se traduit majoritairement par la surveillance et l'enseignement des travaux pratiques, est à la fois une expérience enrichissante et astreignante. Heureusement, nous pouvions compter sur Jessica et Jean-Claude qui ont su nous guider et nous aider tout en gardant le sourire et avec une disponibilité exceptionnelle. J'ai eu beaucoup de plaisir à travailler avec vous !

I was also very lucky to work in the frame of several collaborative works with groups out of University of Geneva. Three main collaborations were part of this PhD work. There was firstly the project with Dr Reto Stöcklin and Daniel Biass from Atheris laboratories (Plan-les-Ouattes), on peptides-rich venoms, which was a great experience! Thanks to the interesting topic and the motivation of Atheris' people, which provided a nice publication. Secondly, the collaboration with Prof. Deniz Tasdemir and Dr Sinikka Rhate in London was a very interesting multidisciplinary project, focused on the quantitation of flavonoids, which also provided a publication. Then, there was the trans-oceanic collaboration with Prof. Dulce Helena S. Silva and Dr Cristiano S. Funari from Araraquara, Brazil, focusing on dereplication that was a big part of my thesis. I would also like to mention the internship in Salvador (Brazil), in the laboratory of Dr Milena Soares, that was a fantastic experience. Another collaboration began as I was finishing this work, with Dr Richard Knochenmuss from TofWerk (Thun) that provided the data shown in the small chapter of the thesis on ion mobility separations. All these collaborations were rich experiences for me and represented a great opportunity to broaden my skills and discover new work environments. I want to warmly thank all these people in Geneva, Thun, London and Brazil for these unique experiences, and Jean-Luc for giving me these chances. Finally, I mainly worked on Waters MS instruments during these five years, and I want to thank all the people from this company that I met: Julien, Bernhard, Pierre-Alain, Stéphane C., Stéphane B., Fabienne, etc. for their help, advices and teaching.

Un dernier groupe de personnes m'a été très précieux durant ces cinq années : je pense bien sûr au groupe de course à pied ! Raimana, Claudia, Jean-Philippe F, Roger, Laurène, Jean-Philippe G, Corentin, Guillaume, Aaron, Nicole, Tiphaine, Renzo, Tobias, Sandra, Laurent, Nadine, Caroline, Vincent, Julien, Antonio, et tous les autres ! Merci de m'avoir aidé à décompresser, à suer sur une piste, et d'avoir partagé des courses (Escalade, Duc, marathon de Genève, 100 km de Bienne, etc.). J'ai également apprécié cette opportunité de rencontrer des personnes d'autres facultés, et de fonctions diverses, étudiants, profs, PAT, etc. Je n'oublierai pas ces bons moments !

Finalement, l'entourage qui joue un rôle prépondérant dans une thèse ne se limite pas au milieu académique, bien au contraire. Car les amis sont très précieux et aident à surmonter les moments difficiles, ce que j'ai spécialement ressenti dans les dernières semaines avant la soutenance. Je ne vais pas vous citer tous ici, mais je voudrais remercier sincèrement tous ceux qui m'ont soutenu d'une manière ou d'une autre – ou qui ont su faire preuve de compréhension durant les moments les plus difficiles.

Je garde bien sûr le meilleur pour la fin : ma famille. Je voudrais vous remercier du fond du cœur pour votre présence durant la soutenance et pour votre aide durant ces cinq années. Ma sœur et mes frères Myriam, Olivier, Patrick, vous êtes pour moi très précieux et je vous remercie pour votre soutien et votre compréhension – particulièrement Olivier qui est déjà passé par une thèse ! Ma marraine Catherine également, ta générosité et ta bonne humeur me touchent, tu es un exemple pour moi. Mes parents finalement, parce que sans vous et sans votre éducation je ne serais pas en train d'écrire ces mots. Merci papa pour tout ce que tu as fait pour moi, pour m'avoir soutenu, donné l'envie d'apprendre, la curiosité et les moyens de réaliser ce parcours dont je suis si heureux. C'est le plus beau des cadeaux. Je vous aime tous très fort !

Appendices

A selection of relevant articles and posters prepared in the frame of this thesis work is presented in the Appendices below.

Annexe I	Scientific communications and collaborations
Annexe II	C.V.
Annexe III	Review article: Ultra high pressure liquid chromatography for crude plant extract profiling
Annexe IV	Article: Detection by UPLC/ESI-TOF-MS of alkaloids in three Lycopodiaceae species from French Polynesia and their anticholinesterase activity.
Annexe V	Article: Advanced methods for natural product drug discovery in the field of nutraceuticals.
Annexe VI	Article: <i>Salvia officinalis</i> for hot flushes: towards determination of mechanism of activity and active principles
Annexe VII	Poster: Rapid log <i>P</i> determination of natural products in crude plant extracts from UHPLC-TOF-MS profiling data - an additional parameter for dereplication and bioavailability.
Annexe VIII	Poster: Log <i>P</i> determination by UHPLC-TOF-MS in natural product analysis: issues and perspectives.
Annexe IX	Poster: Combination of LC retention, high resolution TOF-MS information and web database search as dereplication tools in a chemotaxonomic study of <i>Lippia</i> spp.
Annexe X	Poster: Optimization of <i>Conus consors</i> venom profiling using Ultra-High Pressure Liquid Chromatography (UHPLC)
Annexe XI	Poster: Ion mobility spectrometry in metabolite profiling of complex plant extracts.

Appendix I

Scientific communications and collaborations

This thesis work produced several scientific communications: journal articles, book chapters, oral communications, and posters, listed below. Most of them are reproduced in this thesis.

Most of these communications were prepared by several co-authors from different laboratories, often from different countries. Indeed, today's scientific research is made of collaborations, to share knowledge and get the best of the experience, expertise and instruments of each of the collaborators. Moreover, the majority of the projects are interdisciplinary works, particularly in the pharmaceutical research. It is impossible nowadays to work alone. This thesis, with two co-directors and several collaborations listed below is a good example.

Journal articles

Ho R, Marsousi N, Eugster P, Bianchini JP, Raharivelomanana P, Detection by UPLC/ESI-TOF-MS of Alkaloids in Three Lycopodiaceae Species from French Polynesia and Their Anticholinesterase Activity. *Natural Product Communications* 2009, 4, (10), 1349-1352.

Wolfender JL, Eugster PJ, Bohni N, Cuendet M. Advanced Methods for Natural Product Drug Discovery in the Field of Nutraceuticals. *Chimia* 2011, 65, (6), 400-406.

Eugster PJ, Guillarme D, Rudaz S, Veuthey JL, Carrupt PA, Wolfender J L. Ultra High Pressure Liquid Chromatography for Crude Plant Extract Profiling. *Journal of AOAC International* 2011, 94, (1), 51-70.

Funari CS*, Eugster PJ*, Martel S, Carrupt PA, Wolfender JL, Silva DHS. High resolution ultra high pressure liquid chromatography–time-of-flight mass spectrometry dereplication strategy for the metabolite profiling of Brazilian *Lippia* species. *Journal of Chromatography A* 2012, 1259, 167-178. (* these authors contributed equally to this work).

Eugster PJ, Biass D, Guillarme D, Favreau P, Stöcklin R, Wolfender JL. Peak capacity optimisation for high resolution peptide profiling in complex mixtures by liquid chromatography coupled to time-of-flight mass spectrometry: Application to the *Conus consors* cone snail venom. *Journal of Chromatography A* 2012, 1259, 187-199.

Rahte S, Evans R, Eugster PJ, Marcourt L, Wolfender JL, Kortenkamp A, Tasdemir D. *Salvia officinalis* for Hot Flushes: Towards Determination of Mechanism of Activity and Active Principles. *Planta Medica* 2013, 79, 753-760.

Spaggiari D, Fekete S, Eugster PJ, Veuthey JL, Geiser L, Rudaz S, Guillarme D. Contribution of various types of liquid chromatography-mass spectrometry instruments to band broadening in fast analysis. *Journal of Chromatography A* (submitted).

Eugster PJ, Boccard J, Debrus B, Bréant L, Wolfender JL, Martel S, Carrupt PA. Retention time prediction for dereplication of natural products (C_xH_yO_z) in LC-MS metabolite profiling. *Analytical Chemistry* (submitted).

Book chapters

Eugster PJ, Wolfender JL. UHPLC in Natural Products Analysis. In *UHPLC in Life Sciences*. Editors: Guillarme D, Veuthey JL, RSC Publishing, 2012, p 354.

Eugster PJ, Glauser G, Wolfender JL. Strategies in Biomarker Discovery. Peak annotation by MS and targeted LC-MS micro-fractionation for *de novo* structure identification by micro-NMR. In *Methods in Molecular Biology, Metabolomics Tools for Natural Product Discoveries*. Editors: Roessner U, Dias DA, Humana Press (In press).

Oral communications

Eugster PJ, Guillarme D, Kratou H, Glauser G, Martel S, Rudaz S, Carrupt PA, Wolfender JL. Potential of UHPLC for plant analysis: profiling, dereplication and metabolomics. Oral presentation at the 7th International symposium on chromatography of natural products joined with 6th International symposium of the International Society for the Development of Natural Products (ISCNP&ISDNP 2010), in Lublin, Poland, the 15th of June 2010.

Eugster PJ. Maximising resolution for UHPLC-TOF-MS metabolite profiling of complex natural samples – application to small and large molecules. Oral presentation at the 7th PhD Day of the School of Pharmacy Geneva – Lausanne (EPGL), in Hermance, Switzerland, the 15th of June 2012.

Eugster PJ, Biass D, Guillarme D, Favreau P, Stöcklin R, Wolfender JL. Maximising resolution for UHPLC-TOF-MS metabolite profiling of complex natural samples – application to small and large molecules. Oral presentation at the Fall Meeting of the Swiss Chemical Society (SCS) in Zurich, Switzerland, the 13th of September 2012.

Eugster PJ, Martel S, Carrupt PA, Wolfender JL. Prédiction de la rétention en UHPLC : une dimension supplémentaire en LC-MS pour l'identification rapide de composés bioactifs ou de biomarqueurs dans des extraits naturels. Oral presentation at the Journées scientifiques du médicament in Grenoble, France, the 11th of June 2013.

Posters (only posters presented as first author are listed below)

Eugster P, Martel S, Guillarme D, Carrupt PA, Wolfender JL, Rapid log P determination of natural products in crude plant extracts from UHPLC-TOF-MS profiling data - an additional parameter for dereplication and bioavailability. Abstract in *Planta Medica* 2009, 75, (9), 913-914. Poster presented at the 57th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Products Research (GA), in Geneva, Switzerland, the 18th of August 2009.

Eugster P, Martel S, Guillarme D, Wolfender JL, Carrupt PA, Rapid log P Determination of Natural Products in Crude Plant Extracts from UHPLC-TOF-MS Profiling Data - An Additional Parameter for Dereplication and Bioavailability. Poster presented at the Swiss Pharma Science Day (SPhSD), in Bern, Switzerland, the 2nd of September 2009.

Eugster PJ, Martel S, Guillarme D, Wolfender JL, Carrupt PA. Log P determination by UHPLC-TOF-MS in natural product analysis: issues and perspectives. Poster presented at the Fall Meeting of the Swiss Chemical Society (SCS) in Zurich, Switzerland, the 16th of September 2010.

Eugster PJ, Funari C, Mattioli F, Durigan G, Martel S, Carrupt P, Silva D, Wolfender JL, Combination of LC retention, high resolution TOF-MS information and web database search as dereplication tools in a chemotaxonomic study of *Lippia* spp. Abstract in *Planta Med* 2011, 77, (12), PA49. Poster presented at the 59th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Products Research (GA), in Antalya, Turkey, the 4th of September 2011.

Eugster PJ, Biass D, Guillarme D, Favreau P, Stöcklin R, Wolfender JL. Optimization of *Conus consors* venom profiling using Ultra-High Pressure Liquid Chromatography (UHPLC). Poster presented at the Fall Meeting of the Swiss Chemical Society (SCS) in Lausanne, Switzerland, the 9th of September 2011.

Eugster PJ, Knochenmuss R, Wolfender JL. Ion mobility spectrometry applied to complex plant extract profiling: possibilities, limitations and outlook. Abstract in *Planta Med* 2012, 78, PJ46. Poster presented at the 60th International Congress (Joint Meeting with ASP, AFERP, PSE and SIF) and Annual Meeting of GA, 60th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Products Research (GA), in New York, USA, July 28 - Aug 1, 2012.

Eugster PJ, Knochenmuss R, Wolfender JL. Ion mobility spectrometry in metabolite profiling of complex plant extracts. Abstract in *Chimia* 7-8/2012, Vol. 66. Poster presented at the Fall Meeting of the Swiss Chemical Society (SCS) in Zurich, Switzerland, the 13th of September 2012.

Collaborations

Collaboration with Dr Reto Söcklin and Daniel Biass from Atheris laboratories, Geneva, Switzerland, focusing on the UHPLC analysis of venom samples. An article was published in *Journal of Chromatography A*.

Collaboration with a Lavinia Alexandru from the University of Udine, Italy on the screening of plants from Friuli. This work was part of her PhD thesis.

Collaboration with Dr Dulce Helena Siqueira Silva and Dr Cristiano Soleo Funari from the São Paulo State University, Brazil, focusing on the chemotaxonomy of the *Lippia* genus. An article was published in Journal of Chromatography A.

Collaboration with Dr Richard Knochenmuss from the ToFwerk Company, Thun, Switzerland, aiming at evaluating the use of ion mobility spectrometry for the metabolite profiling of complex plant extracts. A poster was presented at the Fall Meeting of the Swiss Chemical Society in Zurich in 2012.

Collaboration with Dr Deniz Tasdemir and Dr Sinikka Rahte from the University of London, UK, to perform UHPLC-MS experiments on bioactive microfractions from *Salvia officinalis*. An article was published in Planta Medica.

Finally, it is worth mentioning a 4 weeks stay in the team of Dr Milena B.P. Soares, in the Fiocruz center in Salvador, Bahia, Brazil, aiming at setting up and testing a nano-UHPLC-QTOF-MS instrument and teaching students and collaborators.

Appendix II

C.V.

Philippe J. Eugster, born August 11, 1980, in Vevey, Switzerland.

Education

- 2008 – 2013 PhD student, in the Phytochemistry and Bioactive Natural Products Group with Prof. J.L. Wolfender and in the Pharmacochimie Group with Prof. P.-A. Carrupt, School of Pharmaceutical Sciences, University of Geneva, Switzerland.
- 2006 Master in Pharmaceutical Sciences, University of Geneva, Switzerland.
- 2004 Bachelor in Pharmaceutical Sciences, University of Lausanne, Switzerland.
- 1999 Matura AX in Science, Latin and Greek, Gymnase du Bugnon, Lausanne, Switzerland.

Teaching and work experience

- 2008 – 2013 Research assistant in the Phytochemistry and Bioactive Natural Products Group, University of Geneva, Switzerland.
- 2009 – 2013 Teaching assistant for practical work in Phytochemistry for 2nd and 3rd year Bachelor and 1st year Master students, School of Pharmaceutical Sciences, University of Geneva, Switzerland.
- 2012 1-month scientific visitor for nanoUPLC-QTOF-MS setup, protein and small molecules analysis and teaching, Fiocruz, Salvador, Bahia, Brazil.
- 2006 – 2011 Pharmacist, Pharmacie de Copet, Vevey, Switzerland.
- 2008 – 2009 Teaching assistant for practical work in Pharmacochimie for 3rd year Bachelor students, School of Pharmaceutical Sciences, University of Geneva, Switzerland.
- 2005 – 2006 Pharmacist assistant, Pharmacie Saint-Martin, Vevey, Switzerland.

Appendix III

Ultra high pressure liquid chromatography for crude plant extract profiling.

Philippe J. Eugster

Davy Guillarme

Serge Rudaz

Jean-Luc Veuthey

Pierre-Alain Carrupt

Jean-Luc Wolfender

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Review article published in Journal of AOAC International 2011, 94, (1), 51-70.

Appendix IV

Detection by UPLC/ESI-TOF-MS of alkaloids in three Lycopodiaceae species from French Polynesia and their anticholinesterase activity.

Raimana Ho ¹

Niloufar Marsousi ¹

Philippe J. Eugster ¹

Jean-Pierre Bianchini ²

Phila Raharivelomanana ²

¹ School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

² Biodiversité Terrestre et Marine, Université de la Polynésie Française, French Polynesia.

Research article published in Natural Product Communications 2009, 4, (10), 1349-1352.

Appendix V

Advanced methods for natural product drug discovery in the field of nutraceuticals.

Jean-Luc Wolfender

Philippe J. Eugster

Nadine Bohni

Muriel Cuendet

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Research article published in *Chimia* 2011, 65, (6), 400-406.

Appendix VI

Salvia officinalis for hot flushes: towards determination of mechanism of activity and active principles

Sinikka Rahte¹

Richard Evans¹

Philippe J. Eugster²

Laurence Marcourt²

Jean-Luc Wolfender²

Andreas Kortenkamp¹

Deniz Tasdemir¹

¹ School of Pharmacy, University of London, UK.

² School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Research article published in *Planta Medica* 2013 (ahead of print).

Appendix VII

Rapid log *P* determination of natural products in crude plant extracts from UHPLC-TOF-MS profiling data - an additional parameter for dereplication and bioavailability.

Philippe J. Eugster

Sophie Martel

Davy Guillarme

Jean-Luc Wolfender

Pierre-Alain Carrupt

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Poster presented at the Swiss Pharma Science Day (SPhSD), in Bern, Switzerland, the 2nd of September 2009.

Rapid log P Determination of Natural Products in Crude Plant Extracts from UHPLC-TOF-MS Profiling Data

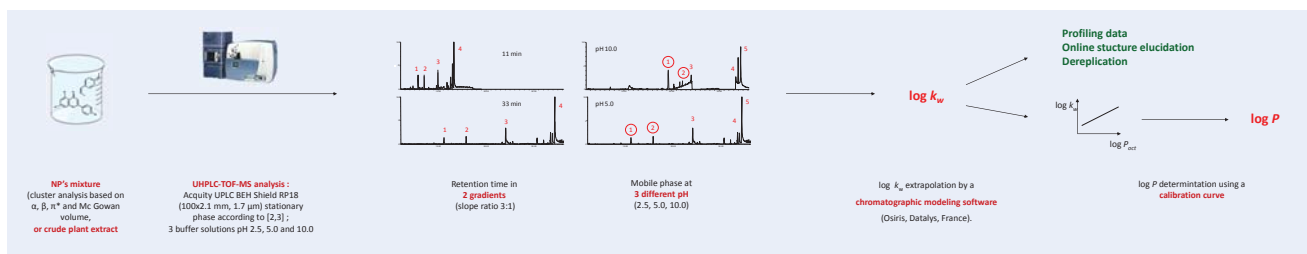
Philippe Eugster, Sophie Martel, Davy Guillaume, Jean-Luc Wolfender, Pierre-Alain Carrupt

School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, Quai E-Ansermet 30, CH-1211 Geneva 4, Switzerland

Introduction

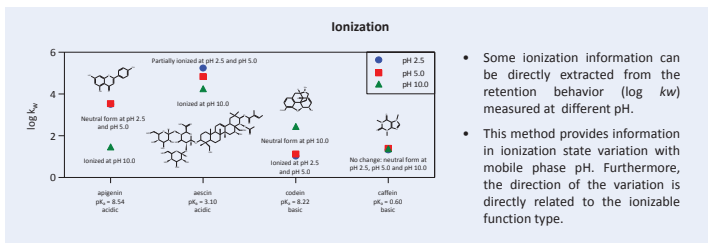
- In phytochemical analysis, HPLC metabolite profiling methods provide a large amount of data on the composition of a given crude plant extracts for both dereplication or rapid on-line structure determination of given natural products (NP's) [1].
- Many physicochemical properties could be extracted from HPLC data, such as lipophilicity.
- Lipophilicity (described by log P) is a key-parameter involved in pharmacokinetic (absorption, distribution, metabolism, elimination and toxicity) and pharmacodynamic processes (ligand-target interactions) and has to be determined as early as possible.
- Liquid chromatography is a fast and low sample consuming technique fully used in log P determination. The method is based on the relationship existing between retention factors and log P using specific chromatographic conditions [2].
- The development of column packed with sub-2µm particles working at high pressure (Ultra High Pressure Liquid Chromatography) also allows higher throughput.
- NP's retention factors extracted from UHPLC-MS metabolite profiling data could then provide log P of compounds of interest prior isolation.

Materials and Methods

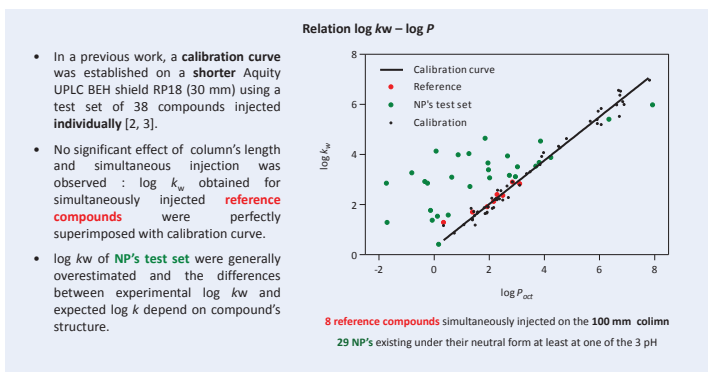


α , β = H-bond donor / acceptor properties ; n^* = polarizability.

Results and discussion



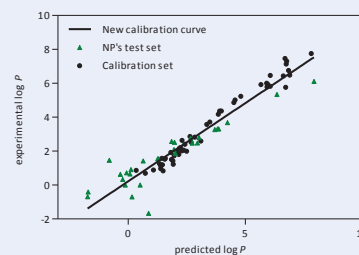
- Some ionization information can be directly extracted from the retention behavior (log kw) measured at different pH.
- This method provides information in ionization state variation with mobile phase pH. Furthermore, the direction of the variation is directly related to the ionizable function type.



- In a previous work, a calibration curve was established on a shorter Acquity UPLC BEH shield RP18 (30 mm) using a test set of 38 compounds injected individually [2, 3].
- No significant effect of column's length and simultaneous injection was observed : log kw obtained for simultaneously injected reference compounds were perfectly superimposed with calibration curve.
- log kw of NP's test set were generally overestimated and the differences between experimental log kw and expected log k depend on compound's structure.

QSPR analysis

- Preliminary graphical analyses demonstrated that the deviation of lipophilicity prediction was not linearly related with a single structural parameter such as molecular volume (or exact mass), polarizability (n^*), the H-bond donor capacity (α) or the H-bond acceptor capacity (β).
- Chromatographic behavior of complex natural compounds in our conditions can be attributed to a complex influence of several inter-molecular interactions between the natural solutes and the chromatographic system.
- Multilinear analyses (MLR) have to be used to identify the combination of structural parameters responsible of the peculiar behavior of NP's. However, the very high correlation (80 – 94 %) between the four solvatochromic parameters for the studied compounds forbid the simultaneous usage of these parameters in a single MLR equation.
- A principal component analysis confirmed these high correlations since a single component PC1 described 88 % of the solvatochromic variation in the chemical space of the 81 reference and natural compounds explored. PC1, the linear combination of molecular mass (24.9 %), α (23.1 %), β (26.2 %) and n^* (25.8 %), can thus be used in a regression analysis.
- Equation demonstrates that PC1 allows to predict log P from the UHPLC measurements for the reference and the natural compounds.



$$\log P = 1.26 (\pm 0.07) \log k_w - 1.08 (\pm 0.37) \text{PC1} - 1.31 (\pm 0.04)$$

$n=81; r^2=0.92; q^2=0.90; F=470$

Conclusions & Perspectives

- The method previously described for small and simple reference compounds was not directly applicable to the log P determination of more complex natural compounds.
- The deviation between measured and expected log kw can be explained by intermolecular interactions. Therefore the model has been adapted and a new equation has been proposed including a linear combination of structural parameters.
- 3 outliers have been identified ; the reason of their deviation has to be investigated.
- A rapid NP's structural parameters determination has to be developed in order to directly apply the new method on non isolated compounds with unknown structure.
- Therefore the new model will be applied to complex matrices such as crude plant extracts and the determined physicochemical properties would be of great value in the identification of new NPs of interest.

References

[1] J.-L. Wolfender, E. F. Queiroz, K. Hostettmann. Expert Opinion on Drug Discovery 1, 237-260 (2006). [2] Y. Henchoz, D. Guillaume, S. Martel, S. Rudaz, J.-L. Veuthey, P.-A. Carrupt. Analytical and Bioanalytical Chemistry, 394 (7), 1919- (2009). [3] A. Guillot, Y. Henchoz, C. Moccand, D. Guillaume, J.-L. Veuthey, P.-A. Carrupt, S. Martel. Chemistry & Biodiversity. In Press. (2009). [4] E. Grata, J. Boccard, D. Guillaume, G. Glauser, P.-A. Carrupt, E. E. Farmer, J.-L. Wolfender, S. Rudaz. Journal of Chromatography B, 871, 261-270 (2008).

Appendix VIII

Log *P* determination by UHPLC-TOF-MS in natural product analysis: issues and perspectives.

Philippe J. Eugster

Sophie Martel

Davy Guillarme

Jean-Luc Wolfender

Pierre-Alain Carrupt

School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

Poster presented at the Fall Meeting of the Swiss Chemical Society (SCS) in Zurich, Switzerland, the 16th of September 2010.

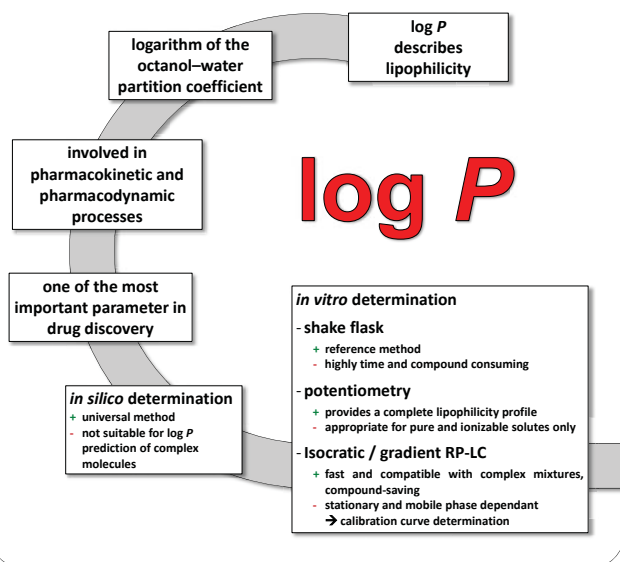
log P determination by UHPLC-TOF-MS in natural product analysis : issues and perspectives

Philippe Eugster, Davy Guillaume, Pierre-Alain Carrupt, Jean-Luc Wolfender, Sophie Martel

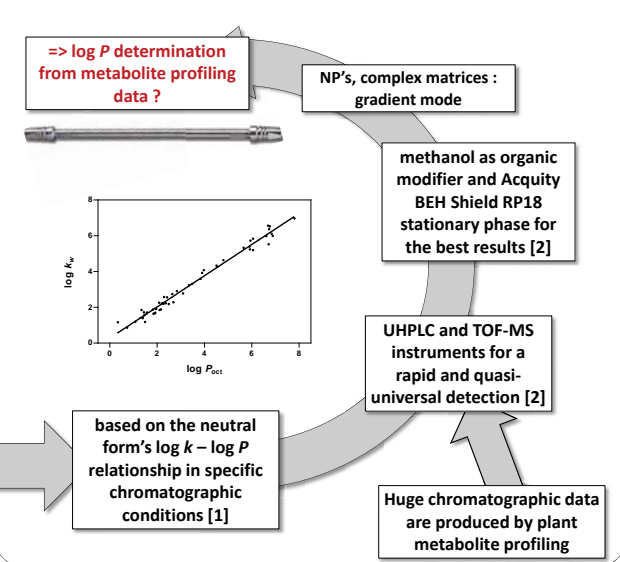
School of Pharmaceutical Sciences, University of Geneva, University of Lausanne, Quai E-Ansermet 30, CH-1211 Geneva 4, Switzerland

philippe.eugster@unige.ch

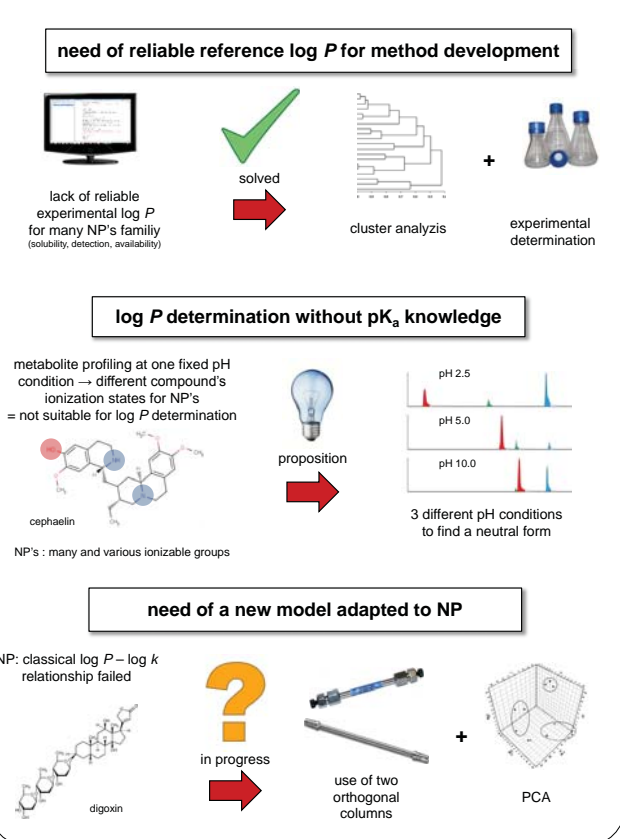
1. Introduction to log P



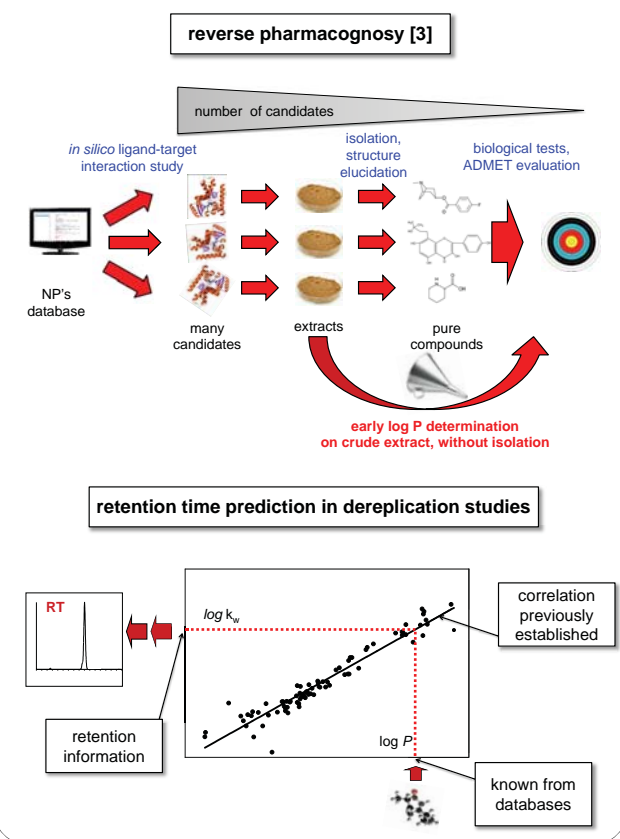
2. log P determination and natural products (NP)



3. Challenges ...



4. ...and applications



Appendix IX

Combination of LC retention, high resolution TOF-MS information and web database search as dereplication tools in a chemotaxonomic study of *Lippia* spp.

Philippe J. Eugster¹

Cristiano S. Funari²

F.M. Mattioli²

Giselda Durigan³

Sophie Martel¹

Pierre-Alain Carrupt¹

Dulce Helena Siqueira Silva²

Jean-Luc Wolfender¹

¹ School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

² NuBBE – Nucleo de Bioensaios, Biossintese e Ecofisiologia de Produtos Naturais, São Paulo State University, Araraquara, Brazil

³ Instituto Florestal, Assis, Brazil.

Poster presented at the 59th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Products Research (GA), in Antalya, Turkey, the 4th of September 2011.

Eugster PJ¹, Funari CS², Mattioli FM², Durigan G³, Martel S¹, Carrupt PA¹, Silva DHS², Wolfender J-L¹

¹ School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, quai E-Ansermet 30, CH-1211 Geneva 4, Switzerland ; ² NuBBE – Núcleo de Bioensaios, Biossintese e Ecofisiologia de Produtos Naturais, Institute of Chemistry, São Paulo State University, Araraquara, SP, CP 355, CEP 14801-970, Brazil ; ³ Instituto Florestal, Floresta Estadual de Assis, CP 104, CEP 19802-970 Assis, SP, Brazil.

Introduction

Dereplication of natural products (NPs) in crude plant extracts represents a key process to rationalize bioactivity guided isolation procedures. [1]


In order to evaluate how far NP annotation can be made from a single LC-MS profiling using high resolution (HR) in both LC and MS dimensions, online molecular formula assignment and LC-retention-based methods were evaluated in the frame of a chemotaxonomic study on various *Lippia* species from Brazil.

Aim of the work

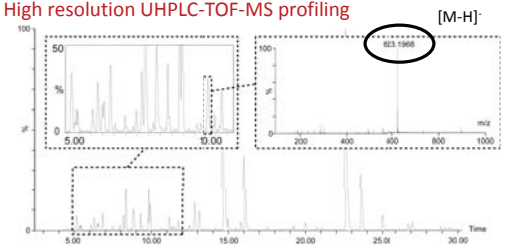
Lippia genus is widely used in ethnobotany as food, medicines, sweetener and beverages flavouring, but relatively few phytochemical investigations have been reported. Furthermore, various taxonomic problems involving some genera from Verbenaceae, including *Lippia*, have been highlighted.

The aim of the work was to compare metabolic fingerprints of fifteen extracts of six *Lippia* species by UHPLC-PDA-ESI-TOF-MS using a comprehensive dereplication strategy. This has involved the rapid isolation of the main NPs, the on line peak annotation of most of the minor NPs in relation with the isolated compounds and bibliographic and chemotaxonomic information. Altogether, this enabled the peak labelling of more than 40 NPs.

We focus here only on the on-line dereplication process.



High resolution UHPLC-TOF-MS profiling



2. Fiehn's 7 Golden Rules [2]

- Chemical rules (e.g. max of C for a given mass)
- Lewis and Senior rules (eg « octet rule »)
- Isotopic abundance pattern
- H / C ratio
- Heteroatom / C ratio
- N+O+P+S sum
- TMS subtraction (GC-MS)

4. Chemotaxonomy

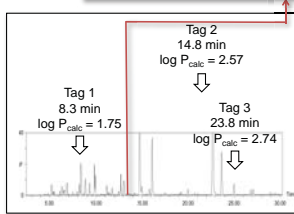
Assuming that the studied compound is already reported in literature, a search in NP's databases for the remaining formula(e) in the studied plants restricts a few candidate compounds.

The higher taxonomic (genus, family) level taken in account for the search has to be chosen with care.

6. log P filter

This method is based on the well-known log P - chromatographic retention [3].

- Neutral form of the analytes is mandatory (here, polyphenols in acidic conditions).
- Method reliability depends on the log P value. Here, log P values are always obtained from the same software in order to eliminate calculation errors.
- Good relative retention information is obtained for homologous compounds.



raw data

623.1968 [M-H]

1. 50 formulae

2. 11 formulae

3. 5 formulae

4. 2 compounds

5. 2 compounds

6. 2 compounds

7. 1 compound

MW calculator

web DB

UV

retention

1. Molecular weight information

Based on the high mass resolution from TOF-MS analyzer (< 5ppm), calculation of all formulae corresponding to 623.1968 m/z is carried on, with a 15 ppm tolerance, and C, H, O, N atoms.

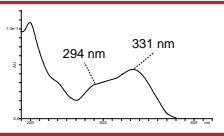
Formula	Mass	Delta	Formula	Mass	Delta
C ₂₅ H ₂₂ O ₁₅	623.1968	0.0000	C ₂₅ H ₂₂ O ₁₅	623.1968	0.0000
C ₂₅ H ₂₄ O ₁₅	623.2018	0.0050	C ₂₅ H ₂₄ O ₁₅	623.2018	0.0050
C ₂₅ H ₂₀ O ₁₅	623.1918	-0.0050	C ₂₅ H ₂₀ O ₁₅	623.1918	-0.0050
C ₂₅ H ₂₂ O ₁₄	623.1902	-0.0066	C ₂₅ H ₂₂ O ₁₄	623.1902	-0.0066
C ₂₅ H ₂₄ O ₁₄	623.1952	-0.0016	C ₂₅ H ₂₄ O ₁₄	623.1952	-0.0016
C ₂₅ H ₂₀ O ₁₄	623.1852	-0.0106	C ₂₅ H ₂₀ O ₁₄	623.1852	-0.0106
C ₂₅ H ₂₂ O ₁₃	623.1886	-0.0082	C ₂₅ H ₂₂ O ₁₃	623.1886	-0.0082
C ₂₅ H ₂₄ O ₁₃	623.1936	-0.0032	C ₂₅ H ₂₄ O ₁₃	623.1936	-0.0032
C ₂₅ H ₂₀ O ₁₃	623.1836	-0.0132	C ₂₅ H ₂₀ O ₁₃	623.1836	-0.0132
C ₂₅ H ₂₂ O ₁₂	623.1870	-0.0098	C ₂₅ H ₂₂ O ₁₂	623.1870	-0.0098
C ₂₅ H ₂₄ O ₁₂	623.1920	-0.0048	C ₂₅ H ₂₄ O ₁₂	623.1920	-0.0048
C ₂₅ H ₂₀ O ₁₂	623.1820	-0.0148	C ₂₅ H ₂₀ O ₁₂	623.1820	-0.0148
C ₂₅ H ₂₂ O ₁₁	623.1854	-0.0114	C ₂₅ H ₂₂ O ₁₁	623.1854	-0.0114
C ₂₅ H ₂₄ O ₁₁	623.1904	-0.0064	C ₂₅ H ₂₄ O ₁₁	623.1904	-0.0064
C ₂₅ H ₂₀ O ₁₁	623.1804	-0.0164	C ₂₅ H ₂₀ O ₁₁	623.1804	-0.0164
C ₂₅ H ₂₂ O ₁₀	623.1838	-0.0130	C ₂₅ H ₂₂ O ₁₀	623.1838	-0.0130
C ₂₅ H ₂₄ O ₁₀	623.1888	-0.0080	C ₂₅ H ₂₄ O ₁₀	623.1888	-0.0080
C ₂₅ H ₂₀ O ₁₀	623.1788	-0.0180	C ₂₅ H ₂₀ O ₁₀	623.1788	-0.0180
C ₂₅ H ₂₂ O ₉	623.1822	-0.0146	C ₂₅ H ₂₂ O ₉	623.1822	-0.0146
C ₂₅ H ₂₄ O ₉	623.1872	-0.0096	C ₂₅ H ₂₄ O ₉	623.1872	-0.0096
C ₂₅ H ₂₀ O ₉	623.1772	-0.0196	C ₂₅ H ₂₀ O ₉	623.1772	-0.0196
C ₂₅ H ₂₂ O ₈	623.1806	-0.0162	C ₂₅ H ₂₂ O ₈	623.1806	-0.0162
C ₂₅ H ₂₄ O ₈	623.1856	-0.0112	C ₂₅ H ₂₄ O ₈	623.1856	-0.0112
C ₂₅ H ₂₀ O ₈	623.1756	-0.0212	C ₂₅ H ₂₀ O ₈	623.1756	-0.0212
C ₂₅ H ₂₂ O ₇	623.1790	-0.0178	C ₂₅ H ₂₂ O ₇	623.1790	-0.0178
C ₂₅ H ₂₄ O ₇	623.1840	-0.0128	C ₂₅ H ₂₄ O ₇	623.1840	-0.0128
C ₂₅ H ₂₀ O ₇	623.1740	-0.0228	C ₂₅ H ₂₀ O ₇	623.1740	-0.0228
C ₂₅ H ₂₂ O ₆	623.1774	-0.0194	C ₂₅ H ₂₂ O ₆	623.1774	-0.0194
C ₂₅ H ₂₄ O ₆	623.1824	-0.0144	C ₂₅ H ₂₄ O ₆	623.1824	-0.0144
C ₂₅ H ₂₀ O ₆	623.1724	-0.0244	C ₂₅ H ₂₀ O ₆	623.1724	-0.0244
C ₂₅ H ₂₂ O ₅	623.1758	-0.0210	C ₂₅ H ₂₂ O ₅	623.1758	-0.0210
C ₂₅ H ₂₄ O ₅	623.1808	-0.0160	C ₂₅ H ₂₄ O ₅	623.1808	-0.0160
C ₂₅ H ₂₀ O ₅	623.1708	-0.0260	C ₂₅ H ₂₀ O ₅	623.1708	-0.0260
C ₂₅ H ₂₂ O ₄	623.1742	-0.0226	C ₂₅ H ₂₂ O ₄	623.1742	-0.0226
C ₂₅ H ₂₄ O ₄	623.1792	-0.0176	C ₂₅ H ₂₄ O ₄	623.1792	-0.0176
C ₂₅ H ₂₀ O ₄	623.1692	-0.0276	C ₂₅ H ₂₀ O ₄	623.1692	-0.0276
C ₂₅ H ₂₂ O ₃	623.1726	-0.0242	C ₂₅ H ₂₂ O ₃	623.1726	-0.0242
C ₂₅ H ₂₄ O ₃	623.1776	-0.0192	C ₂₅ H ₂₄ O ₃	623.1776	-0.0192
C ₂₅ H ₂₀ O ₃	623.1676	-0.0292	C ₂₅ H ₂₀ O ₃	623.1676	-0.0292
C ₂₅ H ₂₂ O ₂	623.1710	-0.0258	C ₂₅ H ₂₂ O ₂	623.1710	-0.0258
C ₂₅ H ₂₄ O ₂	623.1760	-0.0208	C ₂₅ H ₂₄ O ₂	623.1760	-0.0208
C ₂₅ H ₂₀ O ₂	623.1660	-0.0308	C ₂₅ H ₂₀ O ₂	623.1660	-0.0308
C ₂₅ H ₂₂ O	623.1694	-0.0274	C ₂₅ H ₂₂ O	623.1694	-0.0274
C ₂₅ H ₂₄ O	623.1744	-0.0224	C ₂₅ H ₂₄ O	623.1744	-0.0224
C ₂₅ H ₂₀ O	623.1644	-0.0324	C ₂₅ H ₂₀ O	623.1644	-0.0324

3. Chemical web database search

Online check the presence of those molecular formulae in databases (such as PubChem, SciFinder,...) to report known natural products and eliminate non-referenced formulae.

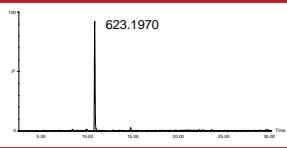
5. UV

Characteristic UV spectra may help eliminates or confirm putative identification (e.g. flavonoids), if concentration is sufficient.



7. confirmation

Clear identification is obtained by injection of the pure compound, if available, in the same conditions and/or by MS/MS experiments.



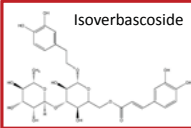
Conclusion

This procedure has been applied in a chemotaxonomic study of Brazilian *Lippia* species and more than 40 compounds were annotated in this way: 14 main compounds isolated, 8 minor compounds identified through the presented method and injected, and 20 minor compounds putatively identified with strong chemotaxonomic evidence. Among them, 12 NPs were detected for the first time in different *Lippia* spp. studied.

This LC-PDA-HR-MS strategy provided high quality structural information leading to reliable peak annotation for crude extract profiling. Complementary information can be obtained by further injection of the standard and/or MS/MS experiments, for complete identification.

Such a fast online dereplication method is helpful to avoid unnecessary isolation and structure elucidation of well-known compounds, and do not require instrument-specific MS/MS databases.

Further development on the LC log P estimation will provide a even better orthogonal filtering.



[1]. Wolfender, J.L. (2009) *Planta Med.* 75:719. [2]. Kind, T., Fiehn, O. (2007) *BMC Bioinformatics* 8:20. [3]. Martel, S. et al. (2008) *Chromatographic Approaches for Measuring log P*, in *Molecular Drug Properties - Measurement and Prediction*. Ed. R. Mannhold, Wiley-VCH, Weinheim, Germany.

Appendix X

Optimization of *Conus* consors venom profiling using Ultra-High Pressure Liquid Chromatography (UHPLC).

Philippe J. Eugster ¹

Daniel Biass ²

Davy Guillarme ¹

Philippe Favreau ²

Reto Stöcklin ²

Jean-Luc Wolfender ¹

¹ School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

² Atheris Laboratories, Geneva, Switzerland

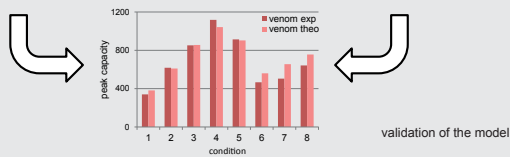
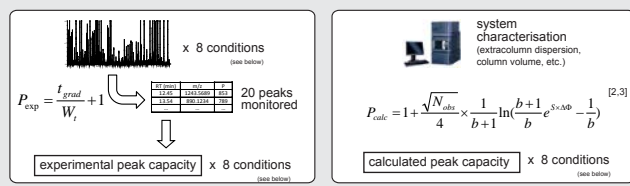
Poster presented at the Fall Meeting of the Swiss Chemical Society (SCS) in Lausanne, Switzerland, the 9th of September 2011.

Introduction

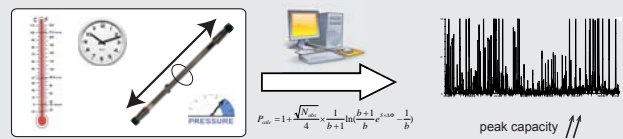
LC-MS has become the reference profiling technique in peptide analysis such as venomomics, providing abundant valuable data [1]. Since an optimized separation provides more resolved peaks and thus, more valuable MS data, UHPLC using sub-2- μm particles packed columns on systems able to work up to 1000 bars is an interesting option. In this study, the aim was to determine suitable conditions to obtain the best separation using UHPLC-TOF-MS. The parameter chosen to describe the quality of the separation was the peak capacity [2]. High temperature and different chromatographic parameters were tested. The results obtained in terms of peak capacity for the venom containing peptides with average MW of 1500 were compared in the same experimental conditions to those of crude plant extract profiling of a medicinal plant *Hypericum perforatum* with constituent having an average MW of 400 to evaluate the influence of the size of the analytes in these conditions.

Strategy

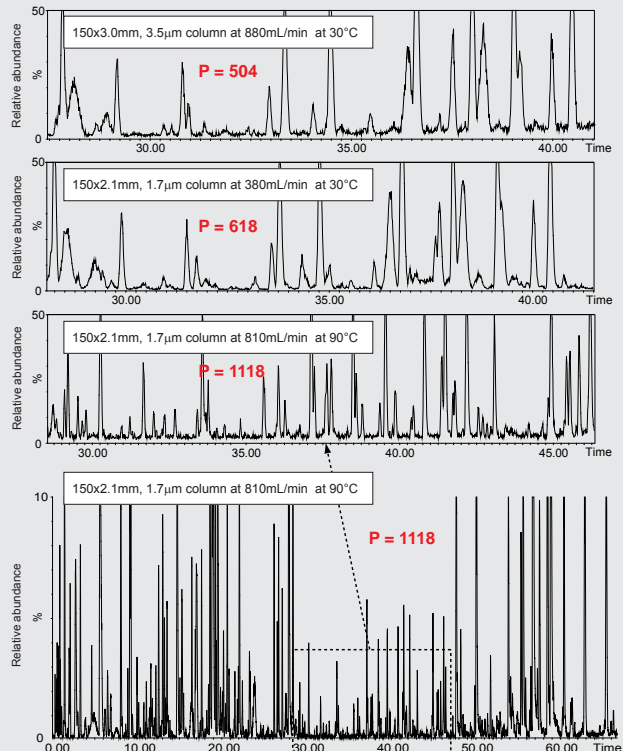
1 experimental and calculated P comparison



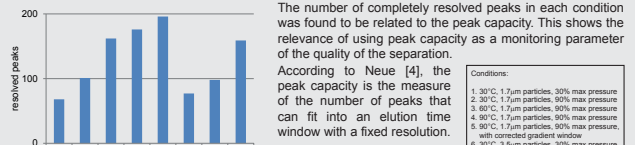
2 optimisation



Results



Relevance of peak capacity as a performance parameter

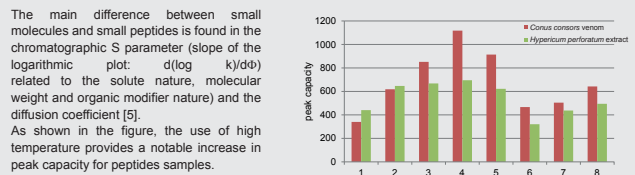


The number of completely resolved peaks in each condition was found to be related to the peak capacity. This shows the relevance of using peak capacity as a monitoring parameter of the quality of the separation. According to Neue [4], the peak capacity is the measure of the number of peaks that can fit into an elution time window with a fixed resolution.

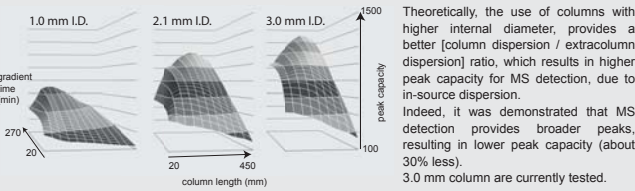
Conditions:

1.	30°C, 1.7 μm particles, 30% max pressure
2.	30°C, 1.7 μm particles, 90% max pressure
3.	60°C, 1.7 μm particles, 90% max pressure
4.	90°C, 1.7 μm particles, 90% max pressure
5.	90°C, 1.7 μm particles, 30% max pressure, with corrected gradient window
6.	30°C, 3.5 μm particles, 30% max pressure
7.	30°C, 3.5 μm particles, 90% max pressure
8.	60°C, 3.5 μm particles, 90% max pressure

Difference between peptides and small molecules

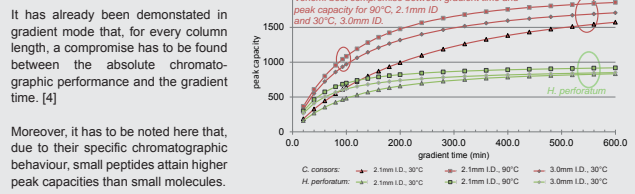


Modelling of column internal diameter (I.D.) effect



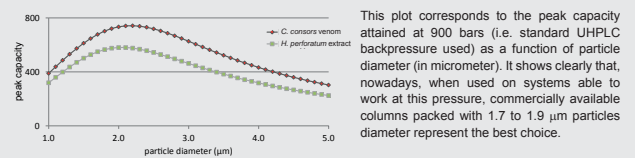
Theoretically, the use of columns with higher internal diameter, provides a better [column dispersion / extracolumn dispersion] ratio, which results in higher peak capacity for MS detection, due to in-source dispersion. Indeed, it was demonstrated that MS detection provides broader peaks, resulting in lower peak capacity (about 30% less). 3.0 mm column are currently tested.

Modelling of gradient time



Moreover, it has to be noted here that, due to their specific chromatographic behaviour, small peptides attain higher peak capacities than small molecules.

Modelling of particle diameter



This plot corresponds to the peak capacity attained at 900 bars (i.e. standard UHPLC backpressure used) as a function of particle diameter (in micrometer). It shows clearly that, nowadays, when used on systems able to work at this pressure, commercially available columns packed with 1.7 to 1.9 μm particles diameter represent the best choice.

Final results and conclusion

Optimal *Conus consors* venom profiling conditions were found at 90°C, using a 150x2.1mm, 1.7 μm particles column, with a gradient slope of 1%/min from 5 to 98% acetonitrile, at 810 $\mu\text{L}/\text{min}$. Modelling clearly demonstrated that the use of sub-2- μm particles, high temperature and maximal pressure provides the best separation of the small peptides, which was confirmed by further injections.

It has to be noted that an important loss of peak capacity is related to the MS source dispersion, that can be reduced by using 3.0 internal diameter columns. Such columns are currently tested.

Because of the high temperature used in this method, stability of these peptides at temperatures up to 90°C is currently tested. An alternative to high temperature is the use of columns of 3.0mm I.D., providing similar results at temperature between 30°C and 60°C.

Such an efficient separation is useful in online peptides identification, and has already been tested on a QTOF-MS. Fast and efficient identification of conotoxins was carried out thanks to this platform.

[1] D. Blass, S. Dutertre, A. Gerbault, J.L. Menu, R. Offord, P. Favreau, R. Stöcklin, J. Proteomics 2009, 72, 210.
[2] U.D. Neue, J. Chromatogr. A 2008, 1184, 107.
[3] E. Grata, D. Guillaume, G. Glauser, J. Boccard, P.A. Carrupt, J.L. Veuthey, S. Rudaz, J.L. Wolfender, J. Chromatogr. A 2009, 1216, 5660.
[4] D. Guillaume, E. Grata, G. Glauser, J.L. Wolfender, J.L. Veuthey, S. Rudaz, J. Chromatogr. A 2009, 1216, 3232.
[5] J. Ruta, D. Guillaume, S. Rudaz, J.L. Veuthey, J. Sep. Sci. 2010, 33, 2465.

Appendix XI

Ion mobility spectrometry in metabolite profiling of complex plant extracts.

Philippe J. Eugster¹

Richard Knochenmuss²

Jean-Luc Wolfender¹

¹ School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, Switzerland

² Tofwerk AG, Thun, Switzerland.

Poster presented at the 60th International Congress (Joint Meeting with ASP, AFERP, PSE and SIF) and Annual Meeting of GA, 60th International Congress and Annual Meeting of the Society for Medicinal Plant and Natural Products Research (GA), in New York, USA, July 28 - Aug 1, 2012.

Ion mobility spectrometry applied to complex plant extract profiling: possibilities, limitations and outlook

Philippe J. Eugster¹, Richard Knochenmuss², Jean-Luc Wolfender¹

¹School of Pharmaceutical Sciences, EPGL, University of Geneva, University of Lausanne, quai Ernest-Ansermet 30, CH-1211 Geneva, Switzerland;

²TOFWERK AG, Feuerwerkerstrasse 39, 3602 Thun, Switzerland

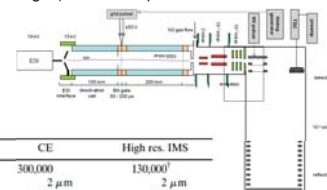
1. Introduction

One of the main challenges when profiling crude plant extracts is to obtain as much information as possible in one single LC-MS analysis for a detailed estimation of its composition. Indeed, plant extracts and other complex natural samples are composed of hundreds of compounds, which vary widely in structure, physicochemical properties and concentration. Today, (U)HPLC-QTOF-MS is the preferred technique for metabolite profiling, providing high performance separation and detection [1], and thus detailed information on the composition of the samples.

Still, there is an on-going need for better resolution in both LC and MS dimensions to further improve the number of features detected. In this respect, ion mobility (IM) that provides a separation mechanism different from both LC and MS is worth to be evaluated.

2. Ion mobility spectrometry (IMS)

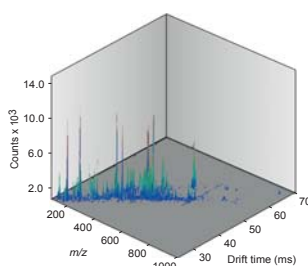
IMS allows separation in the gas phase of the compounds prior to the MS analysis, based on their chemical and physical interactions with a gas (the drift gas). When coupled to a TOF-MS, IM offers an additional high speed separation dimension (milliseconds), based on the mass, charge, size and shape [2,3] of the analytes.



	HPLC	GC	CE	High res. IMS
# of theoretical plates (N)*	25,000	120,000	300,000	130,000 ¹
Efficiency (HETP)	10 μm	400 μm	2 μm	2 μm
Resolving power (R _s) ²	65	145	230	150
Time of separation	30 min	20 min	10 min	50 ms
Plates per second	14	100	500	2,600,000

3. Metabolite profiling of a *Ginkgo biloba* extract by infusion IMS and UHPLC-TOF-MS

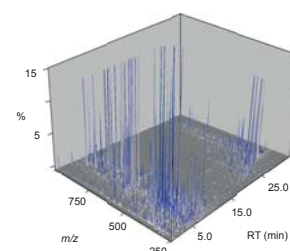
IM-TOF-MS 10 min
363 detected features



Direct infusion ESI-PI-IM-TOF-MS 3D plot of the *Ginkgo biloba* standardised extract after blank subtraction. Analysis time: 10 min. Amount of sample used: 1.0 μg. Drift gas: nitrogen.

A *Ginkgo biloba* extract was analysed using both IM-TOF-MS and UHPLC-TOF-MS, for comparison purpose, since both techniques represent a 2D separation (drift time x m/z and RT x m/z, respectively). The 3D plots highlight clearly the difference in separation, based on different molecular properties.

UHPLC-TOF-MS 30 min
1064 detected features



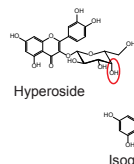
UHPLC-ESI-PI-TOF-MS 3D map of a *Ginkgo biloba* standardised extract, using a 150x2.1 mm, 1.7 μm C18 column with a 5-95% ACN in 30 min. Sample used: 10.0 μg.

Comparison of the IM- and LC-TOF-MS separation

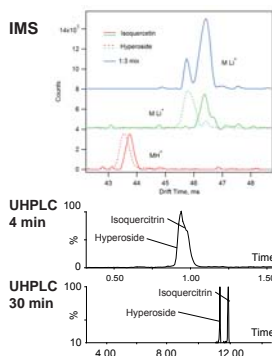
	IMS	LC-MS
ion suppression	+	+
isomer separation	++	+
sample consumed	very low	low
analysis time	very fast	fast
transfer to pre-preparative scale	no	yes
selectivity modification possibilities	cationising reagent, drift gas, ESI solvent temperature	mob. phase stat. phase pH temperature
sensitivity	low	high

4. Separation of closely related stereoisomers by infusion IMS and UHPLC-TOF-MS

Stereoisomers are often present in natural samples and their separation is often challenging.



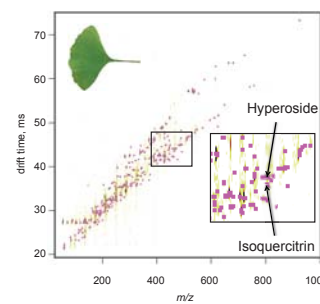
In order to evaluate the separation efficiency obtained using IMS and LC, two flavonoids present in *G. biloba* were selected.



The direct infusion of a mixture of these stereoisomers in IMS did not provide their separation. This was however strongly improved by addition of Li⁺.

Thus, the direct infusion IMS of the *G. biloba* extract under these conditions revealed a significant drift of many features out of the diagonal of the 2D plot. This indicated that IMS provided an additional separation dimension to that only related to the molecular weight. A zoom into this 2D plot revealed that a clear separation was obtained for the M+Li⁺ adduct of the two flavonoids considered.

The LC separation of such isomers in short UHPLC condition was found less efficient. A very long profiling however enabled their separation.



Hyperoside and isoquercitrin highlighted in a *G. biloba* IMS separation. Pink dots represent peaks detected after blank subtraction.

5. Conclusion and outlook

- IMS is able to separate stereoisomers in crude plant extracts.
- IMS and LC provide separation selectivities based on different mechanisms.
- Thus, IMS can be considered as a fast and additional separation technique complementary to LC.

Based on these results, and on previous works [5], hyphenation of LC with IMS seems to be a very promising method for tri-dimensional LC x IM x TOF-MS high resolution profilings of crude extracts. Its practical implementation will be studied in more depth in the frame of forthcoming metabolomic studies.

6. References

- [1] Eugster P.J. et al., J. AOAC Int. 2011, 94, 51-70.
- [2] Kaplan K. et al., Anal. Chem. 2010, 82 (22), 9336-43.
- [3] Dwivedi P. et al., Int. J. Mass Spectrom. 2010, 298, 78-90.
- [4] Ashbury G.R. & Hill H.H., J. Microcolumn Sep. 2000, 12, 172-8.
- [5] Valentine S.J. et al., Expert Rev Proteomics 2005, 2, 553-65.