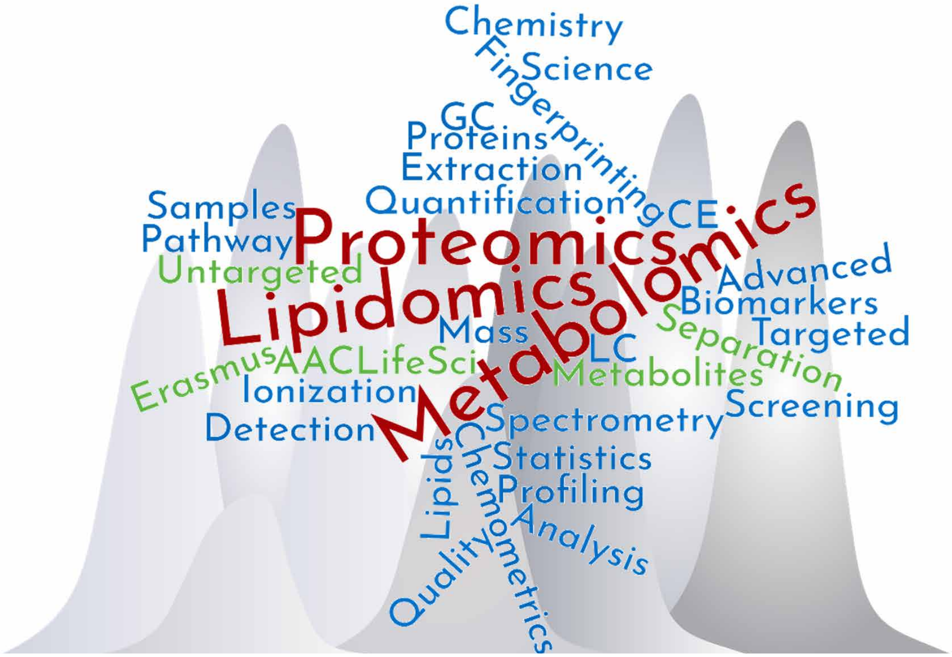
AACLifeSci Course Companion Manual

Advanced Analytical Chemistry for Life Sciences

Edited by
Pedro Domingues, Antonia García, Elżbieta Skrzydlewska



AACLifeSci

Erasmus+

Advanced Analytical Chemistry for Life Sciences

Edited by
Pedro Domingues, Antonia García, Elżbieta Skrzydlewska

AACLifeSci
AACLifeSci Course companion manual

This project has been funded with support from the European Commission.
This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

ISBN 978-83-951534-7-1

Contents

Preface: AACLifeSci Syllabus

Pedro Domingues[1], Antonia García[2], Elżbieta Skrzydlewska[3]

[1] *Mass Spectrometry Center, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal, QOPNA (Química Orgânica de Produtos Naturais e Agroalimentares)*

[2] *Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad San Pablo CEU; Boadilla del Monte, 28668 Madrid, Spain*

[3] *Department of Analytical Chemistry, Medical University of Białystok, 15-089 Białystok, Poland*

I. Rationale

The exponential growth of genomics research in this decade would suggest that related MS-omics research would rapidly develop. Indeed, although there has been a rapid development in proteomics research, other omics, such as lipidomics and metabolomics, have not yet been applied extensively. This differential development is due to different factors, such as the insufficient development of bioinformatics tools, and the lack of skilled professionals in the field. It is in this context that the AACLifeSci course has been developed, aiming to give advanced training to Ph.D. students in the areas of biomolecular and health sciences in mass spectrometry and hyphenated chromatographic technics and also in lipidomics, metabolomics and proteomics.

Nowadays, the rapid development of civilization diseases is considered the leading cause of increased morbidity and mortality in Europe. In this context, quantitative and qualitative research designed to understand, identify biomarkers and discover new therapeutic approaches is one of the possible remedies. Omics, an emerging field of study in biology that englobes other fields, such as genomics, proteomics, lipidomics or metabolomics, is an essential tool that is used nowadays for explaining the biochemical basis of an organism's functioning through characterization and quantification of pools of biological molecules. As single compounds are ineffective biomarkers, omics give new opportunities for a more detailed understanding of diseases. This understanding improves efficiency, monitoring, and leads to personalized therapy.

Omics teaching and learning in European universities is imbalanced. Omics subjects are usually delivered to students through curricular units that can be collectively designed as Advanced Analytical Chemistry (AAC), which delivers the instrumental basis in Ph.D. biochemical related courses. However, the number of specialized units and teaching teams with access to dedicated equipment needed for a comprehensive approach is limited. There is also a lack of AAC-qualified academics, didactic materials or a system for experience exchange at all levels. Thus, only a few Ph.D. students can be appropriately trained in AAC, instrumental analysis or data handling, or acquire the skills necessary to integrate AAC with the needs of health service, hospital diagnostic labs, business or labor market.

Advanced Analytical Chemistry for Life Sciences (ACCLifeSci), was an educational and scientific project coordinated by Doctor Elżbieta Skrzydlewska, the Dean of MUB's Faculty of Pharmacy with the Division of Laboratory Medicine and financed by Erasmus+ Programme and implemented by a partnership of three universities:

- Medical University of Białystok, Poland, Faculty of Pharmacy with the Division of Laboratory Medicine, Department of Analytical Chemistry (Doctor Elżbieta Skrzydlewska)
- Aveiro University, Portugal, Department of Chemistry, Mass Spectrometry Center (Doctor Pedro Domingues)
- University San Pablo-CEU, Spain, (Doctor Antonia García)

This project has been funded with support from the European Commission AACLifeSci. The project aimed to adapt the study programs of the three Partner Universities to the needs of the biomolecular related industries and healthcare service in the use of omics-related Advanced Analytical Chemistry techniques through an exchange of experiences at the European level. This project involved preparing the academic staff of the 3 Partner Universities for educating PhD students in the use of AAC techniques in metabolomics, lipidomics, and proteomics, by introducing modifications in study programs and by creating educational resources, including a companion coursebook and e-materials for students. The project activities included partnership-based learning/training/teaching activities, and involved an exchange of experiences and good practices, creating a stable and active network of cooperation, as well as personal and social development of project participants.

The full course is organized into four modules, which allows structured learning of the contents and facilitates the organization of lectures and practical classes. The course will cover, in the first module, the main aspects of biomolecular mass spectrometry, including the most recent developments in instrumental design and their advantages, along with an understanding of mass spectra analysis for structural elucidation. Fundamentals of the high-performance separation techniques coupled with mass spectrometry, including separation mechanisms and instrumentation will also be covered. This module will provide the analytical basis so that students can further develop their knowledge in the MS-related omics fields. The following modules will cover the fundamentals of metabolomic, proteomic and lipidomic studies. These modules will present the analytical approaches, data processing and data analysis methods and processes for each of the omics methodologies. Some of the bioinformatics resources used today will also be described, and students will have the opportunity to test them in practical lessons.

II. Course Aims and Outcomes

Aims

The course aims to provide students with theoretical and practical chemical, biochemical instrumental and bioinformatics skills in separation techniques and mass spectrometry for life sciences, metabolomics, lipidomics, and proteomics. Students get to experience the different areas of omics and how they are applied in real-life scientific research. Introductory theoretical information about aims and methods of chromatography and mass spectrometry and each of the omics is provided and will be thoroughly discussed. The explanation of how these methods are applied in science, of their practical advantages and limitations, challenges and scientific questions they address is provided. The practical skills developed will be mainly focused on using open-access databases and freeware software tools for data processing and analysis when using these approaches.

By the end of this course, students should appreciate the scientific problems involved in the post-genome era and know how to obtain and how to perform a simple analysis of mass spectrometric and omics data. Understanding the principles, methods, and uses of the newest OMICS techniques will contribute significantly to the students' future in academy and industry.

Specific Learning Outcomes

By the end of this course, students will:

1. critically review the available types of mass analyzers and ionization methods with their advantages and disadvantages, including tandem mass spectrometry analysis.
2. recognize fragmentation patterns of organic molecules mainly from small metabolites, peptides, and lipids.
3. discuss comprehensively the different mechanisms of separation coupled with MS for metabolomics, proteomics, and lipidomics.
4. know and understand the different methods of sample treatment and their limitations for each of the omics applications.
5. identify all stages of an omics study and their different approaches in metabolomics, lipidomics, and proteomics.
6. discuss the use of public software in reprocessing data and pathways analysis.
7. acquire necessary skills in the use of databases and other free access resources.
8. explain to non-specialists how the "omics" disciplines can be expected to provide valuable information in different areas of Life Sciences.

III. Format and Procedures

This course is structured in four independent, but closely related modules:
Module 1 - Separation techniques and Mass Spectrometry for Life Sciences;
Module 2 - Metabolomics;
Module 3 - Lipidomics;
Module 4 - Proteomics.

Each of these modules examines the principal, theoretical, and technical aspects that are essential to an advanced understanding of the OMICs approaches to solve the problems of real-life sciences. The knowledge and skills gained in this course were considered appropriate for a level of knowledge equivalent to a Ph.D. degree.

The course is organized into workshops. Each module will comprise, at least, 1-hour workshop and 2 hours of tutored practical classes where students can acquire practical knowledge and application skills essential for each module. Each module is independent. However, knowledge on the subjects of Module 1 - Separation techniques and Mass Spectrometry for the Life Sciences is essential for understanding the following modules.

MUB Ph.D. program organization and how AACLifeSci will be implemented

The Faculty of Pharmacy with Division of Laboratory Medicine, Medical University of Bialystok offers two different Ph.D. courses: one in the area of medical sciences in the field of medical biology and the other in the area of pharmaceutical sciences.

The Ph.D. program in **Medical sciences** is organized into four years. The second year includes the course of *Modern analytical techniques in biomedical sciences* that has been extended from 15 to 30 hours based on the contents of AACLifeSci. This course corresponds to 3 ECTS.

This course aims to provide students with the knowledge of the strategy and methods of modern and specialized biochemical analysis used mainly in research labs.

At the end of this course students should:

- have acquired knowledge of the sample preparation techniques employed in metabolomics, lipidomics and proteomics;
- be familiar with and able to describe contemporary separation techniques, including chromatographic and electromigration techniques;
- know the research methodology used in the classic and omic analysis;
- demonstrate knowledge of the concepts and practical uses of biostatistical assessment of research results;
- be able to propose an analytical technique for solving a concrete scientific problem from the field of biomedical sciences;
- be able to carry out metabolomic/lipidomic/proteomic analysis;
- be able to interpret the obtained results using statistical methods;
- be able to apply and draw conclusions from research in order to solve problems.

The Ph.D. program in **Pharmaceutical Sciences** is also organized into four years. The second year includes the course *Advanced analytical techniques in omics research* that is attained within 15 hours based on the contents of AACLifeSci. This course corresponds to 1 ECTS. However, this course is preceded by the course *Elements of modern pharmaceutical analysis* which includes knowledge of separation techniques/ cchromatography and mass spectrometry as well as target

MUB /Medical sciences/	Total hours	Mass Spectrometry and Separation Techniques	Metabolomics	Lipidomics	Proteomics
		Modern analytical techniques in biomedical sciences	Advanced analytical techniques in omics research		
Lectures and workshops	6h	2h	1h	1h	2h
Practical Lessons	24h	10h	4h	5h	5h
Student centered learned	57h	15h	10h	14h	18h
Total student effort	87h	27h	15h	20h	25h

analysis. Therefore, the second course includes information only about omics analysis. Totally, 30 hours is dedicated for AACLifeSci, which corresponds to 4 ECTS.

This course aims to provide students with the knowledge of the strategy and methods of omics analysis used in research labs.

At the end of this course students should:

- know the methods of biological sample preparation of analytical and omic analysis;
- know research methodology used in the classic and omic analysis;
- demonstrate knowledge of concepts and practical uses of biostatistical assessment of research results;
- be able to carry out metabolomic/lipidomic/proteomic analysis;
- be able to interpret the obtained results using statistical methods;
- be able to apply and draw conclusions from research in order to solve problems.

MUB /Pharmaceutical sciences/	Total hours	Mass Spectrometry and Separation Techniques	Metabolomics	Lipidomics	Proteomics
		Elements of modern pharmaceutical analysis	Advanced analytical techniques in omics research		
Lectures and workshops	8h	5h	1h	1h	1h
Practical Lessons	22h	10h	4h	4h	4h
Student centred learned	30h	14h	3h	5h	6h
Total student effort	58h	29h	8h	10h	11h

CEU Ph.D. program organization and how AACLifeSci will be implemented

The Faculty of Pharmacy at the University San Pablo-CEU offers a master's degree course named DRUG DISCOVERY, for postgraduate students in Chemistry, Pharmacy, and other Health Sciences. This inter-university master's degree course is organized as a result of the collaboration among three universities: University Complutense, San Pablo-CEU and Alcalá University, all located in Madrid province. These universities have been collaborating in the MEDICAL CHEMISTRY, a Ph.D. program with Quality Mention: Towards the Excellence. Over the course of 2016-17, the fourth edition of this Master Degree was implemented.

The master's degree program covers one academic course with 60 ECTS. In the second semester, it includes an optional subject: ADVANCED BIOANALYTICAL TECHNIQUES, comprising 3 ECTS with 30 h of attendance, modality on-campus.

This course aims to provide students with knowledge and skills on modern techniques for bioanalysis, mainly based on mass spectrometry coupled with high-resolution separation techniques. Besides, knowledge of purification and preconcentration methods is provided. It comprises learning of theory and practice by working with real samples in a modern laboratory.

At the end of this course students should:

- be able to choose and apply the appropriate analytical methods, considering the fundamentals, instrumentation, scope, and applications of the methods currently used for characterization and analysis of pharmaceuticals and biological samples.
- be able to choose and apply the appropriate analytical method for the analysis of active ingredients and their metabolites in biological samples and the analysis of active ingredients at very low concentrations in the final pharmaceutical product.
- identify all stages of the study and their different approaches in metabolomics.

CEU /DRUG DISCOVERY/	Total hours	Mass Spectrometry and Hyphenations	Chromatography and CE in bioanalysis	Metabolomics	Sample purification and concentration
Advanced Bioanalytical Techniques					
Lectures and workshops	24h	11h	8h	3h	2h
Practical Lessons	6h	2h	2h		2h
Student-centered learning	60h				
Total student effort	90h				
ECTS	3				

The subject ADVANCED BIOANALYTICAL TECHNIQUES will be taught based on the contents of AACLifeSci in the academic course, in the second semester and covering the same 3 ECTS.

Classes will be organized as follows:

CEU	Total hours	(1) Mass Spectrometry & (2) Separation techniques	Metabolomics	Lipidomics	Proteomics
Lectures and workshops	22h	6h LC: 2h GC: 1h CE: 1h	4h	4h	4h
Practical Lessons	8h	2h	2h	2h	2h
Student centered learning	60h	30h	10h	10h	10h
Total student effort	90h	42h	16h	16h	16h
ECTS	3	1.4	0.53	0.53	0.53

Organization of UA Ph.D. program in biochemistry and how AACLifeSci will be implemented

The Ph.D. program in Biochemistry of the University of Aveiro is organized into four years. The first year includes seven curricular units, four of which are optional. Two of these optional Curricular Units (Advanced courses in biochemistry I and Advanced courses in biochemistry II) will be based on the contents of AACLifeSci. These UCs are composed of three optional modules each corresponding to 3 ECTS. To complete each of these UCs students must complete two of these modules. The evaluation will be performed on each module, and the final grade will be obtained by a weight average of the evaluation obtained in the two modules. The courses and the respective modules are:

Advanced courses in biochemistry I:

This course aims to provide students with advanced skills in the strategy and methods of modern and specialized biochemical analysis used in research and clinical laboratories and industries. At the end of this course, students should:

- understand the fundamentals and applications of mass spectrometry and hyphenated chromatographic techniques.
- be able to critically evaluate the methods used in lipidomics and metabolomics studies.
- be able to plan lipidomics or metabolomics experiments.
- be able to analyze lipidomics or metabolomics data.

UA	Total hours	Chromatography and Mass Spectrometry for Life Sciences	Lipidomics	Metabolomics
Lectures and workshops	20h	10h	10h	10h
Practical Lessons		5h	5h	5h
Student-centred learning	150h	75h	75h	75h
Total student effort	180h	90h	90h	90h
ECTS	6	3	3	3

Advanced courses in Biochemistry II

This course aims to provide students with advanced skills in strategy and methods of modern and specialized biochemical analysis in Omics, used for proteomics, glycomics, and bioinformatics.

At the end of the course the student must be able to:

- Critically evaluate which methods should be used to solve specific problems of proteomics and glycomics.
- Plan proteomics and glycomics experiments.
- Analyze the data acquired in proteomics and glycomics studies.
- Know how to select and how to use adequate algorithms and tools in these biochemistry applications.

UA	Total hours	Proteomics	Glycomics	Bioinformatics
Lectures and workshops	20h	10h	10h	10h
Practical Lessons		5h	5h	5h
Student centred learning	150h	75h	75h	75h
Total student effort	180h	90h	90h	90h
ECTS	6	3	3	3

Overall, AACLifeSci will be fully integrated into four modules of the Ph.D. program in Biochemistry of the University of Aveiro, each corresponding to 3 ECTS, with a total of 12 ECTS.

Module 1- Separation techniques and Mass Spectrometry for the for Life Sciences (3 ECTS)
Module 2- Metabolomics (3 ECTS)
Module 3- Lipidomics (3 ECTS)
Module 4- Proteomics (3 ECTS)

IV. Additional Resource Readings

Module 1

1. H.J. Hübschmann. Handbook of GC-MS: Fundamentals and Applications (3rd Edition). Somerset, NJ, USA: Wiley, 2015. ProQuest ebrary. Web. 8, 2015.
2. L.R. Snyder, J.J. Kirkland, J.W. Dolan. Introduction to modern liquid chromatography. Wiley, 2011.
3. E. De Hoffmann, V. Stroobant. Mass spectrometry: principles and applications. Wiley, 2007.
4. D.R. Baker. Capillary electrophoresis. Wiley, 1995.
5. https://www.agilent.com/cs/library/usermanuals/public/G7100-90001_CEMSAnalysis_ebook.pdf

Module 2

1. A. García, J. Godzien, Á. López-Gonzálvez, C. Barbas. Capillary electrophoresis mass spectrometry as a tool for untargeted metabolomics. *Bioanalysis*, 2017, 9(1):99-130.

Module 3

1. L. Feng, G.D. Prestwich (ed.) Functional Lipidomics. Taylor & Francis, Boca Raton, 2006.
2. D. Armstrong. Lipidomics: Volumes 1e 2: Methods and Protocols (Methods in Molecular Biology). Humana Press, 2009, 1.
3. U. Loizides-Mangold. On the future of mass-spectrometry-based lipidomics. *FEBS J.*, 2013, 280(12):2817-29.
4. R. Harkewicz, E.A. Dennis. Applications of mass spectrometry to lipids and membranes. *Ann. Rev. Biochem.*, 2011, 80:301-25.

Module 4

1. J. Lovric. Introducing Proteomics: From concepts to sample separation, mass spectrometry and data analysis, Wiley. ISBN: 978-0-470-03523-8, 2011.
2. R. Aebersold, M. Mann. Mass-spectrometric exploration of proteome structure and function, *Nat. Rev. Mol. Cell. Biol.*, 2015, 16(5):269-80.
3. A.M. Silva, R. Vitorino, M.R. Domingues, C.M. Spickett, P. Domingues. Post-translational Modifications and Mass Spectrometry Detection, *Free Radic. Biol. Med.*, 2013, 65:925-41.

Module 1
Separation techniques and Mass Spectrometry for the Life Sciences

Wojciech Łuczaj[1], Antonia García[2], Agnieszka Gęgotek[1], Katarzyna Bielawska[1], Elżbieta Skrzydlewska[1]

[1]*Department of Analytical Chemistry, Medical University of Białystok, 15-089 Białystok, Poland*

[2]*Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad San Pablo CEU; Boadilla del Monte, 28668 Madrid, Spain*

I. Rationale

The first module of the ACCLifeSci course concerns the main aspects of biomolecular mass spectrometry. It includes an overview of the most recent developments in instrumental design and their advantages, along with an understanding of mass spectra analysis for structural elucidation. The fundamentals of high-performance separation techniques coupled with mass spectrometry will also be described, including the separation mechanisms and instrumentation. This module intends to provide students with the essential analytical tools for further deepening of their knowledge on the mass spectrometry-related "omics" fields.

II. Course Aims and Outcomes

Aims

The module aims to provide students with theoretical and practical chemical, biochemical, instrumental and bioinformatics skills in chromatography and mass spectrometry applied to the life sciences. An introductory theoretical description of application of these methods in science, in particular, omics investigation, their advantages and limitations, the challenges and scientific questions they address will be discussed. The practical skills to be developed will mainly consist of using open-access databases and freeware software tools within these approaches.

Specific Learning Outcomes:

At the end of this course, students should be able to:

1. describe the principles of ionization, separation, and detection of molecules by mass spectrometry;
2. describe the principles of tandem mass spectrometry, the instrumentation, and its application;
3. be able to identify small molecules using mass spectrometry data;
4. describe the principles of separation, optimization, and detection of compounds using liquid

chromatography, gas chromatography and capillary electrophoretic methods;

5. predict how changes in experimental conditions may influence the separation with these methods;
6. describe the principle and the design of the components of separation and mass spectrometry analytical instrumentation;
7. choose the most appropriate technique, regarding both separation and detection objectives;
8. select the appropriate instrumentation for each "omics" applications, on the basis of the performance and requirements that are expected.

III. Course contents

Module 1 – Separation techniques and Mass Spectrometry for the Life Sciences

1. Mass spectrometry
    1.1. Introduction to mass spectrometry
    1.2. Ionization techniques in biological mass spectrometry
    1.3. Mass analyzers in mass spectrometry
    1.4. Tandem mass spectrometry
    1.5. Interpretation of mass spectra
2. Separation techniques coupled with mass spectrometry
    2.1. Liquid chromatography – mass spectrometry (LC-MS)
        2.1.1. Introduction to LC-MS
        2.1.2. Theoretical principles of liquid chromatography
            2.1.2.1. The chromatographic process
            2.1.2.2. Normal phase chromatography (NP)
            2.1.2.3. Reversed-phase chromatography (RP)
            2.1.2.4. Hydrophilic interaction chromatography (HILIC)
            2.1.2.5. Ion-Exchange Chromatography (IEC
        2.1.3. Instrumentation for LC-MS
    2.2. Gas chromatography-mass spectrometry (GC-MS)
        2.2.1. Introduction to GC-MS
        2.2.2. Theoretical principles of GC-MS
            2.2.2.1. The chromatographic process
            2.2.2.2. Column phases
            2.2.2.3. GC×GC/MS
            2.2.2.4. Sample pre-treatment and derivatization
        2.2.3. Coupling GC-MS
    2.3. Capillary electrophoresis-mass spectrometry (CE-MS)
        2.3.1. Introduction to CE-MS
        2.3.2. Theoretical principles of CE-MS
            2.3.2.1. Capillary zone electrophoresis
            2.3.2.2. Micellar electrokinetic chromatography

1. Mass spectrometry

1.1. Introduction to mass spectrometry

Since the 1960s, mass spectrometry has become a standard analytical tool in the analysis of organic compounds, although, in the beginning, its applications in biological fields were scarce. Mass spectrometers have become much easier to use in the last 15 years, with enormous progress in instrumentation and nowadays it represents a key analytical technology in the study of proteins, lipids, and metabolites. The current role of MS in biochemistry and life sciences results from the development of new ionization techniques, the improvement of analyzers, fragmentation methods, as well as the outstanding development of separation technologies and methods. This course presents an overview of the most important features of modern MS and its coupling with different separation techniques.

MS is an analytical technique that involves the study of gas phase ions from the analyte, which are separated according to their mass-to-charge ratio (m/z) and their abundance measured. The main components of a mass spectrometer are the ionization source, the mass analyzer, and the detector. The analyzer and the detector are always under high vacuum obtained by using several pumps (pneumatic and turbopumps) in order not to influence the fragmentation pattern of molecules and not to increase the background noise.

Briefly, the MS workflows consist of the ionization of the analyte sample in an ion source, separation of the ionized molecules and their ionized fragments according to their mass-to-charge ratio (*m/z*) using an analyzer, detection of the ions, and the analysis and interpretation of the resulting mass spectrum. Tandem mass spectrometers have the additional capability of selecting ions and inducing their fragmentation to obtain detailed structural information of the selected species (MS/MS).

1.2. Ionization techniques in biological mass spectrometry

There are several ionization sources that allow the analysis of organic compounds. The traditional source, used for the analysis of small volatile molecules, is Electron Ionization (EI). This source requires the analyte to be volatile and thermostable and, currently, it is employed particularly in combination with gas chromatography (GC). In this source, gaseous molecules collide with energetic electrons, leading to their ionization and further fragmentation. The characteristic fragmentation pattern is very reproducible, being frequently used for identification purposes by using spectral databases. In EI ionization, as the result of a single electron removal, a molecular ion radical cation ($M^{+\bullet}$) is generated, with *m/z* that corresponds to the nominal mass (integer mass) of the compound (Figure 1.1).

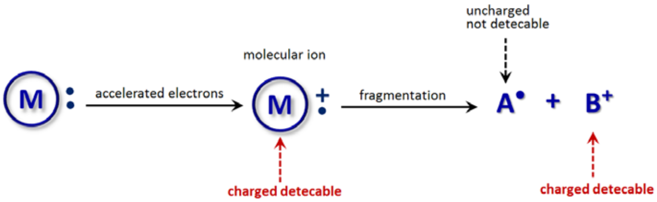


Figure 1.1. Electron Ionization (EI) and subsequent fragmentation.

The main disadvantage of the EI source is that it requires the analyte to be in gas phase before ionization and therefore, polar, non-volatile and thermally labile compounds are excluded unless they can be derivatized to modify their properties. Another problem is the lack of the molecular ion in the spectrum of some compounds.

Mass spectrometry became easily accessible to non-volatile and thermally unstable compounds with the development of new ionization systems in the late 80´s. By then, electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI), had been developed. Both ionization techniques are soft, with universal applications in the analysis of involatile, thermally unstable and even high-mass biological molecules. Nowadays, ESI and MALDI are the most widely used ionization techniques for LC-MS analysis of biological molecules such as peptides, proteins or intact lipids. Both ESI and MALDI are soft ionization techniques since they produce mainly molecular ions with very little fragmentation. Both are desorption sources, being able to operate with liquid or solid phases, and usually, they do not require chemical derivatization in contrast to GC–MS analysis.

The electrospray ionization technique was first reported by John Fenn in 1984. In the electrospray, ESI source, an atmospheric pressure ionization method, a liquid containing the analyte flows through a metal capillary to which a high voltage is applied to produce ions. A spray of charged droplets is formed, and subsequently, they are subjected to desolvation, usually with the use of an N_2 current. The drops become smaller and a phenomenon called "Coulomb fission" occurs when the superficial tension of the liquid is overcome by repulsion between charges in the surface of the droplets. When the droplets are small enough, gas phase ions of analyte are generated, usually by an ionic evaporation process (Figure 1.2).

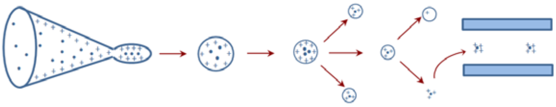


Figure 1.2. The mechanism of electrospray ionization (ESI).

In ESI, the ionization occurs directly from a solution, so thermally labile molecules may be ionized without degradation. The response of the analyte is concentration dependent, so it is possible to work with low flows in LC (between 5 and 100 μL min^{-1}), or even sub-microliter flow in nanoelectrospray sources, and using a very narrow column. Typical solvents are water, methanol, and acetonitrile. In general solvents with low superficial tension and viscosity, and high conductivity are used. It is possible to add modifiers to promote the ionization of the analyte in the solution. Typically, 0.1% formic or acetic acid is used for positive ionization mode, and ammonium formate or acetate are used in negative mode since pure water, or nonpolar organic solvents are less effective.

MALDI (Matrix Assisted Laser Desorption Ionization) was developed by Karas and Hillenkamp and Tanaka in 1988. It is a soft method of ionization used mainly for biomolecules such proteins, peptides, oligonucleotides, polysaccharides and synthetic polymers. In this method, ionization is achieved by mixing the sample solution with a large molar excess of a matrix material. Then, the solvent is evaporated and crystals of the sample–matrix are irradiated with a short pulse laser beam. The matrix can absorb a high amount of energy of the laser radiation and transfer it to the sample molecules that are desorbed as intact gas-phase ions. The ionization mechanism involves desorption and ionization of the analyte by proton transfer reaction with molecules of the matrix (Figure 1.3).

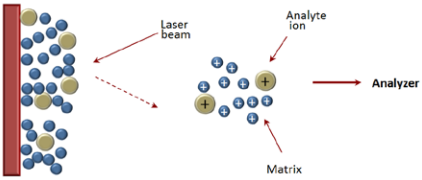


Figure 1.3. The mechanism of matrix-assisted laser desorption/ionization (MALDI) technique.

The matrices commonly used for MALDI are small organic compounds such as 2,5-dihydroxybenzoic acid (DHB), nicotinic, sinapinic or 3,5-dimethoxy-4-hydroxycinnamic acids. The selection of the matrix depends on the properties of the analyte and the laser, which is usually a nitrogen laser (337 nm). The importance of the type of matrix should be emphasized. For example, DHB is suitable for lipid mixtures analyzed in positive and negative mode, while *p*-nitroaniline enables more sensitive detection of phosphatidylethanolamines in a negative ion mode.

MALDI and ESI ionization methods have common (Calibri)advantages, among which the major one is high sensitivity, usually requiring merely a few picomoles of the analyte per analysis. Moreover, the simplicity of operation makes these techniques widely used for rapid screening of a sample. When using MALDI, a significant drawback is that matrix compounds are usually ionized, generating high background signal, resulting in difficulties in analyzing the low *m/z* region (below 800 Da).

Atmospheric-pressure chemical ionization (APCI) is an additional soft ionization technique that can be used for the ionization of neutral non-volatile analytes. APCI is an ionization method related to ESI, where the stream of liquid is dispersed into small droplets, as discussed above. In the case of APCI, this process involves a heater and a high potential on a *corona discharge* electrode which ionizes the nebulizer gas, thus forming plasma. Sample molecules are then ionized in the plasma by proton transfer processes by the eluent. APCI is, in general, better suited for lipophilic compounds and can be successfully combined with HPLC using analytical columns when ESI does not give suitable ionization. Table 1.1 summarizes the main features and compares different ionization techniques.

Table 1.1. Characterization of different ionization techniques.

Ionization technique	Typical analytes	Mass range	Sample introduction	Advantages	Disadvantages
EI	compounds: relatively small, non-polar, thermostable	<1 kDa	GC or liquid/solid	non-polar analytes, no ion suppression, easily coupled with GC, spectrum libraries	hard ionization technique, needs volatile samples, needs thermal stability, low molecular weight compounds
ESI	polar compounds e.g. peptides, proteins, sugar nucleotides	<200 kDa	LC or solution	thermolabile compounds, high MW compounds, sensitivity, easy to interface with LC, soft ionization technique, multi-charged ions	ionizable analytes, sensitive to salts, ion suppression
MALDI	polar compounds e.g. peptides, proteins, sugar nucleotides	<500 kDa	sample mixed with a solid matrix	thermolabile compounds, high MW compounds, sensitivity, less sensitive to salts, soft ionization method	a wide range of matrices, difficulties in quantitative analysis, ion suppression
APCI	neutral compounds	<1 kDa	LC or solution	allows for large scale flow rates, easily to interface with LC, thermostable compounds, soft ionization methods	needs solubility in polar solvents, sensitive to salts, ion suppression

1.3. Mass analyzers in mass spectrometry

A mass spectrum consists in a plot of many ions, presented as *m/z* values, which were produced after the ionization of the analyte. To obtain a mass spectrum, it is necessary to separate the ions of different *m/z* ratios, and then measure the relative intensities of each ion. In MS, mass analyzers are the components that separate ions according to their mass-to-charge (*m/z*) ratio. There are several different mass analyzers that vary in resolution, mass accuracy, dynamic range and capability to perform tandem-MS experiments. These concepts can be defined as:

Resolution - the term is used in many areas of analytical interest and, most frequently, refers to the ability to differentiate between closely related signals. These signals, in MS, are considered as *m/z* ratios of ions and the equations for resolution R and resolving power RP are described below:

$$R = \Delta m/m$$

$$RP = m_1/(m_2\text{-}m_1),$$

where *m1* is the lighter ion and *(m2-m1)* is the difference between two consecutive ions. Another expression is *RP* = *m*/D*m,* where *m* is the measured *m/z* ratio and D*m* (in Da) is usually measured using the peak width at a specific percentage of the peak height, usually the full width at half maximum (FWHM).

Accuracy - is the proximity of the experimental mass (accurate mass) to the true value (exact mass). In mass spectrometry it is determined in ppm using the following equation:

$$Error = (monoisotopic\ exact\ mass - measured\ accurate\ mass)\ /monoisotopic\ exact\ mass\ x\ 10^6$$

An instrument that can measure masses, differentiating only integer masses displays nominal mass accuracy or unit mass accuracy. Higher resolution correlates with higher mass accuracy, and this increases the identification capabilities of an MS instrument.

Linear dynamic range - is defined as the range over which the ion signal is directly proportional to the analyte concentration. The linear range is crucial for accurate measurements, especially for quantification analysis.

There are many different mass separation devices, known as analyzers, used in MS, and each has its advantages and disadvantages. The analyzers most frequently applied to biological analysis are quadrupoles, ion traps, time-of-flight analyzers and, more recently, orbitraps.

The *quadrupole (Q)*, as the name suggests, consists of four parallel rods, as shown in Figure 1.4. Both direct-current (DC) and radiofrequency (RF) potentials are applied to the opposite rod pairs of the quadrupole. If specific DC and RF values are chosen, ions with a particular *m/z* ratio will have stable trajectories and pass through the analyzer to the detector, while all other ions will not be transmitted. These conditions are the basis for the single ion monitoring (SIM) mode, which is most common quadrupole mode used for targeted quantitative analysis, as the signal to noise ratio is highly improved. On the other hand, the quadrupole may be considered as a scanning instrument when operating in full-scan mode. In this mode, by scanning DC and RF in different ratios, all ions can be transmitted, one m/z ratio at a time, to the detector. As a result of full-scanning mode, complete unit resolution mass spectra can be obtained.

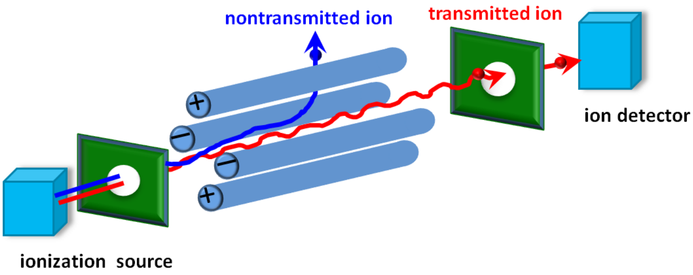


Figure 1.4. Scheme of quadrupole mass analyzer (Q).

The *ion-trap (IT)* mass analyzer, as the name suggests, traps the ions using quadrupolar fields. There are two types of such analyzers: 2D and 3D ion-traps, which trap ions in two and three dimensions, respectively. The 3D ion-trap, also named quadrupole ion-trap, consists of four electrodes: two ring electrodes and two ellipsoid end-cap electrodes, as shown in Figure 1.5. All ions introduced into the ion-trap are trapped by using oscillating electric three-dimensional quadrupole fields. However, when a specific voltage is applied to the ring electrode, the ions with specific *m/z* value will be transmitted to the detector. By changing the voltage value, different ions are transmitted to the detector. In the 2D ion-trap, also known as linear ion-trap, ions are trapped by a two-dimensional radio frequency field, provided by a four-rod quadrupole. Inside the 2D ion-trap, stopping potentials are applied to electrodes at the entrance and at the end of the quadrupole to confine the ions. Both ion-trap analyzers are low-resolution instruments. 2D ion-trap, however, has higher ion storage and scanning rate capacities. In both technologies, helium gas is present to control the trajectory of the ions inside the trap and to perform MS^n experiments.

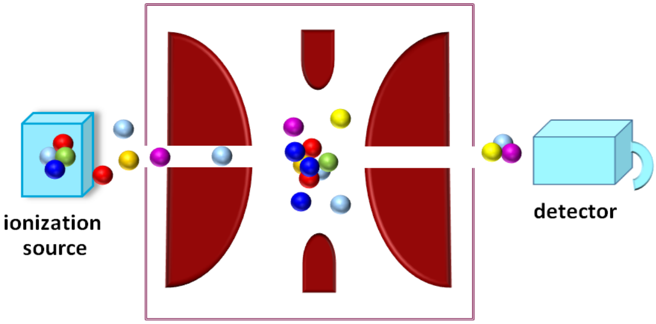


Figure 1.5. Scheme of 3D ion trap mass analyzer (IT).

The *time-of-flight (TOF)* analyzer (Figure 1.6) relies on the fact that when all the ions produced in the source of a mass spectrometer are accelerated with a high voltage, they will possess the same kinetic energy (E_k). As a consequence, the velocity of each ion is related to the m/z (E_k=1/2m/z·v^2). As such, the time necessary for the ions to travel across the flight tube (tube at very high vacuum) of the mass spectrometer will also be related to the *m/z* of the ion. After that time, in which all ions reach the detector, a complete mass spectrum is obtained.

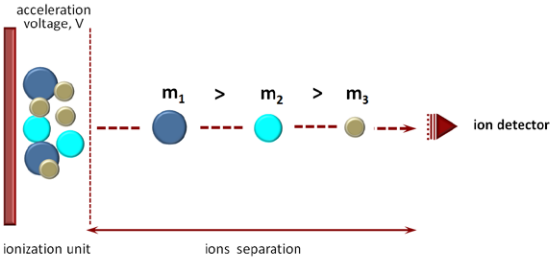


Figure 1.6. The mechanism of ions separation in the time-of-flight (TOF) mass analyzer.

One of the major advantages of the TOF instrument, comparing with quadrupole and ion trap instruments, is that the *m/z* ratios of ions can be determined with high resolution and accuracy. Higher resolutions can be obtained by increasing the length of the flight tube that will lead to a longer time needed for ions to travel from the source to the detector. Another solution is using ion mirrors in TOF instruments, the so-called reflectrons, which reflect the ion beam. A reflectron also increases resolution by diminishing the kinetic energy dispersion. Also, TOF instruments are fast-scanning, being limited by the time in which the heaviest ion reaches the detector. The other advantage of the TOF analyzer is that linear TOF has no limit of the mass range, being most useful for analyzing intact proteins.

More recently, a new analyzer has been developed: the Orbitrap, consists of a barrel-like electrode. Generated in a source ions are radially trapped around a central electrode, and the m/z values are calculated by fast Fourier transform from the oscillation frequencies of the trapped ions. This detector is characterized by its high resolution, up to 500,000 FWHM, high accuracy (<1pp with internal calibration) and sensitivity. In summary, a comparison of the different MS analyzers and their main features are presented in (Table 1.2).

Table 1.2. Characterization of different MS analyzers.

Analyzer	Q	IT	TOF	Orbitrap
Acquisition speed (Hz)	2-10	2-10	10-100	1-18
Mass accuracy (ppm)	low	low	1-10[a] ppm	1-5 ppm
Masss range (*m/z*)	<3000	<6000	<100,000[a] unlimited	<6000
Resolutionb	unit	unit	<50,000[a]	<500,000[b]
Advantages	low cost, easily interfaced to various ionization techniques higher dynamic range	low cost, easily interfaced to various ionization techniques, MS^n	fast scanning, high mass range, high mass accuracy	high mass accuracy, fast polarity switch
Disadvantages	low resolution, low mass accuracy, low mass range, low scanning speed, MS/MS requires multiple analyzers	low resolution, low mass accuracy, low mass range, low scanning speed	high cost, lower dynamic range than Q	lower scanning rate than QTOF, high cost, lower dynamic range than Q

a) in hybrid/reflectron configuration b) working resolution (FWHM)

1.4. Tandem mass spectrometry

Tandem mass spectrometry (MS/MS) is defined as the technique in which two stages of MS are involved, and a step of fragmentation of ions occurs between the two stages. MS/MS can be carried out with a tandem mass spectrometer, consisting of two mass analyzers separated by a collision cell containing a collision gas, usually where collision-induced dissociation (CID) takes place. Tandem-MS experiments can be used for targeted quantitative approaches, in which high resolution and mass accuracy are usually not required or for identification purposes in which high resolution and mass accuracy are usually required.

The *triple quadrupole (QqQ)* is probably the most widely used MS/MS instrument for targeted quantitative approaches. This device, as the name suggests, comprises three quadrupoles in a series (Figure 1.7). The second quadrupole (Q2) is not used as a mass separation unit but as a collision cell (q), where collision-induced dissociation (CID) takes place. Thus, in the Q2 module, the ions transmitted by the first quadrupole (Q1) are fragmented, and the generated product ions are subsequently transmitted into the third quadrupole (Q3).

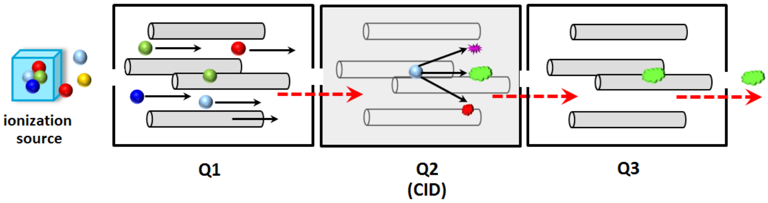


Figure 1.7. The tandem mass spectrometry based on triple quadrupole operation.

Besides triple quadrupoles, other frequently used hybrid MS/MS instruments using different analyzers are quadrupole time-of-flight (QqTOF) and QqOrbitrap instruments. Other instruments that are used for tandem mass spectrometry are the ion trap instruments, in which the three stages of analysis occur in the same analyzer (the ion trap). Also, other fragmentation methods can be used like Electron-Transfer Dissociation (ETD) or Higher-energy Collisional Dissociation (HCD), each having specific applications.

Four main different MS/MS experiments can be carried out with tandem mass spectrometers, although three of them can only be performed using triple quadrupole instruments: *Product-ion scan*: the first module of the tandem mass spectrometer (Q1) is used to isolate the ion of interest (precursor ion). The selected ion is then fragmented in a collision cell (q) by collision with ultrapure gas. The third mass spectrometer (Q3) produces a mass spectrum by scanning the product-ions generated from the precursor ion in the collision cell. This experiment can be performed in all types of tandem mass spectrometers. *Precursor-ion scan*: the third module of the mass spectrometer (Q3) allows only for the transmission of a selected product (fragment) ion, with a known *m/z* value, while the first module (Q1) scans over the mass range of interest, allowing all precursor ions to be transmitted sequentially and fragmented in the collision cell (q). The signal from the detector is recorded if the ions transmitted by Q1 originate the product ion monitored in Q3. This experiment can only be performed in triple quadrupole instruments.

Neutral loss scan mode: In this mode, both mass spectrometers Q1 and Q3 are linked-scanning in a way that the difference of m/z in Q1 and m/z in Q3 corresponds to the mass of a neutral fragment that is lost in the fragmentation in the collision cell (q). Product ions generated as a result of the loss of any other mass are not transmitted by Q3. This experiment can only be performed in triple quadrupole instruments.

Selected reaction monitoring (SRM) and multiple reactions monitoring (MRM): Both are common MS/MS experiments used in quantitative analytical approaches. SRM mode monitors the fragmentation of a selected precursor ion into a selected product ion. This is carried out by setting each of the modules of the tandem mass spectrometer to transmit a single ion, *i.e.*, the precursor ion by Q1 and the product ion by Q3. MRM mode is selected reaction monitoring applied to multiple product ions from one or more precursor ions. In this approach, several pairs of "precursor-product" are monitored. Combination of liquid chromatography separation with SRM/MRM identification can be used to increase the specificity of the measurement, by removing interfering transitions of the "precursor-product". A significant disadvantage of this approach is that it is not suitable for global analysis as the specific pairs of transition ions could not be provided. Regarding the MS detection, MRM is preferable to SIM since MRM mode offers higher sensitivity, higher selectivity, and reproducibility, and, thus, it is currently the most popular type of experiment setup in triple-quadrupoles.

1.5. Interpretation of mass spectra

This section describes the main principles of mass spectra interpretation regarding the ionization techniques described in section 1.2. Mass spectra produced by ESI, MALDI, and APCI ionization methods can originate both positive ions and negative ions.

ESI ionization mode can produce multiple charged ions. The charge number depends on the structure of the compound, particularly the number of possibilities of gaining or losing protons, in positive and negative ionization modes, respectively. This property is particularly useful for interpretation of mass spectra of high molecular weight compounds, such as intact proteins (Figure 1.8), while low molecular weight compounds usually have less potential sites for protonation/deprotonation and, therefore, their spectra show less multiple charged ions.

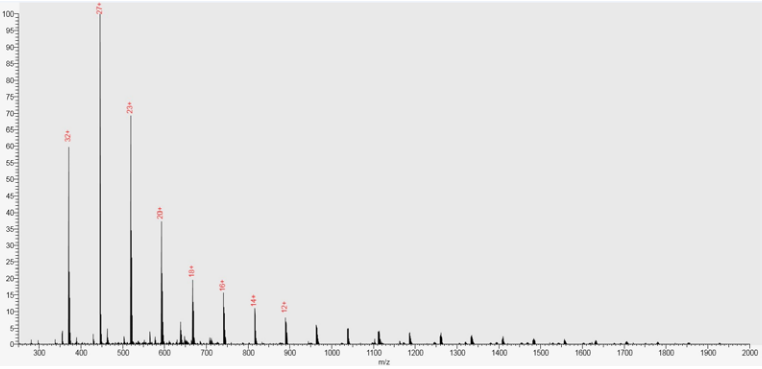


Figure 1.8. ESI mass spectrum of thioredoxin (12 kDa) as an example of a typical ESI mass spectrum with multiple charged ions [Quadrupole-Orbitrap, Thermo Fisher, Q Exactive Plus].

In the case of low molecular weight compounds with only one polar group, their ESI spectra predominantly show molecular ions, usually $[M+H]^+$ or $[M-H]^-$, thus, the molecular weight of the individual species is easy to determine (Figure 1.9). Because some analytes ionize better in the positive mode than in the negative mode, while others behave in the opposite way, sometimes it is worth analyzing the samples using both ionization modes. Additionally, special consideration should be given to the presence of adducts of the analyte with species such as Na^+, K^+, NH_4^+ in the positive mode, and anions from the mobile phase such as formate or acetate in the negative ionization mode. APCI and MALDI mass spectra, similarly to ESI, consist predominantly of ionized molecular species, although, fewer multiple charged ions are observed.

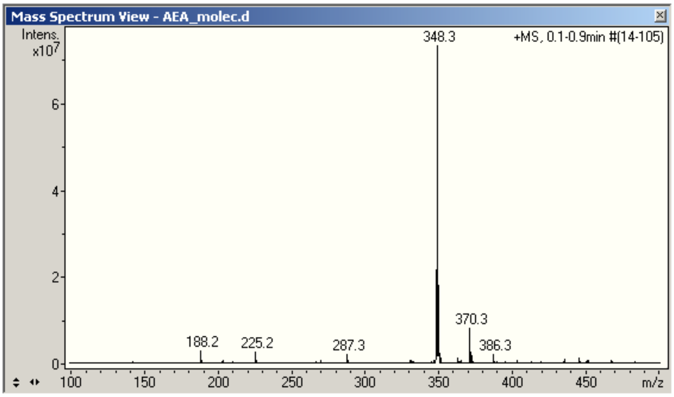


Figure 1.9. ESI mass spectrum of anandamide ionized in the positive mode; the signal at *m/z* 348.3 corresponds to the positively charged molecular ion $[M+H]^+$ of anandamide [IonTrap, Bruker Daltonics, Esquire 6000].

Finally, a few useful aspects in the interpretation of spectra will be discussed. Several types of ions can be observed in the mass spectra other than the molecular ions. These ions are called fragment ions, and their abundance depends on several variables like the ionization method, the fragmentation methods, the presence of labile groups in the molecule, etc. Soft ionization methods usually produce very little or even no fragmentation. In these cases, precursor ions are usually fragmented in a collision cell when working in the tandem mass spectrometry mode. On the other hand, EI usually produces extensive and reproducible fragmentation. Usually, the fragmentation pattern of molecules of the same family (e.g., peptides, lipids, etc) is similar, and knowledge of it is essential to identify these classes, and it will be discussed later. For instance, when using EI as ionization source, compounds containing hydrocarbon chains give rise to a series of ions with a mass difference of 14 Da between each other, corresponding to ($-CH_2-$). Water loss is observed by an 18 Da loss that is common for compounds containing hydroxyl groups. There are many other common fragmentation patterns like these that mass spectrometry scientist use for identifying compounds.

2. Separation techniques coupled with mass spectrometry

2.1. Liquid chromatography-mass spectrometry (LC-MS)

Traditionally, in liquid chromatography, compounds were identified by their retention time (t_R) in a chromatogram. However, in the case of samples with compounds characterized by similar t_R, or of very complex samples, the chromatographic separation does not allow for their identification. As such, further information is usually required, obtained by using an additional structural elucidation technique. For this reason, application of hyphenated techniques, such as LC-MS is desirable, since MS provides one the most sensitive detection methods currently available, while it simultaneously allows for the identification of the compounds. One of the major advantages of LC-MS is the high number of detectable species and the lower matrix effect due to attenuation of ion suppression by the separation step. MS provides information about the molecular weight and compound structure, from picograms or even femtograms of examined compounds. Additionally, it ensures the highest selectivity by allowing the monitoring of ions or characteristic fragment ions from the compound to be analyzed. Thus, LC-MS is an ideal analytical technique for both qualitative and quantitative analysis.

2.1.1. Introduction to LC-MS

The chromatographic separation of compounds from a mixture, in combination with the identification capability of the mass spectrometer, provides several advantages. LC-MS allows differentiating many compounds with similar t_R, but with different m/z or fragmentation pattern. The unique specificity and high selectivity of LC-MS allows quantitation, which chromatography alone does not. "Soft" ionization techniques (section 1.2), as most widely used ionization techniques in LC-MS, allow studying a wide range of compounds, from low molecular weight drugs and metabolites (below 1000 Da) to high molecular weight biopolymers (over 100 000 Da). A mass spectrometer can give the molecular weight of the analyte to be determined, together with the structural information that could be generated, thus helping in the identification. Additionally, the high selectivity and high sensitivity provided by mass spectrometry allow very accurate and precise quantitation.

Nevertheless, despite many advantages of LC-MS analyses, there are some issues that should be kept in mind. For instance, different species have different ionization efficiencies leading to dissimilarities in sensitivity, despite their close concentrations. Also, LC-MS analysis can be complicated by the phenomenon known as "ion suppression" (described in detail in section 1.2) resulting in the reduction of the signal of a particular analyte due to co-eluting substances affecting the ionization process.

A typical chromatographic system contains four major components: sample injector, mobile phase, stationary phase, and detector. Several chromatographic techniques have been used, and the importance of these four components may be different. Generally, in liquid chromatography, the role of the injector is the introduction of the analytes into the flowing solvent system. Currently, a six-port valve with a loop injector is the most widely used. The mobile and stationary phases, as the most important parts of a chromatographic system, are responsible for the separation process. In high performance/pressure liquid chromatography (HPLC), the mobile phase is a solvent or system of solvents delivered with constant flow rate under high pressure (usually up to 400 bar), while the stationary phase is packed inside a chromatographic column in a manner enabling maintaining this high pressure.

2.1.2. Theoretical principles of liquid chromatography

Chromatography is an instrumental separation technique whose aim is separation of the components from a mixture, allowing their identification and quantitation. Identification is usually based on the chromatographic retention characteristics. The definition of chromatography by the International Union of Pure and Applied Chemistry (IUPAC) is as follows: "Chromatography is a physical method of separation in which the components to be separated are distributed between two phases, one of which is stationary (the stationary phase), while the other (the mobile phase) moves in a definite direction". A mobile phase is described as "a fluid which penetrates through or along the stationary layer in a definite direction". It may be a liquid, a gas or a supercritical fluid, while the stationary phase may be a solid, a gel or a liquid. If a liquid, it may be distributed on a solid which may or may not contribute to the separation process".

In a chromatographic separation, analyte molecules are distributed between the mobile phase and the stationary phase to the extent that depends mainly on their chemical structure. The velocity of the sample components in the column is a function of the distribution of these components in two phases remaining in equilibrium. The components with a higher affinity to the stationary phase move more slowly than the components having a greater affinity for the mobile phase. Therefore, separation is the result of different migration rates caused by different values of the partition coefficient (K_s). The Nernst equation can express the partition coefficient:

$$K_S = C_L/C_G$$

where C_L and C_G represent the concentration of the substances in the stationary and mobile phases, respectively. The higher affinity for the stationary phase material, the higher value of K_s and the higher the value of the retention time (t_R).

The R*etention time* (t_R), is the time taken by an analyte to elute from a chromatographic column with a mobile phase and reach the detector. Because column length and the flow rate of a mobile phase affects the t_R, another parameter has been introduced, the *retention factor (k)*. This factor relates the retention time of an analyte to the time taken by an unretained compound, which does not interact with the stationary phase and elutes from the column at a specific time (t_0). The t_0 is sometimes called the dead time of the column. The difference between the t_R and t_0 is often called the reduced or adjusted retention time. Therefore, the retention factor is defined by the following equation:

$$k = (t_R - t_0)/t_0$$

Compounds of a mixture are separated when they interact differently with the mobile and stationary phase. So, the time needed for substances to cross the distance from the point of sample injection to the detector is different. There are two extremes of such interaction. If all the analytes have a total affinity for the mobile phase and do not interact with the stationary phase, they will not be separated and will reach the detector very quickly and at the same time (void or dead time). If all the analytes strongly interact with the stationary phase and do not interact with the mobile phase, they will be retained in the column and will not reach the detector. Therefore, the knowledge about the chemical properties of the analyzed compounds is crucial if a chromatographic separation process is to be considered. If the properties are known, by changing the chromatographic conditions, *e.g.*, properties of the stationary and mobile phases, it is possible to achieve the desired separation. The successful separation of compounds present in a mixture is related to a series of parameters shown in Table 1.3.

Table 1.3. The principal parameters of chromatographic separation.

Parameter	Symbol	Definition
Retention time	t_R	time taken by an analyte to elute from a chromatographic column and reach the detector
Dead time of the column	t_0	time taken by an unretained compound, which does not interact with the stationary phase,
Reduced retention time	$(t_R - t_0)$	the subtraction between retention time and dead time of the column
Retention factor	k	the ratio of reduced retention time to the dead time of the column
Selectivity (separation factor)	α	the ratio of retention factors of two separated compounds
Resolution	R	ability to separate two chromatographic peaks
Plate number	N	the number of small layers, along the chromatographic column, in which an equilibration of the sample distribution between stationary and mobile phase takes place

It should be noted that to obtain sufficient resolution within a reasonable time of the analysis, k values should be in a range from 1 to 10. The efficiency of separation of two compounds (A and B), is defined by the *separation factor* (α) that represents the ratio of their retention factors k_B and k_A (the distribution coefficient for the first two eluted components):

$$\alpha = k_B/k_A$$

The separation of two compounds takes place, when $\alpha>1$. However the larger the value of α, the greater the separation. This possibility is described by another parameter called *resolution* (R), given by the equation:

$$R= (t_B-t_A)/0.5(w_A+ w_B)$$

in which w_A and w_B represent the peak widths of two immediately adjacent peaks corresponding to compounds A and B.

The chromatogram is the graph of the detector response versus the elution time, and the graphical description of the parameters mentioned above is represented in Figure 1.10.

Regarding the chromatographic column, its performance is a major parameter affecting chromatographic separation. Column performance, also known as efficiency, is related to the number of theoretical plates (N). The number of theoretical plates is a measure of the "goodness" of the column. This parameter refers to the number of "small layers", along the chromatographic column, in which an equilibration of the sample distribution between stationary and mobile phase takes place. The number of plates depends on the column length (L), while the second parameter, *plate height* (H), defined as the efficiency of the column in the theoretical plate length, does not. The general rule in chromatography is that the smaller height of the plate, the better for the resolution. Similarly, a high number of plates is also a desirable criterion. The relation between N and H is described below:

$$H = L/N$$

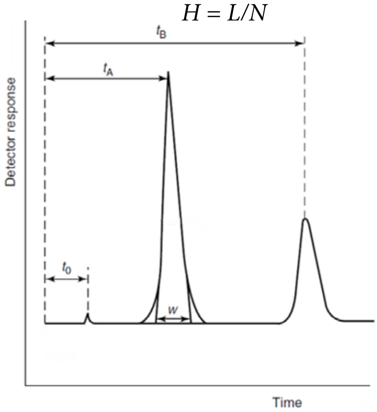


Figure 1.10. A chromatogram showing t_0, dead retention time, t_A and t_B, retention times of compounds A and B and w, peak width.

2.1.2.1. The chromatographic process

A variety of retention mechanisms occur in HPLC. However, particular interactions are based on the relative polarities of separated compounds. The properties of the mobile phase are essential for chromatographic separation. If the mobile phase provides insufficient separation of polar analytes, non-polar analytes will be retained longer on the column, and vice-versa. However, it is not always possible to achieve an adequate separation by using a mobile phase containing a single solvent. Generally, in such situation, a mixture of solvents is used. A separation obtained with a mobile phase of constant composition is defined as *isocratic elution*, while the composition of the mobile phase is changed during the analysis it is defined as *gradient elution*. In gradient elution, a solvent known to elute better the longer-retained compounds is added in increasing amounts during the time of analysis.

Even though there are many different chromatographic conditions, it is not always possible to obtain sufficient separation of all components in a mixture. Therefore, depending on the type of analyte, and on the polarity of stationary and mobile phases, different liquid chromatography modes are used (Table 1.4).

Table 1.4. Comparison of different modes of liquid chromatography.

LC mode	Mobile phase	Stationary phase	Type of separated compounds
NP	organics: dichloromethane, ethyl acetate	silica, amino, cyano, diol	organic compounds not soluble in water
RP	water/organic with or without additives	C18, C8, C4, cyano, amino	neutrals, weak acids, weak bases
HILIC	acetonitrile with water, ionic additives	polar, pure silica	polar compounds
IEC	buffered aqueous solutions	anion or cation, exchange resin	ionic, inorganic ions

2.1.2.2. Normal phase chromatography (NP)

Chromatographic separation in normal-phase results from polar interactions of separated compounds with the stationary phase of a column. The term ‘normal’ refers to the system in which the stationary phase is polar, while the mobile phase is non-polar. In this mode, the polar components are more retained and, by increasing the polarity of the mobile phase, the retention may be decreased. The more polar is the mobile phase, the faster the analytes will be eluted from the column.

Unmodified silica, with its free silanol groups as functional groups, is the most common stationary phase used in NP liquid chromatography. The mobile phase in NP usually relies on hydrocarbons, dichloromethane, ethyl acetate, or another water-immiscible solvent. From the particle point of view, the major reason for using NP chromatography is to increase the retention of polar components and elute hydrophobic compounds. It should be mentioned that NP chromatography is also useful for separating geometric and positional isomers.

2.1.2.3. Reversed-phase chromatography (RP)

Reversed-phase (RP) chromatography is the most common of all the methods used in HPLC. In this mode, the separation is based on the interactions of analytes with a nonpolar stationary phase and a polar liquid mobile phase. In RP, the mobile phase is more polar than the stationary phase. Thus, more polar compounds elute first, followed by the less polar ones. Also, in RP, the chromatographic separation is based on differences in the hydrophobicity of the components of the mixture.

There are many nonpolar stationary phases used in reversed-phase HPLC. The stationary phase of the most commonly used RP columns contains long-chain hydrocarbons covalently bonded to the silica surface. The most popular and universal RP stationary phase is C18 or ODS, containing octadecylsilyl groups. The typical solvent system in RP comprises two main components, a water or buffer solution, and organic solvents, among which the most frequently used are methanol or acetonitrile. RP chromatography is quite a comprehensive technique, suitable for separation of non-polar and ionizable compounds even during one analysis.

NP-LC and RP-LC have both been used for different purposes in biological samples analysis. For instance, an NP column could be applied for separating individual lipid classes based on the polar head groups, while the separation of lipid species on RP column relies on their different hydrophobicities (fatty acyl chains). Moreover, RP gradient chromatography is also useful for metabolite profiling in metabolomics studies.

2.1.2.4. Hydrophilic interaction chromatography (HILIC)

In addition to NP chromatography hydrophilic interaction liquid chromatography (HILIC) is another elution mode for polar compounds. It utilizes hydrophilic stationary phases, with a mobile phase consisting of a solvent system typical for RP, most frequently acetonitrile, with a small amount of water. Salts such as ammonium acetate or ammonium formate are often added to the mobile phase to increase polarity and ion strength. HILIC columns typically contain silica polar surfaces or it can be derivatized to amino or amide bonded phases.

The mechanism of HILIC separation is complex and relies on a liquid/liquid extraction system with water layer formation on the surface of the polar stationary phase and organic mobile phase. Therefore, in HILIC, polar compounds interact with the stationary phase, while less polar ones are distributed in the mobile phase. Since HILIC separation involves high organic mobile phases, the technique is easily adaptable to MS. Moreover, the use of organic solvents increases MS sensitivity due to a decrease of ion suppression.

2.1.2.5. Ion-Exchange Chromatography (IEC)

Ion-exchange chromatography (IEC) involves the separation of ionic and ionizable compounds. It uses packing materials containing ionic functional groups, usually with opposite charges than that of the analytes. In IEC, an anionic stationary phase is used for separation of cations or positively charged molecules, while for negatively charged molecules or anions, the cationic stationary phase is used. In this chromatographic mode, the mobile phase is highly polar, most frequently based on water, with some buffer or salts. Increasing the ionic strength of the mobile phase is the main factor responsible for the elution of compounds from the IEC column.

IEC is useful both for large and small biomolecule separations, such as of amino acids, carboxylic acids or amines. It should be noted that, due to the ion suppression phenomenon, IEC, which utilizes high ion strength in the mobile phase, is relatively difficult to be directly coupled with a mass spectrometer.

2.1.3. Instrumentation for LC-MS

A block diagram of an LC system, illustrating its major components, is shown in Figure 1.11.

a) Pump

The major task of the pump is providing a stable flow, which varies depending on the interface being used in the LC-MS and the parameters of the chromatographic column. For example, in the case of ESI source, when using a 2.1 mm diameter of the column, a flow rate in a range 0.2 – 0.4 mL/min is normally chosen, while the flow rate of 1 mL/min is preferable for APCI with a conventional 4.6 mm column.

Also, when combining LC with MS, the pump should not generate any pulse but provide a constant flow rate of the mobile phase. The pulsation of the flow makes the interpretation of the obtained results difficult. Consequently, to prevent the formation of bubbles in the mobile phase, a degasser is included in every LC-MS system.

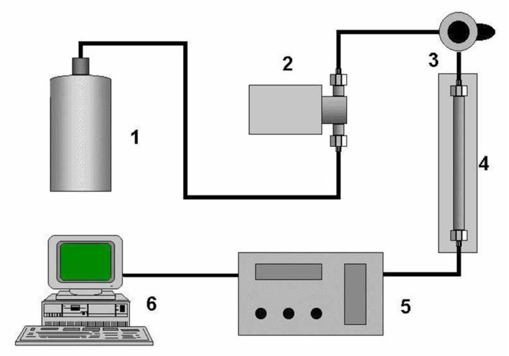


Figure 1.11. Block diagram of an LC system (1 – mobile phase container, 2 - pump, 3 - injector, 4 – chromatographic column, 5 - detector, 6 – computer).

b) Injector

A single type of injector is used almost exclusively in LC and is known as the *loop injector* (or six-port valve injector). In this injector, the sample is introduced, using a micro-syringe, into a mobile phase that fills a loop of a nominal volume. While the loop is filled, the mobile phase is pumped through the valve into the column to keep the column in equilibrium with the mobile phase. At the moment of injection, mobile phase flows through the loop, flushing its contents onto the column. Most importantly, injectors should allow injections with high reproducibility and accuracy, avoiding the presence of air bubbles or pulses.

c) Columns

The format of an HPLC column refers to the column length, column diameter, and particle size of the stationary phase. Typical HPLC columns, capable of maintaining pressures of 400 bars, have a diameter of 4.6 mm, their length varies from 100 to 250 mm and the particle size is usually in the range of 3-5 µm. However, the development of ultra-high performance liquid chromatography (UPLC) pumps, capable of pumping the mobile phase at a pressure of 1200 bars, allows using a reduced particle size in chromatographic columns to improve chromatographic resolution. Besides, the decrease of the mobile phase's flow rate, as a result of the decrease in column's particle size and diameter, confers higher sensitivity and lower limits of detection provided by nano-LC systems, which require only trace amounts of a sample for analysis. Various types of HPLC columns are available, in different formats, as enumerated in Table 1.5.

Table 1.5. HPLC column formats and their nomenclature.

Description	Dimension (i.d.)	Optimum flow-rate
Nanobore column HPLC	0.075 mm	0.3 µL/min
Capillary column HPLC	0.5 mm 1 mm	10 µL/min 50 µL/min
Microbore column HPLC	2.1 mm	250 µL/min
Narrow(small)-bore column HPLC	3 mm	500 µL/min
Normal-bore column HPLC	4.6 mm	1250 µL/min

To know more:

G. Rozing. Trends in HPLC column formats - microbore, nanobore and smaller. *LC GC EUROPE*, 2003 16.6A:14-19.

S.R. Wilson, T. Vehus, H.S. Berg, E. Lundanes. Nano-LC in proteomics: recent advances and approaches. *Bioanalysis*, 2015, 7(14):1799-1815.

N. Gray, M.R. Lewis, R.S. Plumb, I.D. Wilson, J.K. Nicholson. High-throughput microbore UPLC–MS metabolic phenotyping of urine for large-scale epidemiology studies. *J. Proteome Res.*, 2015, 14(6):2714-2721.

d) Detectors

Detectors used in LC are characterized by several parameters including selectivity, detection limit, and sensitivity. The term *selectivity* defines the ability of a detector to determine an analyte of interest without interferences derived from the matrix, the solvents or other substances present in the system. The term *sensitivity* is often used instead of the term *limit of detection,* which refers to the smallest concentration of an analyte that is sufficient for its detection with a set probability. However, the term *sensitivity* refers to the detector response that is related to the concentration of an analyte which reaches the detector. These parameters are fundamental considering the overall performance of an analytical method. However, it should be kept in mind that the highest sensitivity does not provide the lowest limit of detection/determination because interferences present in the sample can also give strong signals.

Many detectors may be used in combination with LC, including the UV, fluorescence, electrochemical, conductivity, refractive index, and MS detectors. Each of them has their specific advantages and disadvantages, but the most widely used LC detectors are UV and MS, with a UV detector frequently used as a supporting one. When using a UV detector, if the wavelength of the maximum absorption of the analyte (λ_{max}) is known, it can be monitored, and the detector may be selective for that analyte. Unfortunately, the UV detector does not have high selectivity because many organic molecules absorb UV radiation, at the same wavelength. In this sense, undoubtedly, MS is a much more reliable technique for obtaining analytical information, especially in complex mixtures. MS detection requires only picogram amounts of a sample to provide accurate information about the molecular weight and the structure of the compound. Mass spectrometers are ideal detectors for both qualitative and quantitative analytical approaches, so currently, this detector is considered the detector of choice for coupling with LC.

2.2. Gas chromatography-mass spectrometry (GC-MS)

2.2.1. Introduction to GC-MS

Gas chromatography coupled with mass spectrometry (GC-MS) is a combination of two advanced instrumental techniques for the analysis of organic compounds. GC-MS is the most effective technique for the analysis of volatile organic compounds in complex matrices in a wide range of concentrations. It is particularly advantageous for the determination of compounds of low molecular weight, medium or low polarity, thermostable, with boiling points below 350–400°C, and concentrations ranging from ppb to ppm. The development of column technology has been significant for the analysis of high-boiling compounds (up to 500°C). Analytical possibilities of GC-MS depend, mainly, on the mass spectrometer performance. With the decreasing detection limits and the growing number of searched compounds, data analysis can be troublesome. Therefore, new developments are needed. This technique is characterized by high selectivity and sensitivity, providing a wide range of applications. It is an essential technique for monitoring a wide range of applications, from environment analysis to biochemistry. GC-MS-based metabolomic approaches are extensively used to understand the processes leading to the development of diseases such as cancers or cardiovascular disease.

2.2.2. Theoretical principles of GC-MS

2.2.2.1. The chromatographic process

In a gas chromatographic system (GC), the sample to be analyzed could be a gas, a liquid or molecules adsorbed on a surface after solid-phase microextraction (SPME). The principle of separation in GC depends on the transfer of a substance (as steam) using carrier gas (mobile phase) through a capillary column frequently coated internally with a thin polymer film. During the transfer into the GC, sample compounds are volatilized by a rapid exposure to a zone kept at a relatively high temperature (200-300°C) and mixed with a stream of pure carrier gas (Ar, He, N_2 or H_2).

A diagram of a GC system, illustrating its major components, is shown in Figure 1.12 depicting the carrier gas, the injector port, the oven, the column and the detector.

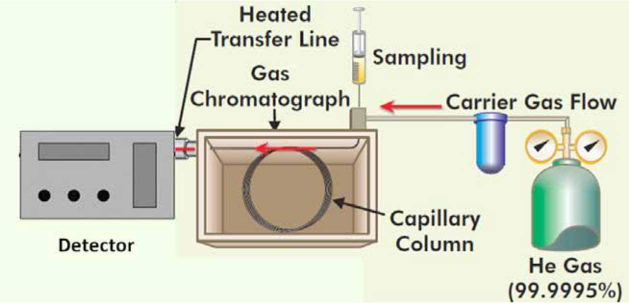


Figure 1.12. Diagram of the GC system (modified from *de.leco-europe.com)*.

Helium is the most common carrier gas in GC-MS, and it is compatible with the detector. There are three main injection modes, including: split mode, which is special for capillary columns, splitless, which is best for trace levels of high-boiling analytes in low-boiling solvents, and on-column, used for thermally unstable compounds.

The components of the sample separated on the column, successively reach the detector, generating electric signals which are proportional to the amount of the separated compound. Two elution modes can be described according to the temperature in the oven: isothermal mode (constant temperature over the analysis) and temperature programming. In the latter, the temperature is increased during the separation to increase analyte vapor pressure and decrease elution times of the most retained compounds.

The traditional detectors used in GC respond to differences in physicochemical properties of the carrier gas with and without the substances eluted from the column, *e.g.*, flame ionization, electron capture or thermal conductivity. Registered changes may be proportional either to the concentration or the mass flow rate of the analyte in a carrier gas. Important features of the detector include high sensitivity, detectability (detection limit), high signal stability, reproducible baseline signal and a wide range of linearity. The resulting data are processed to generate the chromatogram.

2.2.2.2. Column phases

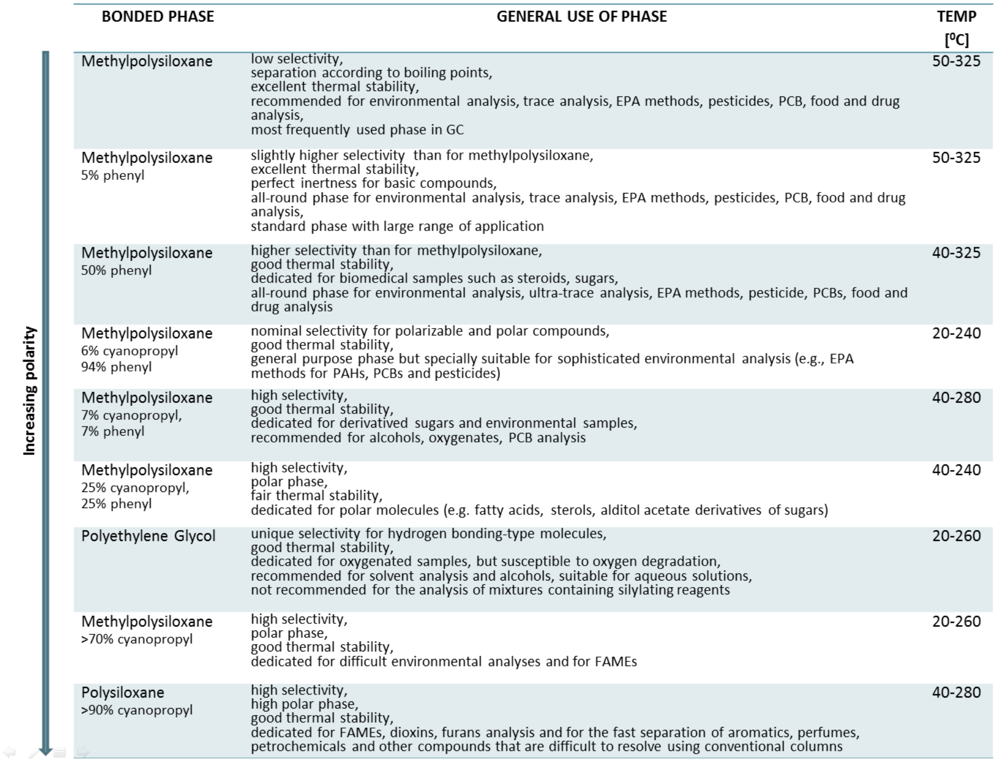
Choosing the type of column has a decisive influence on the quality of the separation of mixture components, which is the result of the chromatographic analysis. There are two types of columns: packed (analytical, micro packed), and open tubular ones, often called capillaries (capillary, microcapillary). Packed columns are filled with particles and capillary columns are open tubular capillary tubes. In packed columns, solid particles (carbon adsorbents, silica, alumina, molecular sieves, porous synthetic polymers, among others) serve as a support or are embedded with a liquid (silicones, squalene, polyethylene glycol, among others), acting as a stationary phase as a means of separating components within the GC.

Adsorbents are less common stationary phases due to the lower reproducibility of the results, the longer retention time, and the appearance of "tails" with much lower separation efficiency. Currently, capillary columns coated with a liquid stationary phase are frequently used. Liquid stationary phases should be chemically inert, capable of dissolving separated components, highly selective for the components of the mixture, should have low volatility and have thermal stability under the operating conditions of the column. They are particularly suitable for the separation of gaseous components with high separation efficiency.

The most important feature of most liquid stationary phases that influences the separation properties is polarity. The separation efficiency of the capillary column also depends on the following parameters and properties: column dimensions (length-L and inner diameter-ID) and phase thickness. Standard commercial length of capillary columns is 15, 30 and 60 m. The inner diameters of capillary columns have standard dimensions: 0.15; 0.25; 0.32 and 0.53 mm. The columns for GC-MS are especially suited for this detection system and characterized by low bleeding.

Liquid stationary phase polarity is usually chosen by taking into account the polarity of the analytes in the mixture. As a general rule, for the separation of non-polar compounds, non-polar phase should be used, and for the separation of the polar compounds, a polar phase should be used.

Table 1.6. Classification of GC stationary phases.



Increasing polarity

BONDED PHASE	GENERAL USE OF PHASE	TEMP [^{0}C]
Methylpolysiloxane	low selectivity, separation according to boiling points, excellent thermal stability, recommended for environmental analysis, trace analysis, EPA methods, pesticides, PCB, food and drug analysis, most frequently used phase in GC	50-325
Methylpolysiloxane 5% phenyl	slightly higher selectivity than for methylpolysiloxane, excellent thermal stability, perfect inertness for basic compounds, all-round phase for environmental analysis, trace analysis, EPA methods, pesticides, PCB, food and drug analysis, standard phase with large range of application	50-325
Methylpolysiloxane 50% phenyl	higher selectivity than for methylpolysiloxane, good thermal stability, dedicated for biomedical samples such as steroids, sugars, all-round phase for environmental analysis, ultra-trace analysis, EPA methods, pesticide, PCBs, food and drug analysis	40-325
Methylpolysiloxane 6% cyanopropyl 94% phenyl	nominal selectivity for polarizable and polar compounds, good thermal stability, general purpose phase but specially suitable for sophisticated environmental analysis (e.g., EPA methods for PAHs, PCBs and pesticides)	20-240
Methylpolysiloxane 7% cyanopropyl, 7% phenyl	high selectivity, good thermal stability, dedicated for derivatived sugars and environmental samples, recommended for alcohols, oxygenates, PCB analysis	40-280
Methylpolysiloxane 25% cyanopropyl, 25% phenyl	high selectivity, polar phase, fair thermal stability, dedicated for polar molecules (e.g. fatty acids, sterols, alditol acetate derivatives of sugars)	40-240
Polyethylene Glycol	unique selectivity for hydrogen bonding-type molecules, good thermal stability, dedicated for oxygenated samples, but susceptible to oxygen degradation, recommended for solvent analysis and alcohols, suitable for aqueous solutions, not recommended for the analysis of mixtures containing silylating reagents	20-260
Methylpolysiloxane >70% cyanopropyl	high selectivity, polar phase, good thermal stability, dedicated for difficult environmental analyses and for FAMEs	20-260
Polysiloxane >90% cyanopropyl	high selectivity, high polar phase, good thermal stability, dedicated for FAMEs, dioxins, furans analysis and for the fast separation of aromatics, perfumes, petrochemicals and other compounds that are difficult to resolve using conventional columns	40-280

In non-polar columns (*e.g.*, squalene, 100% methyl-polysiloxane, (5%-phenyl)-methylpolysiloxane), elution of analytes is carried out according to their volatility, *i.e.*, by the increasing order of their boiling points. In polar columns, *e.g.*, polyethylene glycol (Carbowax), (70% cyanopropyl) phenyl-methylpolysiloxane), the determining factors for analyte separation are both their boiling point and their polarity (the dipole moment-μ) as well as the strength of interactions with the stationary phase (dipole, hydrogen bond). Polar columns, such as cyanopropylphenyl-polysiloxane, are also characterized by high selectivity, which results from the high content of phenolic and cyano groups (Table 1.6). The maximum allowed temperature corresponds to non-polar stationary phases.

Other phases may be used, depending on the purpose: alcohols, saccharides, nitro compounds, aliphatic amines, salts of inorganic acids (*e.g.*, silver nitrate), among others. There are also other known phases, whose separating mechanism involves various interactions of the compounds with the individual particles of the stationary phase or an ordered structure, thereby achieving significant separation selectivity. Such phases include chiral stationary phases, ionic liquid phases, etc.

2.2.2.3. GC×GC/MS

Gas chromatography plays an important role among chromatographic techniques due to its high-resolution capability, flexibility, and a wide range of applications. However, it fails in the analysis of very complex samples because of the limited peak capacity (the number of chromatographic peaks which can be completely separated using just one chromatography column). As the compounds are migrating through the column, chromatographic bands are widened, resulting in a limited number of bands that can be completely separated at the exit of the column. This limitation cannot be overcome by modifying the chromatographic parameters. The only effective solution is to use two-dimensional separation (GC×GC), where the components partially separated on the first chromatography column are separated again, based on a different mechanism, on the second GC column.

Since its introduction in 1991, there has been a rapid development in the GC×GC technique. Today, GC×GC is commonly used in many different fields, covering a wide range of applications. Comprehensive two-dimensional gas chromatography with mass spectrometry (GC×GC/MS) is an emerging technology that provides a two-dimensional separation and a full mass spectral profile, based on the retention time coordinates of compounds, in the two-dimensional separation space. Separation mechanisms in both dimensions must be completely independent of each other, *i.e.*, orthogonal. GC×GC has all the advantages of the one-dimensional ('normal') gas chromatography technique: it is completely automated and provides a sensitive analysis of detected compounds. Also, separations in 2D offer better information and compounds with a similar structure are eluted together in the 2D chromatogram.

At first, the introduced sample is separated on a high-resolution capillary GC column (1D), as in the classical GC. After leaving the first column, the eluent is not moved directly into the detector, but to the modulator, where it is focused in small fractions, at regular short intervals, in a period less than that of a primary peak width, according to the modulation ratio employed. Subsequently, the collected material is injected onto a second column for the fast second dimension (2D) separation (Figure 1.13). After the injection of the sample, the collection of another fraction in the modulator restarts. At the same time, the previous components of the fractions are separated by the second column. The mechanism of this separation is independent of the separation on the first column. It is then possible to separate the components that co-eluted at the end of the first column. Compounds leaving the second column are directed to the detector, resulting in a record of a series of very short (*e.g.*, 3s) chromatograms. The process of collecting fractions and dosing the sample to the 2D column is repeated until the end of the analysis. This allows the separation of peaks to be achieved in the first column, simultaneously providing additional separation of analytes on the second column.

Chromatographic columns in both dimensions should be characterized by different retention mechanisms and different selectivity. For a non-polar/polar (NP/P) combination, co-eluting compounds may consist of similar boiling points at the time of elution but have different 'polarity.' The 2D separation is based on the differences in activity coefficients between the solutes and the polar phase, allowing compounds of different polarity to be separated. In addition to the increased peak capacity, the advantage of the approach is that the extra separation dimension reduces the overload of the eluted peaks, thereby allowing the resolution of lower abundant species.

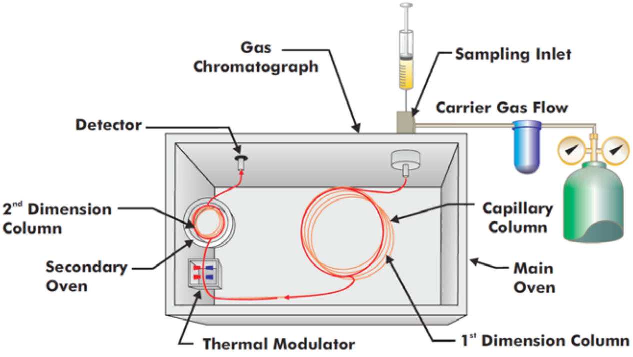


Figure 1.13. Diagram of the GC×GC system (*de.leco-europe.com*).

The chromatograms obtained using GC×GC contain much more information than 1D, allowing more straightforward and accurate identification of unknown substances. Consequently, it requires a very efficient detector. The detector must be able to collect data at a frequency of, at least, 50 Hz. To reproduce accurately the shape of a single peak, the peak should be recorded at least ten times, and peaks eluted from the second column have a minimal base width of 150-400 ms. The choice of chromatographic detectors that could satisfy this criterion is insufficient. The most commonly used detectors in GC×GC chromatography are FID (max. 300 Hz), μ-ECD (50-100 Hz) and TOF (max. 500 Hz).

In conclusion, a retention plane has higher peak capacity than a retention axis and, therefore, can accommodate and resolve highly complex mixtures associated with, for instance, profiling of complex biological samples and metabolite profiling.. In general, the usage possibilities of GC×GC can be divided into three areas: fingerprints of very complex matrices, target analyses, and identification of unknown compounds. An example of a 2D GC chromatogram is shown below in Figure 1.14.

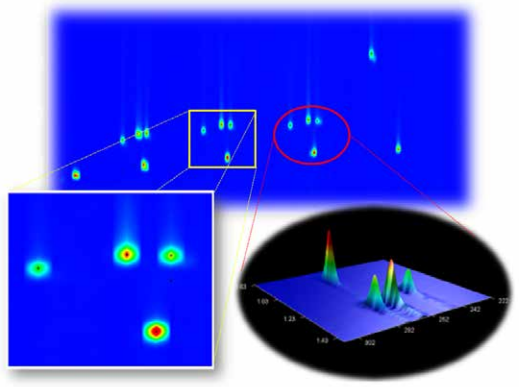
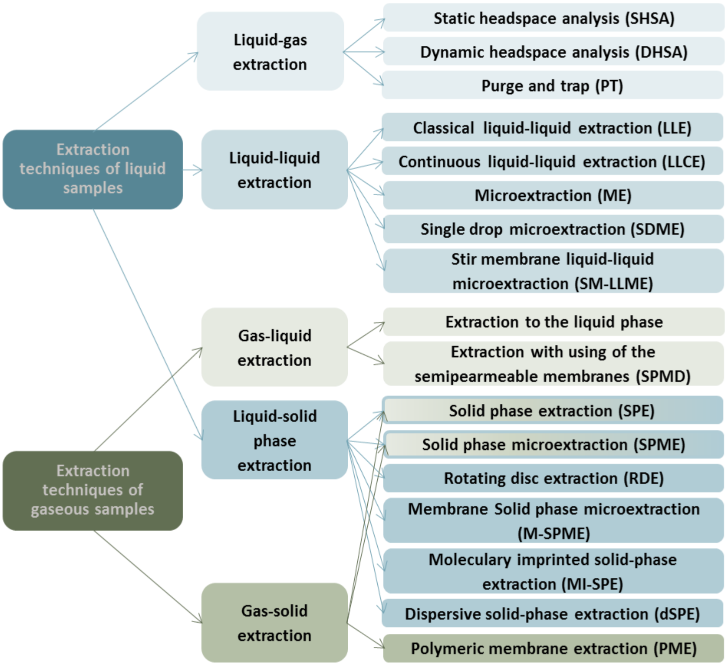


Figure 1.14. Chromatogram of the mixture of volatile compounds (alkanes, alcohols, aldehydes, ketones, amines) analyzed in 2D GC [Pegasus 4D; Leco].

2.2.2.4. Sample pre-treatment and derivatization

It is the most important and challenging step in the GC analysis, which is decisive for the accuracy of the result, particularly for samples from complex matrixes. Sample pre-treatment for GC analysis usually includes the following steps: cleaning, changing the matrix, extraction, derivatization, and enhancement of concentration. The obvious challenges of GC and GC-MS lie in the sample content of non-volatile components and the sample cleanup, which is associated with enrichment of trace ingredients. There is no single, universal method for sample preparation, which would be appropriate for all the analyzed materials. The sample preparation depends on many factors, mainly whether it is gaseous, liquid or solid.

Different extraction methods of volatile and semi-volatile compounds from biological material are employed before GC analysis: classical liquid-liquid extraction (LLE), continuous liquid-liquid extraction (LLCE), supercritical fluid extraction (SFE), pressurized liquid extraction (PLE), microwave assisted extraction (MAE), and ultrasound-assisted extraction (USE) (Figures 1.15 and 1.16), and most of them have been fully automated. Among those, solid phase extraction (SPE) allows a significant reduction of the time of sample preparation, along with the reduced volume of used solvents and of hazardous waste. In this method, analytes are extracted in a liquid-solid system, according to the distribution of the analyte between the liquid sample and the solid sorbent.



Extraction techniques of liquid samples
Extraction techniques of gaseous samples
Liquid-gas extraction
Liquid-liquid extraction
Gas-liquid extraction
Liquid-solid phase extraction
Gas-solid extraction
Static headspace analysis (SHSA)
Dynamic headspace analysis (DHSA)
Purge and trap (PT)
Classical liquid-liquid extraction (LLE)
Continuous liquid-liquid extraction (LLCE)
Microextraction (ME)
Single drop microextraction (SDME)
Stir membrane liquid-liquid microextraction (SM-LLME)
Extraction to the liquid phase
Extraction with using of the semipearmeable membranes (SPMD)
Solid phase extraction (SPE)
Solid phase microextraction (SPME)
Rotating disc extraction (RDE)
Membrane Solid phase microextraction (M-SPME)
Moleculary imprinted solid-phase extraction (MI-SPE)
Dispersive solid-phase extraction (dSPE)
Polymeric membrane extraction (PME)

Figure 1.15. Extraction methods of liquid and gaseous samples.

The principle of separation by SPE depends mainly on the nature of the sorbent. The interactions between the analyte and the polar solid adsorbents (*e.g.*, silica gel, aluminum oxide) are determined by hydrogen bonds as well as dipole-dipole, induced dipole-dipole and dispersion forces (van der Walls forces). This type of extraction is based on the same principles as adsorption chromatography. When the sorbent is silica with a chemically bonded polar group, *e.g.*, amino group (normal phase system), or non-polar, *e.g.*, octadecyl group (reversed-phase), the separation principle is the same as in partition chromatography. The alternatives to SPE are microextraction techniques, namely SPME (solid phase microextraction) and LPME (liquid phase microextraction). SPME extraction involves sorption of micro quantities of organic compounds on a thin, cylindrical layer of the stationary phase coating glass or quartz fiber. The LPME, in turn, is a solvent-minimized sample pre-treatment procedure, in which only several microliters of solvent are required. An essential advantage of microextraction methods is the low limit of detection (5-50 ppt) and short time of sample preparation (*ca.* 10 minutes). For the analysis of volatile compounds, headspace (HS) and purge and trap (PT) techniques are also used.

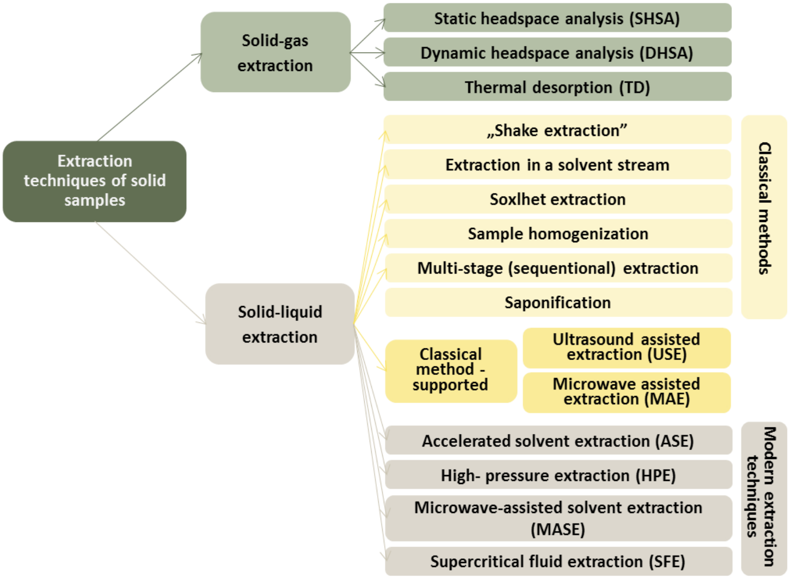

Extraction techniques of solid samples
Solid-gas extraction
Static headspace analysis (SHSA)
Dynamic headspace analysis (DHSA)
Thermal desorption (TD)
Solid-liquid extraction
„Shake extraction"
Extraction in a solvent stream
Soxlhet extraction
Sample homogenization
Multi-stage (sequentional) extraction
Saponification
Classical methods
Classical method - supported
Ultrasound assisted extraction (USE)
Microwave assisted extraction (MAE)
Accelerated solvent extraction (ASE)
High- pressure extraction (HPE)
Microwave-assisted solvent extraction (MASE)
Supercritical fluid extraction (SFE)
Modern extraction techniques

Figure 1.16. Extraction methods of solid samples.

The direct analysis of mixtures of compounds in GC is complicated because of interactions between the compounds or between the compounds and the GC column stationary phase. These interactions can lead to poor peak resolution and peaks tailing, which make proper peak identification impractical or burdensome. Conversion to derivative products will reduce the interactions interfering with the analysis. Derivatization is also used to change the volatility, to increase the thermal stability, to avoid degradation during the chromatographic process, to increase the sensitivity or specificity of the assay by introducing appropriate functional groups or by blocking the functional groups of the analytes. Organic compounds are used for the derivatization of analytes, usually by performing alkylation, acylation or silylation reactions on compounds containing functional groups with active hydrogens, *e.g.*, -COOH, -OH, -NH and –SH (Table 1.7).

Silylating reagents are the most commonly used in GC. These reagents include, among others, trimethylchlorosilane (TMCS), bistrimethylsilyltrifluoro-acetamide (BSTFA), N-methyl-trimethylsilyltrifluoroacetamide (MSTFA), N-methyl-N-t-butyldimethylsilyltrifluoro-acetamide (MTBSTFA). Compounds containing several different functional groups may need multiple derivatization steps.

Table 1.7. Derivatization for GC.

Procedure	Functional group - Compound type	Derivative	Reagent
Silylation	**-OH** -alcohols, phenols	Trimethylsilyl ethers	Bistrimethylsilylacetamide (BSA)
	-CO -ketones, oximes, steroids		Bistrimethylsilyltrifluoroacetamide (BSTFA)
	-COOH -amino acids, fatty		Hexamethyldisilzane (HMDS)
	acids, cannabinols, steroids		N- methyl-N-t-butyldimethylsilyl-trifluoroacetamide
	-(CH_2OH)n -sugars		(MTBSTFA)
	-NH, -NH_2 -amines, urea		N- methyltrimethylsilyltrifluoroacetamide (MSTFA)
	nucleosides,		Trimethylchlorosilane (TMCS)
	-SH -mercaptans		Trimethylsilyldiethylamine (TMS -DEA)
	-CONH, -$CONH_2$ -imides,	Trimethylsilyl amides	Trimethylsilylimidazole (TMSI)
	proteins		Halo- methylsilyl reagents
Alkylation	**-OH** -alcohols, phenols	Methyl esters (DMF)	Benzylbromide
	-CO -aldehydes	Trifluoroacetates (TFAA)	Boron trifluoride (BF_3) in methanol or butanol
	-COOH -amino acids, fatty	Methyl esters (BF_3-methanol)	Diazomethane (N_2CH_2)
	acids		Dimethylformamide (DMF)
	-NH, -NH_2 -amines, amino	Pentafluorobenzyl ethers (PFBBr)	Pentafluorobenzyl bromide (PFBBr)
	sugars		Pentafluorobenzyl- hydroxylamine hydrochloride (PFBHA)
	-CONH –amides	Methyl amides (TMAH)	Tetrabutylammonium hydroxide (TBH)
	-SH -mercaptans	Methyl esters (DMF)	Trifluoroacetic anhydride (TFAA)
			Trimethylanilinum hydroxide (TMAH)
Acylation	**-OH** -alcohols, phenols	Pentafluoropropionates (PFPA)	Heptafluorobutyric anhydride (HFBA)
	-(CH_2OH)n -sugars	Trifluoroacetamides (TFAI)	N-Methyl-bis(trifluoroacetamide) (MBTFA)
	-NH, -NH_2 -amines	Trifluoroacetamides (MBTFA)	Pentafluorobenzoyl chloride (PFBCl)
	-CONH -amides	Trifluoroacetamides (TFAA)	Pentafluoropropanol (PFPOH)
	-SH -mercaptans	Trimethylsilyl ethers (MBTFA)	Pentafluoropropanylimidazole (PFPI)
			Pentafluoropropionic anhydride (PFPA)
			Trifluoroacetic anhydride (TFAA)
			Trifluoroacetylimidazole (TFAI)

2.2.3. Coupling GC-MS

In GC-MS, the essence of the analysis depends on the process of ionization of the molecule, with the following fragmentation. The most commonly used ionization method in GC-MS is electron ionization (EI), already described in section 1.2. To obtain reproducible MS spectra, an EI energy of 70 eV is used. In the EI source, the molecules of analyte lose valence electrons to generate radical cations, and then undergo fragmentation according to some fragmentation rules:

$$M + e^- \rightarrow M^{+\bullet} - 2e^-$$

Besides the conventional EI ion source for GC-MS, the chemical ionization (CI) ion source can also be used. This ionization method is relatively mild, and usually, the molecular ion of the analytes can be observed. However, it is less universal, and fewer compounds ionize when using this ionization method, so its use is not so widespread. Ions formed in the ionization chamber are extracted by a series of electrodes that focus the ions and accelerate them to a mass analyzer.

As mass analyzers separate charged ions according to their *m/z* ratio, recording the mass and abundance of several ions, many types of mass analyzers may be used in the GC-MS equipment, including tandem mass spectrometers, as discussed earlier. GC-MS is an essential tool for metabolite identification and quantification because of reproducible molecular fragmentation patterns produced. Even though the application of GC is limited to volatile compounds, a significant portion of small molecular metabolites is within the range of GC separation. However, since most metabolites have polar functional groups, their peak shapes and recoveries are often unacceptable due to column absorption. Therefore, protection of those functional groups from metabolites with chemical derivatization is usually necessary.

To know more:

H.J. Hübschmann. Handbook of GC-MS: Fundamentals and Applications (3rd Edition). Somerset, NJ, USA: Wiley, 2015. ProQuest ebrary. Web. 8, 2015.

X. Guo. Advances in Gas Chromatography, Chapter 4: "Gas Chromatography in Metabolomics Study" by Y. Qiu and D. Reed. InTech, 2014.

2.3. Capillary electrophoresis-mass spectrometry (CE-MS)

2.3.1. Introduction to CE-MS

Capillary electrophoresis coupled to Mass Spectrometry, CE-MS, is the result of the hyphenation of high efficacy and resolution separation technique to the powerful capabilities of mass spectrometry and it is considered an orthogonal technique to GC or LC. The technique was introduced in the late eighties, and it represents just the newest coupling that can give relevant information about the composition of small compounds such as polar and charged metabolites. As a result of that, it has been used in various research fields in metabolomics such as biomedical, clinical and plant metabolomics. Samples of blood (serum and plasma), urine and other biofluids, cell cultures, tissues, and plants can be analysed by CE-MS with minimal sample treatment. Many different types of analytes including amino acids, short chain organic acids (SCOA), fatty acids derivatives (acyl-carnitines), nucleotides, sugar phosphates, nucleotides, and compounds related to polyamine metabolism, are frequently analyzed by CE-MS. Moreover, small peptides and proteins (after hydrolysis or native proteins) also can be analyzed.

Advantages of CE-MS

CE-MS is well suited for aqueous samples. Moreover, it is very appropriated for samples with high concentration of salts such as culture media or cell extracts.

- Low sample volume is required. That is very important in many experiments when a minimal amount of sample is available (small animals, tears, cerebrospinal fluid and other similar fluids, etc.).
- Adequate for hydrophilic analytes.
- Very efficient separation and an orthogonal mechanism for chromatographic techniques.
- Once the analyte is detected, the capillary can be rinsed very easily and be conditioned prior to the next analysis.

- A green analytical technique. Consumption of reagents and organic solvents is very low and lower than required in LC.

Disadvantages of CE-MS

- Lower reproducibility and robustness of CE regarding migration times and detector response.
- Low sensitivity due to the amount of sample introduced in the system, around nL.
- It is considered the least robust hyphenation.
- No spectral libraries are available. Identification of unknown compounds based on in-house libraries & accurate mass databases.

2.3.2. Theoretical principles of CE-MS

The separation by capillary electrophoresis is based on an electro-driven separation inside narrow-bore silica capillaries typically of 50 μm ID & 100 cm in length. The movement of analytes along the capillary filled with a conductive liquid medium under the influence of the electric field is based on the charge-to-mass ratio, and that gives this technique its unique properties, with their strengths and drawbacks. Bare fuse silica as well as coated silica (neutral or positive coatings) are usually used for different applications although the former is more common.

Migration times instead of retention times and electropherograms (plot of detector response versus migration times) are terms commonly used. The plot of the detector response versus migration time is called electropherogram.

PRINCIPLES OF SEPARATION

Two forces regulate the CE analysis: Electroosmotic flow (EOF) and Electrophoretic mobility (EM). The electroosmotic flow is the consequence of the surface charge on the inner wall of the capillary. With bare capillaries, the silanol groups (Si-OH) on the surface are partially ionised to negatively charged silanoate groups (Si-O$^-$) at pH>2. These negative charges attract the positive charges of the buffer in the capillary forming two inner layers as shown in Figure 1.17. The "fixed layer" is composed of positive charges closer to negative ions while the "mobile layer" is comprised of more dispersed charges.. When the electrical field is applied, the mobile layer of cations is pulled towards the negative electrode, the cations are solvated, and they drag all the buffer solution with them. The result is a net flow of buffer solution in the direction of the negative electrode causing the electroosmotic flow. This type of analysis is called *normal capillary electrophoresis.* There is, also, another mode called *reverse capillary electrophoresis,* in which the polarity of the electrodes is swaped. In this case, the inlet electrode is the cathode, and the flow should be suppressed or reversed, for example by the addition of quaternary amines, or by using positive charge coated capillaries.

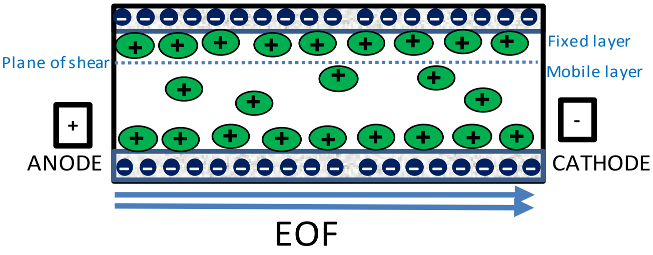


Figure 1.17. Representation of electroosmotic flow in a capillary.

Between the two layers, there is a *plane of shear* that is characterized by a potential difference, called zeta potential (ζ), due to an electrical imbalance of the charges. The zeta potential is proportional to the electroosmotic flow and the thickness of the double layer and inversely proportional to the dielectric constant of the buffer as described by the following equation:

$$\zeta = 4\pi\delta e/\varepsilon$$

where δ is the thickness of the diffuse double layer, *e* is the charge per unit surface area, and ε is the dielectric constant of the buffer.

Different parameters influence the EOF velocity such as the *pH of the buffer*, *the voltage*, and the *temperature*. An increase of *buffer pH* produces a rise in EOF because of the dissociation of Si-OH to Si-O   on the inner capillary wall that increases the zeta potential (ζ). The *electric field* (voltage per length unit) has a significant effect on the electroosmotic flow velocity. The higher the *electric field*, the higher the speed of the electroosmotic and electrophoretic flow and producing shorter analysis time. However, a high voltage can produce the Joule heating that increases the *temperature* in the capillary. It is important to control this parameter because high temperature can destroy and denature the samples and, besides, buffer viscosity gradient could be formed perpendicularly to the capillary thus increasing diffusion and dispersion. The Smoluchowski equation defines the electroosmotic flow velocity:

$$\nu_{eof} = \varepsilon\zeta E/4\pi\eta$$

where ν_{eof} is the electroosmotic flow velocity, ε is the dielectric constant of the buffer, E is the applied electric field in volts/cm, η is the viscosity of the buffer, ζ is the zeta potential measured at the plane of shear.

The buffer has electroosmotic mobility (μ_{eof}) that depends on the buffer characteristics as the dielectric constant and viscosity, and it is independent of the applied electric field as described in the next equation:

$$\mu_{eof} = \varepsilon\zeta/4\pi\eta$$

μ_{eof} is the electroosmotic mobility of the buffer, ε is the dielectric constant, η is the viscosity of the buffer, ζ is the zeta potential measured at the plane of shear.

The electroosmotic flow profile is different from that of a fluid moving under pressure (as in HPLC). In the Figure 1.18, the first diagram shows the flat shape profile of the flow in CE that is due to the movement of all the charges toward the electrodes. The second diagram illustrates the parabolic profile of the flow in HPLC due to the friction from the column wall that slows the flow down closer to the sides of the column. The advantage of the flat flow profile is that all the solute molecules have the same velocity inside of the capillary, thus giving narrow peaks with high efficiency. In the parabolic flow profile in HPLC, the solute molecules in the center of the tube move faster than those closer to the wall of the column giving relatively broad peaks.

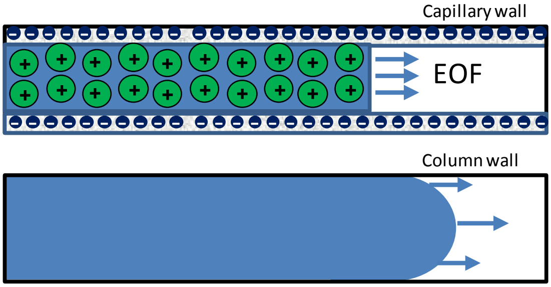


Figure 1.18. Diagrams of flow in CE and HPLC.

Electrophoretic mobility is an additional force that is present only in charged molecules and can drag the ions toward the electrode with the opposite charge. The velocity with which solute moves according to the charge and under an applied electric field is given by the equation:

$$\nu_{ep} = \mu_{ep}E$$

where ν_{ep} is the electrophoretic velocity of the solute, μ_{ep} is the electrophoretic mobility of the solute and E is an applied electrical field.

The electrophoretic mobility (μ_{ep}) is directly proportional to the charge of the solute and inversely proportional to the molecular size and also to the viscosity of the electrophoretic medium as described by the next equation:

$$\mu_{ep} = Q/6\pi r\eta$$

where μ_{ep} is the electrophoretic mobility of the solute, Q is the charge of the solute, r is the ionic radius of the solute, η is the viscosity of the medium.

The last equation proves that the reason why the small highly charged molecules move through the capillary faster than large molecules with a lower charge is their different electrophoretic mobility (μ_{ep}).

The total mobility of the solute (ν_{tot}) in the capillary is defined by the equation:

$$\nu_{tot} = \nu_{ep} + \nu_{eof}$$

In capillary electrophoresis with normal polarity, all molecules have the same electroosmotic flow velocity (ν_{eof}), as it is shown in the Figure 1.19, but they elute at different times. *Cations* elute first because they have electrophoretic velocity (ν_{ep}) headed to the same electrode of the electroosmotic flow, *neutral molecules* have the same velocity of the electroosmotic flow because they are not charged, and finally, *anions* elute last because the direction of electrophoretic velocity (ν_{ep}) is towards the opposite electrode (anode).

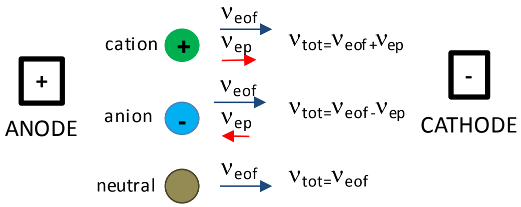


Figure 1.19. Total mobility of the solute in the capillary.

Numerous factors influence on the electrophoretic separations: pH of the separation buffer that influences the charge and the radius of the solvated molecules, the concentration of the buffer as EOF decreases with the square root of the buffer concentration, the viscosity of the separation medium, the temperature, the applied electrical field, etc.

Different separation modes can be distinguished in CE: capillary zone electrophoresis (CZE), micellar electrokinetic capillary chromatography (MEKC), capillary electrochromatography (CEC), capillary gel electrophoresis (CGE), among others. CZE is suitable for polar and charged compounds and it is the most common mode of CE mostly when it is coupled with MS. CE is also suitable for neutral compounds using micelles in the background electrolyte (MECK). Other additives can modify the selectivity of the separation, for example, cyclodextrins for chiral separations.

2.3.2.1. Capillary zone electrophoresis

It is the most straightforward mode and the most widespread mode, but only charge compounds can be separated. The previous section dealt almost entirely with this form. Analytes move from one end of the capillary to the other according to the vector sum of electrophoresis and electroosmotic mobility.

2.3.2.2. Micellar electrokinetic capillary chromatography

Introduced by Terabe in 1984 for separation of neutral compounds instead of CZE. MEKC includes suitable charged detergents in the separation buffer in a concentration sufficiently high to form micelles. These detergents could be cationic, anionic, non-ionic or zwitterionic compounds but when coupled with CE-MS, they must be volatile. These macromolecular aggregates with hydrophobic inner core and at polar surface can interact with the analyte by another mechanism such as partition and columbic forces. More hydrophobic compounds interact more strongly with the micelle and their speed through the capillary is lower. When not interacting with the micelle, the analyte molecule will migrate with the EOF. In all cases, variations in buffer concentration, pH, temperature, or use of additives such as urea, metal ions, or chiral selectors can also be used to affect selectivity. As well as in CZE, modifiers, such as methanol, acetonitrile, and 2-propanol have all been used successfully.

2.3.2.3. Capillary electrochromatography

It is a form of miniaturized liquid chromatography. Capillaries are filled with HPLC type silica-based reversed phase particles, 1-5 µm in size, as a stationary phase. Besides, conventional mobile phases for RP-type separations, such as organic solvent/aqueous buffer mixtures, are used. A mechanism of partitioning of the solutes between mobile and stationary phase is involved, and with charged metabolites, additional electrophoretic velocity will modify the separation. An additional pump for pushing the mobile phase through the filled capillary is required in the system. The generated EOF in the particle packed capillary has a velocity flow profile identical to that in an open-tubular capillary (CZE).

2.3.2.4. Capillary gel electrophoresis

It is widely employed in molecular biology research for molecules only differentiated by size and not by their mass/charge ratio such as proteins and nucleic acids DNA. The capillary is filled with a gel or viscous solution. Polymers in CGE can be covalently crosslinked, hydrogen bonded or just dissolved in buffer solutions EOF is often suppressed so that the migration of the analytes is solely by electrophoresis. In a viscous or polymer network medium, the movement of the analytes is hindered, as longer molecules are retarded more than shorter molecules. For obtaining more uniform charge-mass ratio with proteins, it is common to perform a pretreatment with the SDS detergent. Typically, proteins bind to a constant number of SDS molecules per unit length similarly to SDS-polyacrylamide gel electrophoresis.

2.3.3. Instrumentation

The capillary electrophoresis system consists of a capillary, inlet vial, outlet vial, sample vial, a cooling system able to eliminate the heat produced inside the capillary, two electrodes (an anode and a cathode), a power supply and the detector, as it is shown in Figure 1.20. The capillary, the inlet vial and the outlet vial are filled with the same buffer, which usually consists of an aqueous solution. The anode and cathode are usually a palladium electrode, connected with the power supply, and the detector is usually placed near the end of the capillary (close to the outlet vial).

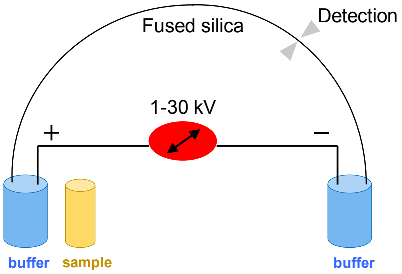


Figure 1.20. The scheme of capillary electrophoresis.

A fused silica capillary covered with polyimide coating is used for CZE in CE-MS. This capillary can be a bare capillary or the inner wall can be coated with polymers (neutral or positive) to reduce electroosmosis to very low levels, or even reverse it. The capillary is placed into the two vials that also contain the electrode, and the sample is loaded. Loading occurs by replacing one of the vials with sample vial and injecting the sample by using gravity, applying an external pressure (hydrodynamic) or applying an electric field (electrokinetic injection). Subsequently, the capillary is placed into the buffer vial, and the electric field is applied for the separation of the components. Different cooling systems are available that are able to reduce the heat produced inside the capillary due to Joule effect while applying the potential.

The most widely used detection systems in capillary electrophoresis are Ultraviolet/Visible detector (UV-VIS), Photodiode Array Detector (PDA), Laser-Induced Fluorescence (LIF) and Mass Spectrometry (MS) (Table 1.8). By far, the highest capabilities are demonstrated by CE-MS, as information on the migration time under the analysis conditions and the mass spectrum of the compounds reaching the detector can be obtained. When using mass spectrometric detectors, the most common ionization source is electrospray ionization ESI.

Table 1.8. Differences between detectors in CE.

Detector	**Advantages**	**Characteristics**	**Limit of Detection**
UV/vis absorbance	• The possibility of direct and indirect detection • Very common detector	- universal	**-10^{-3}** **-10^{-6} for the detection of aromatic compounds**
LIF (Laser-induced fluorescence)	• Highly sensitive and highly selective • Used for fluorescent compounds or derivatives	- selective	**-10^{-6} - 10^{-9}**
MS	• Qualitative and quantitative information • Highly sensitive and highly selective	- universal - selective	**-≈10^{-5} (it depends on the type of MS and metabolites)**

When combined with other detection techniques, such as UV or fluorescence, the inlet, and outlet of the capillary are immersed in vials with the background electrolyte (BGE). However, when using MS as a detector, the outlet vial is replaced by the entrance to the MS. In this case, it is necessary to use an interface that matches the two voltages and keep a stable spray with a very low volume of fluid (Figure 1.21). In CE-ESI-MS, the volume introduced in the capillary is measured in nanoliters and, to maintain a constant flow and spray, it is common to use a *sheath flow or sheath liquid (SL)*. The sheath flow, added in the range of nl/min to ml/min, allows a stable spray and ion formation and is usually composed of a 1:1 mixture of water-methanol with 0.1% acetic acid or formic acid. Achieving stable electrospray operation often depends on balancing of multiple parameters such as capillary position, sheath liquid flow rate and composition, gas sheath flow rate, and ESI conditions. Despite the numerous developments of sheath-less interfaces with proved higher sensitivity, robustness is still compromised.

Regarding the mass analyzer, all commercially available instruments can be connected to CE, being time-of-flight (TOF) analyzer instruments those with the highest number of published applications. This is mainly because of the high scanning speed (duty cycle) necessary to obtain enough number of points along the peak, but also because of the accurate mass and high resolution of these instruments.

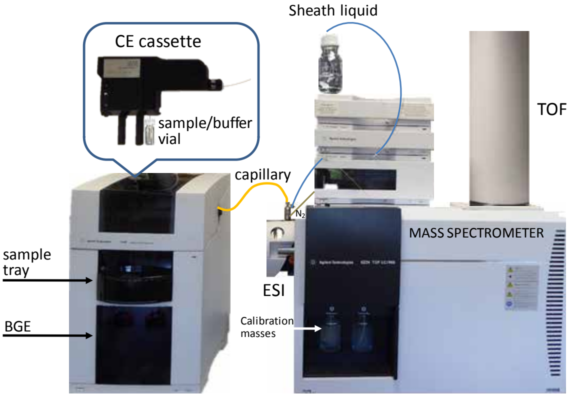


Figure 1.21. Graphical representation of CE coupled to MS, detailing the most essential parts in both pieces of equipment.

To know more:

D.R. Baker. Capillary electrophoresis. Wiley, 1995.

S. Terabe, K. Otsuka, T. Ando. Electrokinetic chromatography with micellar solution and open-tubular capillary. *Anal. Chem.*, 1985, 57 (4):834-841.

T. Soga, Y. Ohashi, Y. Ueno, H. Naraoka, M. Tomita, T. Nishioka. Quantitative metabolome analysis using capillary electrophoresis mass spectrometry. *J Proteome Res.*, 2003, 2(5):488-94.

A. García, S. Naz, C. Barbas. Metabolite fingerprinting by capillary electrophoresis-mass spectrometry. Methods Mol Biol, 2014, 1198:107-23.

A. García, J. Godzien, Á. López-Gonzálvez, C. Barbas. Capillary electrophoresis mass spectrometry as a tool for untargeted metabolomics based on GC-(EI)-Q-MS. *Anal. Chim. Acta*, 2015, 900:21-35.

3. Practical sessions

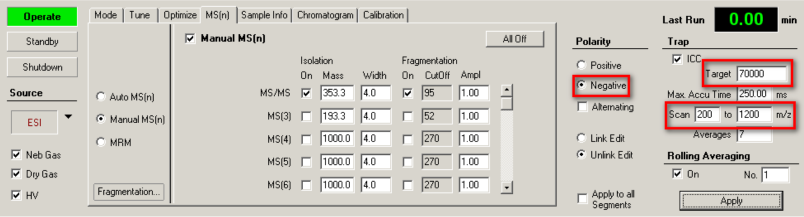
3.1. Acquisition of an ESI-MS mass spectrum

Following the presentation of the operating principles of an ion trap mass spectrometer with an ESI ion source, a solution of the sample will be analyzed using a direct injection infusion pump. A sample of rutin, a low molecular antioxidant with known molecular mass (or a solution of another commercially available standard) will be analyzed to obtain an MS spectrum. The acquisition will be performed both in the positive and negative mode to choose the best approach, depending on the rutin properties.

Additionally, following the presentation of the operating principles and basic construction of Orbitrap mass spectrometer with ESI ion source, a thioredoxin sample will be analyzed by direct infusion. This step allows introducing to students mass spectra of high-molecular-weight compounds, with the possibility to observe multiply charged ions. The obtained ESI-MS spectra will be recorded and used for interpretation in point 3.3.

ESI-MS Experiment:

1. Prepare a DMSO solution of rutin (molecular mass 610.52) at a concentration of 1mg/mL.
2. Dissolve with ACN/H_2O (60/40 V/V) to obtain 10ug/mL solution of rutin (in 1% DMSO).
3. Why is the solution of only 1% DMSO required?
4. Introduce the sample solution into ESI source with syringe pump set at a flow rate of 150µl/h .



Source: ESI

☑ Neb Gas
☑ Dry Gas
☑ HV

○ Auto MS(n)
● Manual MS(n)
○ MRM

☑ Manual MS(n)

	Isolation			Fragmentation		
	On	Mass	Width	On	CutOff	Ampl
MS/MS	☑	353.3	4.0	☑	95	1.00
MS(3)	□	193.3	4.0	□	52	1.00
MS(4)	□	1000.0	4.0	□	270	1.00
MS(5)	□	1000.0	4.0	□	270	1.00
MS(6)	□	1000.0	4.0	□	270	1.00

Last Run 0.00 min

Polarity
○ Positive
● Negative
□ Alternating

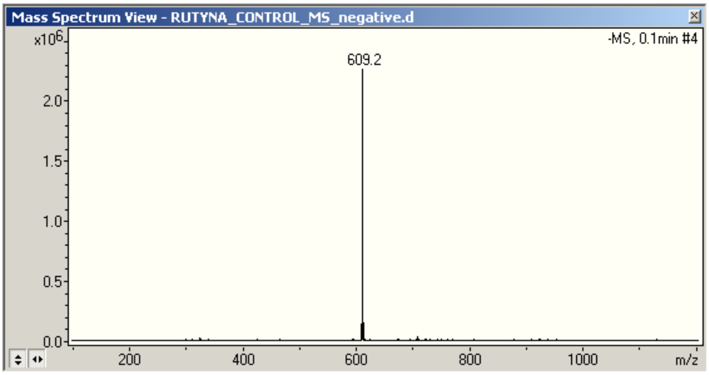
○ Link Edit
● Unlink Edit

□ Apply to all Segments

Trap
☑ ICC
Target 70000
Max. Accu Time 250.00 ms
Scan 200 to 1200 m/z
Averages 7

Rolling Averaging
☑ On No. 1

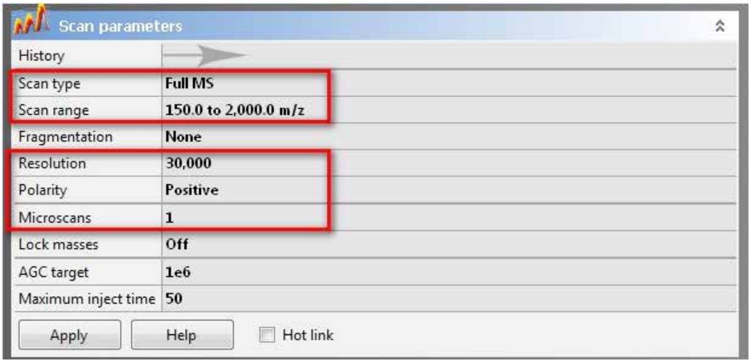
5. Record the mass spectrum during 30 s using the parameters given below on the figure.
6. Find rutin molecular ion and explain m/z value corresponding.



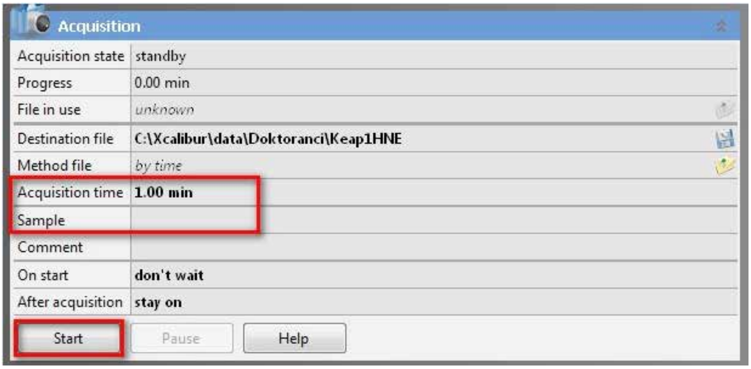
ESI-MS spectra of rutin in negative mode

ESI-OrbiTrap MS Experiment:

1. Prepare a solution of thioredoxin (molecular mass 12kDa) at a concentration of 10µM.
2. Dissolve with ACN/H_2O (50/50 V/V) + 0.1% FA to obtain a 10pM solution of thioredoxin.
3. Why is a solution of 0.1% FA required?
4. Introduce solution into ESI source with syringe pump set at 2µL/min flow rate.

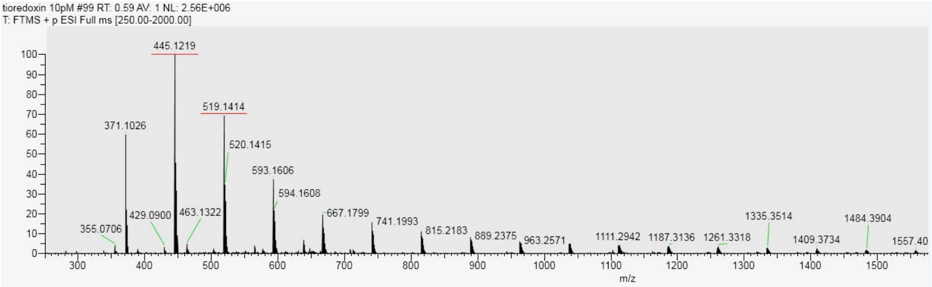


History	
Scan type	Full MS
Scan range	150.0 to 2,000.0 m/z
Fragmentation	None
Resolution	30,000
Polarity	Positive
Microscans	1
Lock masses	Off
AGC target	1e6
Maximum inject time	50



Acquisition state	standby
Progress	0.00 min
File in use	unknown
Destination file	C:\Xcalibur\data\Doktoranci\Keap1HNE
Method file	by time
Acquisition time	1.00 min
Sample	
Comment	
On start	don't wait
After acquisition	stay on

5. Record the mass spectrum during 60 sec. using the parameters given below in the figure.
6. Open the obtained results. Knowing the mass of the whole thioredoxin (12kDa), calculate the charge of each obtained ion.



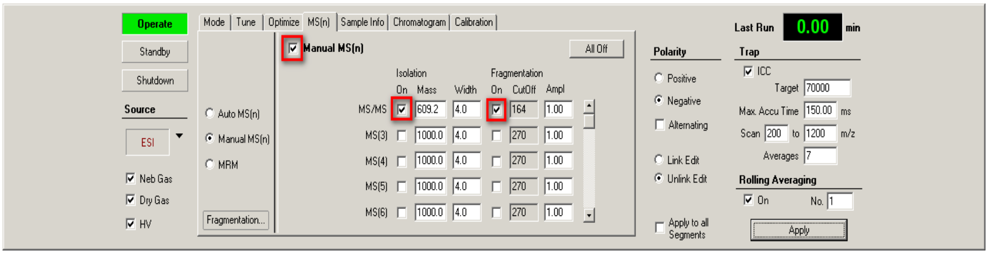
e.g.	ion 445.12	ion 519.14
	m/z = (M+z)/z	m/z = (M+z)/z
	445.12 = (12000+z)/z	519.14 = (12000+z)/z
	z=12000/444.12	z=12000/518.12
	z=27	z=23

3.2. Acquisition of an ESI-MS/MS mass spectrum

In this part of the practical session, the same rutin sample as used in point 3.1 will be analyzed by direct infusion ESI-MS/MS via ESI-ion trap system. In this step, students will learn which m/z ratios of ions (in respect to molecular ion and fragment ions) should be selected to be applied in typical quantitative approach with the use of MRM mode. Then, using specialist software, they will program the selection of a precursor ion, fragmentation of this ion, and measurement of the mass-to-charge ratio of the product ions formed. The ESI-MS/MS spectra will be recorded and used for further interpretation, described in point 3.3.

ESI-MS/MS Experiment:

1. Introduce solution of rutin from point 2 into ESI source with syringe pump set at 150µL/h flow rate.
2. Tick the isolation and fragmentation option in the MS(n) tab, type in m/z value corresponding to rutin molecular ion.



Operate | Standby | Shutdown

Source: ESI

☑ Neb Gas
☑ Dry Gas
☑ HV

Mode | Tune | Optimize | MS(n) | Sample Info | Chromatogram | Calibration

○ Auto MS(n)
● Manual MS(n)
○ MRM

Fragmentation...

☑ Manual MS(n)   All Off

	Isolation			Fragmentation		
	On	Mass	Width	On	CutOff	Ampl
MS/MS	☑	609.2	4.0	☑	164	1.00
MS(3)	□	1000.0	4.0	□	270	1.00
MS(4)	□	1000.0	4.0	□	270	1.00
MS(5)	□	1000.0	4.0	□	270	1.00
MS(6)	□	1000.0	4.0	□	270	1.00

Polarity
○ Positive
● Negative
□ Alternating

○ Link Edit
● Unlink Edit

□ Apply to all Segments

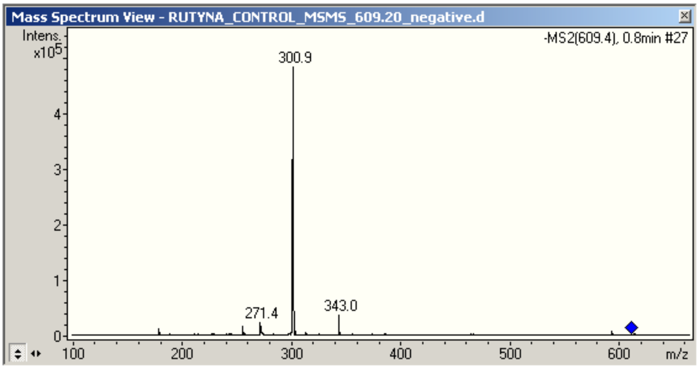
Last Run 0.00 min

Trap
☑ ICC
Target 70000
Max. Accu Time 150.00 ms
Scan 200 to 1200 m/z
Averages 7

Rolling Averaging
☑ On   No. 1

Apply

3. Record the mass spectrum during 30 s with the parameters given above in the figure.



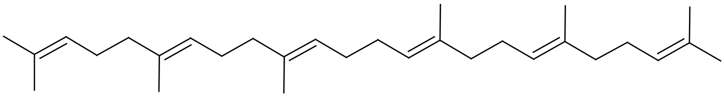
4. ESI-MS/MS spectra of rutin in negative mode.
5. Look carefully at the obtained ESI-MS/MS spectrum.
6. Which ion should be selected for quantitative analysis of rutin in MRM mode?
7. Why?

3.3. Interpretation of the mass spectrum

The students will observe the mass spectra (obtained in point 3.1 and 3.2) and identify the peaks corresponding to the m/z of the molecular ion and its characteristic fragment ions, after detailed examination of the structure of the analyzedcompounds. Moreover, the spectra will be critically investigated regarding the presence of possible adducts with sodium or potassium and multiply charged ions.

3.4. Observation of a GC-MS run and configuration of the selected ion monitoring acquisition

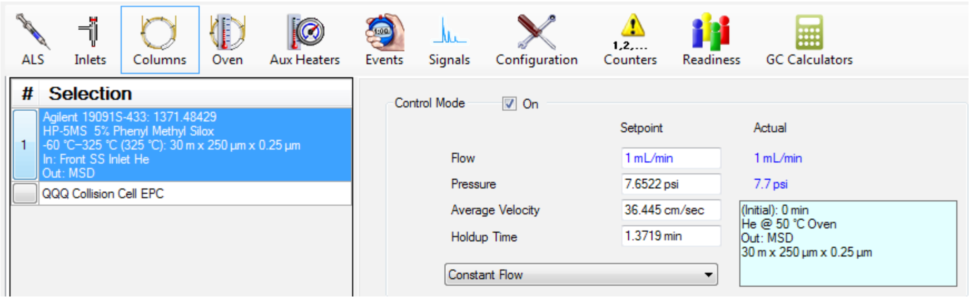
The aim of this exercise is to present to students a practical overview of the gas chromatography-mass spectrometry analytical method. Therefore, qualitative and quantitative determination of a typical nonvolatile compound (e.g., squalene as an ingredient of edible oil samples of amaranth oil, sea buckthorn oil, primrose oil) will be performed by using GC-SIM MS method.



Squalene – $C_{30}H_{50}$, all-trans-Squalene, (6*E*,10*E*,14*E*,18*E*)-2,6,10,15,19,23-hexamethyl-tetra-cosa-2,6,10,14,18,22-hexaene

GC-MS Experiment:

1. Prepare a series of squalene standards in concentration of: 0.5 mg /mL, 1.0 mg/mL, 2 mg /mL, 5 mg /mL and 10 mg /mL in hexane.
2. Transfer 10μL of an oil sample to 2mL GC vial.
3. Add 40 μL of the silylating reagent (BSTFA:TMCS; 99:1) into the vial.
4. Cap the vial and heat at 80^{0}C for 1 hour.
5. Let it cool down to room temperature.
6. Inject 1μL of the sample onto HP-5ms capillary column (0.25-mm; 0.25-μm, 30-m) of GC/MS system with 7890A GC – 7000 quadrupole MS/MS (Agilent Technologies, Palo Alto, CA).
7. Before running the samples, set up in the software the values of crucial parameters of GC-MS analysis as given below:



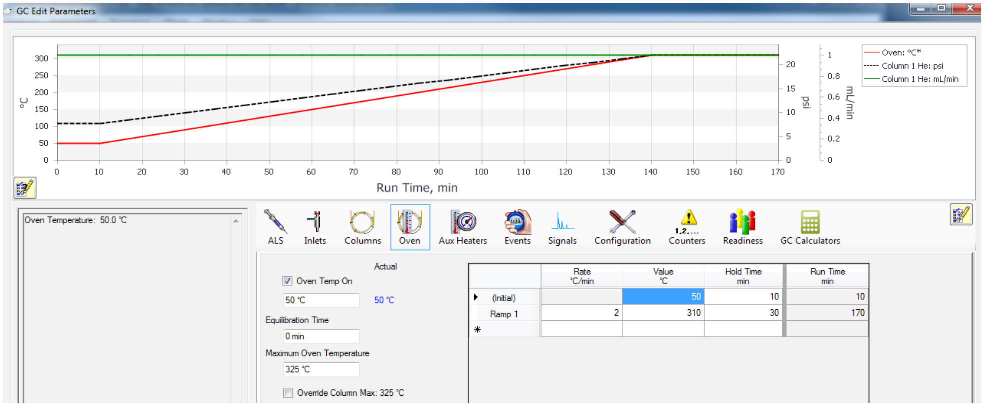
#	Selection
1	Agilent 19091S-433: 1371.48429 HP-5MS 5% Phenyl Methyl Silox -60 °C–325 °C (325 °C): 30 m x 250 µm x 0.25 µm In: Front SS Inlet He Out: MSD
	QQQ Collision Cell EPC

Control Mode ☑ On

	Setpoint	Actual
Flow	1 mL/min	1 mL/min
Pressure	7.6522 psi	7.7 psi
Average Velocity	36.445 cm/sec	(Initial): 0 min He @ 50 °C Oven Out: MSD 30 m x 250 µm x 0.25 µm
Holdup Time	1.3719 min	

Constant Flow

Oven programming: 50^{0}C (10 min), rate 2^{0}C/min to 310^{0}C (30min)



Oven Temperature: 50.0 °C

Actual

☑ Oven Temp On

50 °C 50 °C

Equilibration Time

0 min

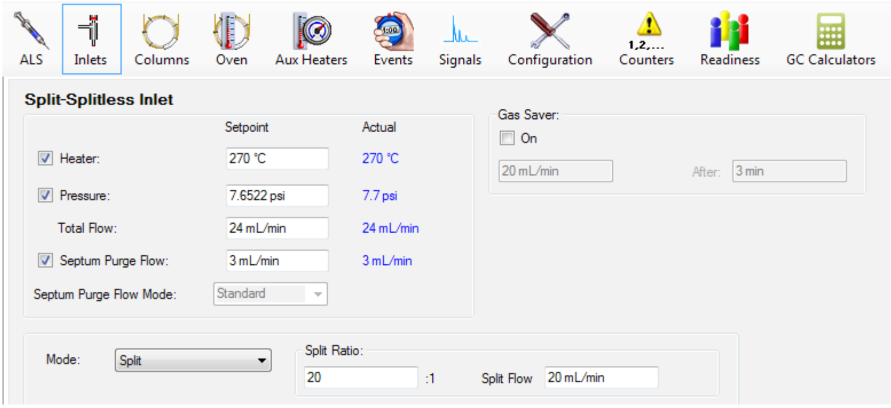
Maximum Oven Temperature

325 °C

□ Override Column Max: 325 °C

	Rate °C/min	Value °C	Hold Time min	Run Time min
(Initial)		50	10	10
Ramp 1	2	310	30	170
*				

Inlet (S/SL): temperature 270^{0}C, split – 20:1



Split-Splitless Inlet

	Setpoint	Actual
☑ Heater:	270 °C	270 °C
☑ Pressure:	7.6522 psi	7.7 psi
Total Flow:	24 mL/min	24 mL/min
☑ Septum Purge Flow:	3 mL/min	3 mL/min
Septum Purge Flow Mode:	Standard	

Gas Saver:

□ On

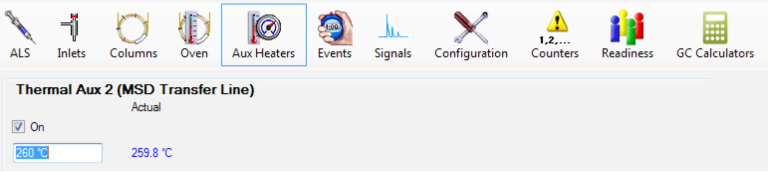
20 mL/min After: 3 min

Mode: Split

Split Ratio:

20 :1 Split Flow 20 mL/min

Transfer line temperature: 260^{0}C



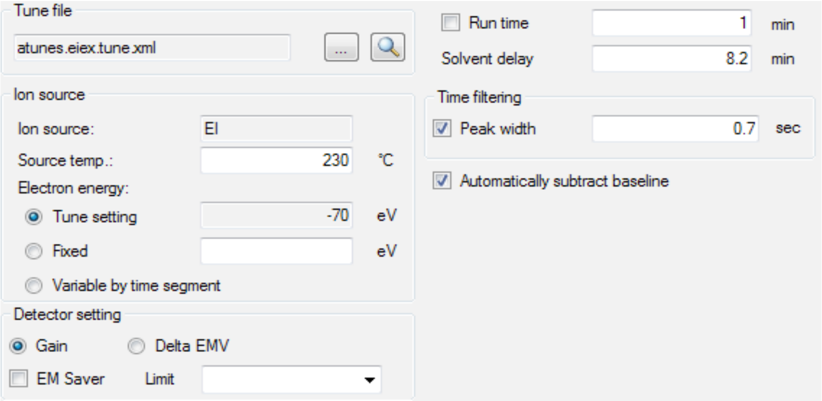
Thermal Aux 2 (MSD Transfer Line)

Actual

☑ On

260 °C 259.8 °C

Ion source (EI): 230^{0}C, Source energy : - 70eV, Solvent delay: 8.2 min



Tune file

atunes.eiex.tune.xml

Ion source

Ion source: EI

Source temp.: 230 °C

Electron energy:

● Tune setting -70 eV

○ Fixed eV

○ Variable by time segment

Detector setting

● Gain ○ Delta EMV

□ EM Saver Limit

□ Run time 1 min

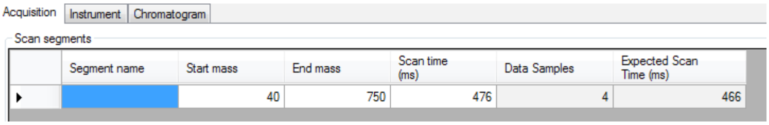
Solvent delay 8.2 min

Time filtering

☑ Peak width 0.7 sec

☑ Automatically subtract baseline

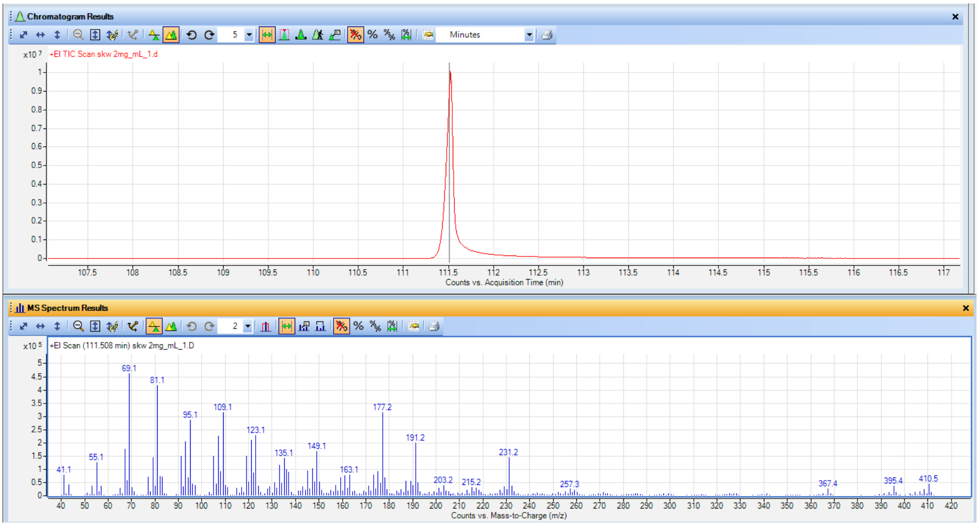
8. In the “MS-SIM/Scan Parameters” window chose “Scan mode’ to obtain a total ion chromatogram (TIC) in the range of m/z (40-750 m/z).



Scan segments

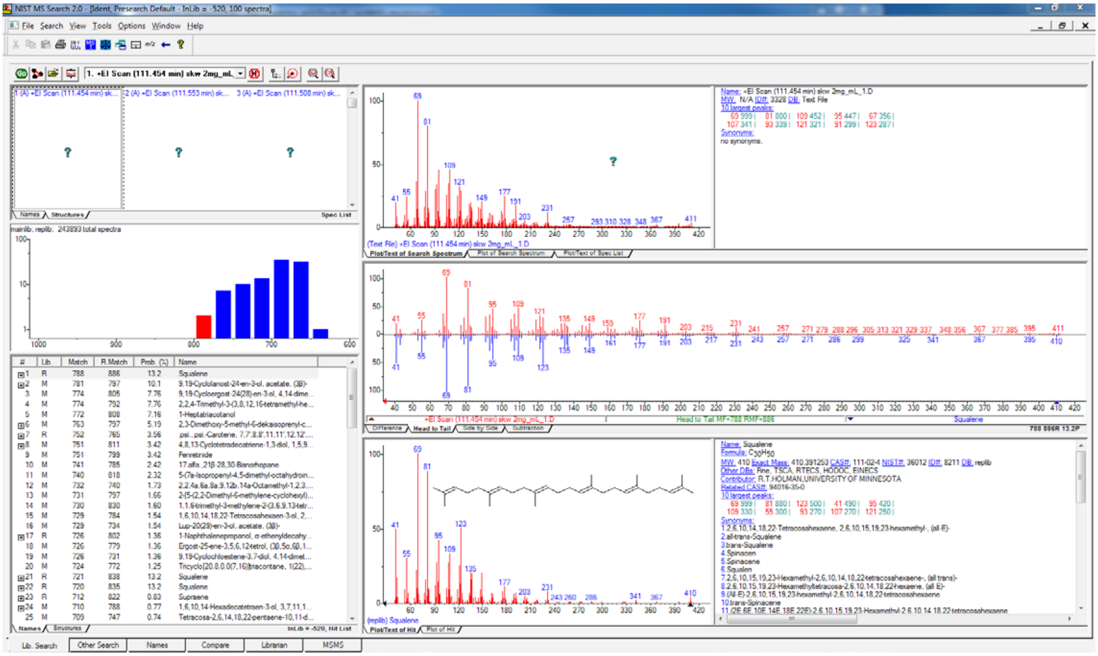
	Segment name	Start mass	End mass	Scan time (ms)	Data Samples	Expected Scan Time (ms)
▶		40	750	476	4	466

9. Perform qualitative analysis using the above scan method and record the chromatogram of the squalene standard at concentration of e.g. 2 mg/mL (in hexane).



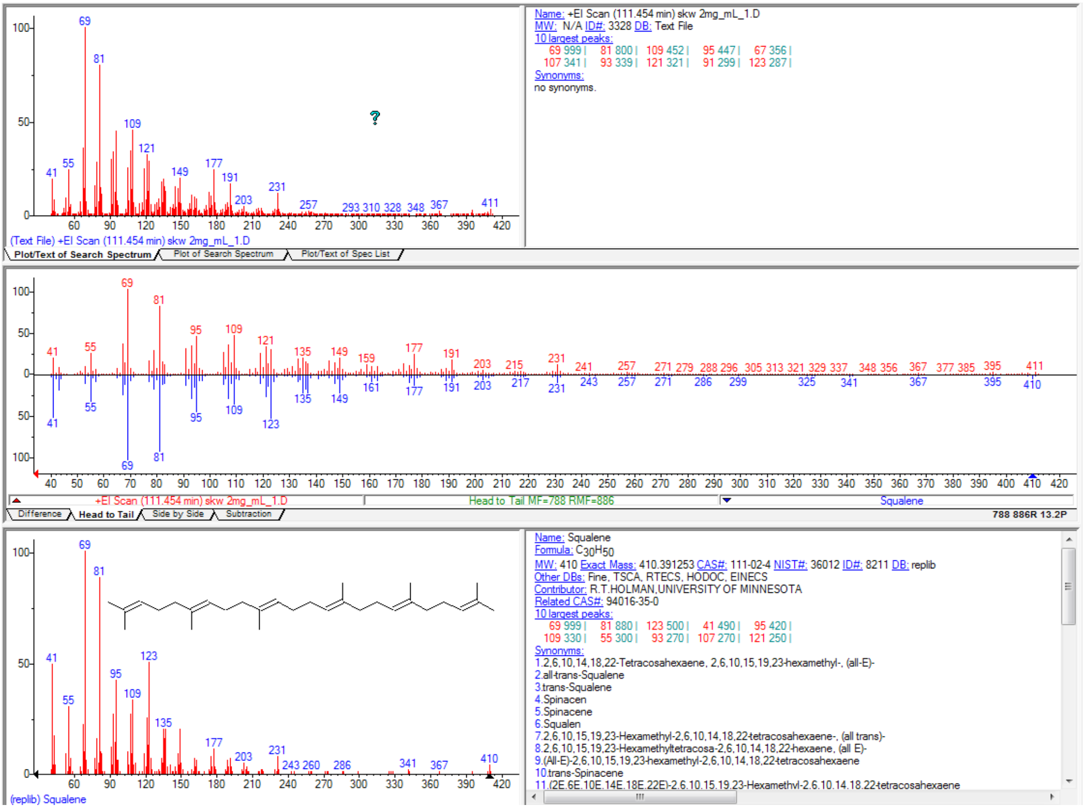
Chromatogram and mass spectrum of squalene standard.

10. Search the library (NIST MS) for the obtained spectrum; use the command "Search Using NIST Program".


NIST MS Search 2.0 - [Ident, Presearch Default - InLib = -520, 100 spectra]
File Search View Tools Options Window Help
1. +EI Scan (111.454 min) skw 2mg_mL_
Names Structures Spec List
Plot/Text of Search Spectrum Plot of Search Spectrum Plot/Text of Spec List
Difference Head to Tail Side by Side Subtraction
Plot/Text of Hit Plot of Hit
Lib. Search Other Search Names Compare Librarian MSMS
InLib = -520, Hit List

mainlib; replib: 243893 total spectra

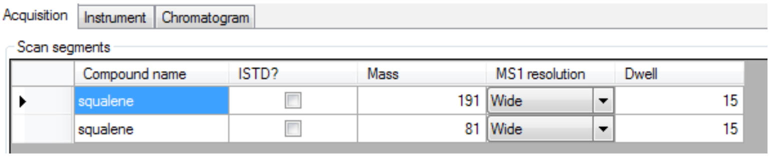
#	Lib.	Match	R.Match	Prob. (%)	Name
1	R	788	886	13.2	Squalene
2	M	781	797	10.1	9,19-Cyclolanost-24-en-3-ol, acetate, (3β)-
3	M	774	805	7.76	9,19-Cycloergost-24(28)-en-3-ol, 4,14-dime....
4	M	774	792	7.76	2,2,4-Trimethyl-3-(3,8,12,16-tetramethyl-he....
5	M	772	808	7.16	1-Heptatriacotanol
6	M	763	797	5.19	2,3-Dimethoxy-5-methyl-6-dekaisoprenyl-c....
7	R	752	765	3.56	.psi.,.psi.-Carotene, 7,7',8,8',11,11',12,12',....
8	M	751	811	3.42	4,8,13-Cyclotetradecatriene-1,3-diol, 1,5,9....
9	M	751	799	3.42	Fenretinide
10	M	741	785	2.42	17.alfa.,21β-28,30-Bisnorhopane
11	M	740	818	2.32	5-(7a-Isopropenyl-4,5-dimethyl-octahydroin....
12	M	732	740	1.73	2,2,4a,6a,8a,9,12b,14a-Octamethyl-1,2,3,....
13	M	731	797	1.66	2-[5-(2,2-Dimethyl-6-methylene-cyclohexyl)...
14	M	730	830	1.60	1,1,6-trimethyl-3-methylene-2-(3,6,9,13-tetr...
15	M	729	784	1.54	1,6,10,14,18,22-Tetracosahexaen-3-ol, 2,....
16	M	729	734	1.54	Lup-20(29)-en-3-ol, acetate, (3β)-
17	R	726	802	1.36	1-Naphthalenepropanol, α-ethenyldecahy....
18	M	726	779	1.36	Ergost-25-ene-3,5,6,12-tetrol, (3β,5α,6β,1....
19	M	726	731	1.36	9,19-Cyclochloestene-3,7-diol, 4,14-dimet....
20	M	724	772	1.25	Tricyclo[20.8.0.0(7,16)]triacontane, 1(22),....
21	R	721	838	13.2	Squalene
22	R	720	835	13.2	Squalene
23	R	712	822	0.83	Supraene
24	M	710	788	0.77	1,6,10,14-Hexadecatetraen-3-ol, 3,7,11,1....
25	M	709	747	0.74	Tetracosa-2,6,14,18,22-pentaene-10,11-d....


(Text File) +EI Scan (111.454 min) skw 2mg_mL_1.D
Plot/Text of Search Spectrum
Plot of Search Spectrum
Plot/Text of Spec List
+EI Scan (111.454 min) skw 2mg_mL_1.D
Head to Tail MF=788 RMF=886
Squalene
Difference
Head to Tail
Side by Side
Subtraction
788 886R 13.2P
(replib) Squalene

Name: +EI Scan (111.454 min) skw 2mg_mL_1.D
MW: N/A ID#: 3328 DB: Text File
10 largest peaks:
69 999 | 81 800 | 109 452 | 95 447 | 67 356 |
107 341 | 93 339 | 121 321 | 91 299 | 123 287 |
Synonyms:
no synonyms.

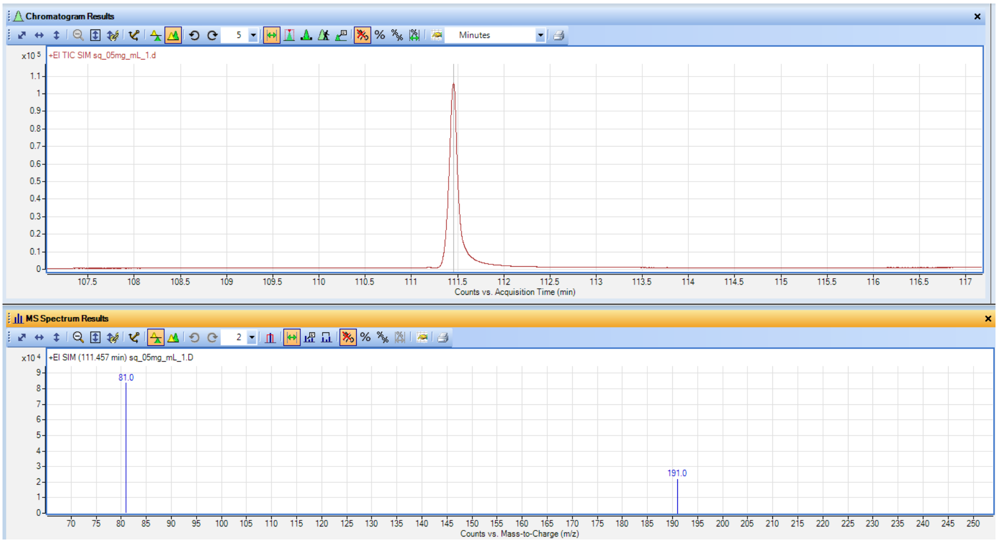
Name: Squalene
Formula: $C_{30}H_{50}$
MW: 410 Exact Mass: 410.391253 CAS#: 111-02-4 NIST#: 36012 ID#: 8211 DB: replib
Other DBs: Fine, TSCA, RTECS, HODOC, EINECS
Contributor: R.T.HOLMAN,UNIVERSITY OF MINNESOTA
Related CAS#: 94016-35-0
10 largest peaks:
69 999 | 81 880 | 123 500 | 41 490 | 95 420 |
109 330 | 55 300 | 93 270 | 107 270 | 121 250 |
Synonyms:
1.2,6,10,14,18,22-Tetracosahexaene, 2,6,10,15,19,23-hexamethyl-, (all-E)-
2.all-trans-Squalene
3.trans-Squalene
4.Spinacen
5.Spinacene
6.Squalen
7.2,6,10,15,19,23-Hexamethyl-2,6,10,14,18,22-tetracosahexaene-, (all trans)-
8.2,6,10,15,19,23-Hexamethyltetracosa-2,6,10,14,18,22-hexaene, (all E)-
9.(All-E)-2,6,10,15,19,23-hexamethyl-2,6,10,14,18,22-tetracosahexaene
10.trans-Spinacene
11.(2E,6E,10E,14E,18E,22E)-2,6,10,15,19,23-Hexamethyl-2,6,10,14,18,22-tetracosahexaene

11. For quantitative analysis, create MS-selected ion monitoring (MS-SIM) method by selecting, in the software, ions at m/z 191 and 81, corresponding to the characteristic fragments of squalene structure.



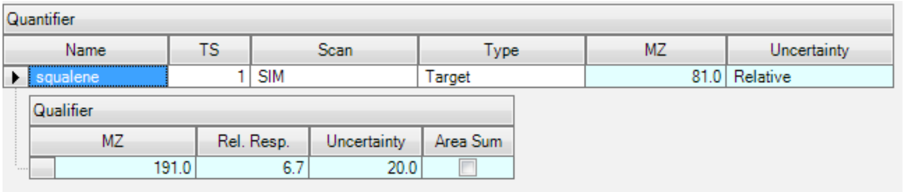
	Compound name	ISTD?	Mass	MS1 resolution	Dwell
▶	squalene	□	191	Wide	15
	squalene	□	81	Wide	15

12. Using the created SIM method, record the chromatogram of the squalene standard at a concentration of, e.g.: 0.5 mg/mL (in hexane).



Selected ion monitoring chromatogram and "SIM - mode spectrum" of squalene standard. An ion at m/z 191 is used as "qualifier ion," and an ion at m/z 81 is used as "quantifier ion" (the qualifier is used solely to determine an observed qualifier ratio).

13. In the next step, record the chromatograms of the squalene standards at concentration of: 0.5 mg/mL, 1.0 mg/mL, 2 mg/mL, 5 mg/mL and 10 mg/mL in hexane.

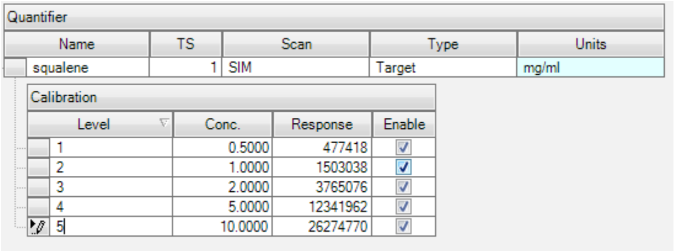

Quantifier

	Name	TS	Scan	Type	MZ	Uncertainty
▶	squalene	1	SIM	Target	81.0	Relative

Qualifier

	MZ	Rel. Resp.	Uncertainty	Area Sum
	191.0	6.7	20.0	□

14. Create calibration levels in the software.



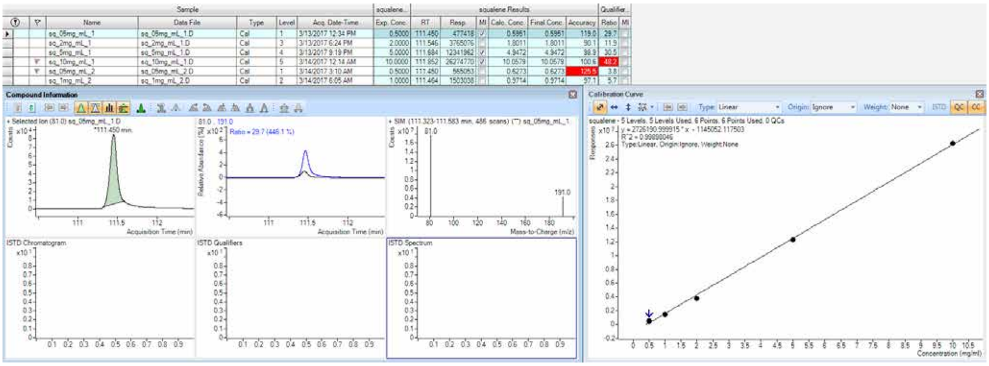
Quantifier

Name	TS	Scan	Type	Units
squalene	1	SIM	Target	mg/ml

Calibration

Level	Conc.	Response	Enable
1	0.5000	477418	☑
2	1.0000	1503038	☑
3	2.0000	3765076	☑
4	5.0000	12341962	☑
5	10.0000	26274770	☑

15. After applying this method to the batch (in Quantitative Analysis software part), generate the Calibration Curve using "Analyze Batch" command from the menu.



Sample							squalene...	squalene Results						Qualifier...	
		Name	Data File	Type	Level	Acq. Date-Time	Exp. Conc.	RT	Resp.	MI	Calc. Conc.	Final Conc.	Accuracy	Ratio	MI
▶		sq_05mg_mL_1	sq_05mg_mL_1.D	Cal	1	3/13/2017 12:34 PM	0.5000	111.450	477418	✓	0.5951	0.5951	119.0	29.7	□
		sq_2mg_mL_1	sq_2mg_mL_1.D	Cal	3	3/13/2017 6:24 PM	2.0000	111.546	3765076	□	1.8011	1.8011	90.1	11.9	□
		sq_5mg_mL_1	sq_5mg_mL_1.D	Cal	4	3/13/2017 9:19 PM	5.0000	111.684	12341962	✓	4.9472	4.9472	98.9	30.5	□
	▼	sq_10mg_mL_1	sq_10mg_mL_1.D	Cal	5	3/14/2017 12:14 AM	10.0000	111.852	26274770	✓	10.0579	10.0579	100.6	48.2	□
	▼	sq_05mg_mL_2	sq_05mg_mL_2.D	Cal	1	3/14/2017 3:10 AM	0.5000	111.450	565053	□	0.6273	0.6273	125.5	3.8	□
		sq_1mg_mL_2	sq_1mg_mL_2.D	Cal	2	3/14/2017 6:05 AM	1.0000	111.464	1503038	□	0.9714	0.9714	97.1	5.7	□

Calibration Curve

Type: Linear Origin: Ignore Weight: None

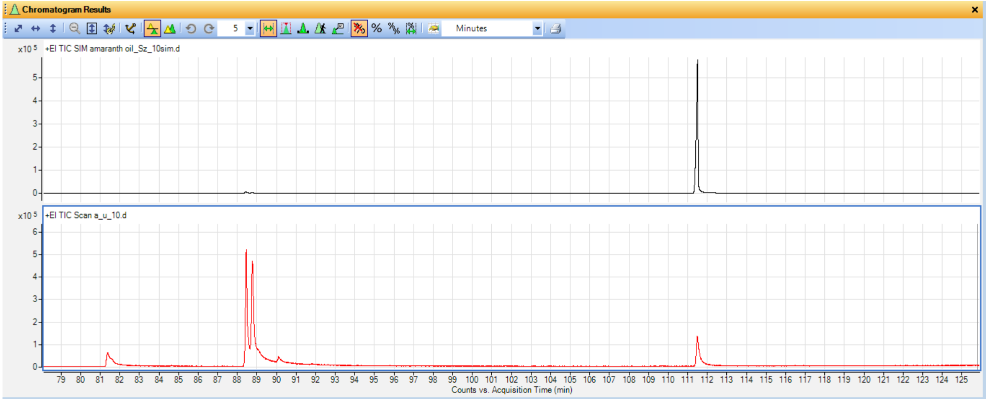
squalene - 5 Levels, 5 Levels Used, 6 Points, 6 Points Used, 0 QCs

y = 2726190.999915 * x - 1145052.117503

R^2 = 0.99890046

Type:Linear, Origin:Ignore, Weight:None

16. Using the above SIM method, record the chromatograms of the silylated oil samples (amaranth oil, sea buckthorn oil, primrose oil).



A comparison of amaranth oil chromatograms recorded in SIM and Scan mode.

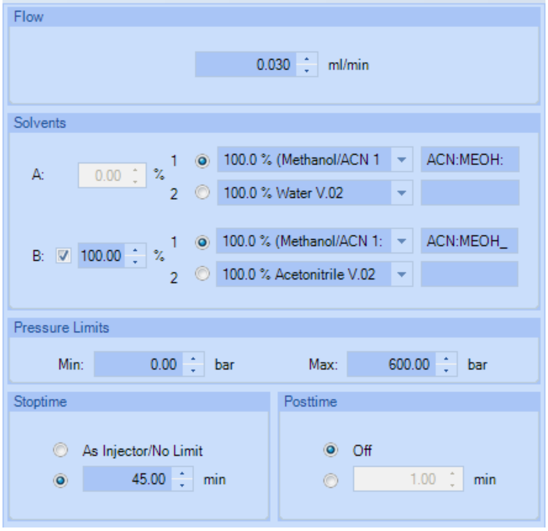
17. Select all the recorded chromatograms of oil samples from the pathway, and then click OK to add them to the previously created batch file.

18. Next click the "Analyze Batch" icon to obtain a table with quantitative results.

3.5. Observation of an HPLC-MS run and configuration of the data-dependent MS/MS acquisition

A plasma lipid extract will be analyzed in data-dependent MS/MS mode using QTOF LC-MS system following prior optimization of the critical parameters for this type of analysis, e.g. the background level, the number of selected precursors or the number of scans. During practice, the operating principles and the basic construction of HPLC system and QTOF, as a hybrid mass spectrometer, will also be provided. The significance of the preferred and excluded compound mass list concerning the data-dependent acquisition will also be discussed during this part of the practical session.

1. Run an analysis of a plasma lipid extract in data-dependent mode using the method of "Phospholipid_HILIC_negative.m" from the main context menu of Mass Hunter Acquisition Agilent software. Parameters of LC part of the method are given below:



Flow

0.030 ml/min

Solvents

A: 0.00 %
1 ● 100.0 % (Methanol/ACN 1 ACN:MEOH:
2 ○ 100.0 % Water V.02

B: ☑ 100.00 %
1 ● 100.0 % (Methanol/ACN 1: ACN:MEOH_
2 ○ 100.0 % Acetonitrile V.02

Pressure Limits

Min: 0.00 bar Max: 600.00 bar

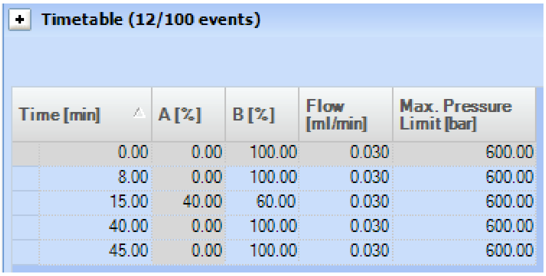
Stoptime

○ As Injector/No Limit
● 45.00 min

Posttime

● Off
○ 1.00 min

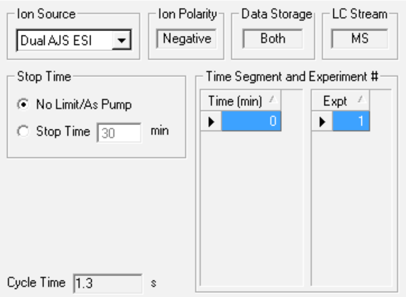
Mobile phase A consisted of 25% water, 50% acetonitrile and 25% (v/v) methanol, with 10 mM ammonium acetate. Mobile phase B consisted of 60% acetonitrile and 40% methanol with 10 mM ammonium acetate.



+ Timetable (12/100 events)

Time [min]	A [%]	B [%]	Flow [ml/min]	Max. Pressure Limit [bar]
0.00	0.00	100.00	0.030	600.00
8.00	0.00	100.00	0.030	600.00
15.00	40.00	60.00	0.030	600.00
40.00	0.00	100.00	0.030	600.00
45.00	0.00	100.00	0.030	600.00

2. In the meantime, consider and critically discuss the values of major ESI ionization parameters.



Ion Source: Dual AJS ESI

Ion Polarity: Negative

Data Storage: Both

LC Stream: MS

Stop Time

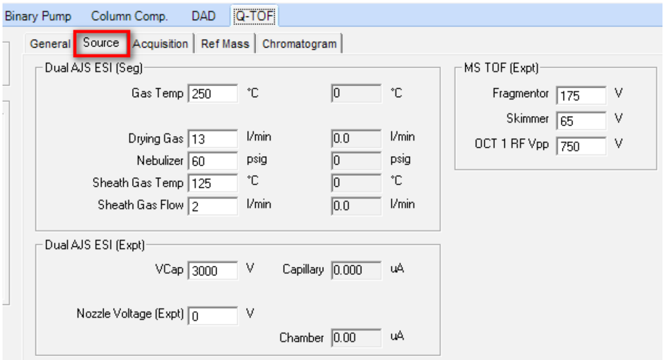
● No Limit/As Pump

○ Stop Time 30 min

Time Segment and Experiment #

Time (min)	Expt
0	1

Cycle Time 1.3 s



Dual AJS ESI (Seg)

Gas Temp	250	°C	0	°C
Drying Gas	13	l/min	0.0	l/min
Nebulizer	60	psig	0	psig
Sheath Gas Temp	125	°C	0	°C
Sheath Gas Flow	2	l/min	0.0	l/min

MS TOF (Expt)

Fragmentor 175 V

Skimmer 65 V

OCT 1 RF Vpp 750 V

Dual AJS ESI (Expt)

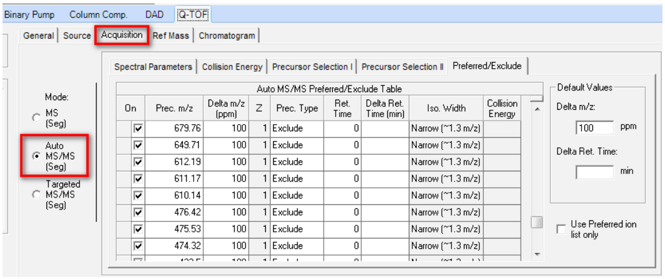
VCap 3000 V

Capillary 0.000 uA

Nozzle Voltage (Expt) 0 V

Chamber 0.00 uA

Data-dependent type of analysis should be selected in the acquisition tab of Mass Hunter Agilent software.



Mode:

○ MS (Seg)

● Auto MS/MS (Seg)

○ Targeted MS/MS (Seg)

Auto MS/MS Preferred/Exclude Table

On	Prec. m/z	Delta m/z (ppm)	Z	Prec. Type	Ret. Time	Delta Ret. Time (min)	Iso. Width	Collision Energy
☑	679.76	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	649.71	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	612.19	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	611.17	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	610.14	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	476.42	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	475.53	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	474.32	100	1	Exclude	0		Narrow (~1.3 m/z)	

Default Values

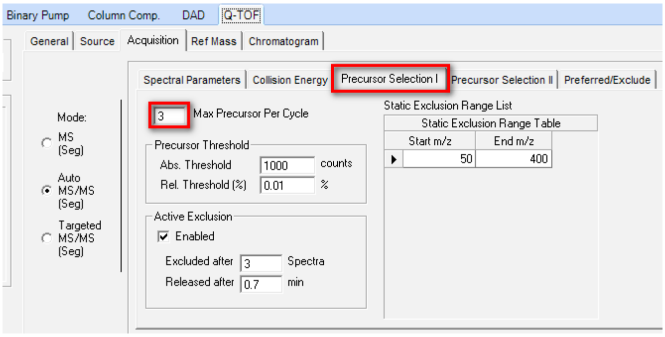
Delta m/z: 100 ppm

Delta Ret. Time: min

□ Use Preferred ion list only

3. Critically discuss important parameters for data dependent analysis such as:

- number of selected precursors



Mode:

○ MS (Seg)

● Auto MS/MS (Seg)

○ Targeted MS/MS (Seg)

3 Max Precursor Per Cycle

Precursor Threshold

Abs. Threshold 1000 counts

Rel. Threshold (%) 0.01 %

Active Exclusion

☑ Enabled

Excluded after 3 Spectra

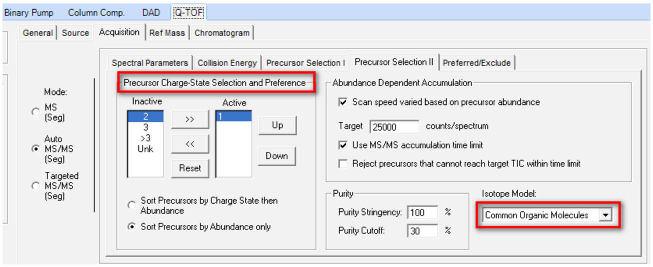
Released after 0.7 min

Static Exclusion Range List

Static Exclusion Range Table

Start m/z	End m/z
50	400

- isotopic model



Mode:
○ MS (Seg)
● Auto MS/MS (Seg)
○ Targeted MS/MS (Seg)

Precursor Charge-State Selection and Preference

Inactive	Active
2 3 >3 Unk	1

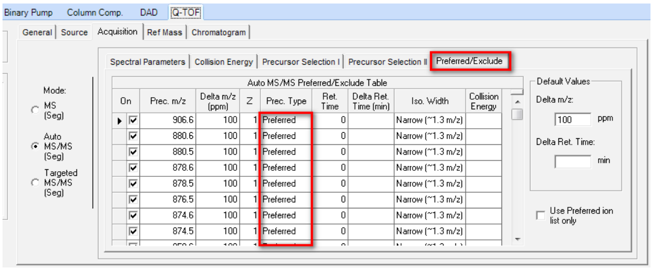
○ Sort Precursors by Charge State then Abundance
● Sort Precursors by Abundance only

Abundance Dependent Accumulation
☑ Scan speed varied based on precursor abundance
Target 25000 counts/spectrum
☑ Use MS/MS accumulation time limit
☐ Reject precursors that cannot reach target TIC within time limit

Purity
Purity Stringency: 100 %
Purity Cutoff: 30 %

Isotope Model: Common Organic Molecules

- creation of preferred and excluded m/z values lists

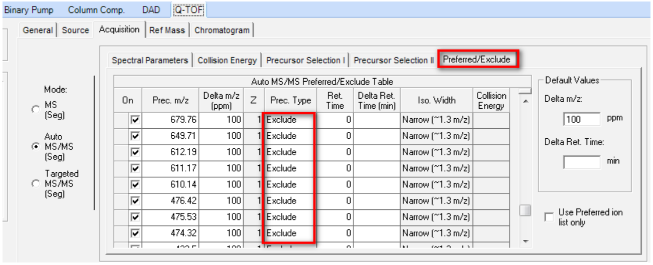


Mode:
○ MS (Seg)
● Auto MS/MS (Seg)
○ Targeted MS/MS (Seg)

Auto MS/MS Preferred/Exclude Table

On	Prec. m/z	Delta m/z (ppm)	Z	Prec. Type	Ret. Time	Delta Ret. Time (min)	Iso. Width	Collision Energy
☑	906.6	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	880.6	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	880.5	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	878.6	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	878.5	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	876.5	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	874.6	100	1	Preferred	0		Narrow (~1.3 m/z)	
☑	874.5	100	1	Preferred	0		Narrow (~1.3 m/z)	

Default Values
Delta m/z: 100 ppm
Delta Ret. Time: min
☐ Use Preferred ion list only



Mode:
○ MS (Seg)
● Auto MS/MS (Seg)
○ Targeted MS/MS (Seg)

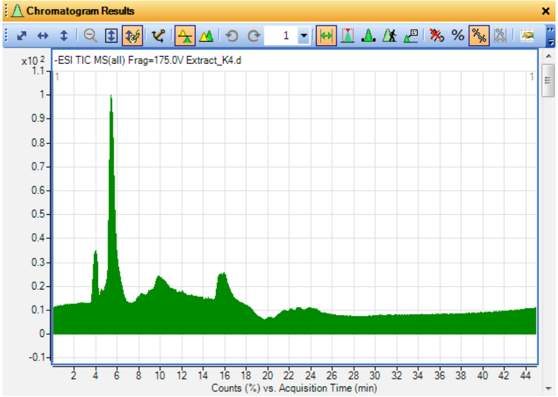
Auto MS/MS Preferred/Exclude Table

On	Prec. m/z	Delta m/z (ppm)	Z	Prec. Type	Ret. Time	Delta Ret. Time (min)	Iso. Width	Collision Energy
☑	679.76	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	649.71	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	612.19	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	611.17	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	610.14	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	476.42	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	475.53	100	1	Exclude	0		Narrow (~1.3 m/z)	
☑	474.32	100	1	Exclude	0		Narrow (~1.3 m/z)	

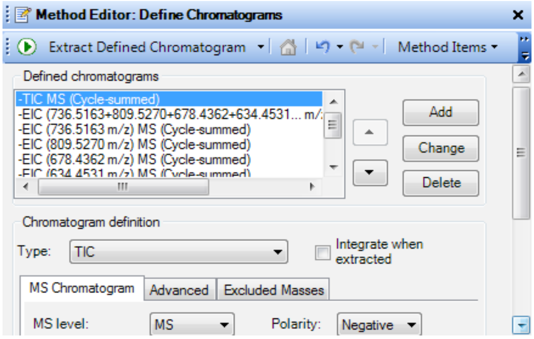
Default Values
Delta m/z: 100 ppm
Delta Ret. Time: min
☐ Use Preferred ion list only

What is the importance and meaning of the above lists in the case of data dependent analysis?

4. Generate Total Ion Chromatogram (TIC) to obtain phospholipid profile of the analyzed plasma sample.



5. Generate Extracted Ion Chromatograms (EIC) using "Extract defined chromatogram" command in the Mass Hunter software for each phospholipid class internal standard added to the sample to estimate the retention time ranges, enabling assignment of identified species to a particular class.

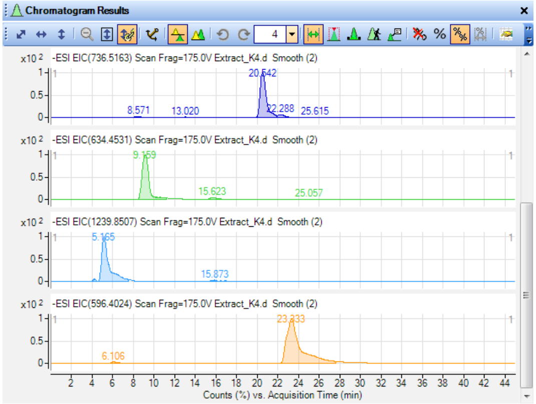


-TIC MS (Cycle-summed)
-EIC (736.5163+809.5270+678.4362+634.4531... m/
-EIC (736.5163 m/z) MS (Cycle-summed)
-EIC (809.5270 m/z) MS (Cycle-summed)
-EIC (678.4362 m/z) MS (Cycle-summed)
-EIC (634.4531 m/z) MS (Cycle-summed)

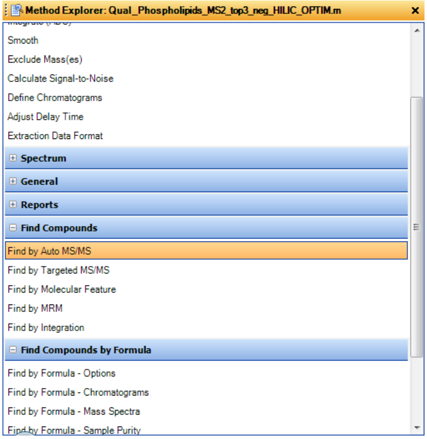
Type: TIC

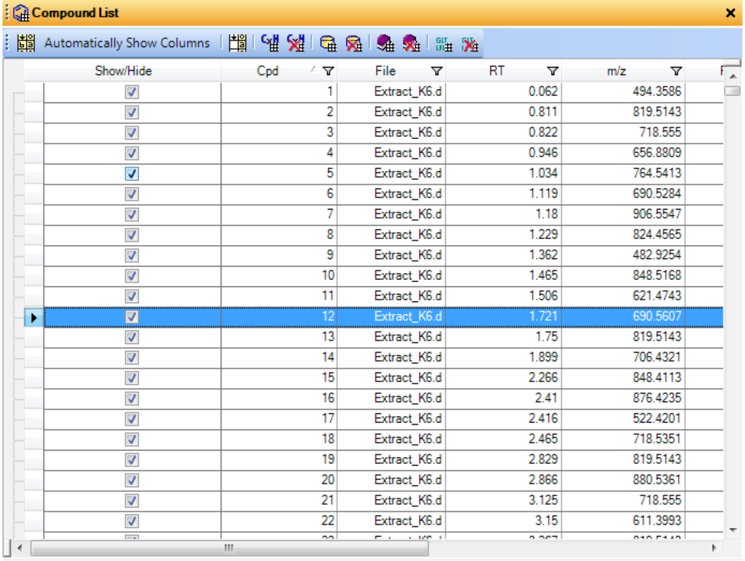
MS level: MS

Polarity: Negative



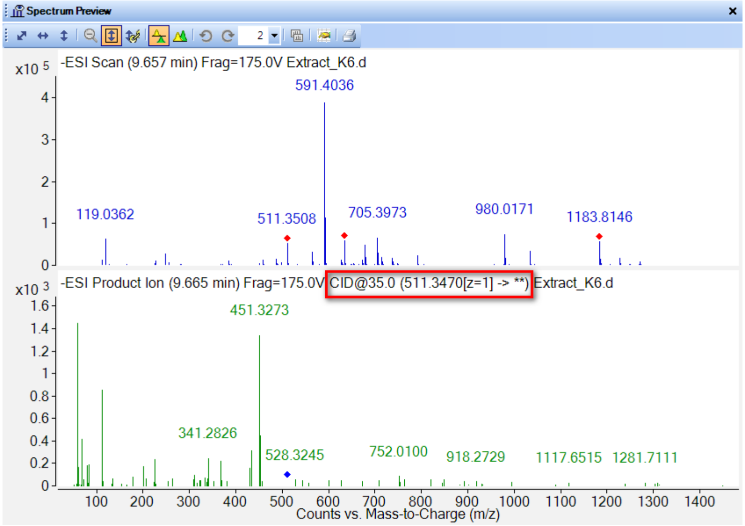
6. Generate a list of compounds using "find compound by "Auto MSMS" function from "Find compound" context menu.


Method Explorer: Qual_Phospholipids_MS2_top3_neg_HILIC_OPTIM.m
Smooth
Exclude Mass(es)
Calculate Signal-to-Noise
Define Chromatograms
Adjust Delay Time
Extraction Data Format
Spectrum
General
Reports
Find Compounds
Find by Auto MS/MS
Find by Targeted MS/MS
Find by Molecular Feature
Find by MRM
Find by Integration
Find Compounds by Formula
Find by Formula - Options
Find by Formula - Chromatograms
Find by Formula - Mass Spectra
Find by Formula - Sample Purity



Show/Hide	Cpd	File	RT	m/z
☑	1	Extract_K6.d	0.062	494.3586
☑	2	Extract_K6.d	0.811	819.5143
☑	3	Extract_K6.d	0.822	718.555
☑	4	Extract_K6.d	0.946	656.8809
☑	5	Extract_K6.d	1.034	764.5413
☑	6	Extract_K6.d	1.119	690.5284
☑	7	Extract_K6.d	1.18	906.5547
☑	8	Extract_K6.d	1.229	824.4565
☑	9	Extract_K6.d	1.362	482.9254
☑	10	Extract_K6.d	1.465	848.5168
☑	11	Extract_K6.d	1.506	621.4743
☑	12	Extract_K6.d	1.721	690.5607
☑	13	Extract_K6.d	1.75	819.5143
☑	14	Extract_K6.d	1.899	706.4321
☑	15	Extract_K6.d	2.266	848.4113
☑	16	Extract_K6.d	2.41	876.4235
☑	17	Extract_K6.d	2.416	522.4201
☑	18	Extract_K6.d	2.465	718.5351
☑	19	Extract_K6.d	2.829	819.5143
☑	20	Extract_K6.d	2.866	880.5361
☑	21	Extract_K6.d	3.125	718.555
☑	22	Extract_K6.d	3.15	611.3993

7. Find compounds at m/z of, for example, 591.4036.
8. Assign the retention time for each one of these compounds.
9. Assign the m/z values of fragment ions characteristic for each one of these compounds from point 7 and use them.



To know more:

E. de Hoffmann, V. Stroobant. Mass spectrometry: principles and applications. Wiley, 2007.

H.J. Hübschmann. Handbook of GC-MS: Fundamentals and Applications (3rd Edition). Wiley, 2015.

J. Kool, W.M. Niessen. Analyzing biomolecular interactions by mass spectrometry. Wiley, 2015.

Module 2
Metabolomics

Coral Barbas, Danuta Dudzik, Mª Fernanda Rey-Stolle, Francisco J. Rupérez, Antonia García

Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad San Pablo CEU; Madrid, Spain

I. Rationale

The module aims for postgraduate students to develop advanced knowledge and skills in the most recent "omic" discipline. Metabolomics is the systematic analysis of the complete set of intra and extra-cellular metabolites in a biological sample. The final aim of metabolomics is to provide insights into metabolite alterations and molecular mechanism of an underlying disease state and discovery of novel diagnostic and prognostic clinical biomarkers. This module is divided into four sections covering the main aspects of metabolomics: 1) Aims and definitions, 2) Approaches, experimental design, special requirements, analytical tools, quality control and workflows in metabolomics, 3) Data pre-treatment, data processing, statistical analysis and identification of biomarkers and, in the last part, 4) Biochemical interpretation, pathway analysis and biomarker validation along with several applications.

The course will be taught through a combination of lectures, workshops, tutorials and drop-in sessions including demonstration/practical hands-on lessons with real data by using free available bioinformatics resources applied nowadays. All these teaching resources will allow for consolidating the specific outcomes.

II. Course Aims and Outcomes

Aims

This course aims to provide students with an understanding of the core concepts and approaches for metabolomics studies. Students will acquire basic theoretical and practical knowledge of non-targeted metabolomics including the pipeline, steps and requirements, analysis of metabolomics data, preprocessing, statistics, identification, and annotation. Besides, there is a need for biomarker validation and future investigation of the established biological hypothesis.

Learning outcomes:

By the end of this course, students will:

1. Define and apply common metabolomics terminology.
2. Discuss comprehensively the different mechanisms of separation coupled to MS for metabolomics.

3. Identify all steps of the metabolomics study and their different approaches.
4. Be able to choose between the different approaches to solve a specific problem.
5. Know and understand the different methods of sample treatment and their limitations in metabolomics.
6. Know and understand the different steps in data processing.
7. Be able to use public software in data processing and pathways analysis.
8. Know different approaches for statistical analysis applied to metabolomics.
9. Acquire basic skills in the use of metabolite databases as free access resources.
10. Explain to non-specialists how this "omics" discipline can be expected to provide valuable information in different areas of Life Science.
11. Communicate and justify conclusions clearly and unambiguously to both specialist and non-specialist audiences.
12. Continue the learning process, largely, autonomously.

III. Course content

Module 2 – Metabolomics

1. Introduction to metabolomics

2. Analytical approaches in metabolomics

2.1. Workflow of the metabolomics study

2.1.1. Experiment design

2.1.2. Sample collection

2.1.3. Metabolite extraction

2.1.4. Separation and detection

2.2. Quality Control and Quality Assurance Procedure in Metabolomics

3. Data processing and identification of metabolites

3.1. Data processing pipeline

3.1.1. Raw data processing

3.1.2. Data pre-processing

3.1.3. Data pre-treatment

3.2. Metabolite identification

3.2.1. Identification in GC-EI-MS

3.2.2. Metabolite Identification in an LC or CE-(ESI)MS metabolomics experiment: Working with Databases

3.2.3. Metabolite centered, simple MW search compound databases

3.2.4. Metabolite centered, spectral databases

3.2.5. Mediators

3.3. Statistical analysis

3.3.1. PCA

3.3.2. PLS-DA

1. Introduction to metabolomics

Metabolomics refers to the systematic analysis of the complete set of intra and extra-cellular metabolites in a given biological system (including microbial, plant and mammalian systems). Metabolites are low-molecular weight (typically < 1500 Da) intermediates that constitute the building blocks for many other biological components (e.g., proteins, DNA, RNA). They participate in general metabolic reactions and are essential for the regulation, growth and normal function of the cell. Metabolites, a downstream of transcriptional, translational and post-translational processes, serve as the most proximal deponents of alternations that occur in the body in response to a pathological process. Therefore, metabolomics gives a unique advantage to provide a direct picture of the current state and phenotype of an organism. Understanding of the metabolome, defined as the total collection of metabolites in a cell, tissue, or organism is critical for obtaining a more comprehensive view of the functioning of cells, tissues, and organisms. In most cases, metabolomics consists on a differential study of metabolic fingerprints (metabolomics fingerprinting) generated from "control" and "test" subgroups of observations to find differences in their profiles in response to external stimuli (pathologies, effects of biochemical or environmental stresses, food processing, etc.). A summary of specific terms inherent to the metabolomic field is given in Table 2.1.

Table 2.1. Metabolomics related definitions.

Term	Definition
Metabolomics	The study of the complete set of metabolites varying according to the physiology, developmental or pathological state of the cell, tissue, organ or organism.
Metabolites	Small molecules, low-molecular weight (< 1500 Da) intermediates, building blocks for other biological components, essential for the regulation, growth and normal function of the cell.
Metabolome	The final downstream product of the genome, the complete set of all low molecular-weight metabolites present in a cell or organism. The metabolome is divided into exometabolome (metabolites outside the cell) and endometabolome (intracellular metabolites).
Metabotype	The metabolic phenotype.
Metabolic fingerprinting	Unbiased, global, high-throughput approach to classifying samples based on metabolite patterns or "fingerprints" that change in response to disease, environmental or genetic perturbations with the ultimate goal of identifying discriminating metabolites.
Metabolic profiling	Identification and quantification or semi-quantification of a specific group of metabolites which are related to a specific metabolic pathway.
Metabolic footprinting	Analysis of the (exo)metabolites secreted/excreted by an organism under controlled conditions.
Untargeted metabolomics	Global in scope, hypothesis-free, with the aim of simultaneously measuring as many metabolites as possible from biological samples.
Targeted metabolomics	Involves using one particular analytical method for the determination and quantification of a small set of known metabolites.
Metabolic pathway	A set of chemical reactions inside a cell.

Metabolomics has already proven to be an effective, well-established and valuable method of investigating human health and disease, aging, lifestyle, and drug discovery and development. Therefore, the ultimate goal of metabolomics is to provide insight into metabolite alterations and molecular mechanism underlying disease state and discovery of novel diagnostic and prognostic clinical biomarkers. The number of studies involving metabolome analysis together with improvements in analytical technology has been recently increasing year-on-year.

2. Analytical approaches in metabolomics

Depending on the objectives of metabolomics study, two analytical strategies can be applied: targeted and untargeted approach (Figure 2.1).

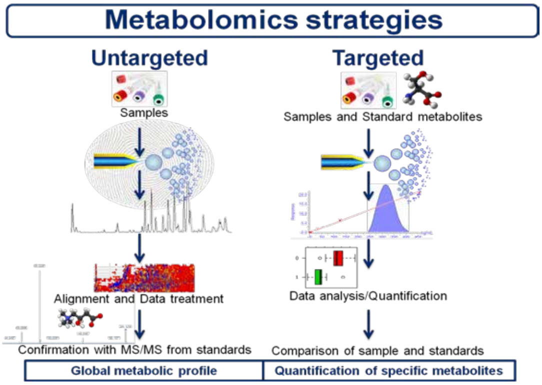

Metabolomics strategies
Untargeted
Targeted
Samples
Samples and Standard metabolites
Alignment and Data treatment
Data analysis/Quantification
Confirmation with MS/MS from standards
Comparison of sample and standards
Global metabolic profile
Quantification of specific metabolites

Figure 2.1. The untargeted and targeted analytical approach to metabolomics studies.

Targeted approach

Targeted metabolomics refers to studies aiming to measure specific known molecules (typically <20), focusing on one or more related metabolic pathways which have been defined as biologically relevant in previous studies. Therefore, methods with high levels of specificity, precision, and accuracy to define absolute amounts of each metabolite are applied such as, for example, quantitation of amino acids using selected reaction monitoring methods with triple quadrupole (QqQ) MS. Absolute quantification of the targeted compound is performed by the incorporation of appropriate internal standards and the comparison of data to the constructed calibration curves.

Untargeted approach

Untargeted (or non-targeted) metabolomics, referred as metabolomics fingerprinting, is global in scope with the main objective of simultaneously measuring as many metabolites as possible in each sample without bias, within a single experimental run. In the untargeted strategy, the metabolites to be identified are not known before the study. Untargeted metabolomics is applied to hypothesis-free studies, which is discovery phase with the objective to define novel and previously unobserved changes in metabolome. That can be linked to the biological function of an organism and mechanisms underlying altered metabolic changes related to the development of a disease state. This approach leads to the generation of a new hypothesis that should be verified in the subsequent validation study, including targeted approach.

2.1. Workflow of the metabolomics study

A typical workflow of untargeted metabolomics analysis is shown in (Figure 2.2) and is directly related to a proper experimental design. Crucial points of the workflow include the (1) experimental design, (2) sample collection, (3) extraction protocol, (4) data acquisition, (5) data processing and analysis and (6) metabolite identification that allows biological interpretation. Each step of the metabolomics workflow is of great importance. Therefore, quality assurance procedures should be applied to reduce unwanted pre-analytical and experimental variation and to guarantee plausibility of the metabolomics studies.

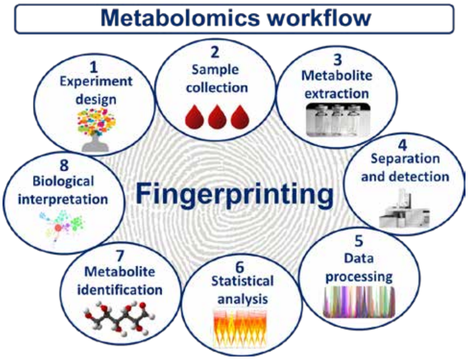


Figure 2.2. Typical workflow for untargeted metabolomics studies.

2.1.1. Experimental design

An experimental design describes in details how the study will be performed. It should cover all relevant aspects of the workflow from sampling to final data analysis. An experimental design means that all the samples to be studied are chosen according to a well-considered manner. If the samples have already been collected (e.g., sample cohorts form epidemiological studies) it is important to know (1) how they were collected and stored; (2) how control and treatment samples can be matched; and (3) if there is a clear phenotype between the control and experimental groups.

The experimental design should include a clear description of the number of samples, replicates and studied groups to ensure obtaining data with sufficient statistical power. Studies based on an experimental model, e.g., animal or cell culture systems, in a well-controlled laboratory environment usually require low sample size (typically 6-12) to identify changes that are statistically robust. However, studies based on general human population are more complex and include much more variables that could introduce bias into the designed study. Factors related to lifestyle (e.g., diet, exercise), demography (e.g., ethnicity) and physiology (e.g., gender, age, BMI) and clinical data (e.g., medical background, medications) should be considered and will allow to identified and avoid confounding variables. Efficient and careful experimental design is required to ensure that even small biologically relevant variations in the metabolome can be accurately measured and will be significantly greater than alterations introduced by experimental bias.

2.1.2. Sample collection

Follow the required features established in the experiment design. Collect all samples with the same appropriated containers, close them tightly, identify them and store at -80^{0}C whenever possible. Defrost just on the analysis day and keep them on ice.

2.1.3. Metabolite extraction

Metabolite extraction is another step that has to be well controlled for reliable study outcomes. Samples are usually analyzed without major preparation and method of extraction depends on the sample matrix and the analytical system applied.

LC-MS. For LC-MS based analysis, samples containing proteins require a deproteinization step that is a process of removal of the protein complement. Hence, typically cold organic solvent (e.g., methanol, acetonitrile) or a mixture of solvents (e.g., methanol/ethanol (1:1, v/v) is added to precipitate the protein, which can be then separated by further centrifugation. The resulting supernatant contains extracted metabolites and is suitable for LC-MS analysis.

GC-MS. For GC-MS, besides further deproteinization, more exhaustive procedures need to be performed as GC is incompatible with non-volatile and thermally labile compounds. They include ketone functional group protection by methoximation and derivatization to increase metabolites volatility. The process of derivatization decreases the boiling point of many endogenous metabolites, which makes them volatile enough to enable chromatographic separations. Compounds containing functional groups with active hydrogens such as –SH (thiols), -OH (alcohols, polyols, phenols, enols), -NH (amines, amides) and –COOH (carboxylic acids) are of primary concern in GC-MS analysis, because of the tendency of these functional groups to form intermolecular hydrogen bonds. These intermolecular hydrogen bonds affect the inherent volatility of compounds, their tendency to interact with column packing materials and their thermal stability. Thus, after deproteinization, the samples should be dried before addition of any derivatizing agents. Methoximation utilizes *O*-Methoxyamine hydrochloride in pyridine to stabilize carbonyl moieties by suppressing keto-enol tautomerism and the formation of multiple acetal- or ketal- structures. It also helps to reduce the number of derivatives of reducing sugars and generates only two forms of the –N=C< derivative, syn/anti. The subsequent silylation process involves replacement of active hydrogen atoms with an alkylsilyl group, for example, trimethylsilyl (–$SiMe_3$). Modification of the functional group of a molecule by chemical derivatization enables the analysis of many polar compounds, increasing their volatility, thermal stability, and chromatographic peak symmetry.

CE-MS. Typical sample preparation for CE-MS analysis usually consists of two steps, i.e. sample dilution, and deproteinization. For better sensitivity, ionic strength in sample solution should be lower than background electrolyte filling the capillary.

2.1.4. Separation and detection

The choice of analytical platform can significantly influence the coverage of the metabolome acquired in untargeted studies. Mass spectrometry (MS) and nuclear magnetic resonance (NMR) are the most frequently employed methods of detection in the analysis of the metabolome. NMR is very useful for structure characterization of unknown compounds, however, it is also characterized by low sensitivity, and, additionally, much higher equipment costs compared to

MS-based techniques. Whereas, the most important advantages of MS are its high sensitivity, and high-throughput. The view of it consists of extremely diverse chemical compounds that make it virtually impossible to simultaneously capture the complete metabolome present in the sample in a single analytical platform. Therefore, the combination of separation techniques (liquid chromatography, gas chromatography, and capillary electrophoresis) with MS tremendously expands the capability of the chemical analysis of highly complex biological samples and allows to obtain a complete metabolite coverage. GC-MS offers the capacity to analyze low-polarity volatile metabolites of fats and esters, and high-polarity metabolites of amino acids and organic acids converted into volatile derivatives, while non- and semi-polar high molecular weight metabolites (e.g., sphingolipids, glycerophospholipids) are typically studied applying LC-MS. Additionally, CE-MS offers a complementary approach to LC-MS for analyzing anions, cations, and neutral particles. Those analytical methods have been developed and validated to provide reproducible and robust data. Table 2.2 depicts the main characteristics of separation techniques coupled to MS for metabolomics studies.

Table 2.2. Separation techniques coupled to mass spectrometry for metabolomics.

Analysis Technique	**Application**	**Advantages**	**Disadvantages**
GC-MS	Separation, identification, and quantification of volatile and thermally stable less polar metabolites.	High chromatographic resolution, availability of wide spectrum of libraries for metabolites identification.	Inability to analyze thermo-labile and high molecular-weight metabolites, the requirement of derivatization for non-volatile metabolites.
LC-MS	Separation, identification, and quantification of very broad groups of metabolites, depending on the type of column and mobile phase.	High sensitivity, large sample capacity, derivatization not required, ability to analyze thermo-labile compounds.	Limited availability of commercial libraries, restriction on LC eluents, matrix effect, limited potential in identification unless an MS-MS technique is used.
CE-MS	Separation, identification, and quantification of polar and ionized metabolites, using reduced sample volumes.	High resolution and rapid analysis, utility for complex biological samples, even if in a small volume.	Limited availability of commercial libraries. Buffer incompatibility, detection limits. Limited potential for identification unless an MS-MS technique is used.

To know more:

S.G. Villas-Bôas, S. Mas, M. Åkesson, J. Smedsgaard, and J. Nielsen, Mass spectrometry in metabolome analysis. Mass spectrometry reviews 24:613-646, 2005

2.2. Quality Control and Quality Assurance Procedure in Metabolomics

Quality assurance (QA) is required for all analytical strategies (targeted and untargeted), but especially for large-scale studies. In targeted studies applying MS platforms, quality control samples (QC) are routinely applied to quantitatively define accuracy and precision of the analytical method. In the context of metabolomics, QC samples are applied to (1) control the performance of the analytical system; (2) equilibrate the system in order to achieve fully reproducible conditions; (3) correct small levels of drift in the measured signal over analysis (within batches) and between analytical batches; (4) integrate the data from different analytical batches; (5) calculate metabolite measurement precision of replicate injections of QC.

QC samples are analyzed at the beginning and at the end of an analytical run as well as at regular intervals through the analysis to monitor the stability and reproducibility of the analytical process. QC samples should be representative of the qualitative and quantitative composition of the samples being analyzed in the study. The assumption is that the QC sample will contain a mean concentration of all of the components that are present in the samples under investigation. Random and systematic sources of variation will affect the reproducibility of the data within the experiment. Several post data acquisition methods based on QC performance have been developed for quality assurance of metabolomics experiments. Data quality assessment include overview of the analytical precision based on (1) examination of raw data acquired; (2) plotting the sum of metabolic feature intensities for every experimental and QC sample; (3) checking for signal drift, sensitivity loss or variation in QC predicted in unsupervised multivariate model (Principal Component Analysis, PCA-X); (4) calculation of the coefficient of variation CV (% RSD) for every metabolic feature detected in the QC samples. Calculating the CV for each metabolic feature (possible metabolite) for all the QC samples across the run, provide a quantitative measure of the variability. Acceptance of CV limits, lower than 20% for LC-MS and CE-MS data and lower than 30% for GC-MS, will allow to filtrate the complex data matrix and remove metabolites showing unacceptable variation.

Figure 2.3 illustrates the possible outcome of the prediction of QC samples in multivariate PCA-X.

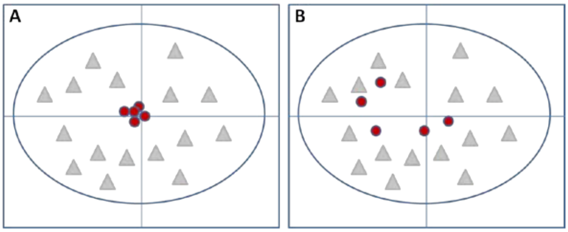


Figure 2.3. Schematic representation of the QC samples prediction in the multivariate PCA-X model. Panel 3.A. Tight clustering of QC sample could be observed, indicating the precise analytical outcome. Panel 3.B. The dispersion of QC samples indicates the loss of the system stability and reproducibility within the analytical run.

3. Data processing and identification of metabolites

Metabolomics research leads to complex data sets involving hundreds to thousands of metabolites. Correct data handling is a fundamental step that has an enormous impact on extent and quality of the metabolomics results. However, metabolomics data processing, data treatment and identification of metabolites are probably also some of the most challenging steps of this approach due to some intrinsic characteristics: data generated from instrumental signals are noisy and contain a lot of highly correlated variables compared to the number of individuals. However, there are several advanced tools for metabolomics data preprocessing and analysis, both commercially provided (e.g., Agilent Technologies) or freely available online platforms (e.g., XCMS, MZmine, MetAlign, MetaboAnalyst, MeltDB or the recently developed W4M).

3.1. Data processing pipeline

In metabolomics studies, it is important to separate interesting biological variation from unwanted sources of variability introduced to the experiment. Therefore, comprehensive analysis of metabolomics data requires exhaustive data analysis strategy that is often unique and requires specialized data analysis software that enables chemometric analysis, bioinformatics, and statistics. A general workflow of metabolomics data analysis is based on four main steps related to (1) raw data processing; (2) data pre-processing; (3) data pre-treatment and (4) data treatment (Figure 2.4).

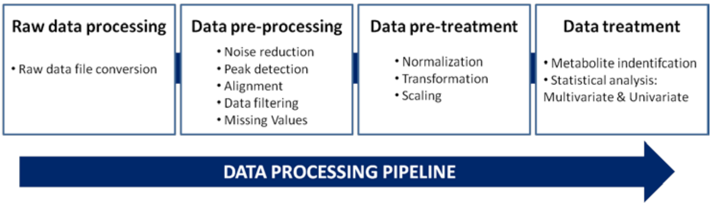


Figure 2.4. The main steps in the metabolomics data analysis workflow.

Untargeted metabolomics studies are characterized by the simultaneous measurement of a large number of metabolites from each sample. Consequently, a large amount of data is generated. The major issues to be considered in untargeted metabolomics studies are related to two key factors introducing biases to the data: biological variance and technical errors.

3.1.1. Raw data processing

Once the data has been generated, the output has to be organized to extract data out of the raw data files, and to turn measurements into available scientific data. Raw data can be converted from commercial formats to open formats both using software that is provided by instrumental vendors and by freely available and open source online platforms for metabolomics data preprocessing and analysis (e.g., XCMS).

3.1.2. Data pre-processing

Data pre-processing refers to all actions directly affecting raw data before further data analysis. The main issue is to improve signal quality and reduce possible biases present in the raw data. This step includes: (1) noise reduction; (2) peak detection and deconvolution; (3) alignment; (4) missing values imputation techniques and (5) data filtration.

(1) Noise reduction

A baseline correction, filtering the noise, is used to remove low-frequency artifacts and apparent differences that are generated by experimental protocol and instrumental variation. This step facilitates the peak detection and potentially reduces the detection of false positive features. Some software has integrated the filtering and peak detection algorithms into a single function.

(2) Peak detection and deconvolution

The purpose of peak detection and peak picking (deconvolution) is to identify and quantify the signals in the spectra coming from compounds detected in the sample.

Each peak in the chromatogram or electropherogram appears plotting the intensity (abundance) in the y-axis and the retention time (RT) or migration time (MT) in the x-axis. Those intensities are measured in every scan, in a set of several thousands, in the established mass range over the detection time window. Besides, the graph obtained by MS, the mass spectrum, offers dual information besides the separation time. In complex samples, co-elution of different compounds giving a single peak is frequent, and moreover, the mass spectrum along the peak is the sum of all the co-eluting compounds. Deconvolution is the process of computationally separating co-eluting components and creating a pure spectrum for each component by extracting ions from a complex total ion chromatogram (TIC), even with the target compound signal at trace levels.

The peaks are detected across the spectrum, and their areas are integrated to provide quantification of the underlying metabolite. As a result of this process, complex 3D MS data matrices are converted into lists of annotated metabolic features acquired at their corresponding RT or MT in the form of a mass-to-charge ratio (m/z) and relative intensity of the measured compounds.

Peak detection method should identify the true signals and avoid the false positives. The methodology differs for hard and soft ionization sources, GC-EI-MS or LC-ESI-MS, CE-ESI-MS, respectively.

GC-EI-MS

Working with EI source, the mass spectrum obtained is essentially a fingerprint for the molecule that can be used to identify any compound. Eventually, GC/MS gives a 3D graph which has both chromatogram and a spectrum for each separated component. The first peak separation is due to chromatographic resolution, where analytes in the sample mixture are physically separated by boiling point order plus their selective interaction with the stationary phase of the analytical column and elute at different retention times, although there are analytes eluting at similar retention times. The second separation is due to spectral resolution, where analytes eluting at similar retention times (that do not have a total chromatographic resolution) are separated by their different (unique) mass to charge ratios by the mass analyzer.

The total ion current (TIC) chromatogram represents the summed intensity across the entire range of masses being detected at every point in the analysis. In an extracted-ion chromatogram (EIC), one or more m/z values representing one or more analytes of interest are recovered ('extracted') from the entire data set for a chromatographic run. Figure 2.5 shows the 3D data obtained by GC/MS.

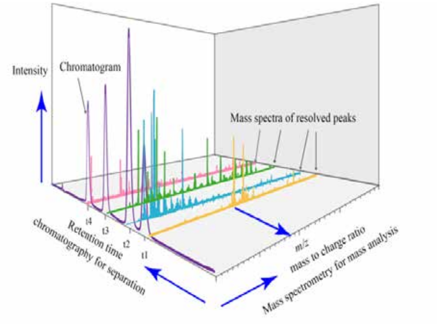


Figure 2.5. 3D Data of GC/MS chromatogram: Retention time, MS spectrum and Intensity

The utility of metabolomics studies largely depends on the number of identified metabolites and the links to their biological interpretation. The fragmentation of metabolites during EI is highly characteristic of the chemical structure, allowing these mass spectra to be used for the identification of compounds by mass spectral libraries. This procedure can be performed succesfully if the spectrum corresponds to the pure component without contamination by masses belonging to other sources (co-elution, background, column bleeding). As in complex samples, multiple analytes elute simultaneously, obscuring individual species, and the TIC chromatogram often provides limited information.

When extracting mass spectra to identify unknown components, it can be difficult to see if there is any co-elution with other components. A co-elution would produce a mixed mass spectrum and when a library searched that mixed mass spectrum, it would most likely result in a poor quality match, no potential matches or incorrect identification. There are ways to check for co-elutions, and this can be achieved manually by either clicking across the peak to see if the mass spectrum changes along a single peak or by extracting key ions to produce EICs, which when overlaid, should show the same shapes and retention times indicating that they all belong to the peak of interest.

Deconvolution of spectra in GC-EI-MS

There are ways to improve the quality of the mass spectrum before identification by spectral library searching, the most common one is by background subtraction on both sides. However, in most software performance, both background subtraction and manual extraction of ions can be a lengthy and time-consuming process, specially when looking at many peaks in the chromatogram, and when there are many samples. To identify and extract the quantitative information of the corresponding metabolites, the spectrum for every single metabolite has to be constructed. This spectrum construction step is called *deconvolution* in GC-EI-MS data processing. Therefore, as it was explained before, deconvolution is the process of computationally separating co-eluting components and creating a pure spectrum for each component. Specifically, for each observed EIC that results from two or more components, deconvolution calculates the contribution of each component to the EIC.

The free software for extracting ions even at trace level used for this technique is AMDIS (Automated Mass Spectrometry Deconvolution and Identification System) developed by NIST (National Institute of Standards and Technology). Figure 2.6 shows the result of the deconvolution process.

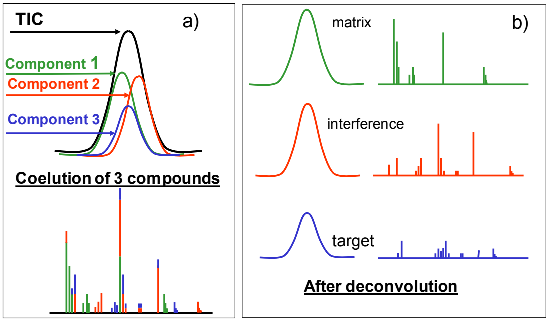


Figure 2.6. a) Before and b) After the deconvolution process (adapted from https://www.agilent.com/cs/library/Support/Documents/f05017.pdf)

The overall deconvolution process in AMDIS consists of four sequential steps: 1) Noise analysis, 2) Component perception, 3) Model shape and 4) Spectrum deconvolution.

In the first step, the noise characteristics for a GC-MS data file are extracted by calculating the noise factor to be used for representing signal magnitude in noise units. In the second step, component finding, individual chromatographic components are perceived. The rationale behind component perception is that a component exists when a sufficient magnitude of ions maximizes together. In the third step, the model peaks to be used in the next step for deconvolution are determined. The model shape for each perceived component is taken as the sum of the individual ion chromatograms that maximize together and whose sharpness values are within 75% of the maximum value for this component. In the last step, the proper deconvolution, 'purified' spectra from individual ion chromatograms for each component are extracted using the model shapes and the least-squares method.

AMDIS considers the peak shapes of all extracted ions and their apex retention times (RT). In the example of Figure 2.7, only some of the extracted ion chromatograms (EICs) are overlaid for clarity with the apex spectrum. Ion 160 EIC has the same RT as ions 50, 170 and 280, but it has a different peak shape. Ion 185 has a different peak shape and an earlier RT. Ions 75 and 310 have similar peak shapes, but they have different RTs.

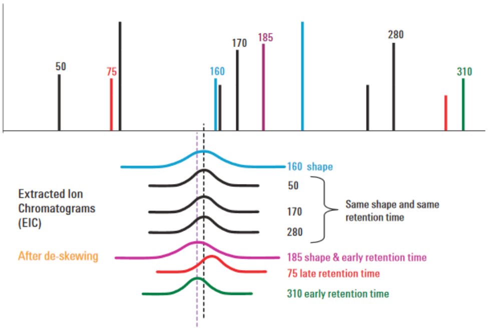

50
75
160
170
185
280
310
Extracted Ion Chromatograms (EIC)
After de-skewing
160 shape
50
170
280
Same shape and same retention time
185 shape & early retention time
75 late retention time
310 early retention time

Figure 2.7. Extracted Ion Chromatograms of determined ions (image provided by Agilent).

Figure 2.8 shows the EICs after the different peak shapes or RTs are eliminated from Figure 2.7. Ions 50, 170 and 280 remain.

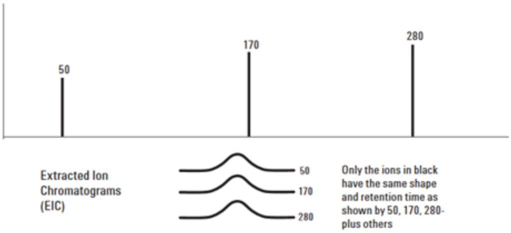

50
170
280
Extracted Ion Chromatograms (EIC)
50
170
280
Only the ions in black have the same shape and retention time as shown by 50, 170, 280- plus others

Figure 2.8. Ions with the same shape and retention time (image provided by Agilent).

Figure 2.9 shows all the ions with similar peak shapes and RTs, within the criteria set earlier by the analyst. These are grouped and referred to as a component by AMDIS

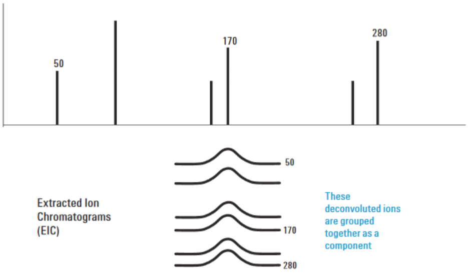

50
170
280
Extracted Ion Chromatograms (EIC)
50
170
280
These deconvoluted ions are grouped together as a component

Figure 2.9. Mass spectrum of a component after deconvolution (image provided by Agilent).

Deconvolution of spectra in LC-ESI-MS and CE-ESI-MS

Peak-based methods are the most common algorithms for feature detection in MS studies. Such algorithms, e.g., Molecular Feature Extractor (Agilent), consider the accuracy of the mass measurements to group-related ions by charge-state envelope, isotopic distribution, and possible chemical relationships when determining whether different ions are from the same metabolic feature. It can consider related ions like adducts: proton, sodium, potassium and ammonia adducts in the positive ionization or loss of a proton, adducts with formate, chloride, etc. in the negative ionization mode. The chromatogram before and after deconvolution is shown in Figure 2.10.

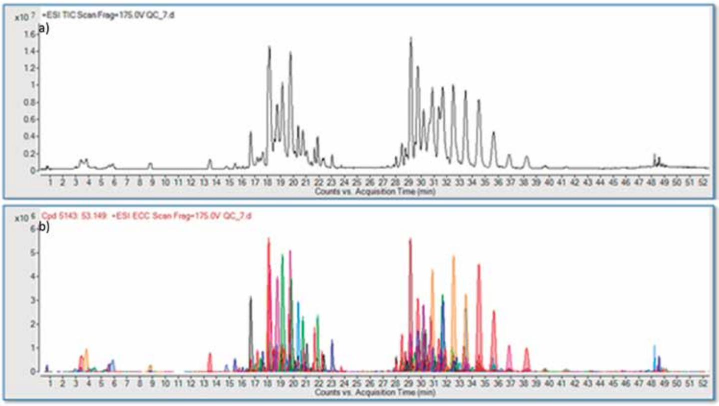


Figure 2.10. Comparison of two chromatograms from the same analysis of a complex mixture: a) Total Ion Chromatogram and b) Chromatograms from every single compound obtained after deconvolution. b) Chromatograms from every single compound obtained after deconvolution.

(3) Alignment

Spectral alignment is one of the main pre-processing steps in metabolomics studies involving multiple sample analysis. When analyzing multiple spectra, the position of the peaks corresponding to the same metabolic feature may be affected by non-linear shifts**.** In MS-based studies, peak shifts are mainly observed across the RT axis. Therefore, spectral alignment (matching peak by m/z and RT among all samples) must be applied to correct that variability in the analyzed samples as it can profoundly affect the quality of the study. Spectral alignment algorithms can be divided into two main groups: (1) those in which data are aligned before peak detection and (2) peak-based alignment methods, where the detected spectral peaks are aligned across samples.

(4) Missing values imputation

Unfortunately, not all of pre-processing data methods generate a complete data set and dealing with missing values in data matrix is a considerable challenge in metabolomics studies. There are several sources of missing values, such as: (1) limits in computational detection; (2) low signal intensity; (3) measurement error; (4) deconvolution/alignment errors; (5) errors in the identification of the signals from the background; or (6) simply the absence of the metabolite in the sample. In general, three types of missing values could be identified: (1) data missing completely at random (MCAR), when the missing data is not related to any observed variable or sample; (2) data missing at random (MAR), when the missing data is related to one or more observed variables but not to the sample; (3) data missing not at random (MNAR), when the missing data is related to the sample itself. Missing values possess a serious problem for further data processing and statistical analysis. Missing values are due to several factors including concentrations below detection limits on the MS systems but also misidentification by the software. Therefore, several algorithms for missing values imputation (MVI) have been developed. Those methods include: (1) replacement by zero; (2) replacement by mean/median, minimum value, half of minimum value; (3) k-means nearest neighbour (kNN); (4) Probabilistic PCA (PPCA); (5) Random Forest (RF); (6) Bayesian PCA (BPCA) method.

(5) Data filtration

Metabolomics data matrix is typically constructed with thousands of metabolite features across multiple samples. Before statistical analysis, this data matrix must undergo further filtration to reduce the number of variables. There are different manners of data filtering, but general consideration is to remove metabolic features with poor repeatability. Two main strategies for data filtering are used in metabolomics studies: (1) filtering by presence, based on the percentage of the samples present one of the sample group; (2) filtering based on quality assurance procedures and the threshold of RSD value calculated for each metabolic feature detected in QC samples. Additionally, data should be cleaned from experimental bias by blank background subtraction.

3.1.3. Data pre-treatment

Several factors (e.g., unwanted experimental & biological variations and technical errors) may cause difficulties with the identification of meaningful differences in the metabolic profiles. To remove specific types of unwanted variations, the signal drift correction and the scaling methods are adopted.

a) Normalization

Normalization is applied to correct for unwanted peak intensity differences and to stabilize the variance within the dataset. Unwanted experimental variation can originate from human error (e.g., sample extraction and preparation), within instrument variation (e.g., temperature changes within the instrument, sample degradation or loss of performance of the instrument during long run of samples), between instrument variation, different batches, different laboratories, and different analytical platforms. Unknown biological variation (e.g., number of cells, the concentration of biofluid) is also common and can be confounded with the factors of interest. Normalization of those interferences can be performed with or without applying an internal standard as a reference to calculate the observed analytical errors. Normalization methods that are not based on internal standards often apply the sum, mean or the median of the responses of all metabolites across a sample as a normalization factor. Other methods include, e.g., probabilistic quotient normalization (PQN), median fold change normalization (FC), quantile normalization, or locally weighted scatterplot smoothing (LOESS) normalization. The choice of normalization technique relies on a set of assumptions regarding metabolic measurements and biological variability of a data set.

b) Transformation

Abundances of metabolites in a data matrix usually have a right-skewed distribution, therefore a log transformation is often applied before statistical analysis.

c) Scaling

Scaling is performed to adjust for differences in fold change between metabolites which may be caused by large differences in the variation of the measured responses. Scaling methods divide each variable by a factor, the scaling factor, which is different for each variable. A range of scaling methods has been applied in metabolomics including auto-scaling, Pareto scaling, range scaling and VAST scaling.

To know more:

O. Fiehn. Metabolomics--the link between genotypes and phenotypes. *Plant Mol. Biol.*, 2002, 48(1-2):155-71.
W.B. Dunn, D.I. Broadhurst, H.J. Atherton, et al. Systems level studies of mammalian metabolomes: the roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem. Soc. Rev.*, 2011, 40(1):387-426.
W.B. Dunn, I.D. Wilson, A.W. Nicholls, and D. Broadhurst. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis*, 2012, 4:2249-2264
A. Alonso, S. Marsal and , A. Julià. Analytical methods in untargeted metabolomics: state of the art in 2015. *Frontiers Bioeng. Biotech.*, 2015, 3:23.
R. Di Guida, J. Engel, J.W. Allwood, et al. Non-targeted UHPLC-MS metabolomic data processing methods: a comparative investigation of normalisation, missing value imputation, transformation and scaling. *Metabolomics*, 2016, 12(5):1-14.

3.2. Metabolite identification

Often, one of the most challenging steps is the identification of metabolites. According to Donald Rumsfeld (Secretary of Defense, USA): "…as we know, there are known knowns; there are things we know we know. We also know there are known unknowns; that is to say, we know there are some things we do not know. However, there are also unknown unknowns – the ones we don't know we don't know."

3.2.1. Identification in GC-EI-MS

In GC-EI-MS, putative identification of metabolites could be done by using two criteria: searching the pure spectrum obtained by deconvolution against commercial spectral libraries and by RT matching with a spectral library with locked RT for all the entries (constant retention time for the same analyte obtained by the same analysis method). Depending on the match factor from the search, target compounds can be identified or flagged in a complex TIC. Each component now will be searched against a retention time library (called target library). There are metabolomics target spectra libraries, RTL libraries that have been specifically developed to facilitate the identification of metabolites. Due to its high reproducibility, chromatographic peak resolution, and the existence of libraries of mass spectra, GC-EI-MS is regarded as the gold standard for metabolomics research.

The power of deconvolution is appreciated while comparing the top two spectra in Figure 2.11. The raw scan or original non-deconvoluted scan is shown on top. The clean scan, which is the deconvoluted component, is shown in the middle. The bottom scan is the identified compound in a spectral library. Without deconvolution, the analyst would visually compare the background subtracted raw scan and library scans for confirmation. Using that type of comparison, it would be very difficult, if not impossible, to say that Fenbuconazole, the target compound in this example, is present.

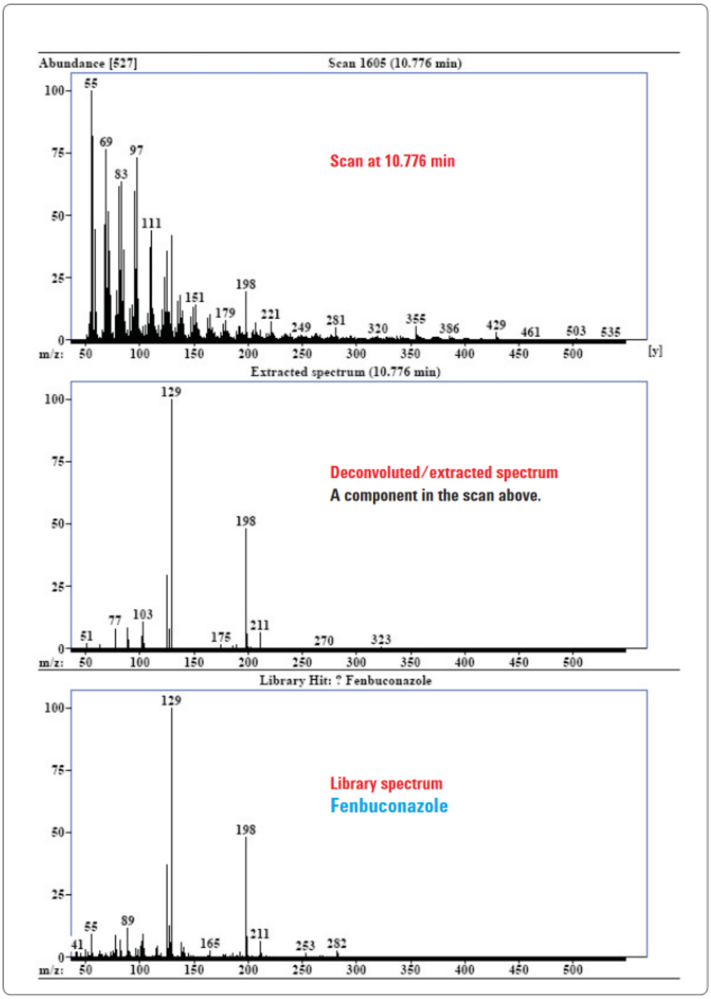


Figure 2.11. Comparison of raw, deconvoluted and library spectra (image provided by Agilent Technologies).

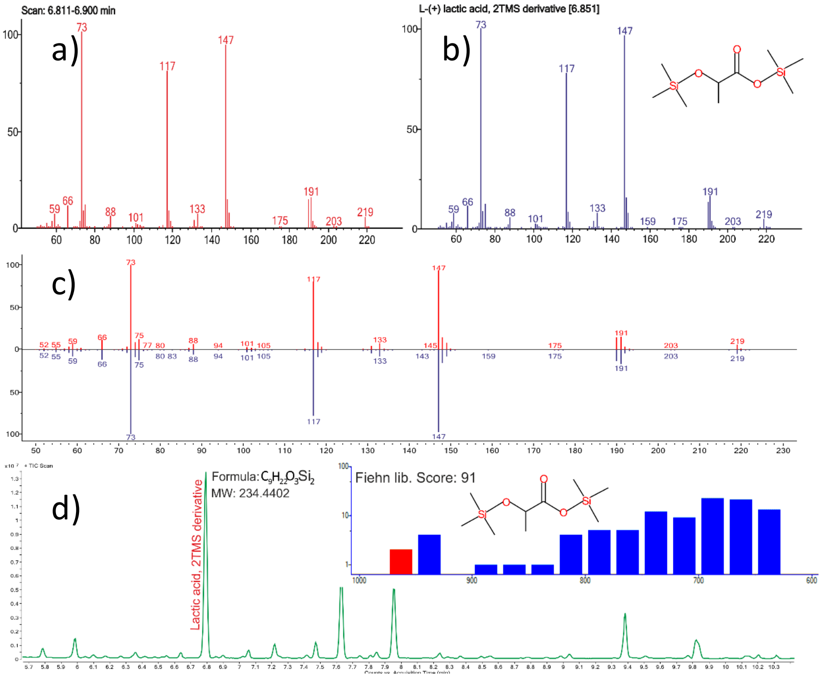


Figure 2.12. Identification of a metabolite in a GC-EI-MS chromatogram by matching RT and spectrum against a target library. a) Obtained spectrum, b) Spectrum form a target library, c) Plot of overloading spectra, d) TIC chromatogram with the searching score over 100.

Experiments performed on accurate mass instrumentation while using custom-made accurate mass databases would invariably increase the number of compounds identified. Accurate mass instrumentation would also facilitate the identification of unknowns by formula generation of molecular peaks and their fragments.

3.2.2. Metabolite Identification in an LC or CE-(ESI)-MS metabolomics experiment: Working with Databases

It is commonly accepted that in a typical metabolomics experiment, there are 3 levels of metabolite identification, from the least identified to the most: approximately 50% of the compounds will be unidentified compounds ("unknown"), but among the other 50% there can be compounds putatively identified in a compound class thanks to a query-match process using different databases; this identification can be refined with more instrumental data such as the retention time or the MS/MS fragmentation; compounds can be furthermore positively identified when a known standard is analysed under the same conditions and results from the unknown and the standard match.

Databases for metabolomics have experienced a deep evolution in recent years, in extension (number of metabolites), depth (fields of information per metabolite), and functionality (types of queries). As an example, the Human Metabolome DataBase (HMDB) was first released in 2006 with 2,180 metabolites, and it currently includes more than 100 fully searchable fields of information for each compound in a list that comprises more than 42,000 metabolites.

Nevertheless, the work with databases in the metabolomics field is far from being as standardized as in the other "omics", and researchers in this field must work with different databases and combine the information provided by them because a list of metabolomics-related databases can expand up to more than 50 different online resources. That is the reason why several mediators have appeared in recent years, to enable researchers to manage information from several sources.

The databases that can be employed in metabolomics can be classified according to their type of information as pathway centered and metabolite centered, and the latter can be further subdivided into spectral databases or compound databases. This classification cannot be considered as a closed boundaries organization, because they are continuously modified, and part of the information can be found in different databases.

The following section tries to summarize some of the most important characteristics of some of the most metabolomics-relevant databases available nowadays. It is important to remark that, although they are used in metabolomics research, the primary purpose of some databases is not the metabolomics field; such is the case for PubChem or KEGG, for example.

3.2.3. Metabolite-centered, simple MW search compound databases

ChEBI (Chemical Entities of Biological Interest) is an initiative from the European Bioinformatics, part of the European Molecular Biology Laboratory (EMBL). It includes compounds from different sources to provide further standardised descriptions of molecular entities (close to 45,000 have received the "3 stars" qualification according to curation) that enable other databases to annotate their entries consistently. ChEBI focuses on high-quality manual annotation, non-redundancy, and provision of chemical ontology. It is searchable by name, formula or structure.

PubChem is the NIH database of many millions (close to 100) of compounds (<1000 atoms) and substances (more than 200 million). Data come from 80 different vendors/depositors. Substances are "impure/duplicates", whereas compounds are single entities (PubChem Compound ID - CID gives a unique compound). Entries include synonyms, chemical properties, the chemical structure including SMILES and InChI strings, bioactivity, and links to structurally related compounds and other NCBI databases like PubMed. It is searchable by name, formula, MW range, structure, H-bond donor/acceptor count or XlogP.

ChemSpider is a free chemical structure database providing fast access to over 58 million structures, properties, and associated information. By integrating and linking compounds from ~500 data sources, ChemSpider enables researchers to obtain the most comprehensive view of freely available chemical data from a single online search. It is owned by the Royal Society of Chemistry. It is searchable by name, synonym, InChi, structure, registry #, SMILES, calculated properties (but not by formula or mass). The data includes names, synonyms, Wikipedia articles, descriptions, data sources, suppliers, patents, articles, properties, MESH headings, pharmacology links, and spectra (UV, IR, NMR, MS) sourced from other sites.

3.2.4. Metabolite-centered, spectral databases

These databases support not only MW or MW range searches, but also support parent ion searches (positive, negative, neutral), peak list searches (from MS or MS/MS data) as well as MS/MS spectral matching. These DBs, rather than the simple MW search tools, are more intended for MS-based metabolomics and compound identification.

BioCyc, developed by SRI International (Menlo Park, California), is a collection of curated databases for different organisms. Databases are organised according to the level of manual updates they have received. Tier-1 databases such as EcoCyc (for E. coli) and HumanCyc are highly curated, while most BioCyc databases (Tier 2 and 3) have been computationally derived. These databases are particularly applicable to organism-specific metabolite identification and metabolic reconstructions using the pathway search.

HMDB is a database devoted to human metabolism which has been developed with support from the Canadian Institutes of Health Research, Alberta Innovates - Health Solutions and The Metabolomics Innovation Centre. For each data entry, information is given on the chemical, biological and clinical characteristics as well as references to the literature including reported disease associations, related enzymes, and transporters, in addition to links to external databases such as KEGG.

Komic Market (Kazusa Omics Data Market) is a database of metabolite annotations from MS peaks detected in metabolomics studies. It comes from the project "Development of Fundamental Technologies for Controlling the Material Production Process of Plants", supported by the New Energy and Industrial Technology Development Organisation, Japan.

LipidBank is the official database of the Japanese Conference on the Biochemistry of Lipids (JCBL). This database is devoted to neutral lipids. It covers several different classes, and all molecular information is manually curated and approved by experts in lipid research. Each entry includes a lipid name, molecular structure, spectral information, and literature references.

LipidMaps is funded by a large-scale collaborative research grant ("Glue Grant") from the NIH National Institute of General Medical Sciences. It aims to provide identification and quantitation of mammalian lipids, including the quantification of changes in response to perturbation. LipidMaps Proteome Database (LMPD) is also included in this resource.

MassBank is a public repository of mass spectral data based on sharing identifications and structure elucidations of chemical compounds detected by mass spectrometry. MassBank is accessible through two domains: Japanese (http://massbank.jp) and European (http://massbank.eu) (NORMAN MassBank). The tool is deployed in both domains, but some functions are only provided in the Japanese one.

METLIN is a trademark of the Scripps Research Institute, which develops and applies mass spectrometry-based technologies for understanding metabolism. It includes cloud-based data processing informatics (XCMS), and nanostructure imaging mass spectrometry (NIMS). With almost

1,000,000 real compound entries (not from prediction), this is one of the largest databases available. Entries in METLIN include metabolites, lipids, steroids, plant and bacterial metabolites, small peptides and exogenous drug metabolites and toxicants. IsoMETLIN - A module for isotope-based metabolomics is also included.

MycompoundID is a web-based resource developed at the University of Alberta for identification of compounds based on chemical properties, including accurate mass. Different searches are possible including MS, MS2, PEP searches of unlabeled and dimethyl labeled peptides and chemical isotope labeled MS data. Searches are performed across an evidence-based metabolome library (EML) which consists of 8,021 known human endogenous metabolites and their predicted metabolic products including 375,809 compounds from one metabolic reaction and 10,583,901 from two reactions. In silico predicted compounds are generated from HMDB entries.

3.2.5. Mediators

CEU MassMediator, a collaborative development from the CEMBIO and the Bioengineering Laboratory of Polytechnic School at Universidad CEU San Pablo, Madrid, is a tool which performs an automated search across external data sources (HMDB, KEGG, LipidMaps, METLIN and MINE) and provides possible identifications for a given mass (unifying similar hits obtained from more than one database into a single hit).

CSI:FingerID is a database specific for MS^n identification. It supports further research on peaks unidentified at the MS level. It is a collaborative development between Friedrich Schiller University, Germany and Helsinki Institute for Information Technology at Aalto University, Finland, that combines fragmentation tree computation and machine learning to improve both the total percentage of identified molecules and the precision of identification.

MAGMa is an annotation tool developed within the eMetabolomics project, funded by the Netherlands eScience Center at Wageningen University in collaboration with the Netherlands Metabolomics Centre. MS^n data can be uploaded as a hierarchical tree of fragment peaks, based on m/z or chemical formulae, and candidate molecules are automatically retrieved from PubChem, KEGG or HMDB. A matching score is calculated based on the quality of explanation of the fragment peaks.

MassTRIX is an online tool for the annotation of high precision mass spectrometry data. Results are displayed on organism-specific KEGG pathway maps, and any additional genomic or transcriptomic information can be added. The tool was developed at the Helmholtz Zentrum München in collaboration between Philippe Schmitt-Kopplin and Karsten Suhre.

3.3. Statistical analysis

After obtaining the matrix with cleaned data containing the list of possible compounds and their abundances in each sample, the next step is to perform the statistical analysis using chemometric tools, which provides model-based descriptions of the biological variation in the system under study. Two different statistical analyses are used in non-targeted metabolomics data analysis: univariate and multivariate data analysis (MVDA).

The univariate data analysis assumes that only one factor influences the response of variables. In a complex disease, different biologic pathways simultaneously governed by multiple variables are involved and compromised. The traditional statistical analysis tends to transform all problems into univariate problems, even those that are inherently multivariate. For this reason, MVDA is a proper statistical tool for the interpretation of data coming from a metabolomics study. It summarizes data tables with many variables and few observations and works by reducing the number of variables and classifying the data. MVDA provides statistical models which specifically single out representatives of metabolites of interest (annotated peaks), and which can further be definitively identified, chemically or structurally. The first step of MVDA is the creation of the X matrix, where the samples are arranged in rows and all the variables (the compounds) in columns. Subsequently, it is always necessary to perform data pre-treatment (scaling and transformation) to improve relevant information. After this, the data are analyzed through the unsupervised principal component analysis (PCA), the supervised partial least squares regression-discriminant analysis (PLS-DA) and orthogonal partial least squares-discriminant analysis (OPLS -DA). Both methods compress the original data matrix in such a way that underlying patterns may be revealed.

3.3.1. PCA

It is a dimension reduction method that is widely used for data exploration and visualization. It was first proposed in 1901 by Pearson. PCA aims to reveal underlying patterns by compressing the data, trying to retain as much as possible of the original information. PCA represents the natural starting point for any multivariate data analysis. It is useful to provide a global graphical overview of all samples in the data matrix, revealing outliers, groups, clusters, similarities/dissimilarities, dominating variables. The position of each sample in the model scatter plot is used to relate them to each other: samples that are close to each other have a similar multivariate profile; on the contrary, samples that lie far from each other have dissimilar properties.

PCA and Detection of Outliers

Outliers are defined as data points whose values are out of the majority of the other data points. They influence the validity of the metabolomics results by altering the data variance and distribution and thus reducing the statistical power of the data analysis. Outliers can be biological or analytical, depending on their source. The biological ones are difficult to recognize. They result from random or induced biological variation among the samples which frequently occurs in the study of complex diseases with a heterogeneous group of patients. Besides, information from random biological outliers should be included in the statistical analysis. On the contrary, analytical outliers cause elevated distortion of the biological information and should be excluded from the statistical analysis. Their origin could be different: sampling, storage, sample treatment, analysis, feature finding or deconvolution/identification process. The decision to keep or remove any outliers should always be justified.

Once the PCA model is obtained, outliers can be highlighted by plotting the intensity of the total detected compounds versus the analysis order and observing the Hotelling's T2 Range and the 'Distance to Model' option (available in SIMCA P+). Different graphs and analyses of raw and filtered data are necessary to individuate and eliminate the outliers. In a PCA scores scatter plot, strong outliers are revealed as points far away from the elliptic border representing the 95% confidence intervals of the model variation. Hotelling's T2 Range plot depicts the distance from the origin in the model plane (score space) for each selected observation (represented as a column). If a value is located far above the critical limit, it indicates that it is out of trend in the selected range of components in the score-space. Hence the probability that the observations belong to the same class as the other samples is lower than 5%.

3.3.2. PLS-DA

It is required to discriminate between the groups and maximize group differences. It models the relationship between two or more data classes, using a series of local least square fits. PLS assesses a relationship between a descriptor matrix X and a response matrix Y. The crucial difference between PLS and PCA is that PLS is a "supervised" technique, which uses additional information to produce a statistical model, whereas PCA is "unsupervised" and does not require further data input.

PLS-DA and PCA can be used as dimension reduction methods in predictive models, before linear discriminant analysis (PLS-DA, PCA-DA). The bilinear statistical approach of partial least squares discriminant analysis (PLS-DA) is one of the most popular supervised methods used in metabolomics. In PLS-DA, the X matrix contains the data variables, while the Y matrix contains the class variable for which values are chosen to be the class descriptor.

3.3.3. OPLS-DA

An evolution of PLS, it is a linear regression method that has been employed successfully for prediction modeling in various biological and biochemical applications. One of the advantages provided by the OPLS method is its capacity to model data with both noisy and multi-collinear variables, such as spectral metabolomics data. In simple terms, OPLS uses information in the Y matrix to decompose the X matrix into blocks of structured variation correlated to and orthogonal to Y. OPLS can, analogously to PLS-DA, be used for discrimination (OPLS-DA). The main benefit of OPLS-DA is the possibility to separate predictive variation from non-predictive (orthogonal) one. However, OPLS-DA, contrary to PLS-DA, cannot be used for the analysis of more than two classes at the same time.

When a model (PCA, PLS-DA, etc.) is built, there are two quality parameters of the model (R^2 and Q^2) that are of statistical significance and can be considered the very first step in validation.

-R^2 is the percent of the variation of the training set explained by the model. It measures how well the model fits the data. A good model has large R^2, which means good reproducibility.

-Q^2, which predicts variance, gives information about the ability of the model to predict new data. A large Q^2 indicates good predictability.

To ensure that the model is powerful for diagnostics, R^2 and Q^2 should be high and not vary by more than 0.2–0.3. The disproportion between these two values provides the first warning that the model could have been formed by over-fitting. It is assumed that for biological data acceptable values are: $R^2 > 0.7$ and $Q^2 > 0.4$ [35].

3.3.4. S-plot, Jack-knife interval and Variable Importance in the Projection (VIP)

OPLS-DA is the starting point to perform the **S-plot**, the Jack-Knifed analyses, and VIP for putative biomarkers identification. The S-plot and Jack-Knife confidence intervals are useful tools to select metabolite differentiating groups. Briefly, the S-plot is a visual representation, presented as a scatter plot, of both covariance and correlation between the metabolites and the class designation displayed on a score scatter plot. S-plot can unfold the contributions of the different sources to the distribution of a data sample in a given variable.

The **Jack-Knife** technique, presented as a column plot, is employed to construct the confidence interval for estimated model parameters. The confidence interval reflects the variable stability and uncertainty and should be small. From the Jack-knife analysis, a putative list of the metabolites more relevant in differentiating groups can be created.

The **VIP** score shows the contribution of each predictor variable to the model. In SIMCA P+, VIP plots are sorted based on the importance of variables, thus VIP value is often used for variable selection. To construct an accurate model, the metabolites are usually selected with a VIP score >1.

3.3.5. Validation of the multivariate statistical models

To avoid the risk of overfitting, as the results found after MVDA are sensitive to chance-correlations, the statistical model needs to be validated. Validation tools, such as (a) permutation test and (b) cross-validation, are most widely used to provide an objective assessment of the performance and stability of a model.

(a) Permutation Test: A permutation test is used to assess the significance of a classification. A class assignment can be permuted several times, and for each permutation, a model between the data and the permuted class-assignment can be built. The discrimination between classes of the model based on the permutated class-assignment is compared to the discrimination of the model based on the original classification. The permutation test is more suitable for large datasets covering over 50 samples.

(b) Cross-Validation: It seems to be the most suitable method for medium and small datasets.

In cross-validation, the first step is to divide the samples into several groups. These groups are then sequentially withheld while the remaining samples are used to build cross-validated models. The withheld samples for each model are predicted, and the predictive ability is measured and this process is repeated until all the samples are withheld and predicted. To cross-over the data in successive rounds, using the leave 1/3 out approach, the data set is randomly divided into three parts, and 1/3 of the samples are excluded to build a model with the remaining 2/3 of the samples. The excluded samples are then predicted by this new model by performing a T-predicted score analysis, and the procedure is repeated until all the samples have been predicted at least once. If all the data are predicted correctly, the model can be considered valid.

To know more:

E. Saccenti, H.C. Hoefsloot, A.K. Smilde, J.A. Westerhuis, M.M. Hendriks. Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics*, 2014, 10:361-374.

K. Varmuza, and P. Filzmoser. Introduction to multivariate statistical analysis in chemometrics. CRC press, 2009.

J. Godzien, M. Ciborowski, S. Angulo, and C. Barbas. From numbers to a biological sense: How the strategy chosen for metabolomics data treatment may affect final results. A practical example based on urine fingerprints obtained by LC-MS. *Electrophoresis*, 2013, 34:2812-2826.

J. Trygg, E. Holmes, T. Lundstedt. Chemometrics in metabonomics. *J. Proteome Res.*, 2007, 6:469-479.

4. Data analysis

4.1. From data identification to pathways: Wise but assisted evaluation of data to evaluate biological relevance.

The list or lists of compound identifications (not yet a list of compounds) must be curated, always trying to remove or consolidate the assignations. In the case of LC-MS or CE-MS, the aspects to be taken into account include adducts and multiple charged species, fragments (neutral losses, breakdown products, rearrangements), isotope peaks, and noise peaks.

Also, cautious revision of the identifications must be done, to remove impossible assignments. Two types of possibilities to reject identifications can be established: knowledge about the relationship between the properties of the putatively identified compound (polarity, charge, adducts…) and the experimental conditions (extraction solvent, mobile phase, type of chromatography, ionization conditions…); and knowledge about the design of the experiment (organism, treatment, sample…).

Once the list has been curated, those identifications (and relative abundances, in some cases), can be submitted to a new set of bioinformatic tools that permit to evaluate the biochemical significance of those changes. These tools are complementary to the general biochemistry information collected in the textbooks, combined with new approaches to that knowledge, such as the Roche Biochemical Pathways or the IUBMB-Sigma-Nicholson Metabolic Pathways Chart, online searchable complete metabolic maps.

Besides, the bioinformatics tools related to biochemical pathways can be divided into two groups: the pathways databases, with similar characteristics to those of compounds databases, but being the records of full pathways instead of compounds; and the pathway analysis tools, able to apply mathematical algorithms to evaluate the significance of the changes observed, according to the enrichment of the signals, and the amount and relevance of the metabolites that change.

4.1.1. Pathway Databases

A biochemical pathway can be easily defined as a series of reactions converting a set of substrates into a set of products. This definition is subjective and non-standard, because the selection of metabolites which are included into one particular pathway is arbitrary, and pathways have different size (number of metabolites and reactions involved) and often overlap, because one metabolite can be part of different pathways. Also, the pathways established in one database can be different from those from another one.

Despite the fact that there are intrinsic problems common to all pathway databases, it is true that they constitute a very rich source of biological data that relates metabolites to genes, proteins, diseases, thus signaling events and processes, also providing various tools to permit visualization and gene/metabolite mapping, frequently in multiple species.

Some of the most important pathway databases are:

KEGG (Kyoto Encyclopedia of Genes and Genomes) is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from genomic and molecular-level information. It is a computer representation of the biological system, consisting of molecular building blocks of genes and proteins (genomic information) and chemical substances (chemical information) that are integrated with the knowledge on molecular wiring diagrams of interaction, reaction and relation networks (systems information). It also contains disease and drug information (health information) such as perturbations to the biological system. It includes 508 pathway maps, 17,931 compounds, and 11,015 glycans, as well as 10,476 biochemical reactions.

Recon 2 accounts for 1,789 enzyme-encoding genes, 7,440 reactions and 2,626 unique metabolites distributed over eight cellular compartments. One can explore the content of the reconstruction by searching/browsing metabolites and reactions.

The **Small Molecule Pathway Database** (SMPDB) comprises nearly 900 hand-drawn small molecule pathways, of which 384 are drug pathways, 232 diseases, and 220 general metabolic pathways. It depicts cell compartments, organelles, protein locations, quaternary structures, in order to convert gene, protein or chemical lists into pathways or disease diagnoses.

PathWhiz is a web server designed to permit the creation of colorful, biologically accurate pathway diagrams that are machine readable and interactive, with a Google Maps-style viewer.

4.1.2. Pathway Analysis

With metabolomics data, two types of pathway analysis can be considered: enrichment and topological analysis. The purpose of the metabolite enrichment is to test if there are biologically meaningful groups of metabolites that are significantly enriched in the data. Such meaningfulness is related to the pathways involved, the localization of those metabolites or the disease that can be involved. Topological analysis (Pathway analysis *sensu stricto*) is developed to extend and enhance metabolite set enrichment analysis for pathways by considering pathway structures and supporting pathway visualization.

4.1.3. Metabolite enrichment

IMet-Q (intelligent Metabolomic Quantitation) is an automated tool with friendly user interfaces for quantifying metabolites in full-scan liquid chromatography-mass spectrometry (LC-MS) data. iMet-Q has a complete quantitation procedure for noise removal, peak detection, and peak alignment. Furthermore, it gives the charge states and isotope ratios of detected metabolite peaks to facilitate metabolite identification.

MBRole (Metabolite Biological Role) performs overrepresentation (enrichment) analysis of categorical annotations for a set of compounds of interest. These categorical annotations correspond to biological and chemical information available in a number of public databases and software. Although focused on qualitative annotations, MBRole is general and versatile enough to perform functional enrichment analysis in any metabolomic sample (including additional biological and chemical annotations for human metabolites), and hence complements existing software for the rising field of metabolomics.

BiNChE is an enrichment analysis tool for small molecules based on the ChEBI Ontology. BiNChE displays an interactive graph that can be exported as a high-resolution image or in network formats. The tool provides plain, weighted and fragment analysis based on either the ChEBI Role Ontology or the ChEBI Structural Ontology. BiNChE aids in the exploration of large sets of small molecules produced within Metabolomics or other Systems Biology research contexts. The open-source tool provides easy and highly interactive web access to enrichment analysis with the ChEBI ontology tool and is additionally available as a standalone library.

IMPaLA is a web tool for the joint pathway analysis of transcriptomics or proteomics and metabolomics data. IMPaLA performs over-representation or enrichment analysis with user-specified lists of metabolites and genes using over 3000 pre-annotated pathways from 11 databases. As a result, pathways can be identified that may be dysregulated on the transcriptional level, the metabolic level or both. Evidence of pathway dysregulation is combined, allowing for the identification of additional pathways with a changed activity that would not be highlighted if the analysis were applied to any of the functional levels alone.

MPEA (Metabolite Pathway Enrichment Analysis) is a rapid tool for functional analysis and biological interpretation of metabolic profiling data. MPEA follows the concept of gene set enrichment analysis (GSEA) and tests whether metabolites involved in some predefined pathway occur towards the top (or bottom) of a ranked query compound list. In particular, MPEA is designed to handle many-to-many relationships that may occur between the query compounds and metabolite annotations.

MSEA (Metabolite Set Enrichment Analysis) is a web-based tool to help identify and interpret patterns of metabolite concentration changes in a biologically meaningful context for human and mammalian metabolomic studies. Key to the development of MSEA has been the creation of a library of approximately 1000 predefined metabolite sets covering various metabolic pathways, disease states, biofluids, and tissue locations. MSEA also supports user-defined or custom metabolite sets for more specialized analysis. MSEA offers three different enrichment analyses for metabolomic studies including overrepresentation analysis (ORA), single sample profiling (SSP) and quantitative enrichment analysis (QEA). ORA requires only a list of compound names, while SSP and QEA require both compound names and compound concentrations. MSEA generates easily understood graphs or tables embedded with hyperlinks to relevant pathway images and disease descriptors. For non-mammalian or more specialized metabolomic studies, MSEA allows users to provide their own metabolite sets for enrichment analysis. The MSEA server also supports conversion between common metabolite names, synonyms, and major database identifiers.

4.1.4. (Topological) Pathway Analysis

The structure of biological pathways represents our knowledge about the complex relationships between molecules (activation, inhibition, reaction, etc.). However, neither over-representation analysis or pathway enrichment analysis consider the pathway structure when determining which pathways are more likely to be involved in the conditions under study. It is obvious that changes in the key positions of a network will have more severe impact on the pathway than changes on marginal or relatively isolated positions. There are two well-established node centrality measures to estimate node importance - betweenness centrality and degree centrality. The former focuses on node relative to overall pathway structure, while the latter focuses on immediate local connectivities.

There are many commercial pathway analysis software tools, such as Pathway Studio, MetaCore, Ingenuity Pathway Analysis, etc. Compared to them, the pathway analysis in **MetaboAnalyst** is a free, web-based tool designed for metabolomics data analysis. It uses the high-quality KEGG metabolic pathways as the backend knowledgebase. It integrates many well-established (e.g., univariate analysis, over-representation analysis) methods, as well as novel algorithms/ concepts (GlobalTest, GlobalAncova, pathway topology analysis) into pathway analysis. Also, MetPA implements a Google-Map style interactive visualization system to help users understand their analysis results.

4.2. Biomarker validation: Assigning relevance to metabolites

Once a biomarker (or a group of them) is proposed as a relevant metabolite set that could classify the samples according to the conditions under study, it is necessary to set up a targeted, quantitative experiment, applied to a larger group of samples (patients, in a clinical study, for instance), ideally, independent from those selected for performing the discovery phase. Results from this experiment are analysed to evaluate the validity of such biomarker, and the most commonly used methodology for it is based on the evaluation of the receiver operator characteristic (ROC) curves.

ROC curve analysis is the most objective and statistically valid method for biomarker performance evaluation. It allows building a predictive model from one or more variables, which can be used to classify new subjects into specific groups (*e.g.*, healthy vs. diseased). A ROC curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The true-positive rate is also known as sensitivity, recall or probability of detection in machine learning. The false-positive rate is also known as the fall-out or probability of false alarm and can be calculated as (1 – specificity).

The area under the curve (AUC) ranks from 0.5 to 1.0, and the goal for the researcher would be to maximize AUC under ROC curve while minimizing the number of metabolites used in the biomarker panel.

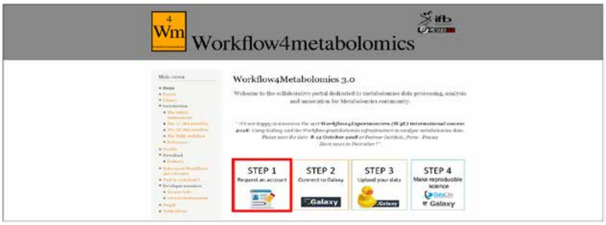
The main tool for building ROC curves was Roccet, which is nowadays implemented into the **MetaboAnalyst** toolbox.

5. Practical sessions on metabolomics

The practical sessions will start with an introductory visit to the laboratory and the identification of the main analytical tools: tissue homogenizers, ultrasounds sources, ultracentrifuges, Speed-Vac & nitrogen stream for concentration, inserts & vials, caps, septa, clamps, etc. along with the analytical instrumentation coupled to mass spectrometry mainly used in metabolomics. GC-MS-Q; LC-MS-QTOF; LC-MS-QqQ; CE-MS-TOF.

5.1. Targeted and non-targeted metabolomics

A protocol for LC-MS based metabolomics with plasma samples will be read and discussed interactively with the students in the laboratory. A tutorial session will show all the steps performed in a professional environment, using only open source, free software. The metabolomics study will consist of two groups: CASE and CONTROL with six samples each plus 4 QC samples analysed using LC-MS with positive and negative ionization.



Workflow4metabolomics

Workflow4Metabolomics 3.0

Welcome to the collaborative portal dedicated to metabolomics data processing, analysis and annotation for Metabolomics community.

5.2. Preprocessing with Workflow for Metabolomics (W4M)

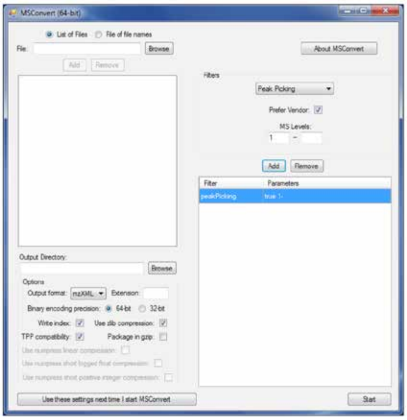
This tutorial will show how to prepare a data set and use Workflow4Metabolomics online infrastructure built on Galaxy environment. W4M is a collaborative project between metabolomics (MetaboHUB French infrastructure) and bioinformatics platforms (IFB: Institut Français de Bioinformatique) and provides an open-source platform for computational metabolomics.

Before starting, each student will need to create an account in the **Workflow4Metabolomics**. W4M: http://workflow4metabolomics.org/. Click on "STEP 1: Request an account" and fill in the form. Credentials are provided within one working day.

5.2.1.MS files conversion

Within W4M environment, MS files must be converted into an open format, such as .mzML, .mzXML, .mzData, and .cdf. If the commercial software (usually supplied with the equipment) does not include an exporting tool, other open source software as ProteoWizard (http://proteowizard.sourceforge.net/) can be used for raw data conversion.

It is recommended to use only letters (lower or upper case), digits, and "_" characters (with no blank) in all the names provided (sample names, file names, folder names, variables, etc.), as well as to convert raw files to **mzXML** in centroid mode (smaller files) so that the files are compatible with the xmcs centWave method.



MSConvert (64-bit)

● List of Files ○ File of file names

Filters: Peak Picking

Prefer Vendor: ☑

MS Levels: 1 -

Filter	Parameters
peakPicking	true 1-

Output Directory:

Options

Output format: mzXML Extension:

Binary encoding precision: ● 64-bit ○ 32-bit

Write index: ☑ Use zlib compression: ☑

TPP compatibility: ☑ Package in gzip: □

Use numpress linear compression: □

Use numpress short logged float compression: □

Use numpress short positive integer compression: □

Recommended parameters for MS File Conversion for W4M

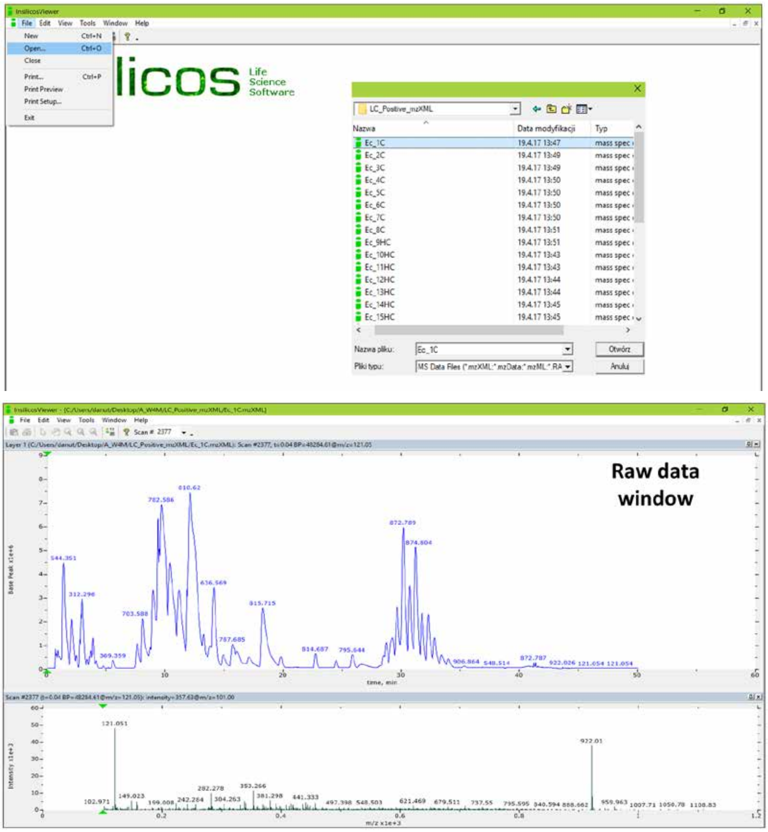
5.2.2. Raw data inspection

Usually, researchers use the software provided by the MS Instrument supplier, but there are also open source alternatives, such as **InsilicosViewer**, a free viewer for displaying mzXML MS data with most of the functionalities offered by its commercial counterparts. This tool includes mzXML/mzData reading software developed by the Institute of System Biology (http://www.systemsbiology.org)

Open your .mzXML format raw data.

1. Select File > Open File.
2. Go to your folder that contains the mzXML format data files.
3. Select your data.
4. Click on Open.

The data file appears in the Raw Data Window. (see Figure)



InsilicosViewer

LC_Positive_mzXML

Nazwa	Data modyfikacji	Typ
Ec_1C	19.4.17 13:47	mass spec
Ec_2C	19.4.17 13:49	mass spec
Ec_3C	19.4.17 13:49	mass spec
Ec_4C	19.4.17 13:50	mass spec
Ec_5C	19.4.17 13:50	mass spec
Ec_6C	19.4.17 13:50	mass spec
Ec_7C	19.4.17 13:50	mass spec
Ec_8C	19.4.17 13:51	mass spec
Ec_9HC	19.4.17 13:51	mass spec
Ec_10HC	19.4.17 13:43	mass spec
Ec_11HC	19.4.17 13:43	mass spec
Ec_12HC	19.4.17 13:44	mass spec
Ec_13HC	19.4.17 13:44	mass spec
Ec_14HC	19.4.17 13:45	mass spec
Ec_15HC	19.4.17 13:45	mass spec

Nazwa pliku: Ec_1C

Pliki typu: MS Data Files (*.mzXML;*.mzData;*.mzML;*.RA

InsilicosViewer - [C:/Users/danut/Desktop/A_W4M/LC_Positive_mzXML/Ec_1C.mzXML]

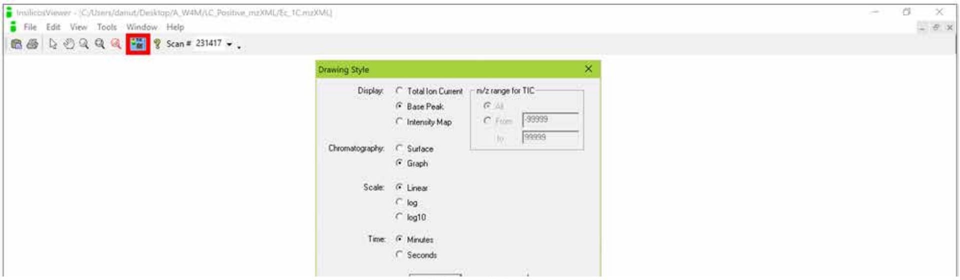
Scan # 2377

Layer 1 (C:/Users/danut/Desktop/A_W4M/LC_Positive_mzXML/Ec_1C.mzXML): Scan #2377, t=0.04 BP=48284.61@m/z=121.05

Scan #2377 (t=0.04 BP=48284.61@m/z=121.05): intensity=357.63@m/z=101.00

Figure. Raw data window

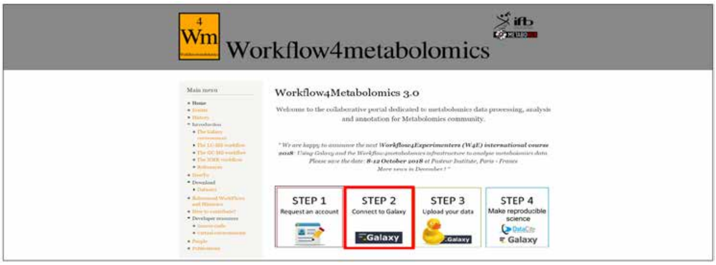
To display more drawing style click on > TIC icon or Tools > Drawing style
To display more drawing styles click on > TIC icon or Tools > Drawing style


InsilicosViewer - [C:/Users/danut/Desktop/A_W4M/LC_Positive_mzXML/Ec_1C.mzXML]
File Edit View Tools Window Help
Scan # 231417

Drawing Style

Display: ○ Total Ion Current ● Base Peak ○ Intensity Map
m/z range for TIC: ● All ○ From -99999 to 99999
Chromatography: ○ Surface ● Graph
Scale: ● Linear ○ log ○ log10
Time: ● Minutes ○ Seconds

5.2.3. Upload data into Galaxy



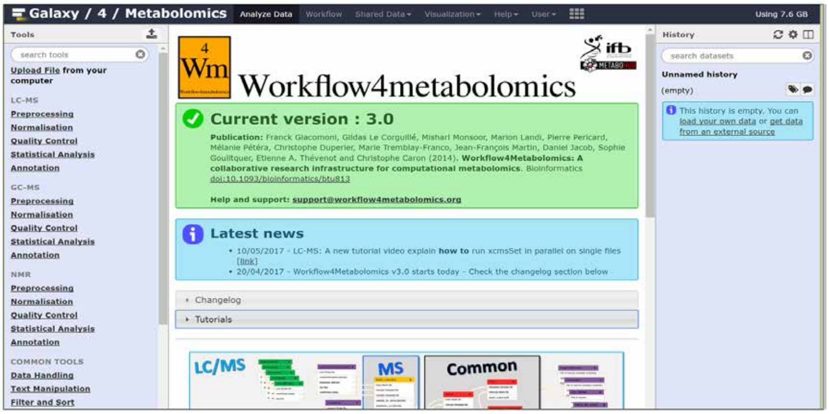
Workflow4Metabolomics 3.0

Welcome to the collaborative portal dedicated to metabolomics data processing, analysis and annotation for Metabolomics community.

" We are happy to announce the next **Workflow4Experimenters (W4E) international course 2018**: Using Galaxy and the Workflow4metabolomics infrastructure to analyse metabolomics data. Please save the date: **8-12 October 2018** at Pasteur Institute, Paris - France. More news in December ! "

STEP 1 Request an account
STEP 2 Connect to Galaxy
STEP 3 Upload your data
STEP 4 Make reproducible science

The main window will be displayed.



This history is empty. You can load your own data or get data from an external source

Current version : 3.0

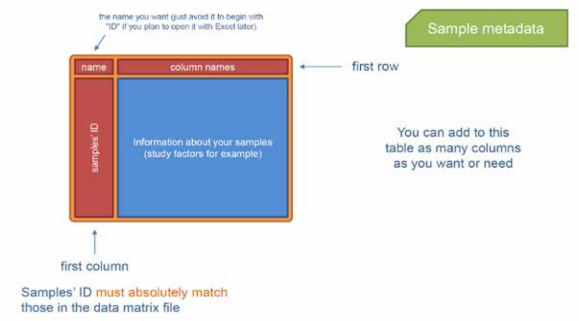
Publication: Franck Giacomoni, Gildas Le Corguillé, Misharl Monsoor, Marion Landi, Pierre Pericard, Mélanie Pétéra, Christophe Duperier, Marie Tremblay-Franco, Jean-François Martin, Daniel Jacob, Sophie Goulitquer, Etienne A. Thévenot and Christophe Caron (2014). **Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics**. Bioinformatics doi:10.1093/bioinformatics/btu813

Help and support: support@workflow4metabolomics.org

Latest news

- 10/05/2017 - LC-MS: A new tutorial video explain **how to** run xcmsSet in parallel on single files [link]
- 20/04/2017 - Workflow4Metabolomics v3.0 starts today - Check the changelog section below

The data (files) to be uploaded must be accompanied by the sample meta datafile: information concerning your samples and this information must be provided to the system, following strict criteria:


the name you want (just avoid it to begin with "ID" if you plan to open it with Excel later)
Sample metadata
name
column names
first row
samples' ID
Information about your samples (study factors for example)
You can add to this table as many columns as you want or need
first column
Samples' ID must absolutely match those in the data matrix file

In the particular case of the samples which will be used in this practical session, the samples information followed the structure described in the table:

Sample Name	class	polarity	Sample Type	batch	Injection Order	diet
QC	one	positive	pool	B1	1	NA
C1	one	positive	sample	B1	7	C
HC3	one	positive	sample	B1	10	HC
BL	one	positive	blank	B1	12	NA
...	…	…	…	…	…	…

The files must be tabulated: TSV files or TXT files with tabulation as a separator.

There are two methods for data upload: with a set of single files or creating first a .zip file, with a predefined structure of subdirectories. We will practice only the first option.:

- Single file (recommended): You can put a single file as the input. That way, you will be able to launch several xcmsSet in parallel and use "xcms.xcmsSet Merger" before "xcms.group"
- Zip file: You can put a zip file containing your inputs: myinputs.zip (containing all your conditions as sub-directories).

Zip file: Steps for creating the zip file

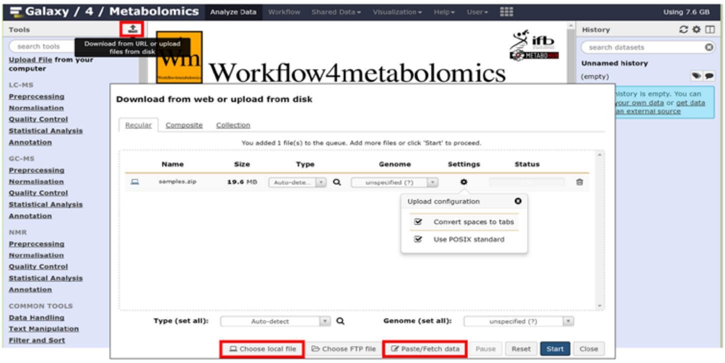
Step1: Creating your directory and hierarchize the subdirectories

VERY IMPORTANT: If you zip your files under Windows, you must use the 7Zip software (http://www.7-zip.org/), otherwise your zip will not be properly unzipped on W4M platform (zip corrupted bug).

Step2: Create your zip file

Data import <2Gb

To import the data use option Paste/Fetch data or Choose local file > From the option Upload configuration click on Convert spaces to tabs > Start > Close



Download from web or upload from disk

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
samples.zip	19.6 MB	Auto-dete...	unspecified (?)		

Upload configuration

☑ Convert spaces to tabs

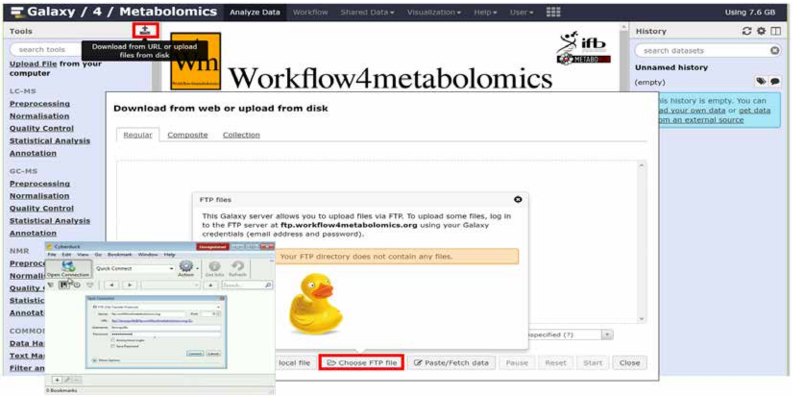
☑ Use POSIX standard

Type (set all): Auto-detect

Genome (set all): unspecified (?)

…istory is empty. You can …your own data or get data …an external source

Data import <2Gb



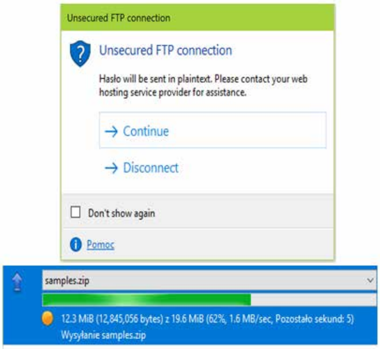
Download from web or upload from disk

FTP files

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **ftp.workflow4metabolomics.org** using your Galaxy credentials (email address and password).

Your FTP directory does not contain any files.

Step 1. Choose a FTP Client > Go to Cyberduck > Open Connection > Server: ftp.workflow4metabolomics.org > and type your Username and password, next Open Connection and copy paste your files.



Unsecured FTP connection

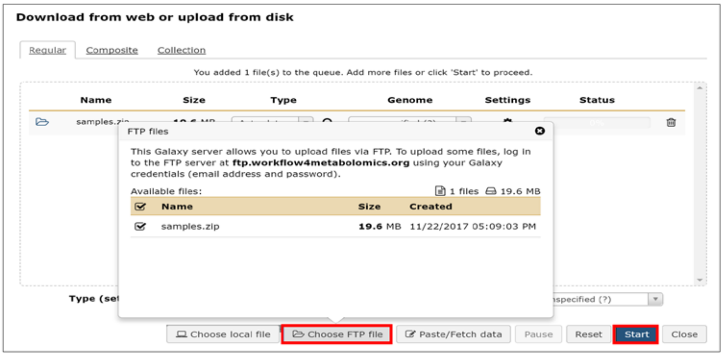
Hasło will be sent in plaintext. Please contact your web hosting service provider for assistance.

□ Don't show again

samples.zip

12.3 MiB (12,845,056 bytes) z 19.6 MiB (62%, 1.6 MB/sec, Pozostało sekund: 5)

Wysyłanie samples.zip



You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

FTP files

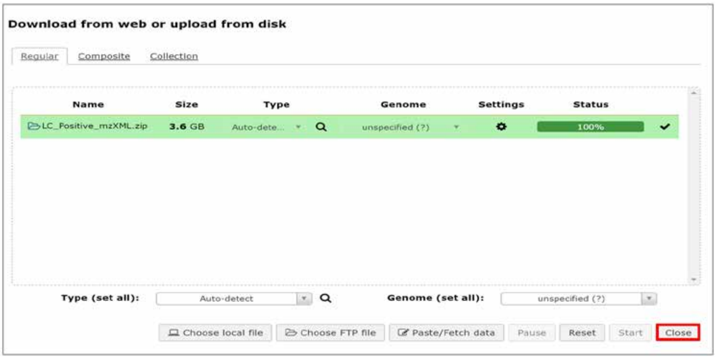
This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **ftp.workflow4metabolomics.org** using your Galaxy credentials (email address and password).

Available files: 1 files 19.6 MB

☑	Name	Size	Created
☑	samples.zip	**19.6** MB	11/22/2017 05:09:03 PM

When the transfer is completed, select again Choose FTP file > select your data > Start.

Your data will be uploaded to the WM platform.

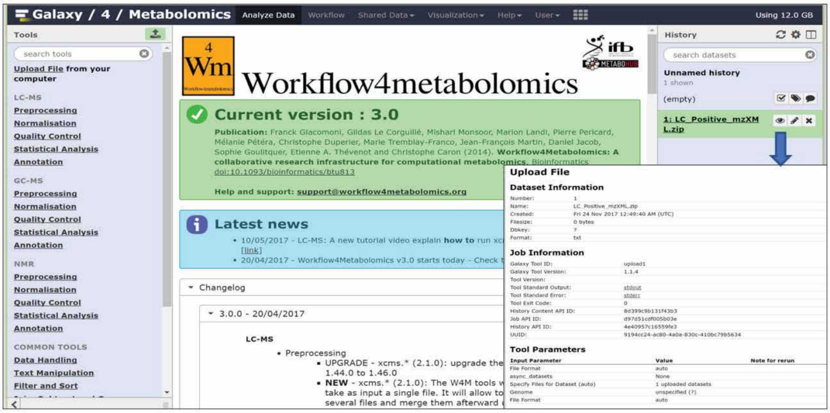


Name	Size	Type	Genome	Settings	Status
LC_Positive_mzXML.zip	**3.6** GB	Auto-dete...	unspecified (?)		100% ✓

Type (set all): Auto-detect

Genome (set all): unspecified (?)

You can view the details of your uploaded data



Current version : 3.0

Publication: Franck Giacomoni, Gildas Le Corguillé, Misharl Monsoor, Marion Landi, Pierre Pericard, Mélanie Pétéra, Christophe Duperier, Marie Tremblay-Franco, Jean-François Martin, Daniel Jacob, Sophie Goulitquer, Etienne A. Thévenot and Christophe Caron (2014). **Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics**. Bioinformatics doi:10.1093/bioinformatics/btu813

Help and support: support@workflow4metabolomics.org

Latest news

- 10/05/2017 - LC-MS: A new tutorial video explain **how to** run xc [link]
- 20/04/2017 - Workflow4Metabolomics v3.0 starts today - Check t

Changelog

3.0.0 - 20/04/2017

LC-MS

- Preprocessing
  - UPGRADE - xcms.* (2.1.0): upgrade the 1.44.0 to 1.46.0
  - **NEW** - xcms.* (2.1.0): The W4M tools w take as input a single file. It will allow to several files and merge them afterward

Upload File

Dataset Information

Number:	1
Name:	LC_Positive_mzXML.zip
Created:	Fri 24 Nov 2017 12:49:40 AM (UTC)
Filesize:	0 bytes
Dbkey:	?
Format:	txt

Job Information

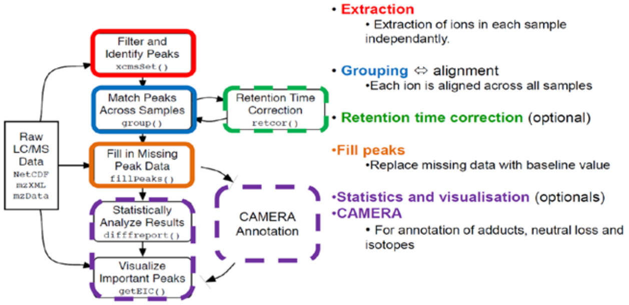
Galaxy Tool ID:	upload1
Galaxy Tool Version:	1.1.4
Tool Version:	
Tool Standard Output:	stdout
Tool Standard Error:	stderr
Tool Exit Code:	0
History Content API ID:	8d399c9b131f43b3
Job API ID:	d97d51cdf005b03e
History API ID:	4e40957c16559fe3
UUID:	9194cc24-ac80-4a0a-830c-410bc79b5634

Tool Parameters

Input Parameter	Value	Note for rerun
File Format	auto	
async_datasets	None	
Specify Files for Dataset (auto)	1 uploaded datasets	
Genome	unspecified (?)	
File Format	auto	

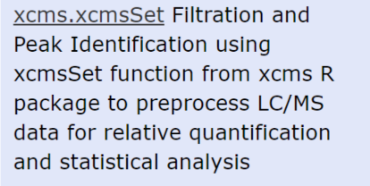
5.2.4. LC-MS pre-processing using XCMS

XCMS is an R software package dedicated to the peak extraction and retention time alignment from mass spectrometry coupled with gas or liquid chromatography (GC-MS and LC-MS) acquisition files.


Filter and Identify Peaks
xcmsSet()
Match Peaks Across Samples
group()
Retention Time Correction
retcor()
Raw LC/MS Data
NetCDF
mzXML
mzData
Fill in Missing Peak Data
fillPeaks()
Statistically Analyze Results
diffreport()
CAMERA Annotation
Visualize Important Peaks
getEIC()

- **Extraction**
    - Extraction of ions in each sample independantly.
- **Grouping** ⇔ alignment
    - Each ion is aligned across all samples
- **Retention time correction** (optional)
- **Fill peaks**
    - Replace missing data with baseline value
- **Statistics and visualisation** (optionals)
- **CAMERA**
    - For annotation of adducts, neutral loss and isotopes

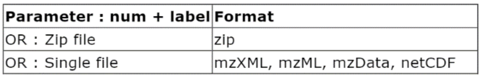
Peak extraction - xcmsSet



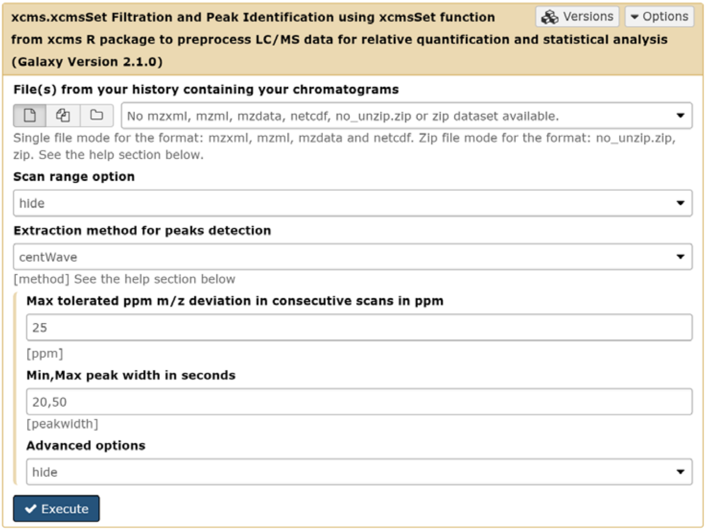
xcms.xcmsSet Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

This tool is used for preprocessing data from multiple LC/MS files (forma ts NetCDF, mzXML, and mzData). It extracts ion from each sample independently, using a statistic model; peaks are filtered and integrated.

Input files:



Parameter : num + label	Format
OR : Zip file	zip
OR : Single file	mzXML, mzML, mzData, netCDF



xcms.xcmsSet Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis (Galaxy Version 2.1.0)

File(s) from your history containing your chromatograms

No mzxml, mzml, mzdata, netcdf, no_unzip.zip or zip dataset available.

Single file mode for the format: mzxml, mzml, mzdata and netcdf. Zip file mode for the format: no_unzip.zip, zip. See the help section below.

Scan range option

hide

Extraction method for peaks detection

centWave

[method] See the help section below

Max tolerated ppm m/z deviation in consecutive scans in ppm

25

[ppm]

Min,Max peak width in seconds

20,50

[peakwidth]

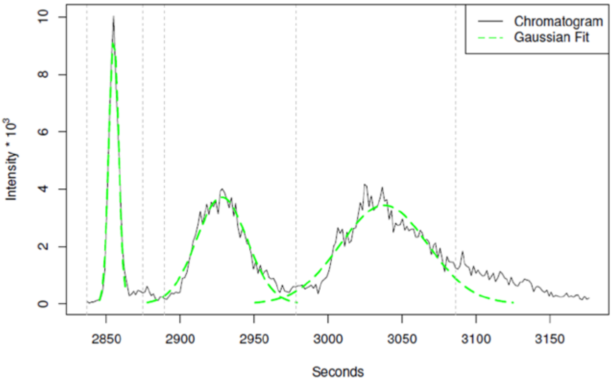
Advanced options

hide

Parameters:

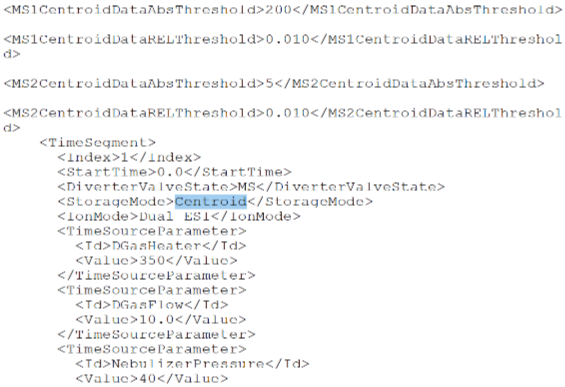
<u>Extraction method for peak detection:</u>

- **Matched Filter:** is dedicated to centroid or profile low-resolution MS data
- **Centwave:** is dedicated to high-resolution centroid data. The algorithm aims at detecting "Mass traces" or "region of interest (ROI)" which are defined as regions with less than a defined deviation of m/z in consecutive scans. This deviation must be lower than the value of the "ppm" parameter.



- **MSW:** Continuous wavelet transform. Wavelet-based, used for direct infusion data, can be used to locate chromatographic peaks on different scales.

You can check the type of your data in your raw data file.



```
<MS1CentroidDataAbsThreshold>200</MS1CentroidDataAbsThreshold>

<MS1CentroidDataRELThreshold>0.010</MS1CentroidDataRELThreshol
d>

<MS2CentroidDataAbsThreshold>5</MS2CentroidDataAbsThreshold>

<MS2CentroidDataRELThreshold>0.010</MS2CentroidDataRELThreshol
d>
    <TimeSegment>
      <Index>1</Index>
      <StartTime>0.0</StartTime>
      <DiverterValveState>MS</DiverterValveState>
      <StorageMode>Centroid</StorageMode>
      <IonMode>Dual ESI</IonMode>
      <TimeSourceParameter>
        <Td>DGasHeater</Td>
        <Value>350</Value>
      </TimeSourceParameter>
      <TimeSourceParameter>
        <Td>DGasFlow</Td>
        <Value>10.0</Value>
      </TimeSourceParameter>
      <TimeSourceParameter>
        <Td>NebulizerPressure</Td>
        <Value>40</Value>
```

Max. tolerated ppm m/z deviation in consecutive scans in ppm (related to m/z)

Fluctuation of m/z value (ppm) from scan to scan. This parameter has to be set according to mass spectrometer accuracy.

Min., Max. peak width in seconds: (related to RT)

The main purpose of the peak width parameter is to roughly estimate the peak width range; this parameter is not a threshold. The wavelets used for peak detection are calculated from this parameter.

Important: Do not choose the minimum peak too small, as it will not increase sensitivity, but will cause peaks to be split.

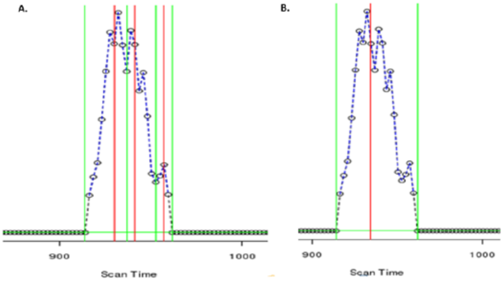
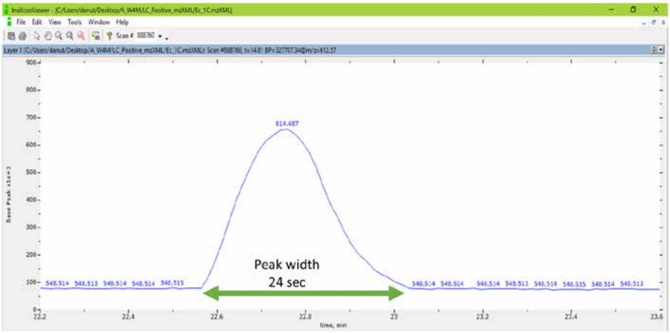
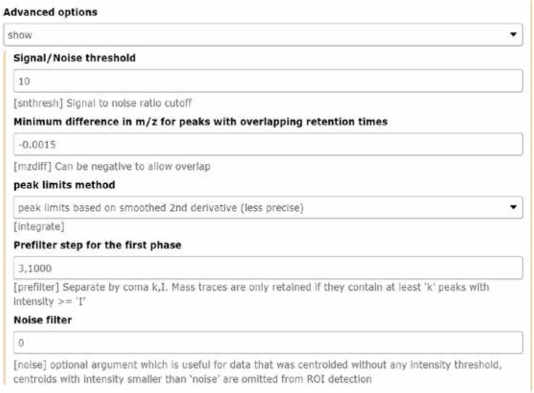


Figure. Example: peak width 45s **A.** Using the peak width (20,60) the peak will be split into three peaks, each detected as a 10s wide separate peak; **B.** Using peak width of (20-120) will keep the peak intact.

You can check that parameter in your raw data. Calculate the peak width for a narrow peak and a wide one.


InsilicosViewer - [C:/Users/danut/Desktop/A_W4M/LC_Positive_mzXML/Ec_1C.mzXML]
File Edit View Tools Window Help
Scan # 888760
Layer 1 [C:/Users/danut/Desktop/A_W4M/LC_Positive_mzXML/Ec_1C.mzXML]: Scan #888760, t=14.81 BP=327707.34@m/z=612.57
900
800
700
600
500
400
300
200
100
0
Base Peak x1e+3
614.687
548.514
548.513
548.515
Peak width
24 sec
22.2
22.4
22.6
22.8
23
23.2
23.4
23.6
time, min

Advanced options:



Advanced options

show ▾

Signal/Noise threshold

10

[snthresh] Signal to noise ratio cutoff

Minimum difference in m/z for peaks with overlapping retention times

-0.0015

[mzdiff] Can be negative to allow overlap

peak limits method

peak limits based on smoothed 2nd derivative (less precise) ▾

[integrate]

Prefilter step for the first phase

3,1000

[prefilter] Separate by coma k,I. Mass traces are only retained if they contain at least 'k' peaks with intensity >= 'I'

Noise filter

0

[noise] optional argument which is useful for data that was centroided without any intensity threshold, centroids with intensity smaller than 'noise' are omitted from ROI detection

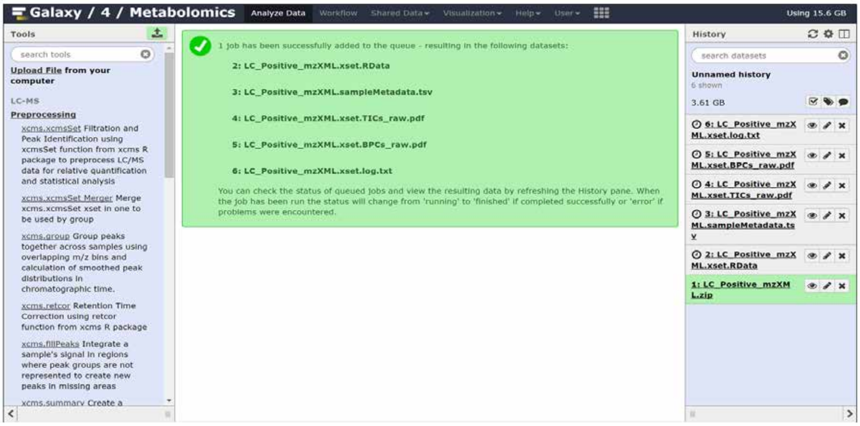
Minimum difference in m/z for peaks with overlapping retention times (related to m/z and RT):

Minimum difference of m/z for peaks with overlapping RT (co-eluting peak). Must be negative to allow overlap.

Prefilter (related to intensity):

A peak must be present in n scans with an intensity greater than k.

After setting all the parameters press **Execute** and wait... This step takes time.



Upload File from your computer

LC-MS

Preprocessing

xcms.xcmsSet Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

xcms.xcmsSet Merger Merge xcms.xcmsSet xset in one to be used by group

xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

xcms.retcor Retention Time Correction using retcor function from xcms R package

xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

xcms.summary Create a

1 job has been successfully added to the queue - resulting in the following datasets:

2: LC_Positive_mzXML.xset.RData

3: LC_Positive_mzXML.sampleMetadata.tsv

4: LC_Positive_mzXML.xset.TICs_raw.pdf

5: LC_Positive_mzXML.xset.BPCs_raw.pdf

6: LC_Positive_mzXML.xset.log.txt

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

6: LC_Positive_mzXML.xset.log.txt

5: LC_Positive_mzXML.xset.BPCs_raw.pdf

4: LC_Positive_mzXML.xset.TICs_raw.pdf

3: LC_Positive_mzXML.sampleMetadata.tsv

2: LC_Positive_mzXML.xset.RData

1: LC_Positive_mzXML.zip

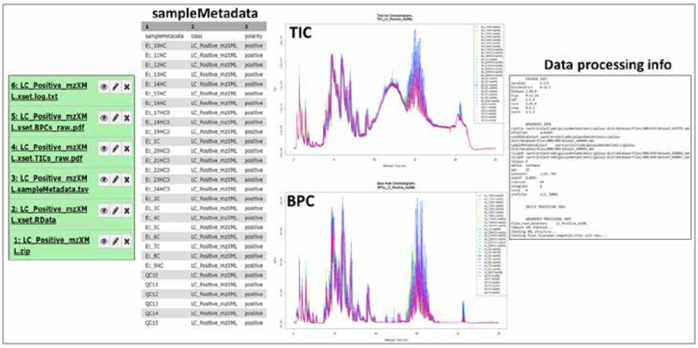
Output files (History):

xset.TICs_raw.pdf "Total Ion Chromatograms" graph in pdf format**.**

xset.BPCs_raw.pdf "Base Peak Chromatograms" graph in pdf format with each class sample opposed.

sampleMetadata.tsv Tabular file that contains information for each sample, it is associated with class and polarity (positive, negative and mixed). This file is necessary for the ANOVA and PCA step of the workflow.

xset.RData: rdata.xcms.raw format. **Rdata file that is necessary for the second step of the workflow "xcms.group".**



6: LC_Positive_mzXML.xset.log.txt

5: LC_Positive_mzXML.xset.BPCs_raw.pdf

4: LC_Positive_mzXML.xset.TICs_raw.pdf

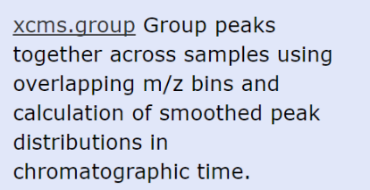
3: LC_Positive_mzXML.sampleMetadata.tsv

2: LC_Positive_mzXML.xset.RData

1: LC_Positive_mzXML.zip

1	2	3
sampleMetadata	class	polarity
Ec_10HC	LC_Positive_mzXML	positive
Ec_11HC	LC_Positive_mzXML	positive
Ec_12HC	LC_Positive_mzXML	positive
Ec_13HC	LC_Positive_mzXML	positive
Ec_14HC	LC_Positive_mzXML	positive
Ec_15HC	LC_Positive_mzXML	positive
Ec_16HC	LC_Positive_mzXML	positive
Ec_17HCD	LC_Positive_mzXML	positive
Ec_18HCD	LC_Positive_mzXML	positive
Ec_19HCD	LC_Positive_mzXML	positive
Ec_1C	LC_Positive_mzXML	positive
Ec_20HCD	LC_Positive_mzXML	positive
Ec_21HCD	LC_Positive_mzXML	positive
Ec_22HCD	LC_Positive_mzXML	positive
Ec_23HCD	LC_Positive_mzXML	positive
Ec_24HCD	LC_Positive_mzXML	positive
Ec_2C	LC_Positive_mzXML	positive
Ec_3C	LC_Positive_mzXML	positive
Ec_4C	LC_Positive_mzXML	positive
Ec_5C	LC_Positive_mzXML	positive
Ec_6C	LC_Positive_mzXML	positive
Ec_7C	LC_Positive_mzXML	positive
Ec_8C	LC_Positive_mzXML	positive
Ec_9HC	LC_Positive_mzXML	positive
QC10	LC_Positive_mzXML	positive
QC11	LC_Positive_mzXML	positive
QC12	LC_Positive_mzXML	positive
QC13	LC_Positive_mzXML	positive
QC14	LC_Positive_mzXML	positive
QC15	LC_Positive_mzXML	positive

The output file is an **xset.RData file**. You can continue your analysis using it in **xcms.group** tool.

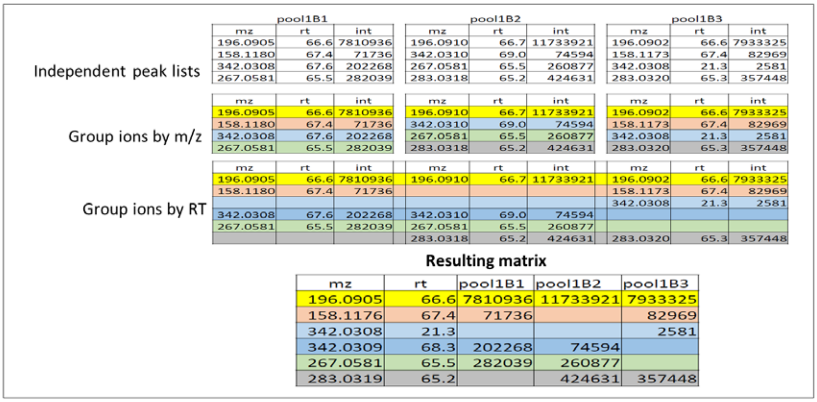


xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

After matching peaks into groups, xcms can use those groups to identify and correct correlated drifts in retention time from run to run.

The aligned peaks can then be used for a second pass of peak grouping which will be more accurate than the first one.

The whole process can be repeated iteratively. Not all peak groups will be helpful for identifying retention time drifts. Some groups may be missing peaks from a large fraction of samples and thus provide an incomplete picture of the drift at that time point. Still, others may contain multiple peaks from the same sample, which is a sign of improper grouping.



Independent peak lists

pool1B1			pool1B2			pool1B3		
mz	rt	int	mz	rt	int	mz	rt	int
196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325
158.1180	67.4	71736	342.0310	69.0	74594	158.1173	67.4	82969
342.0308	67.6	202268	267.0581	65.5	260877	342.0308	21.3	2581
267.0581	65.5	282039	283.0318	65.2	424631	283.0320	65.3	357448

Group ions by m/z

mz	rt	int	mz	rt	int	mz	rt	int
196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325
158.1180	67.4	71736	342.0310	69.0	74594	158.1173	67.4	82969
342.0308	67.6	202268	267.0581	65.5	260877	342.0308	21.3	2581
267.0581	65.5	282039	283.0318	65.2	424631	283.0320	65.3	357448

Group ions by RT

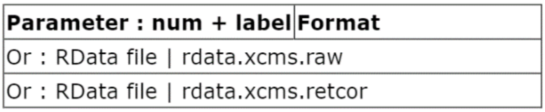
mz	rt	int	mz	rt	int	mz	rt	int
196.0905	66.6	7810936	196.0910	66.7	11733921	196.0902	66.6	7933325
158.1180	67.4	71736				158.1173	67.4	82969
						342.0308	21.3	2581
342.0308	67.6	202268	342.0310	69.0	74594			
267.0581	65.5	282039	267.0581	65.5	260877			
			283.0318	65.2	424631	283.0320	65.3	357448

Resulting matrix

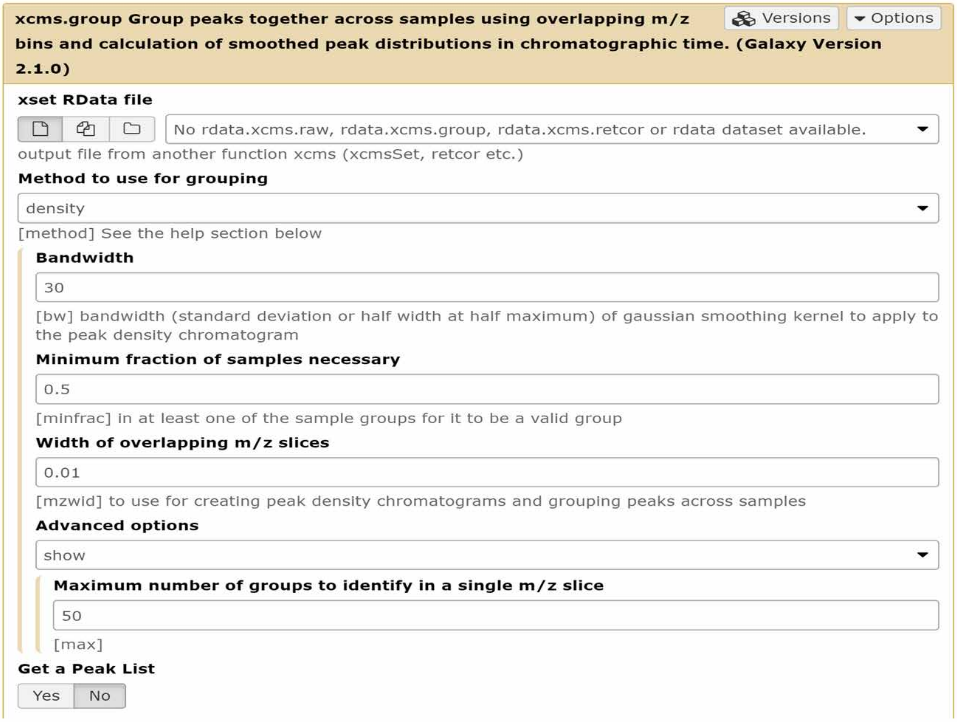
mz	rt	pool1B1	pool1B2	pool1B3
196.0905	66.6	7810936	11733921	7933325
158.1176	67.4	71736		82969
342.0308	21.3			2581
342.0309	68.3	202268	74594	
267.0581	65.5	282039	260877	
283.0319	65.2		424631	357448

After peak identification with xcmsSet, this tool groups the peaks which represent the same analyte across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time. It allows for rejection of features, which are only partially detected within the replicates of a sample class.

Input files:



Parameter : num + label	Format
Or : RData file | rdata.xcms.raw	
Or : RData file | rdata.xcms.retcor	



xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time. (Galaxy Version 2.1.0)

xset RData file

No rdata.xcms.raw, rdata.xcms.group, rdata.xcms.retcor or rdata dataset available.

output file from another function xcms (xcmsSet, retcor etc.)

Method to use for grouping

density

[method] See the help section below

Bandwidth

30

[bw] bandwidth (standard deviation or half width at half maximum) of gaussian smoothing kernel to apply to the peak density chromatogram

Minimum fraction of samples necessary

0.5

[minfrac] in at least one of the sample groups for it to be a valid group

Width of overlapping m/z slices

0.01

[mzwid] to use for creating peak density chromatograms and grouping peaks across samples

Advanced options

show

Maximum number of groups to identify in a single m/z slice

50

[max]

Get a Peak List

Yes No

Parameters:

Method to use for grouping

mzClust. Runs high-resolution alignment on single spectra samples stored in the RData file generated by the **xcmsSet tool**.

Density. Groups peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

Nearest. Groups peaks together across samples by creating a master peak list and assigning corresponding peaks from all samples. It is inspired by the alignment algorithm of mzMine.

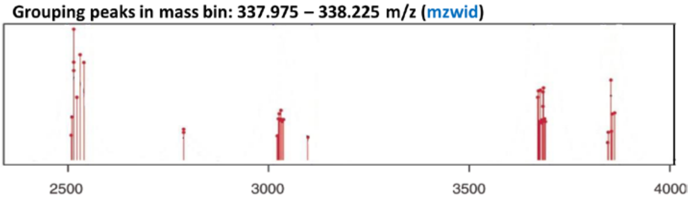
Bandwidth [bw] (related to RT)

The standard deviation of the gaussian metapeak that groups together peaks.

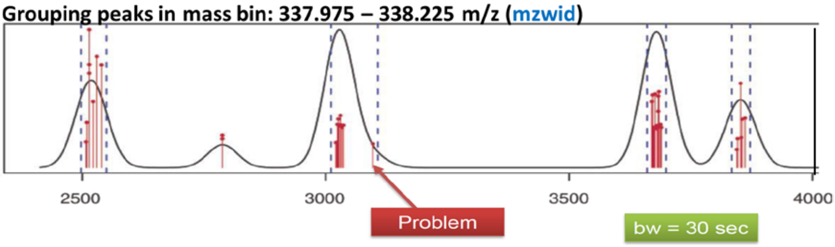
Width of overlapping m/z slices [mzwid] (related to m/z)

Size of m/z slices (bins). Range of m/z to be included in a group. Depends on mass spectrometer accuracy.

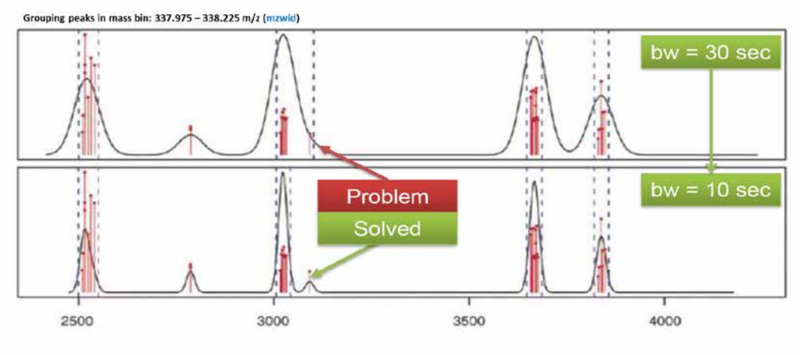
A binning of mass domain is performed. **mzwid** defined the size of the bin. Then for each m/z bin, all ions of all samples are taken into account for all retention times. Kernel density estimator method is used to detect a region of retention time with a high density of ions.



- A Gaussian model groups together peaks with similar RT.
- The inclusiveness of ions in a group is defined by the standard deviation of the Gaussian model (**bandwidth**) corresponding to the **bw** parameter (it could be interpreted as RT window).
- Vertical dash lines indicate that the feature is valid and will be retained in the data matrix.
- To be valid, the number of peaks in a group must be greater than the percentage of the total number of samples. The minfrac parameter defines this threshold.



- Decreasing bw allows separating these two groups.



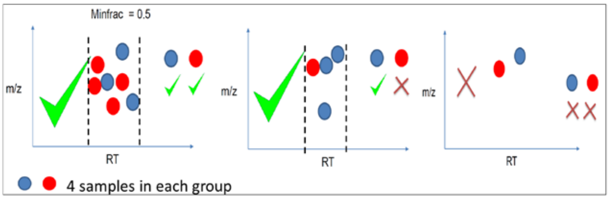
- The resulting m/z and RT of the feature corresponding to the median of m/z and RT of all ions grouped as a single feature.

<u>Minimum fraction of sample necessary [minfrac] (related to samples)</u>

Minimum fraction of the total number of samples for a group to be considered as valid (minimum sample detected in at least one class to be considered as a group). Minfrac=0.5 corresponds to 50%. (n=10 minfrac=0.5 found in at least 10 out of 50).

Max (related to the number of ions)

A maximum number of groups detected in single m/z slices.

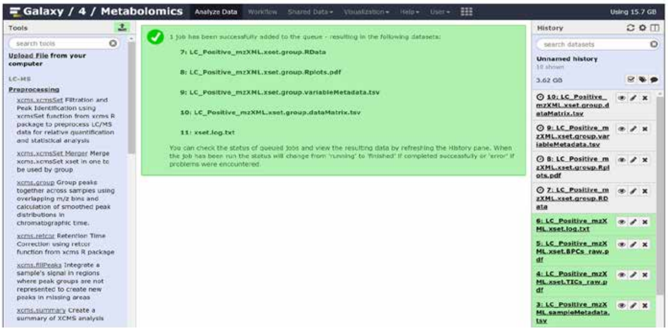


Output files (History)

xset.group.RData: rdata.xcms.group format

xset.group.Rplots.pdf

Rdata file that will be necessary for the third and fourth step of the workflow (xcms.retcor and xcms.fillpeaks).



Upload File from your computer

LC-MS

Preprocessing

xcms.xcmsSet Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

xcms.xcmsSet Merger Merge xcms.xcmsSet xset in one to be used by group

xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

xcms.retcor Retention Time Correction using retcor function from xcms R package

xcms.fillPeaks Integrete a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

xcms.summary Create a summary of XCMS analysis

1 job has been successfully added to the queue - resulting in the following datasets:

7: LC_Positive_mzXML.xset.group.RData

8: LC_Positive_mzXML.xset.group.Rplots.pdf

9: LC_Positive_mzXML.xset.group.variableMetadata.tsv

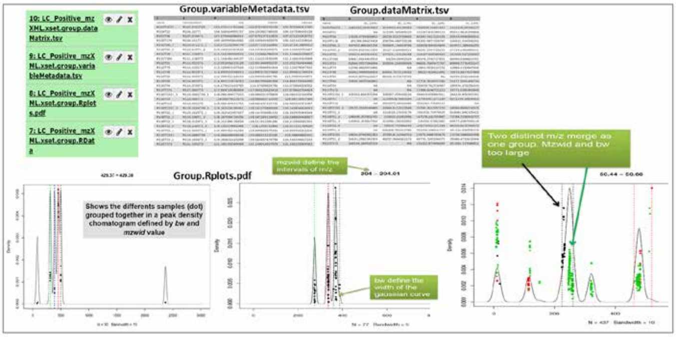
10: LC_Positive_mzXML.xset.group.dataMatrix.tsv

11: xset.log.txt

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Unnamed history
10 shown
3.62 GB

10: LC_Positive_mzXML.xset.group.dataMatrix.tsv
9: LC_Positive_mzXML.xset.group.variableMetadata.tsv
8: LC_Positive_mzXML.xset.group.Rplots.pdf
7: LC_Positive_mzXML.xset.group.RData
6: LC_Positive_mzXML.xset.log.txt
5: LC_Positive_mzXML.xset.BPCs_raw.pdf
4: LC_Positive_mzXML.xset.TICs_raw.pdf
3: LC_Positive_mzXML.sampleMetadata.tsv



10: LC_Positive_mzXML.xset.group.dataMatrix.tsv
9: LC_Positive_mzXML.xset.group.variableMetadata.tsv
8: LC_Positive_mzXML.xset.group.Rplots.pdf
7: LC_Positive_mzXML.xset.group.RData

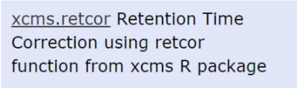
Shows the differents samples (dot) grouped together in a peak density chomatogram defined by *bw* and *mzwid* value

mzwid define the intervals of m/z

bw define the width of the gaussian curve

Two distinct m/z merge as one group. Mzwid and bw too large

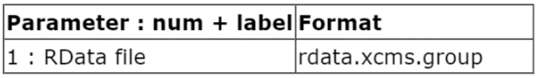
xcms.retcor (Alignment)



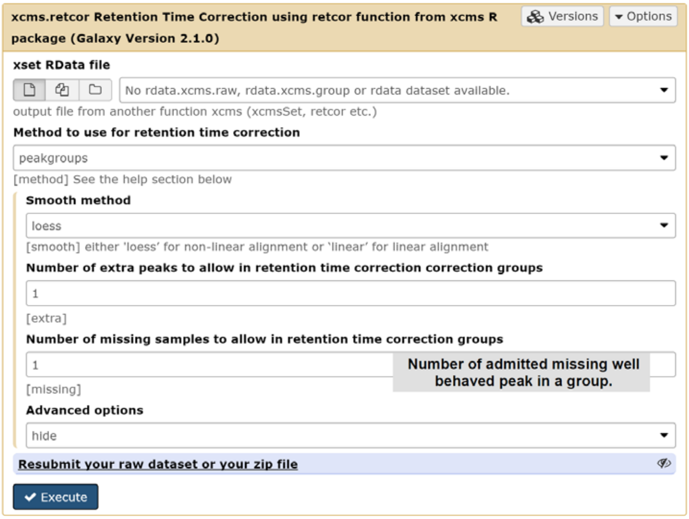
xcms.retcor Retention Time Correction using retcor function from xcms R package

After matching peaks into groups, xcms can use those groups to identify and correct correlated drifts in retention time from run to run. The aligned peaks can then be used for a second pass of peak grouping which will be more accurate than the first one. The whole process can be repeated iteratively. Not all peak groups will be helpful for identifying retention time drifts. Some groups may be missing peaks from a large fraction of samples and thus provide an incomplete picture of the drift at that time point. Still, others may contain multiple peaks from the same sample, which is a sign of improper grouping.

Input files



Parameter : num + label	Format
1 : RData file	rdata.xcms.group



xcms.retcor Retention Time Correction using retcor function from xcms R package (Galaxy Version 2.1.0)

xset RData file

No rdata.xcms.raw, rdata.xcms.group or rdata dataset available.

output file from another function xcms (xcmsSet, retcor etc.)

Method to use for retention time correction

peakgroups

[method] See the help section below

Smooth method

loess

[smooth] either 'loess' for non-linear alignment or 'linear' for linear alignment

Number of extra peaks to allow in retention time correction correction groups

1

[extra]

Number of missing samples to allow in retention time correction groups

1

Number of admitted missing well behaved peak in a group.

[missing]

Advanced options

hide

Resubmit your raw dataset or your zip file

Parameters:

<u>Method</u>

peakgroups

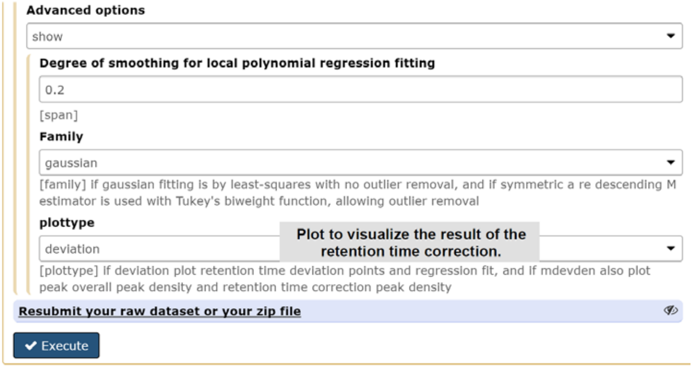
xcms ignores those groups by only considering well-behaved peak groups which are missing at most one sample and have at most one extra peak. (Those values can be changed with the **missing** and **extra** arguments.)

For each of those well-behaved groups, the algorithm calculates a median RT and, for every sample, a deviation from that median. Within a sample, the observed deviation changes over time in a nonlinear fashion. Those changes are approximated using a local polynomial regression technique implemented in the **loess** function. By default, the curve fitting is done using least-squares on all data points. However, it is possible to enable outlier detection and removal by setting the **family** argument to **symmetric.**

obiwarp

Calculate RT deviations for each sample using the obiwarp code at "http://obi-warp.sourceforge.net/". This function can align multiple samples with a center-star strategy.

Ordered Bijective Interpolated Warping (OBI-Warp) aligns matrices along a single axis using Dynamic Time Warping (DTW) and a one-to-one (bijective) interpolated warp function. OBI-Warp harnesses the non-linear, comprehensive alignment power of DTW and builds on the discrete, non-bijective output of DTW to give natural interpolants that can be used across multiple datasets.



Advanced options

show

Degree of smoothing for local polynomial regression fitting

0.2

[span]

Family

gaussian

[family] if gaussian fitting is by least-squares with no outlier removal, and if symmetric a re descending M estimator is used with Tukey's biweight function, allowing outlier removal

plottype

deviation

Plot to visualize the result of the retention time correction.

[plottype] if deviation plot retention time deviation points and regression fit, and if mdevden also plot peak overall peak density and retention time correction peak density

Resubmit your raw dataset or your zip file

Output files (History)

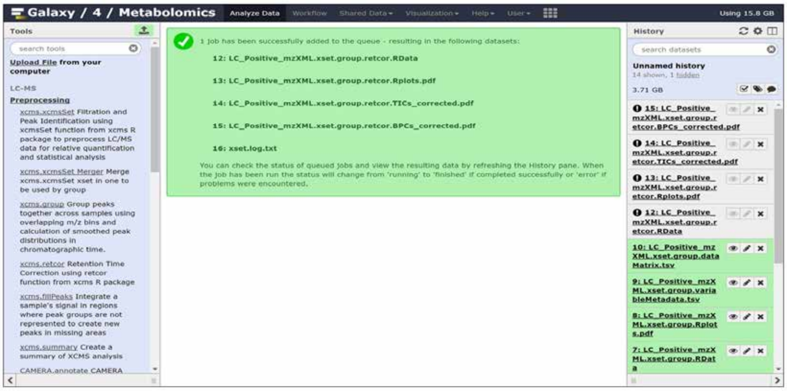
xset.group.retcor.TICs_corrected.pdf TIC graph in pdf format, corrected after a retcor step.

xset.group.retcor.BPCs_corrected.pdf BPC graph in pdf format, corrected after a retcor step.

xset.group.retcor.RData: rdata.xcms.retcor Rdata file that will be necessary for the **xcms.group** step of the workflow.

The output file is an **xset.retcor.RData** file. You can continue your analysis using it in **xcms.fillPeaks** tool.

After a retcor step, it is mandatory to do a group step. Otherwise, the rest of the workflow will not work with the RData file (the initial peak grouping becomes invalid and is discarded).



Upload File from your computer

LC-MS

Preprocessing

xcms.xcmsSet Filtration and Peak Identification using xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

xcms.xcmsSet Merger Merge xcms.xcmsSet xset in one to be used by group

xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

xcms.retcor Retention Time Correction using retcor function from xcms R package

xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

xcms.summary Create a summary of XCMS analysis

CAMERA.annotate CAMERA

1 job has been successfully added to the queue - resulting in the following datasets:

12: LC_Positive_mzXML.xset.group.retcor.RData

13: LC_Positive_mzXML.xset.group.retcor.Rplots.pdf

14: LC_Positive_mzXML.xset.group.retcor.TICs_corrected.pdf

15: LC_Positive_mzXML.xset.group.retcor.BPCs_corrected.pdf

16: xset.log.txt

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Unnamed history

14 shown, 1 hidden

3.71 GB

15: LC_Positive_mzXML.xset.group.retcor.BPCs_corrected.pdf

14: LC_Positive_mzXML.xset.group.retcor.TICs_corrected.pdf

13: LC_Positive_mzXML.xset.group.retcor.Rplots.pdf

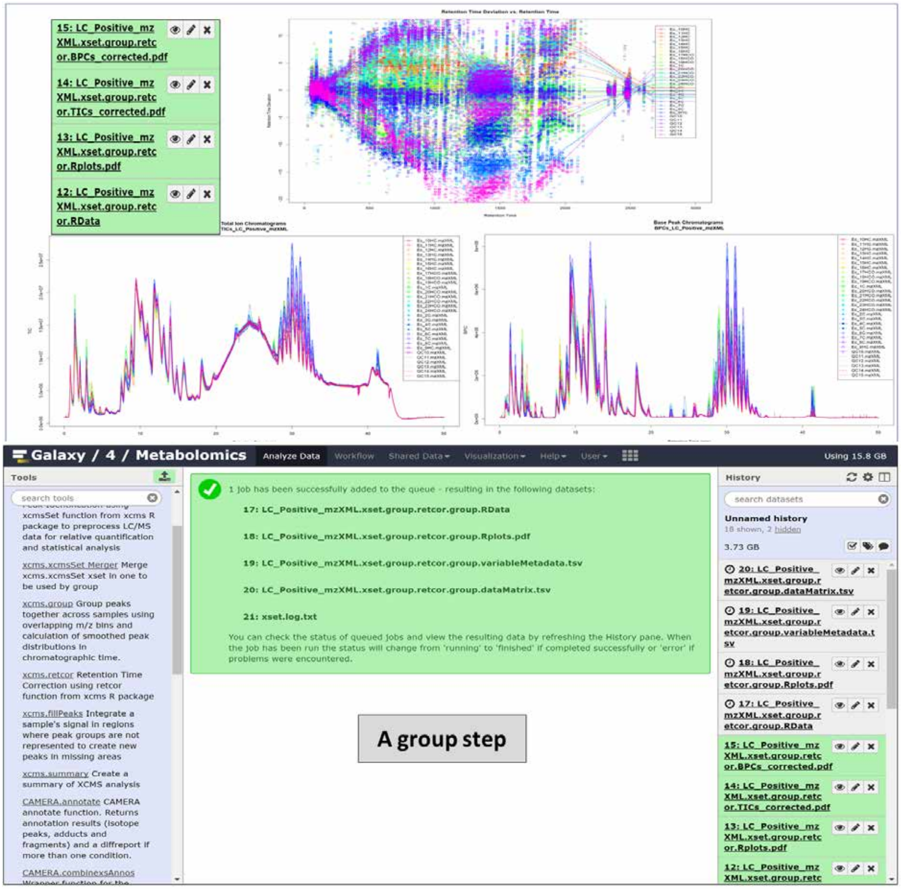
12: LC_Positive_mzXML.xset.group.retcor.RData

10: LC_Positive_mzXML.xset.group.dataMatrix.tsv

9: LC_Positive_mzXML.xset.group.variableMetadata.tsv

8: LC_Positive_mzXML.xset.group.Rplots.pdf

7: LC_Positive_mzXML.xset.group.RData


15: LC_Positive_mzXML.xset.group.retcor.BPCs_corrected.pdf
14: LC_Positive_mzXML.xset.group.retcor.TICs_corrected.pdf
13: LC_Positive_mzXML.xset.group.retcor.Rplots.pdf
12: LC_Positive_mzXML.xset.group.retcor.RData
Retention Time Deviation vs. Retention Time
Total Ion Chromatograms TICs_LC_Positive_mzXML
Base Peak Chromatograms BPCs_LC_Positive_mzXML
Galaxy / 4 / Metabolomics
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 15.8 GB
Tools
search tools
History
search datasets
A group step

xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

xcms.xcmsSet Merger Merge xcms.xcmsSet xset in one to be used by group

xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

xcms.retcor Retention Time Correction using retcor function from xcms R package

xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

xcms.summary Create a summary of XCMS analysis

CAMERA.annotate CAMERA annotate function. Returns annotation results (isotope peaks, adducts and fragments) and a diffreport if more than one condition.

CAMERA.combinexsAnnos

1 job has been successfully added to the queue - resulting in the following datasets:

17: LC_Positive_mzXML.xset.group.retcor.group.RData

18: LC_Positive_mzXML.xset.group.retcor.group.Rplots.pdf

19: LC_Positive_mzXML.xset.group.retcor.group.variableMetadata.tsv

20: LC_Positive_mzXML.xset.group.retcor.group.dataMatrix.tsv

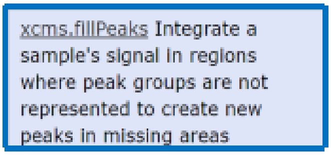
21: xset.log.txt

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Unnamed history
18 shown, 2 hidden
3.73 GB

20: LC_Positive_mzXML.xset.group.retcor.group.dataMatrix.tsv
19: LC_Positive_mzXML.xset.group.retcor.group.variableMetadata.tsv
18: LC_Positive_mzXML.xset.group.retcor.group.Rplots.pdf
17: LC_Positive_mzXML.xset.group.retcor.group.RData
15: LC_Positive_mzXML.xset.group.retcor.BPCs_corrected.pdf
14: LC_Positive_mzXML.xset.group.retcor.TICs_corrected.pdf
13: LC_Positive_mzXML.xset.group.retcor.Rplots.pdf
12: LC_Positive_mzXML.xset.group.retc

2.xcms.fillpeaks

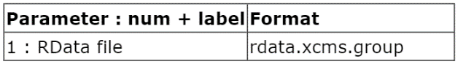


xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

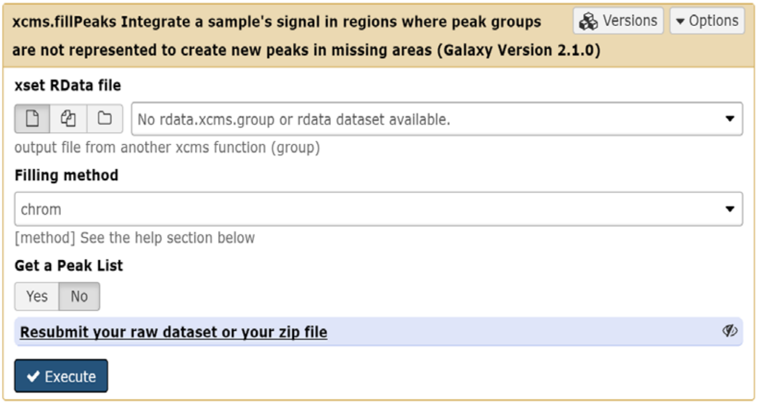
Integrate areas of missing peaks. For each sample, identify peak groups where that sample is not represented. For each of those peak groups, integrate the signal in the region of that peak group and create a new peak.

According to the type of raw-data, there are two different methods available. For filling GCMS/LCMS data the method, "chrom" integrates raw-data in the chromatographic domain, whereas "MSW" is used for peak lists without RT information like those from direct-infusion spectra.

Input files



Parameter : num + label	Format
1 : RData file	rdata.xcms.group



xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas (Galaxy Version 2.1.0)

xset RData file

No rdata.xcms.group or rdata dataset available.

output file from another xcms function (group)

Filling method

chrom

[method] See the help section below

Get a Peak List

Yes [✓ No]

Resubmit your raw dataset or your zip file

Parameters:

Filling Method

- **chrom**

This method produces intensity values for those missing samples by integrating raw data in a peak group region. In a given group, the start and ending RT points for integration are defined by the median start and end points of the other detected peaks. The start and end *m/z* values are similarly determined. Intensities can still be zero, which is a rather unusual intensity for a peak. This is the case if, e.g., the raw data were thresholded, and the integration area contains no actual raw intensities, or if one sample is miscalibrated, such raw data points are (just) outside the integration area.

Importantly, if RT correction data is available, the alignment information is used to more precisely integrate the proper region of the raw data. If the corrected RT is beyond the end of the raw data, the value will be not-a-number (NaN).

- **MSW**

"MSW" is used for peak lists without RT information like those from direct-infusion spectra.

Get a Peak List

If 'true', the module generates two additional files corresponding to the peak list: the variable metadata file (corresponding to information about extracted ions such as mass or RT), and the data matrix (corresponding to related intensities).

Decimal places for [mass or retention time] values in identifiers

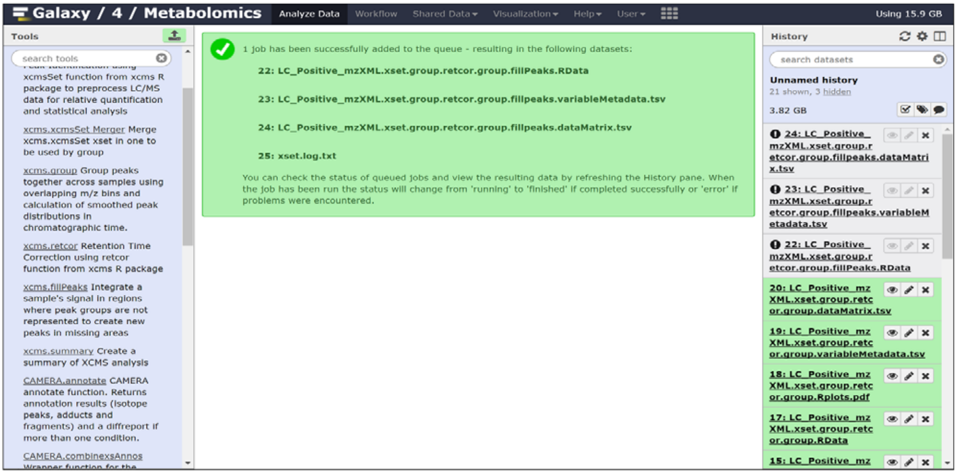
Ions' identifiers are constructed as MxxxTyyy where 'xxx' is the ion median mass and 'yyy' is the ion median RT.

Two parameters are used to adjust the number of decimal places wanted in identifiers for mass and RT respectively. These parameters do not affect decimal places in columns other than the identifier one.

Reported intensity values

This parameter determines which values should be reported as intensities in the dataMatrix table; it corresponds to xcms 'intval' parameter:

- **into:** integrated area of original (raw) peak
- **maxo:** maximum intensity of original (raw) peak
- **intb:** baseline corrected integrated peak area (only available if peak detection was done by 'findPeaks.centWave')



xcmsSet function from xcms R package to preprocess LC/MS data for relative quantification and statistical analysis

xcms.xcmsSet Merger Merge xcms.xcmsSet xset in one to be used by group

xcms.group Group peaks together across samples using overlapping m/z bins and calculation of smoothed peak distributions in chromatographic time.

xcms.retcor Retention Time Correction using retcor function from xcms R package

xcms.fillPeaks Integrate a sample's signal in regions where peak groups are not represented to create new peaks in missing areas

xcms.summary Create a summary of XCMS analysis

CAMERA.annotate CAMERA annotate function. Returns annotation results (isotope peaks, adducts and fragments) and a diffreport if more than one condition.

CAMERA.combinexsAnnos Wrapper function for the

✓ 1 job has been successfully added to the queue - resulting in the following datasets:

22: LC_Positive_mzXML.xset.group.retcor.group.fillPeaks.RData

23: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.variableMetadata.tsv

24: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.dataMatrix.tsv

25: xset.log.txt

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Unnamed history
21 shown, 3 hidden
3.82 GB

24: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.dataMatrix.tsv
23: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.variableMetadata.tsv
22: LC_Positive_mzXML.xset.group.retcor.group.fillPeaks.RData
20: LC_Positive_mzXML.xset.group.retcor.group.dataMatrix.tsv
19: LC_Positive_mzXML.xset.group.retcor.group.variableMetadata.tsv
18: LC_Positive_mzXML.xset.group.retcor.group.Rplots.pdf
17: LC_Positive_mzXML.xset.group.retcor.group.RData
15: LC_Positive_mz

Output files

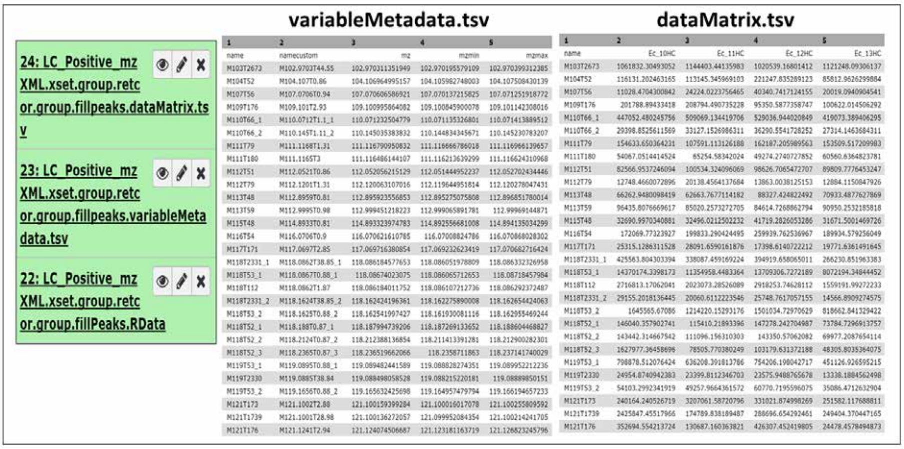
xset.fillPeaks.RData: rdata.xcms.fillpeaks format.

Rdata file that will be used in the **CAMERA.annotate** or **xcms.summary** step of the workflow.

xset.variableMetadata.tsv: tabular format. Table containing information about ions; can be used as one input of **Quality_Metrics** or **Generic_filter* modules.

xset.dataMatrix.tsv: tabular format. Table containing ions' intensities; can be used as one input of **Quality_Metrics** or **Generic_filter* modules.

The output file is a **xset.fillPeaks.RData** file. You can continue your analysis using it in **CAMERA.annotate** or **xcms.summary** tool.



24: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.dataMatrix.tsv

23: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.variableMetadata.tsv

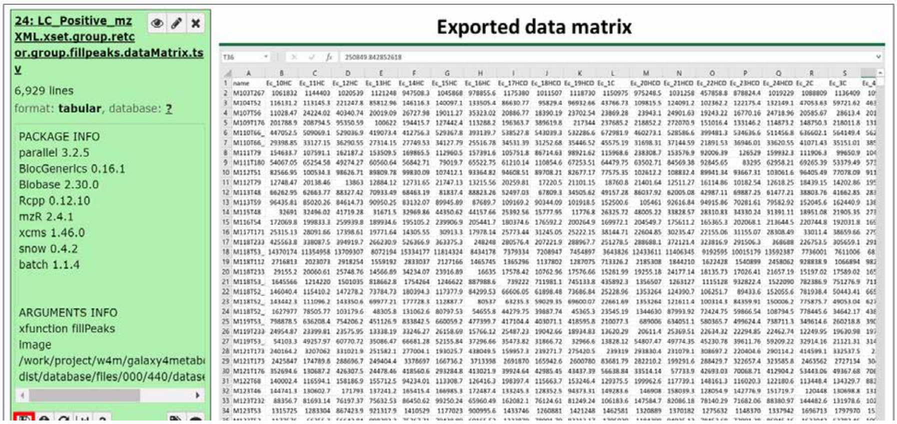
22: LC_Positive_mzXML.xset.group.retcor.group.fillPeaks.RData

variableMetadata.tsv

1	2	3	4	5
name	namecustom	mz	mzmin	mzmax
M103T2673	M102.9703T44.55	102.970311351949	102.970195579109	102.970399312385
M104T52	M104.107T0.86	104.106964995157	104.105982748003	104.107508430139
M107T56	M107.0706T0.94	107.070606586921	107.070137215825	107.071251918772
M109T176	M109.101T2.93	109.100995864082	109.100845900078	109.101142308016
M110T66_1	M110.0712T1.1_1	110.071232504779	110.071135326801	110.071413889512
M110T66_2	M110.145T1.11_2	110.145035383832	110.144834345671	110.145230783207
M111T79	M111.1168T1.31	111.116790950832	111.116666786018	111.116966139657
M111T180	M111.1165T3	111.116486144107	111.116213639299	111.116624310968
M112T51	M112.0521T0.86	112.052056215129	112.051444952237	112.052702434446
M112T79	M112.1201T1.31	112.120063107016	112.119644951814	112.120278047431
M113T48	M112.8959T0.81	112.895923556853	112.895275075808	112.896851780014
M113T59	M112.9995T0.98	112.999451218223	112.999065891781	112.99969144871
M115T48	M114.8933T0.81	114.893323974783	114.892556681008	114.894135034299
M116T54	M116.0706T0.9	116.070621610785	116.07008824786	116.070868028302
M117T171	M117.0697T2.85	117.069716380854	117.069232623419	117.070682716424
M118T2331_1	M118.0862T38.85_1	118.086184577653	118.086051978809	118.086332326958
M118T53_1	M118.0867T0.88_1	118.08674023075	118.086065712653	118.08718457984
M118T112	M118.0862T1.87	118.086184011752	118.086107212736	118.086292372487
M118T2331_2	M118.1624T38.85_2	118.162424196361	118.162275890008	118.162654424063
M118T53_2	M118.1625T0.88_2	118.162541997427	118.161930081116	118.162955469244
M118T52_1	M118.188T0.87_1	118.187994739206	118.187269133652	118.188604468827
M118T52_2	M118.2124T0.87_2	118.212388136854	118.211413391281	118.212900282301
M118T52_3	M118.2365T0.87_3	118.236519662066	118.2358711863	118.237141740029
M119T53_1	M119.0895T0.88_1	119.089482441589	119.088828274351	119.089952212236
M119T2330	M119.0885T38.84	119.088498058528	119.088215220181	119.08889850151
M119T53_2	M119.1656T0.88_2	119.165632425698	119.164957479794	119.166194657233
M121T173	M121.1002T2.88	121.100159399284	121.100016017078	121.100255809592
M121T1739	M121.1001T28.98	121.100136272057	121.099952084354	121.100214241705
M121T176	M121.1241T2.94	121.124074506687	121.123181163719	121.126823245796

dataMatrix.tsv

1	2	3	4	5
name	Ec_10HC	Ec_11HC	Ec_12HC	Ec_13HC
M103T2673	1061832.30493052	1144403.44135983	1020539.16801412	1121248.09306137
M104T52	116131.202463165	113145.345969103	221247.835289123	85812.9626299884
M107T56	11028.4704300842	24224.0223756465	40340.7417124155	20019.0940904541
M109T176	201788.89433418	208794.490735228	95350.5877358747	100622.014506292
M110T66_1	447052.480245756	509069.134419706	529036.944020849	419073.389406295
M110T66_2	29398.8525611569	33127.1526986311	36290.5541728252	27314.1463684311
M111T79	154633.650364231	107591.113126188	162187.205989563	153509.517209983
M111T180	54067.0514414524	65254.58342024	49274.2740727852	60560.6364823781
M112T51	82566.9537246094	100534.324096069	98626.7065472707	89809.7776453247
M112T79	12748.4660072896	20138.4564137684	13863.0038125153	12884.1150847926
M113T48	66262.9480098419	62663.7677114182	88327.424822492	70933.4877627869
M113T59	96435.8076669617	85020.2573272705	84614.7268862794	90950.2532185818
M115T48	32690.9970340881	32496.0212502232	41719.2826053286	31671.5001469726
M116T54	172069.77323927	199833.290424495	259939.762536967	189934.579256049
M117T171	25315.1286311528	28091.6590161876	17398.6140722212	19771.6361491645
M118T2331_1	425563.804303394	338087.459169224	394919.658065011	266230.851963383
M118T53_1	14370174.3398173	11354958.4483364	13709306.7272189	8072194.34844452
M118T112	2716813.17062041	2023073.28526089	2918253.74628112	1559191.99272233
M118T2331_2	29155.2018136445	20060.6112223546	25748.7617057155	14566.8909274575
M118T53_2	1645565.67086	1214220.15293176	1501034.72970629	818662.841329422
M118T52_1	146040.357902741	115410.21893396	147278.242704987	73784.7296913757
M118T52_2	143442.314667542	111096.156310303	143350.57062082	69977.2087654114
M118T52_3	1627977.36458696	78505.770380249	103179.631372188	48305.8035364075
M119T53_1	798878.512076424	636208.391813786	754206.198042717	451126.926595215
M119T2330	24954.8740942383	23399.8112346703	23575.9488765678	13338.1884562498
M119T53_2	54103.2992341919	49257.9664361572	60770.7195596075	35086.4712632904
M121T173	240164.240526719	3207061.58720796	331021.874998269	251582.117688811
M121T1739	2425847.45517966	174789.838189487	288696.654292461	249404.370447165
M121T176	352694.554213724	130687.160363821	426307.452419805	24478.4578494873



24: LC_Positive_mzXML.xset.group.retcor.group.fillpeaks.dataMatrix.tsv

6,929 lines

format: **tabular**, database: ?

PACKAGE INFO
parallel 3.2.5
BiocGenerics 0.16.1
Biobase 2.30.0
Rcpp 0.12.10
mzR 2.4.1
xcms 1.46.0
snow 0.4.2
batch 1.1.4

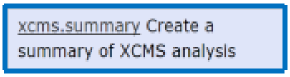
ARGUMENTS INFO
xfunction fillPeaks
image
/work/project/w4m/galaxy4metab
dist/database/files/000/440/datase

Exported data matrix

T36 | 250849.842852618

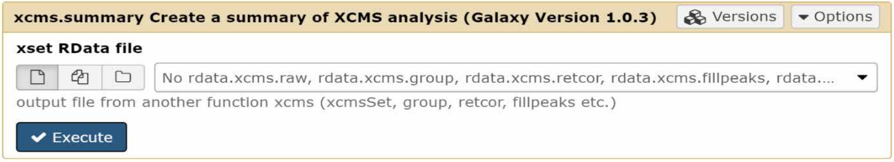
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	name	Ec_10HC	Ec_11HC	Ec_12HC	Ec_13HC	Ec_14HC	Ec_15HC	Ec_16HC	Ec_17HCO	Ec_18HCO	Ec_19HCO	Ec_1C	Ec_20HCO	Ec_21HCO	Ec_22HCO	Ec_23HCO	Ec_24HCO	Ec_2C	Ec_3C	Ec_4
2	M103T267	1061832	1144403	1020539	1121248	947508.3	1045868	978855.6	1175380	1011507	1118730	1150975	975248.5	1031258	457858.8	878824.4	1019229	1088809	1136409	10
3	M104T52	116131.2	113145.3	221247.8	85812.96	146116.3	140097.1	133505.4	86630.77	95829.4	96932.66	43766.73	109815.5	124091.2	102362.2	122175.4	132149.1	47053.63	59721.62	463
4	M107T56	11028.47	24224.02	40340.74	20019.09	26727.98	19011.27	35323.02	20886.77	18390.19	23702.54	23869.28	23943.1	24901.63	19243.22	16770.16	24718.96	20585.67	28613.4	201
5	M109T176	201788.9	208794.5	95350.59	100622	194415.7	127442.4	113288.2	196363.7	389619.8	217344	237685.2	218852.2	272070.9	151016.4	133146.2	114873.2	148750.3	218011.8	131
6	M110T66_	447052.5	509069.1	529036.9	419073.4	412756.3	529367.8	393139.7	538527.8	543039.3	532286.6	672981.9	460273.1	528586.6	399481.3	534636.6	511456.8	636602.1	564149.4	562
7	M110T66_	29398.85	33127.15	36290.55	27314.15	27749.53	34127.79	25516.78	34531.39	31252.68	35446.52	45575.19	31698.31	37144.59	21891.53	36946.01	33620.55	41071.43	35151.01	385
8	M111T79	154633.7	107591.1	162187.2	153509.5	169865.5	112960.5	157391.6	105751.8	86714.63	98921.62	115968.6	238308.7	153576.9	92006.39	126529	159932.3	111906.3	99650.9	104
9	M111T180	54067.05	65254.58	49274.27	60560.64	56842.71	79019.7	65522.75	61210.14	110854.6	67253.51	64479.75	63502.71	84569.38	92845.65	83295	62958.21	69265.39	53379.49	573
10	M112T51	82566.95	100534.3	98626.71	89809.78	99830.09	107412.1	93364.82	94608.51	89708.21	82677.17	77575.35	102612.2	108832.4	89941.34	93667.31	103061.6	96405.49	77078.09	911
11	M112T79	12748.47	20138.46	13863	12884.12	12731.65	21747.13	13215.56	20259.81	17220.5	21101.15	18760.8	21401.64	12511.27	16114.86	10182.54	12618.25	18439.15	14202.86	195
12	M113T48	66262.95	62663.77	88327.42	70933.49	68463.19	81837.4	88823.26	52497.03	67809.3	34505.62	49157.28	86037.92	62005.08	42987.11	69887.25	61477.21	38803.76	41662.85	283
13	M113T59	96435.81	85020.26	84614.73	90950.25	83132.07	89945.89	87689.7	109169.2	90344.09	101918.5	152500.6	105461	92616.84	94915.86	70281.61	79582.92	152045.6	162440.9	138
14	M115T48	32691	32496.02	41719.28	31671.5	32969.86	44350.62	44157.66	25392.56	15777.95	11776.8	26325.72	48005.22	33828.57	28310.83	34330.24	31391.11	18951.08	21905.35	277
15	M116T54	172069.8	199833.3	259939.8	189934.6	195105.2	239906.9	205441.7	180374.6	176592.2	200264.9	169972.1	204549.7	175611.2	165365.3	202068.1	213644.5	220744.8	192031.8	165
16	M117T171	25315.13	28091.66	17398.61	19771.64	14305.55	30913.3	17978.14	25773.44	31245.05	25222.15	38144.71	22604.85	30235.47	22155.06	31155.07	28308.49	33011.4	38659.66	275
17	M118T233	425563.8	338087.5	394919.7	266230.9	526366.9	363375.3	248248	280576.4	207221.9	288967.7	251278.5	288688.1	372121.4	323816.9	291506.3	368688	226753.5	305659.1	291
18	M118T53_	14370174	11354958	13709307	8072194	15334177	11814324	8434178	7379334	7208947	7454897	3643826	12433611	11406345	9192595	10015179	13592387	7736001	7611006	68
19	M118T112	2716813	2023073	2918254	1559192	2833037	2127166	1465745	1365296	1137802	1287075	713326.2	2185308	1844210	1622428	1540899	2458062	928838.9	1066894	982
20	M118T233	29155.2	20060.61	25748.76	14566.89	34234.07	23916.89	16635	17578.42	10762.96	17576.66	15281.99	19255.18	24177.14	18135.73	17026.41	21657.19	15197.02	17589.02	165
21	M118T53_	1645566	1214220	1501035	818662.8	1754264	1246622	887988.6	739222	711981.1	745133.8	435892.3	1356507	1263127	1115128	932822.4	1522090	782386.9	751276.9	711
22	M118T52_	146040.4	115410.2	147278.2	73784.73	180394.3	117377.9	84299.53	66606.05	61898.48	73686.84	25228.96	1353264	124390.7	106251.7	89433.6	152055.6	781938.4	50443.41	665
23	M118T52_	143442.3	111096.2	143350.6	69977.21	177728.3	112887.7	80537	63235.3	59029.35	69600.07	22661.69	1353264	121611.4	100314.3	84359.91	150006.2	775875.7	49053.04	627
24	M118T52_	1627977	78505.77	103179.6	48305.8	131062.6	80797.53	54655.8	44279.75	39887.74	45365.3	23545.19	1344630	87993.92	72424.75	59866.54	108794.5	778445.6	34642.17	438
25	M119T53_	798878.5	636208.4	754206.2	451126.9	833842.5	660059.2	477399.7	417104.4	403071.1	418595.8	210077.3	689006	634051.1	580365.7	499624.4	738711.3	349614.6	260218.8	390
26	M119T233	24954.87	23399.81	23575.95	13338.19	33246.27	26158.69	15766.12	25487.23	19042.66	18934.83	12620.29	20611.4	25369.51	22634.32	22294.85	22462.74	12249.95	19630.98	197
27	M119T53_	54103.3	49257.97	60770.72	35086.47	66681.28	52155.84	37296.66	35473.82	31866.72	32966.6	13828.12	54807.47	49774.35	45230.78	39611.76	59209.22	32914.16	21121.31	314
28	M121T173	240164.2	3207062	331021.9	251582.1	277004.1	193025.7	438049.5	159957.3	239271.7	275420.5	239319	293830.4	231079.1	308697.2	220404.6	290114.2	414599.1	332537.5	2
29	M121T173	2425847	174789.8	288696.7	249404.4	3378697	166736.2	3713398	2691870	165942.6	2600780	83681.79	282210.2	199291.6	288429.7	322657.4	323585.8	2463562	2727134	30
30	M121T176	352694.6	130687.2	426307.5	24478.46	418560.6	293284.8	413021.9	39924.64	42985.45	43437.39	56638.84	33514.14	57733.9	42693.03	70068.71	412904.2	53443.06	49367.68	70
31	M122T68	140002.4	116594.1	158186.9	155712.5	94234.01	113308.7	126416.3	198397.4	115663.7	153246.4	129375.5	199962.6	117739.1	148161.3	116020.3	122180.6	113448.4	134329.7	88
32	M123T46	144741.3	180602.7	171793	137241.2	165415.4	166985.3	172487.4	133245.3	128352.5	94373.31	149283.6	146908	158039.3	128054.9	142776.9	151719.7	120448	130698.8	13
33	M123T232	88356.7	81693.14	76197.37	75632.53	86450.62	99250.24	65960.49	162082.1	76124.61	81249.24	106183.6	147584.7	82086.18	78140.29	71682.06	88380.97	144482.6	131978.6	10
34	M123T53	1315725	1283304	867423.9	921317.9	1410529	1177023	900995.6	1433746	1260881	1421248	1462581	1320889	1370182	1275632	1148370	1337942	1696713	1797970	15

3. xcms.summary



xcms.summary Create a summary of XCMS analysis

This tool provides an HTML summary which summarizes your analysis using the [W4M] XCMS and CAMERA tools.



xcms.summary Create a summary of XCMS analysis (Galaxy Version 1.0.3)

xset RData file

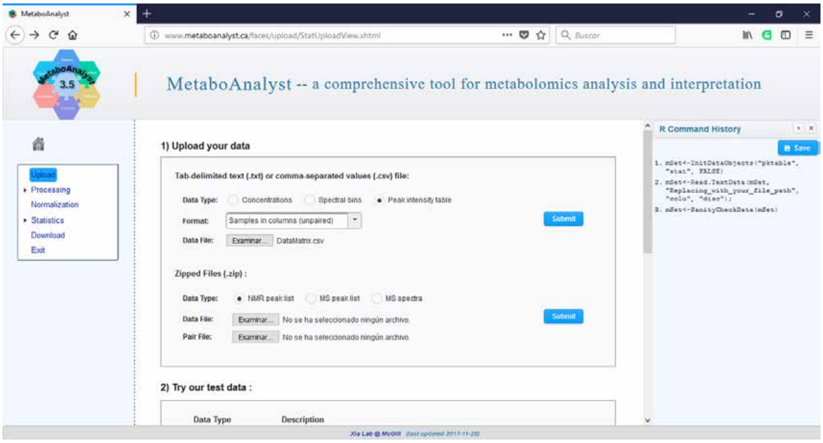
No rdata.xcms.raw, rdata.xcms.group, rdata.xcms.retcor, rdata.xcms.fillpeaks, rdata....

output file from another function xcms (xcmsSet, group, retcor, fillpeaks etc.)

5.3. Statistics

Univariate and Multivariate Statistical Analysis: PCA, PLS-DA will be performed with Metaboanalyst (https://metaboanalyst.ca). The purpose of MetaboAnalyst is to provide a free, user-friendly, and easily accessible tool for analyzing data arising from high-throughput metabolomics data. It is designed to address two common types of problems: 1) to identify features that are significantly different between two conditions (biomarker discovery); 2) to use the metabolomic data to predict the conditions under study (classification). Also, MetaboAnalyst also provides tools for compound identification and pathway mapping for annotating significant features.

Go to the main page and click on >>click here to start<<



MetaboAnalyst -- a comprehensive tool for metabolomics analysis and interpretation

1) Upload your data

Tab delimited text (.txt) or comma separated values (.csv) file:

Data Type: ○ Concentrations ○ Spectral bins ● Peak intensity table

Format: Samples in columns (unpaired)

Data File: DataMatrix.csv

Zipped Files (.zip) :

Data Type: ● NMR peak list ○ MS peak list ○ MS spectra

Data File: No se ha seleccionado ningún archivo.

Pair File: No se ha seleccionado ningún archivo.

2) Try our test data :

Data Type Description

```
1. mSet<-InitDataObjects("pktable",
   "stat", FALSE)
2. mSet<-Read.TextData(mSet,
   "Replacing_with_your_file_path",
   "colu", "disc");
3. mSet<-SanityCheckData(mSet)
```

Select the adequate Data type and Format, and Submit data. As there are no missing values in the data provided, the next step can be skipped.

There are multiple options in each step for statistical analysis. Students are encouraged to try alternatives to those that are going to be proposed within this course.

As it is plasma, no external normalization is required.

We will work after log transformation, and Pareto scaling.

Students will try to find compounds significant after One Way ANOVA & post hoc analysis and after PLS-DA.

Options for clustering will be evaluated.

The results will be combined in an MS Excel spreadsheet, and a list of significant m/z will be obtained.

5.4. Annotation

Identification using CEU Mass Mediator, http://ceumass.eps.uspceu.es/.
Pathway analysis: integrating enrichment analysis and pathway topology analysis.
Visualization of the obtained results for the model organism, with Metaboanalyst.

To know more:

E.G. Armitage, F. J. Rupérez, C. Barbas. Metabolomics of diet-related diseases using mass spectrometry. *Trends Anal. Chem.*, 2013, 52:61-73.

A. Mastrangelo, A. Ferrarini, F. Rey-Stolle, A. García, C. Barbas. From sample treatment to biomarker discovery: A tutorial for untargeted metabolomics based on GC-(EI)-Q-MS. *Anal. Chim. Acta*, 2015, 900:21-35.

A. García, S. Naz, C. Barbas. Metabolite fingerprinting by capillary electrophoresis-mass spectrometry. *Methods Molecular Biology*, 2014, 1198:107-223.

S. Naz, A. García, M. Rusak, C.Barbas. Method development and validation for rat serum fingerprinting with CE-MS: application to ventilator-induced-lung-injury study. *Analytical Bioanalytical Chemistry*, 2013, 405(14):4849-4858.

W4M Workflow4Metabolomics. http://workflow4metabolomics.org/

Module 3
Lipidomics

Elisabete Maciel[1], Eliana Alves[2], Pedro Domingues[2], Rosário Domingues[1,2]

Mass Spectrometry Center, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal

[1] CESAM (Centro de Estudos do Ambiente e do Mar)

[2] QOPNA (Química Orgânica, Produtos Naturais e Agroalimentares)

I. Rationale

Lipidomics can be defined as a large-scale analysis of lipid composition in biological samples. In the last decade, there has been an increasing interest in lipidomics, due to the development of analytical techniques, especially mass spectrometry. Also, the recognition of the crucial role of lipids in the maintenance of cell hemostasis and the development of a large number of pathologies has contributed to these developments. Lipidomic analysis can provide novel insights into biochemical processes to pinpoint new lipid biomarkers of disease and give evidence of modified metabolic pathways that can be useful for innovative therapeutic strategies.

The first part of this module will focus on the main analytical strategies to identify the lipid profile of biological samples. It will include the study of the main technologies used so far, namely lipid extraction protocols, lipid separation/fractionation methods, mass spectrometry approaches to lipid analysis, and bioinformatics platforms for data analysis. The second part of this module will focus on a case study and is designed to familiarize students with lipidomics analysis. The course is based on expository and interactive lectures with students, demonstration/practical sessions, and analysis of practical cases that will allow consolidating concepts and achieving specific learning outcomes.

II. Course Aims and Outcomes

Aims

This course intends to provide students with an advanced and integrated view of the analytical strategies that can be used in the study of the lipidome. Students will acquire knowledge about theory and practice related to the importance of the lipidomics in the context of bioanalytical sciences. It will cover the basics and advanced concepts of the structural features and function of lipids. Practical concepts related to lipidomics methodologies, including the different steps of extraction, fractionation and structural analysis of lipids, focused on mass spectrometry-based approaches will be underlined. It is also intended to contextualize the potentialities of the analysis of the deviations in the lipidome, associated with several pathologies such as cardiovascular pathologies, neurodegenerative diseases, cancer, diabetes, and inflammation, for the development of new biomarkers or therapeutic strategies.

Learning outcomes:

By the end of this course, students should be able to:

1. Define and apply common lipidomics terminology;
2. Recognize the principles of the most common lipidomics techniques;
3. Understand the different mass spectrometry-based lipidomics workflows;
4. Analyze lipidomics data;
5. Communicate and justify conclusions, clearly and unambiguously, to both specialist and non-specialist audiences;
6. Continue the learning process, mostly autonomously.

III. Course contents

Module 3 – Lipidomics

1. Introduction to lipidomics
   - 1.1. Analytical approaches in lipidomics
     - 1.1.1. Sample preparation: lipid extraction and fractionation
     - 1.1.2. MS-based lipid identification: concepts
       - 1.1.2.1. LC-MS Untargeted lipidomics approaches
       - 1.1.2.2. Shotgun Targeted lipidomics approaches
   - 1.2. Quantification
   - 1.3. Data procession and lipid identification/quantification
   - 1.4. Practical sessions
     - 1.4.1. Identification of phospholipidome profile of THP1- monocytes using a lipidomics approach based on HILIC-LC-MS.
   - 1.5. Additional resource readings

1. Introduction to lipidomics

Lipidomics has been defined as “the full characterization of lipid molecular species and of their biological roles with respect to expression of proteins involved in lipid metabolism and function, including gene regulation (in http://lipidlibrary.aocs.org/Analysis/content.cfm?ItemNumber=39284, April 2018).

Lipids are a group of biomolecules that includes a huge diversity of molecular species distributed among different lipid classes, present in all living beings, and with distinct functions. This variety is even greater when considering the structural modifications resulting from lipid peroxidation and glycation. Lipids play different roles, for example, as cell membrane structural units, where they are essential as modulators of protein activity and as cell recognition and signaling molecules. Lipids participate in the regulation of important cellular processes, including cell proliferation, apoptosis, metabolism, and migration. The complete characterization of a lipidome represents a major challenge, not only because of the structural diversity but also due to a large number of molecular lipid species existing at very low concentrations.

Lipid biosynthesis and metabolism are governed by a set of enzymes that allow maintaining of the structural diversity existing in the lipid classes. Lipid biosynthesis occurs predominantly in the endoplasmic reticulum and the mitochondria, and after that, lipids can be translocated to the other membranes and organelles. An imbalance in cellular lipid profile affects the lipid signaling network and can be associated with the onset and development of various diseases in humans. Thus, lipidomics, as an emerging tool with a great potential to decode deviation in lipidome, helps to understand physio-pathological processes in disease, to define new diagnostic biomarkers or to define new therapeutic strategies.

The identification of the lipid profile of a biological sample requires the use of a set of sample treatment procedures and analytical techniques that should be planned according to the objectives of the investigation. In general, a lipidomics workflow starts with the extraction of the total lipids, followed by fractionation, analysis by mass spectrometry approaches and data processing (Figure 3.1).

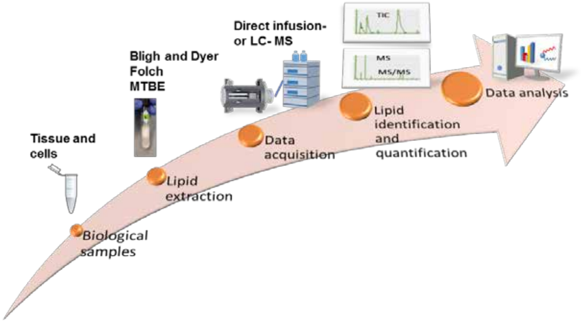


Figure 3.1. Basic lipidomics workflow.

The first step of the lipidomics workflow is the extraction of the total lipids from the biological samples. Different types of samples can be extracted, namely cells, tissues or biofluids. Since a large variety of lipids compose the lipid extract, it is advantageous, in some specific cases, to perform a purification step, for example, by separating the total extract into lipid classes. This fractionation can be done using separation methods such as thin-layer chromatography (TLC), and high-performance liquid chromatography (HPLC) or solid-phase extraction (SPE). Identification of the lipid molecular species in each class can then be performed using mass spectrometry. Identification of the fatty acids esterified to lipids can be performed by gas chromatography coupled with mass spectrometry after hydrolysis and derivatization procedures. Because of the development of mass spectrometry instruments and methods, the identification and characterization of lipid molecular species have undergone tremendous improvements in sensitivity and speed of analysis. Since the 1990s, the development of the electrospray ionization method (ESI) has allowed for coupling of mass spectrometers with HPLC-MS, an essential advance for the study of complex samples. The commercialization of high-resolution mass spectrometers, such as Q-TOF and Q-Orbitrap, allowed data acquisition with high sensitivity. LC-MS based approaches allow to quickly identify and quantify thousands of lipid species, essential for the lipid analysis of biological samples.

Depending on the primary goal of the study, two different general strategies can be considered: untargeted or targeted analysis. Untargeted lipidomics aims at broad coverage, i.e. identification and quantification of as many lipid species as possible. Also, it aims to determine a specific lipid signature of a tissue, cell, organelle or biofluid. It can also be used to prospect disease biomarkers, by comparing lipidomics data from healthy to disease conditions. It usually makes use of LC-MS platforms to analyze total lipid extracts in a single run and generate a high amount of data that requires bioinformatics analysis and further validation. The information gathered from untargeted approaches can be further used to design targeted lipidomics strategies. Targeted lipidomics analysis usually aims at the detection and quantification of a panel of specific lipids. It can be used to identify specific changes in a lipid class, and for that, it uses targeted mass spectrometric approaches such as precursor ion scan (PIS) or neutral loss scan (NLS). Quantification of specific lipid molecular species is usually performed using multiple resonance monitoring (MRM) in QqQ mass spectrometers.

1.1. Analytical approaches in lipidomics

1.1.1. Sample preparation: lipid extraction and fractionation

The lipidomics workflow begins by obtaining a total lipid extract from the biological matrix, such as biofluids, cells, tissues or other matrices, using organic solvents (Figure 3.2). Several common methods can be used for the lipid extraction, by using organic solvents with distinct polarities, which can be adapted depending on the type of samples, such as tissues, cell extracts or biological fluids. The most popular methods of lipid extraction are the Folch method (chloroform/methanol

(2:1, by volume)) and the Bligh and Dyer method (chloroform: methanol (1:2, by volume)). The former is more often used to extract biofluids (plasma, serum, and urine) while the latter is mostly used to extract lipids from tissues and cell pellets. In these methods, lipids are recovered from the lower phase, in the chloroform fraction. Other methods have also been developed, to avoid the use of chlorinated solvents, such as chloroform. For example, the MTBE method uses methyl-*tert*-butyl ether (MTBE)/methanol/water (10:3:2.5, by volume). This method also has the advantage that the organic phase is the upper phase, which makes it easier to recover the lipid extract. Also, this method allows for faster and cleaner lipid recoveries and can be suited for automated shotgun profiling, by using MTBE/methanol (1:1). To analyse lipid extracts rich in neutral lipids such as triglycerides (TG), more non-polar solvents, such hexane, should be chosen.

SPE is another approach that can also be used to separate phospholipids (PL) from plasma. In this procedure, plasma is treated with acidified acetonitrile (ACN) to precipitate proteins, and the upper phase is then separated using an ionic phase SPE cartridge. PL are then eluted with acetonitrile/ammonium hydroxide (5%). Aminopropyl cartridges are used to separate cholesterol, TG and PL fractions, while silica-SPE cartridges are recommended to separate neutral lipids (free fatty acids, cholesterol, and TG) and PL fractions.

Lipid classes can also be obtained using TLC. One-dimension or two-dimension TLC have been used in several lipidomics studies. The solvent system can be chosen depending on the lipid classes to be separated. Usually, PL classes are fractionated by one-dimension TLC using a solvent system composed of chloroform/ethanol/water/triethylamine (35:30:7:35, by volume). The spots are detected after spraying the TLC plate with a solution of primuline in acetone and visualization under a UV lamp. Each PL class is identified by comparison of the retention factors (Rf) with PL standards applied to the same plate. The phosphorus content in each spot can be quantified, which allows for determining the relative content of each PL class in the total lipid extract. Furthermore, spots can be scraped off from the plate and lipids from each spot can be recovered by using appropriate organic solvents and further analyzed. Two-dimensional TLC (2-D TLC) can also be used to separate lipids. The combination of solvent systems for 2D-TLC is chosen based on the lipid classes to be isolated. Although 2-D TLC highly improves the separation performance, it has some disadvantages when compared with the one-dimensional TLC, since only a single sample can be applied on the plate, and thus, the simultaneous application of samples and standards is not possible. Moreover, 2D-TLC is more time-consuming.

The total lipid extracts or fractions obtained from SPE or TLC are then analyzed by targeted or untargeted mass spectrometry-based approaches. There are three most popular approaches using mass spectrometry for the analysis of lipid extract/fractions: direct infusion, using shotgun strategies, or LC-MS platforms. Despite different possibilities, nowadays, the most common approach is the analysis of total lipid extracts by LC-ESI-MS. Both platforms require skills in mass spectrometry data analysis and interpretation to pinpoint each lipid class.

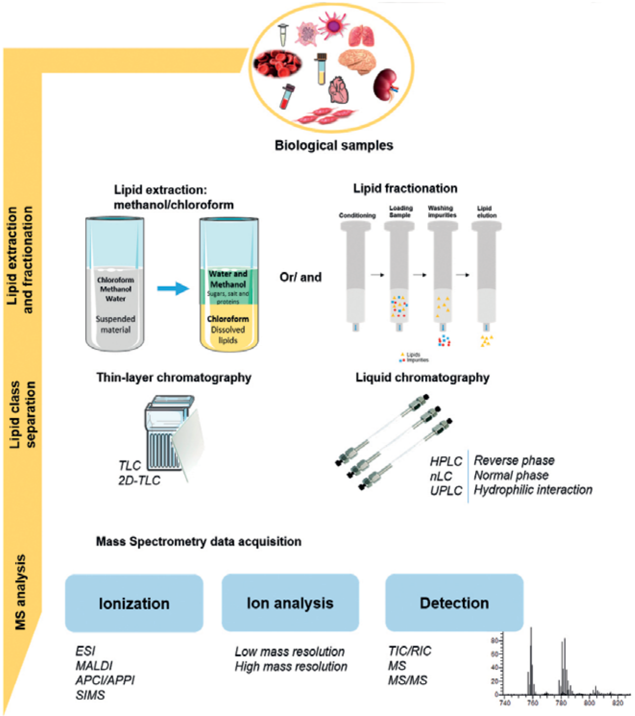


Figure 3.2. Sample preparation in lipidomics, from lipid extraction, through fractionation to analysis.

1.1.2. MS-based lipid identification: concepts

Lipid analysis by MS requires the information obtained by the interpretation of MS and the MS/MS spectra. Phospholipids (PL) are the principal component of the total lipid extracts from cells and tissues. Considering this, the mass spectra are acquired in positive or in negative ionization modes, depending on the structural features of the PL molecular species. Phospholipids such as: phosphatidylcholine (PC), phosphatidylethanolamine (PE) and sphingomyelin (SM) are usually identified in MS spectra acquired in positive ion mode. In this mode, these PL display the corresponding $[M+H]^+$ ions but, in some cases, $[M+Na]^+$ ions can be observed with low abundance. Phosphatidylinositol (PI), phosphatidylserine (PS), phosphatidylglycerol (PG) and cardiolipin (CL) are preferentially identified in MS spectra in negative ion mode and assigned as $[M\text{-}H]^-$ ions. PE can also be seen in the MS data obtained in the negative mode as $[M\text{-}H]^-$. In the case of CL, doubly charged ions ($[M\text{-}2H]^{2-}$) are also observed. In the case of LC-MS approaches, which use eluents with ammonium acetate or ammonium formate, PC and SM can be seen in MS data in the negative mode as adducts with acetate anions or format anions, $[M+CH_3COO]^-$ or $[M+COO]^-$, respectively (Table 3.1). PI can also be seen as $[M+NH_4]^+$ ions. This knowledge allows calculating the molecular weight of each PL molecular species. Nowadays, the acquisition of high-resolution MS data in the Orbitrap mass spectrometers, which allow acquiring MS data with higher resolution, from 70 000 to 500 000, provides information on the exact mass values of ions with a mass accuracy below 2 ppm.

Table 3.1 Type of ions and adducts formed during electrospray ionization and scan conditions for molecular-ion independent analysis of polar lipids. The most abundant ions formed in ESI-MS are shown in bold.

Lipid class	Positive ion mode	Negative ion mode
Phosphatidylcholine (PC)	**$[M+H]^+$**,$[M+Na]^+$,	$[M+Ac-H]^-$
Phosphatidylethanolamine (PE)	$[M+H]^+$,$[M+Na]^+$	$[M-H]^-$
Phosphatidylglycerol (PG)	$[M+NH_4]^+$, $[M+Na]^+$	**$[M-H]^-$**
Phosphatidylinositol (PI)	$[M+NH_4]^+$	**$[M-H]^-$**
Phosphatidylserine (PS)	$[M+H]^+$	**$[M-H]^-$**
Cardiolipin (CL)		**$[M-H]^-$**, $[M-2H]^{2-}$

Tandem mass spectra (MS/MS) of each ion are required to confirm the identity of each molecule. In the MS/MS spectra, it is possible to observe typical fragment ions of each PL class, allowing to identify the polar head group of the PL. Specific neutral losses of the fatty acid such as ketene (loss of R=CO) or acid loss of RCOOH, in the positive mode, or the formation of the carboxylate ions, $RCOO^-$, in the negative mode, allow for the identification of the fatty acyl composition of each PL molecular species (Figures 3.3 and 3.4).

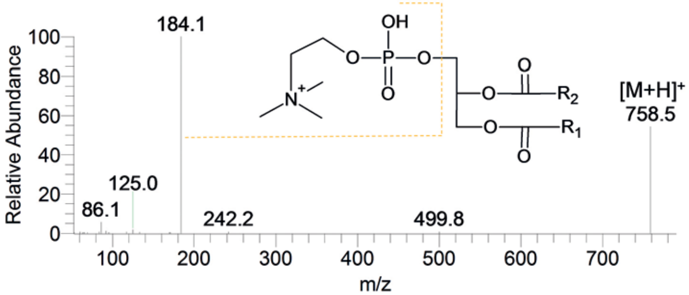


Figure 3.3. MS/MS spectrum of [M+H] $^+$ ions of PC, showing the typical product of PC class at *m/z* 184.

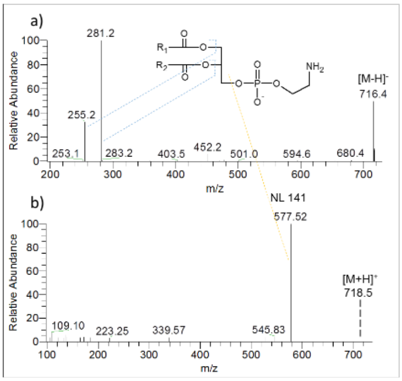


Figure 3.4. MS/MS spectra of phosphatidylethanolamine (PE) in the negative ion mode (**a**), showing the typical carboxylate anions RCOO- that allow to identify the fatty acyl chains in the PE molecules, and positive ion mode (**b**) showing the typical neutral loss of 141 Da, typical of PE class, that, in this case leads to the formation of an abundant fragment ion at *m/z* 577.5.

The most common ions and neutral losses observed in the tandem mass spectra of each PL class are represented in Table 3.2.

Table 3.2. Typical fragmentation of polar lipids observed in the MS/MS spectra.

Phospholipid class	**Detected ions in MS**	**Positive ion**	**Negative ion**
Phosphatidylcholine (PC)	$[M+H]^+$	PIS *m/z* 184	-
Phosphatidylethanolamine (PE)	$[M+H]^+$ $[M-H]^-$	NLS 141 Da	-
Phosphatidylserine (PS)	$[M+H]^+$ $[M-H]^-$	NLS 185 Da	NLS 87 Da
Phosphatidylglycerol (PG)	$[M-H]^-$	-	NLS 74 Da
Phosphatidylinositol (PI)	$[M-H]^-$	-	PIS *m/z* 241
Cardiolipin (CL)	$[M-H]^-$ $[M-2H]^{2-}$		
Sphingomyeline (SM)	$[M+H]^+$	PIS *m/z* 184	-

1.1.2.1. LC-MS Untargeted Lipidomics approaches

In the untargeted lipidomics, the analysis is performed by LC-MS. This approach has the advantage of being able to separate and characterize complex lipid extracts. Profiling of lipids in their representative classes and molecular species relying on mass spectrometry combined with chromatography gives the possibility to separate and concentrate different classes according to their physicochemical properties. The use of LC-MS platforms has improved the resolution and sensitivity in lipidomics analysis.

Through LC-MS platforms, the identification of lipid species is based on the retention time (RT) and the interpretation of MS and MS/MS data (Figure 3.5). LC is performed either by reversed-phase (RP), normal-phase (NP) or by hydrophilic interaction liquid chromatography (HILIC). The RT depends on the experimental conditions such as the type of LC column, flow rate, type of eluent, among others. The capacity and selectivity of the column are variable and depend largely on the column manufacturer, whereas efficiency and resolution can be controlled, to some extent, by the elution conditions used.

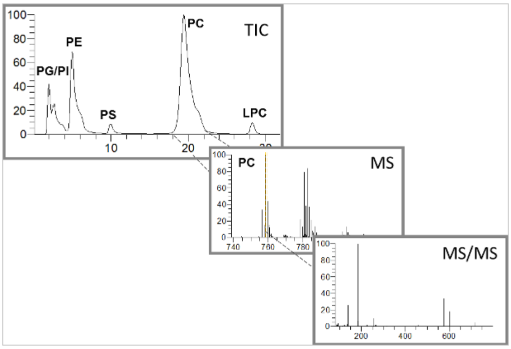


Figure 3.5. LC-MS data from a lipid extract analysis using a HILC-MS platform. Example of an LC Total Ion Chromatogram (**TIC**), and **MS** spectrum of a PC class showing the $[M+H]^+$ ions of all the molecular species of this PL class. The **MS/MS** spectrum of one precursor ion selected from the MS spectrum showing product ions this allow to confirm the structural features of the molecular species.

In lipidomics, the most used columns are NP or HILIC, which allow the separation of the lipid classes in only one run. They have been used to prospect potential markers of disease or to identify deviations in metabolic pathways associated with specific pathologies. These columns elute the lipid species according to their hydrophilicity, which is mainly dependent on the polar head properties. Reversed-phase C18 columns separate lipids based on their hydrophobic properties, which mainly depend on the number of carbons and the degree of saturation of the fatty acyl substituents. Thus, lipids containing longer and saturated fatty acyl chains are eluted later than those containing shorter and polyunsaturated acyl chains. C18 LC approches are usually used after prior fractionation and isolation of the class of interest, or by using targeted mass spectrometry approaches.

Untargeted approaches enable the identification and quantification of a large number of species. Thus, the analysis of these data requires the fast identification and quantification of hundreds of lipid species, which is quite challenging, but bioinformatics tools are being developed for data processing, organization, analysis, and visualization.

1.1.2.2. Shotgun Targeted Lipidomics approaches

As discussed before, targeted approaches make use of typical fragmentation patterns or reporter ions of PL classes to define a specific precursor ion scan (PIS) or neutral loss scans (NLS). For instance, lipid PIS of the ion at *m/z* 184 is used to detect PC species and NLS of 141 Da to detect PE, are typically used in shotgun lipidomics. The selectivity of these analytical approaches is very advantageous and makes use of the specific fragmentation pathways of the unique head group of each PL class (Table 1). These approaches can be performed either by direct infusion in ESI-QqQ or Q-trap mass spectrometers or using LC-MS platforms. Multiple reaction monitoring (MRM) approach, usually performed in triple quadruple spectrometers, is a targeted MS analysis that screens a specific parent ion/fragment ion pairs. This MS/MS mode is commonly used to quantify compounds, usually using a calibration curve.

A major advantage of shotgun lipidomics is that all molecular species of a lipid class can be observed in a single mass spectrum, without separation. This approach is very fast, allowing the analysis of hundreds of samples in a short period of time. The main disadvantage of this approach is the observation of ion suppression phenomena, which can dramatically reduce the sensitivity. Also, in some cases, it is not possible to define specific NLS.

To know more:

U. Loizides-Mangold. On the future of mass-spectrometry-based lipidomics. *FEBS J.*, 2013, 280(12):2817-29.

X. Han. Lipidomics for studying metabolism. *Nature Rev. Endocrin.*, 2016, 12(11):668-679.

L. Yang et al. Recent advances in lipidomics for disease research. *J. Separation Sci.*, 2016, 39:38-50.

M. Holčapek, G. Liebisch, K. Ekroos. Lipidomic Analysis. *Analytical Chemistry*, 2018, 2090:4249-4257.

1.2. Quantification

In lipidomics, different levels of quantification can be achieved. Relative quantification of PL as classes can be performed after separation of each PL class by TLC, identification of each lipid class in comparison with PL standards, and further quantification of phosphorus in each spot.

At the molecular level, relative or absolute quantification can be addressed, either by using LC-MS or shotgun approaches. Relative quantification of molecular species can be carried out to pinpoint the profile of a specific class as well as the variations in the profiles of the same class in different biological conditions. For that, and after identification of each PL molecular species, it is necessary to calculate the peak area of each ion. To determine the relative abundance of each lipid species within the selected class, it is necessary to normalize each peak by the total sum of the areas of all the peaks. Internal standards can be added to the sample (cells, tissues or biofluids) before lipid extraction or they can be added to the total lipid extracts, just prior to the LC-MS run. In this case, peaks can be normalized using the peak area of the lipid standard. Altogether, this enables estimating the relative abundance of each lipid species within a specific lipid class.

To achieve the absolute quantification of each molecular species by LC-MS analysis, it is necessary to build a calibration curve for each species to be quantified. Multiple reaction monitoring (MRM) approach, usually performed in triple quadruple spectrometers, is a target MS analysis that screens specific parent ion/fragment ion pairs. This MS/MS mode is commonly used for absolute quantitation of compounds, usually using a calibration curve.

The drawback of both qualitative and quantitative approaches when using NP or HILIC columns, is that they cannot discriminate PL species within the same classes with the same molecular weight. For instance, PC(36:4) can be PC(16:0/20:4), PC(18:3/18:1), or PC(18:2/18:2). In this case, it can only be assumed that the quantification of PC(36:4) is being made. An alternative approach using pre-purification and C18 columns can be used.

To know more:

X. Han, K. Yang, R.W. Gross. Multi-dimensional mass spectrometry-based shotgun lipidomics and novel strategies for lipidomic analyses. *Mass Spectrom. Rev.*, 2012, 31:134-78.

1.3. Data processing and lipid identification/quantification

The high throughput LC-MS lipidomics generates a considerable amount of data that requires bioinformatic tools to perform the analysis and integrate all information. Also, lipidomics usually requires adequate high-content databases, for supporting the lipid identification. Since there are no universal bioinformatics tools for the automated analysis, interpretation and quantification of LC-MS data remain a challenge.

Lipidomics data analysis should start with the identification of the lipid species, which can be performed based on the identification of the RT and the molecular ion *m/z* values in MS raw data and MS/MS analysis (Figure 3.6).

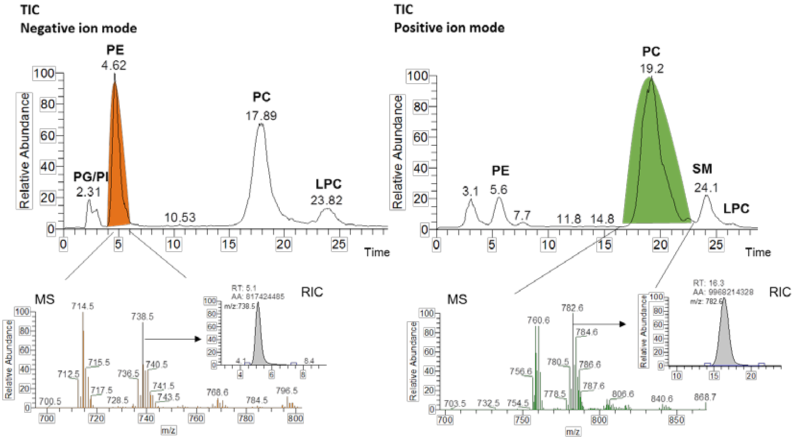


Figure 3.6. LC chromatogram obtained during the lipidomics analysis of the total extract of keratinocytes. In this figure, it is shown the RIC chromatograms for selected ions of PE and PC lipid classes, and the correspondent MS spectra. The area of each peak can be used for quantification.

Lipid identification is usually performed by manual data analysis, aided by consulting accurate mass and fragmentation databases. Currently, there is no universal lipid classification or database of compounds, although there are a few lipid databases that can help in this task, such as LipidBank (http://lipidbank.jp/), LIPIDATA (http://lipidata.in/) and LIPID MAPS (http://www.lipidmaps.org/). Also, nearly every mass spectrometer vendor has developed and commercialized their own software for storing and handling the lipidomics data. The software packages available for lipid MS data analysis, both commercial and freeware, have been developed for specific types of applications and data acquisition modes. Commercialized software available now are LipidView^{TM} (from Sciex), which has been developed for multiple precursor ions and neutral loss scanning, LipidSearch^{TM} (from Thermo Scientific) for LC/MS-based lipidomics data, together with the high-resolution accurate-mass data produced by Orbitrap^{TM}- based mass spectrometers. There are also some free and open-source software tools and libraries for MS lipidomics data analysis. The most frequently used are LipidXplorer (https://wiki.mpi-cbg.de/wiki/lipidx/index.php/Main_Page.), ALEX software (http://mslipidomics.info/contents/?page_id=133), Lipid Blast (http://fiehnlab.ucdavis.edu/ projects/LipidBlast) and MS-DIAL (http://prime.psc.riken.jp/Metabolomics_Software/MS-DIAL/). They allow for the qualitative and quantitative analysis of lipid data, acquired using different approaches and different mass spectrometers. Both LipidXplorer and ALEX have been designed for shotgun lipidomics using high-resolution mass spectrometers. LipidBlast has been developed for polar lipids analysis through MS/MS experiments and mass spectral library search, using either low- or high-resolution instruments. The MS-DIAL has been established to deal with both data dependent and independent MS/MS experiments.

After quantitation, for further data processing, usually, the data are normalized and subjected to statistical analysis, usually using the same approaches as described in the metabolomics chapter of this book.

1.4. Practical sessions

1.4.1. Identification of the phospholipidome profile of THP1- monocytes using a lipidomics approach based on HILIC-LC-MS.

The lipid extract obtained from THP1 human monocytic cell lines were analyzed by LC-ESI-MS and LC-ESI-MS/MS in the positive and negative ion modes. In Figure 3.7, we show the LC-ESI-MS spectra obtained in the positive ion mode, at retention times (RT) of 22 min (A) and of 7 min (B), and in the negative mode, at RT of 22 min (C) and RT of 7 min (D).

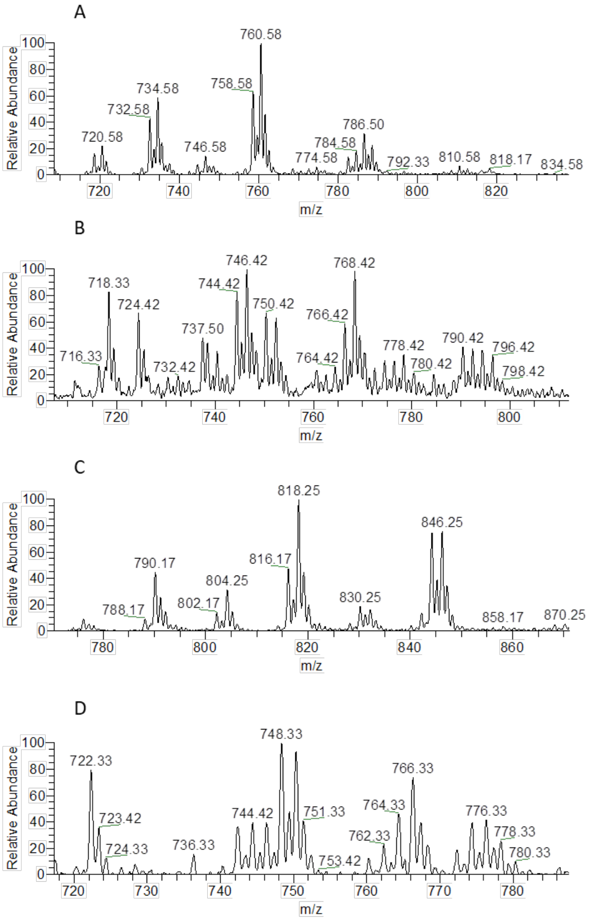


Figure 3.7. The LC-ESI-MS spectra after analysis of the total lipid extract obtained in the positive ion mode with a retention time of 22 min (A) and 7 min (B) and obtained in the negative ion mode with a retention time of 22 min (C) and 7 min (D).

To obtain information about the phospholipid classes, it is necessary to proceed to the study of fragmentation of the selected molecular ions. Figure 3.8 shows the ESI-MS/MS spectra acquired for the ions observed at *m/z* 760.58 (Fig A), at *m/z* 818.58 (Fig B), at *m/z* 746.5(Fig C) and at *m/z* 744.56 (Fig D).

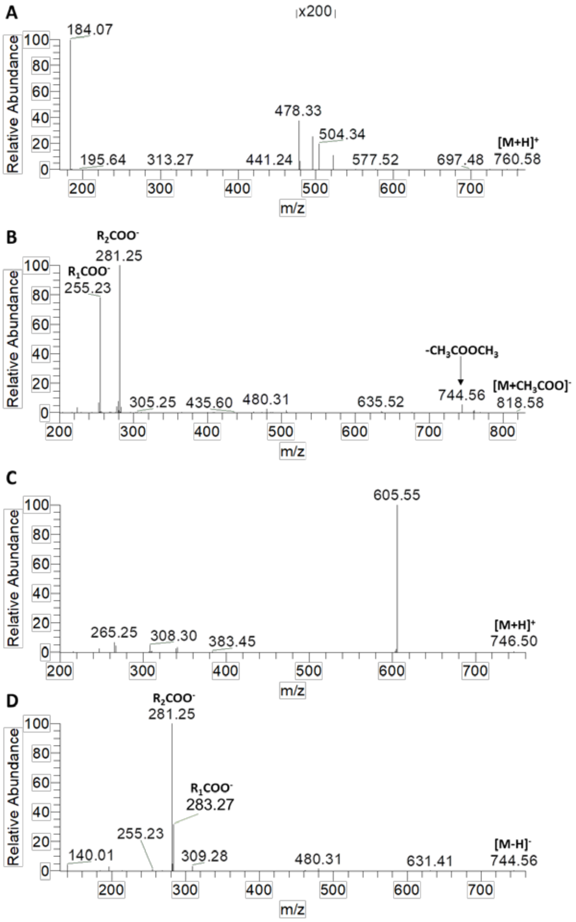


Figure 3.8. The ESI-MS/MS spectra of the ions observed at *m/z* 760.58 (A), *m/z* 818.58 (B), *m/z* 746.50 (C) and *m/z* 744.56 (D). Spectra A and C were acquired in positive ion mode, while spectra B and D were acquired.

Questions

1. Based on the typical fragmentation pathways of phospholipids, what phospholipid classes have the retention time of 7 min and 22 min in the LC-MS chromatogram?
2. Taking into account the ESI-MS/MS spectra obtained at 7 min in both negative and positive mode, propose a structure for this phospholipid species.
3. What other phospholipid classes are to be expected in LC-ESI-MS analysis of the THP1s lipid extract?
4. Describe another approach that allows evaluating the lipid changes in THP1 cell lines.

1.5. Additional Resource Readings

J.D. Martins, E.A. Maciel, A. Silva, I. Ferreira, P. Domingues, B.M. Neves, M.T. Cruz, M.R.M. Domingues. Phospholipidomic profile variation on THP-1 cells exposed to skin or respiratory sensitizers and respiratory irritant. *Journal of Cell Physiology*, 2016, 231(12):2639-51.

M. Pulfer, R.C. Murphy. Electrospray mass spectrometry of phospholipids. *Mass Spectrometry Reviews*, 2003, 22(5):332-64.

Y.H. Rustam, E. Gavin. Analytical Challenges and Recent Advances in Mass Spectrometry Based Lipidomics. *Anal. Chem.*, 2018, 90:374-397.

K. Yang, X. Han. Lipidomics: Techniques, Applications, and Outcomes Related to Biomedical Sciences. *Trends in Biomedical Sciences*, 2018, 41:954-969.

H.C. Lee, T Yokomizo. Applications of mass spectrometry-based targeted and nontargeted lipidomics. *Biochem Biophys Res Commun*, 2018, doi: 10.1016/j.bbrc.2018.03.081.

S. Sethi, E. Brietzke. Recent advances in lipidomics: Analytical and clinical perspectives. *Prosaglandins Other Lipid Mediat.*, 2017, 128-129:8-16.

Thanks are due to Fundação para a Ciência e a Tecnologia (FCT, Portugal), the European Union, the Quadro de Referência Estratégica Nacional (QREN), the Programa Operacional Potencial Humano (POPH), the Fundo Europeu de Desenvolvimento Regional (FEDER) and Programa Operacional Factores de Competitividade (COMPETE) for funding the QOPNA (Química Orgânica de Produtos Naturais e Agroalimentares; UID/QUI/00062/2013) research unit and to RNEM (LISBOA-01-0145-FEDER-022125) for the Portuguese Mass Spectrometry Network.

Module 4
Proteomics

Tânia Melo[1], Rita Ferreira[1], Rosário Domingues[1,2], Pedro Domingues[1],

Mass Spectrometry Center, Department of Chemistry, University of Aveiro, 3810-193 Aveiro, Portugal

[1] QOPNA (Química Orgânica de Produtos Naturais e Agroalimentares)

[2] CESAM (Centro de Estudos do Ambiente e do Mar)

I. Rationale

Proteomics is the large-scale analysis of proteins. This information is, nowadays, fundamental for understanding the molecular mechanisms underlying the role of proteins in pathophysiological conditions, aiming to improve treatment, to identify protein biomarkers and signaling events, as well as protein interactions. The first part of this module will focus on the current methodologies used to analyze and identify proteins. This will include the study of the principal technologies involved, namely protein separation methods, mass spectrometry approaches to protein analysis, and protein database analysis. The second part of this module will focus on a case study and is designed to familiarize students with protein analysis software that is freely available. The course is based on expository and interactive lectures with students and demonstration/practical sessions and analysis of practical cases that will allow consolidating specific outcomes.

II. Course Aims and Outcomes

Aims

This course aims to provide students with an understanding of the core concepts and approaches for the analysis of the proteome. Students shall acquire basic knowledge about theory and practice of proteomics and the use of these methods within biomedical research, including the methods of extraction, purification, identification, and quantification of proteins and the use of proteomic bioinformatics tools.

Learning outcomes:

By the end of this course, students will:

1. Define and apply common proteomics terminology;
2. Recognize the principles of the most common proteomics techniques;
3. Understand the different mass spectrometry-based proteomics workflows;
4. Be able to choose between the different proteomic approaches to solve a specific problem;
5. Be able to use proteomics bioinformatics tools;

6. Interpret the data resulting from the identification of proteins;
7. Communicate and justify conclusions clearly and unambiguously to both specialist and non-specialist audiences;
8. Continue the learning process, largely autonomously.

III. Course contents

Module 4- Proteomics

1. Introduction to Proteomics
2. Analytical approaches in proteomics
   2.1. Sample preparation
   2.2. MS-based protein identification: concepts
   2.3. Mass spectrometry approaches for the identification of proteins
   2.4. Quantitative proteomics
   2.5. Identification of PTMs
   2.6. Detection and quantification of subcellular protein localization
   2.7. Detection and quantification of protein interactions
3. Data processing and identification of proteins
   3.1. Protein sequence databases
   3.2. Protein database search engines
   3.3. Data analysis
4. Practical sessions
   4.1. Protein identification using PMF and PFF approaches with MASCOT
   4.2. Protein identification from an LC-MS experiment
      4.2.1.Transformation of data to mzML with MSConverter in ProteoWizard
      4.2.2. Generation of FASTA database from UniProt (SwissProt)
      4.2.3. Search Engines: SearchGUI
      4.2.4. Generation and evaluation of results: PeptideShaker for peptide and protein visualization, and validation. PTM analysis
      4.2.5. Data analysis: protein information, pathway analysis, and gene ontology

1. Introduction to Proteomics

Briefly, proteomics is defined as the large-scale study of proteins. Proteomics is based on a range of recent new technologies that provide a fast and accurate determination of the proteome. This task is very complex due to both technical and biological constraints. At the biological level, the existence of a broad range of protein concentrations in biological samples constitutes one of the principal difficulties. This high concentration range translates into several technical difficulties, including the need for fractionation and enrichment of less abundant proteins and the importance of developing rapid and sensitive methods for processing large numbers of samples. Also at the biological level, the proteome, defined as the proteins expressed by a genome, cell, tissue or organism at a certain time, is highly dynamic and can change with age, stress conditions, and nutrients, among others. Moreover, the post-translational modifications (PTMs), largely contributes to an increase in the complexity of the problem.

Several different proteomic approaches provide data allowing to address different biological questions. The proteomics-based expression analysis seeks to determine all the proteins that are expressed in a given condition. Quantitative proteomics aims to measure the abundance of expressed proteins, for example, by comparing experimental groups with control samples. Other approaches include the identification of PTMs and the study of protein-protein interactions:

- Detection and quantification of protein levels
- Detection and quantification of protein modifications
- Detection and quantification of subcellular protein localization
- Detection and quantification of protein interactions

For each of these biological questions, we can use different mass spectrometry-based proteomics workflows. Nevertheless, the basic proteomics workflow usually includes the extraction of proteins from the organism or tissue under study, the enzymatic digestion of the protein by proteases and purification of the peptides, the analysis of the peptides by tandem mass spectrometry and the analysis of the data using bioinformatics tools (Figure 4.1).

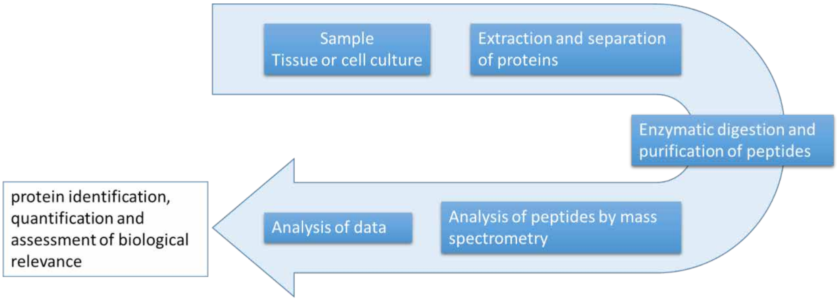

Sample
Tissue or cell culture
Extraction and separation of proteins
Enzymatic digestion and purification of peptides
Analysis of peptides by mass spectrometry
Analysis of data
protein identification, quantification and assessment of biological relevance

Figure 4.1. Basic workflow in proteomics.

Proteomics research approaches can be varied and can be classified, depending on the objective, in discovery proteomics and targeted proteomics. In discovery proteomics, a limited number of samples to be analyzed is used, to optimize protein identification, but it is intended to identify as many proteins as possible. Targeted proteomics strategies limit the number of proteins that will be monitored to achieve the highest sensitivity and throughput of samples. Nevertheless, as a rule, the greater the number of proteins to be analysed, and the higher the dynamic range in a sample, the more difficult it is to identify them and to measure their presence quantitatively.

2. Analytical approaches in proteomics

2.1. Sample preparation

The preparation of the protein sample to be analyzed is critical in proteomic studies. The quality of the protein sample and the reproducibility of the used methods for protein extraction and separation significantly affects the quality of the proteomics data. Protein extraction and purification is a very complex task since proteins are a very heterogeneous group of compounds, with differences in size, charge, solubility, and concentration. The range of protein concentrations in a cell is of several orders of magnitude, and more than eight orders of magnitude in serum. So far, there is no standard method of sample preparation for proteomics studies, which makes the comparison between studies very difficult. However, although protocols can be very different, they all share a common philosophy, as described in Figure 4.2. This figure shows the general sample preparation procedure, highlighting some of the different options that are available.

In the beginning, the classical approach in proteomics used to be two-dimensional gel electrophoresis (2-DE) and MALDI-TOF mass spectrometry (MS) analysis. More recently, the approaches usually include 1D SDS-PAGE followed by LC-ESI-MS/MS (GeLC-MS/MS). However, because of many problems associated with gel-based techniques, which include protein precipitation, and difficulties associated with resolution and reproducibility, in recent years, different gel-free approaches have been developed. These usually include the use of multistep high-resolution liquid chromatography approaches that allow protein pre-fractionation and high sensibility in protein identification.

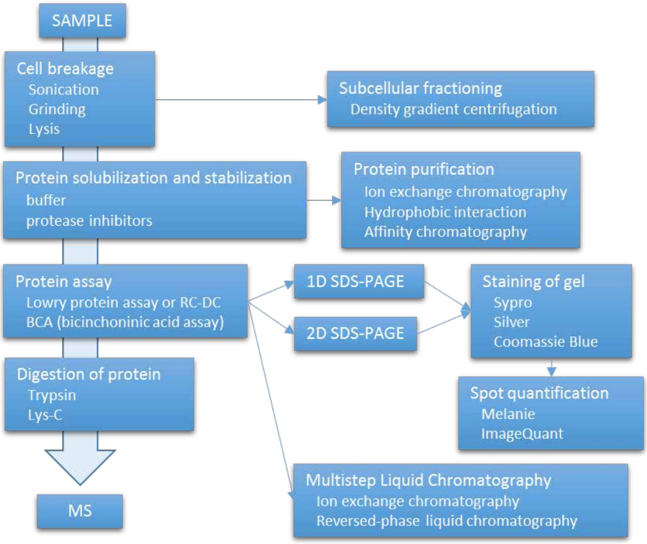


Figure 4.2. General sample preparation procedure, highlighting different options available.

To know more:

J. Lovric. Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis. ISBN: 978-0-470-03524-5, Chapter 2.1, 2011.

2.2. MS-based protein identification: concepts

There are several strategies for protein identification using mass spectrometry. The methods used for identification of proteins are designated by "bottom-up" and "top-down" approaches (Figure 4.3). The "bottom-up" approach is the traditional proteomics approach, in which proteins are enzymatically digested, and the resulting peptides are analyzed using mass spectrometry, and tandem mass spectrometry and the protein is identified based on the identity of the peptides.

For the enzymatic digestion of the proteins, different enzymes or a combination of enzymes can be used, although the most frequently chosen is trypsin. Digestion of the proteins separated using a gel-based approach is usually performed directly in the gel, after cutting plugs that contain the proteins to be identified (in-gel digestion), which is followed by extraction of the peptides. When using a liquid-based separation approach, proteins are usually digested using an in-solution digestion protocol, followed by purification of the peptides before the analysis by MS.

The use of chromatographic separation of the peptides is usually required to increase the number of proteins that are identified in the sample. This fact is due to the existence of the "ion suppression effect" when using soft ionization techniques (MALDI and ESI). The ion suppression effect consists on the suppression of the MS signal of more hydrophilic and low proton affinity peptides. This separation is usually performed using C18 reversed phase nanoLC-MS, and peptides are typically eluted with a binary solvent gradient consisting of water and acetonitrile and an ionpairing agent (for example, formic acid). In complex samples, longer gradients are used, or ion exchange chromatography can be combined with reversed-phase chromatography, either off-line or online, using different columns (2D-LC) or in the same column (Multidimensional Protein Identification Technology, MudPIT). When using a proteomic approach that does not perform any prior separation of the proteins before digestion (shotgun approach), usually the use of 2D-LC separation of the peptides is required for the best performance.

After separation of the peptides, they are ionized, using ESI or MALDI as ionization sources, and analyzed by mass spectrometry (MS) and tandem mass spectrometry (MS/MS). The discrimination effect in the analysis of different peptides in MALDI and ESI contributes to different protein coverage rates when using these different ionization methods. Commonly, the ESI allows a slightly better coverage rate, although MALDI is less sensitive to contamination and to the ion suppression effect.

In the top-down approach, a purified protein is ionized, usually using an ESI source and fragmented by electron capture dissociation (ECD) or electron transfer dissociation (ETD) fragmentation methods. The fragment ions are then analyzed using high-resolution analyzers such as FT-ICR or orbitraps. This approach usually requires a higher degree of purification of the proteins, usually using multidimensional liquid chromatography approaches. Also, it requires a higher amount of proteins and the use of costly instrumentation.

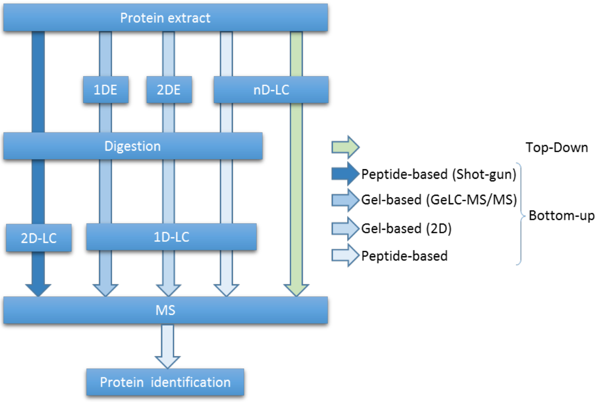


Figure 4.3. MS-based protein identification approaches.

To know more:
Z. Szabo, T. Janaky. Challenges and developments in protein identification using mass spectrometry. *Trends in Analytical Chemistry*, 2015, 69:76-87.

2.3. Mass spectrometry approaches for the identification of proteins

Two main techniques, peptide mass fingerprinting (PMF) or peptide fragmentation fingerprinting (PFF), are used for the identification of proteins using the “bottom-up” approach (Figure 4.4). The PMF technique identifies the proteins by matching the molecular weight of the peptides originated by enzymatic digestion with the theoretical molecular weight of the peptides generated *in silico* from protein or DNA databases. The experimental data is a list of peptide mass values from the digestion of a protein by a specific enzyme, such as trypsin, acquired in a high resolution and high accuracy instrument. This technique can be used in single protein samples since it allows fast analysis, if several tryptic peptides are observed (ideally more than five). The proteins identified by using this method are scored based on different algorithms.

However, these are usually based on the protein sequence coverage, on the observation of long peptides, which are unlikely to be present in multiple proteins, and on the mass accuracy of the observation.

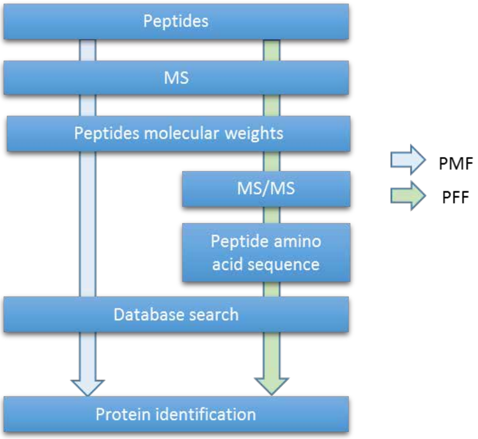


Figure 4.4. Main techniques used for the identification of proteins using the "bottom-up" approach: peptide mass fingerprinting (PMF) or peptide fragmentation fingerprinting (PFF).

The PFF technique identifies the proteins by matching the MS/MS fragmentation pattern of the tryptic peptides with the theoretical fragment spectrum generated *in silico* from a protein or DNA database. This technique is considered the "golden standard" approach for identifying proteins. It identifies proteins by inference and is very useful for the identification of proteins in very complex samples, since it is easily automated for high throughput and can obtain matches from limited data, simultaneously allowing to identify many variable modifications. When using this approach, usually an LC-MS/MS experiment is designed, to obtain the highest sensitivity and a large amount of MS/MS data. To do this, a data-dependent MS/MS experiment is programmed, in which the mass spectrometer automatically acquires MS/MS data from the n-top (n=10-20) most intense peptides observed in the previous MS scan (Figure 4.5).

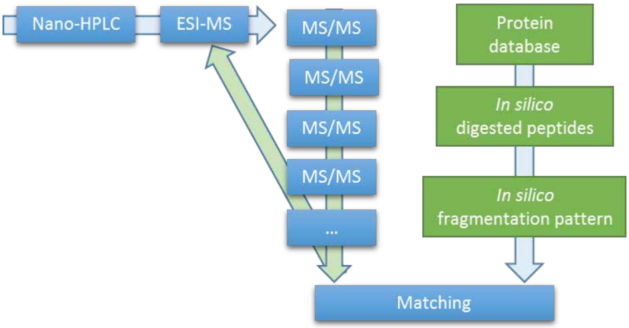

Nano-HPLC
ESI-MS
MS/MS
MS/MS
MS/MS
MS/MS
...
Protein database
In silico digested peptides
In silico fragmentation pattern
Matching

Figure 4.5. Data-dependent MS/MS experiment in peptide fragmentation fingerprinting (PFF) proteomics approach.

To know more:
J. Lovric. Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis. ISBN: 978-0-470-03524-5, Chapters 4.2 and 4.3, 2011.

2.4. Quantitative proteomics

Although the objective of some proteomic experiments is to identify the proteins that are present in a sample, nowadays most experiments are concerned with the study of the changes that occur at the expression level, when the subjects are in different conditions, for example, healthy versus diseased. For these applications, the accurate quantitation of proteins is essential and constitutes a central aspect of proteomics.

Currently, there are various spectrophotometric assays for detection and measuring the amount of an individual protein in a solution. However, these methods rely on the prior purification of the protein of interest. There are also several well-established methods for the quantitation of individual proteins, either in solution or in using a solid-phase assay, which is based on the use of labeled antibodies. However, these rely on the quality of the antibody and the strength of the antigen-antibody binding and are low-throughput methods. Nevertheless, these labeled antibodies can be grouped into analytical protein microarrays, allowing high throughput quantitative analysis. Unfortunately, the most successful microarrays contain only a small number of well-characterized antibodies.

The high-throughput and large-scale quantitative analysis in proteomics is currently performed in two different approaches: based on high-resolution protein separation by two-dimensional (2D) gel electrophoresis or based on the relative abundance of peptide ions on the mass spectrum (Figure 4.6). Although it allows for the simultaneous quantitation of thousands of proteins, the use of 2D gel electrophoresis represents a bottleneck in the large-scale analysis of proteins due to the difficulties with resolution, reproducibility, and spot matching. These difficulties can be somewhat obviated by the use of Fluorescence Difference Gel Electrophoresis (2D DIGE). In 2D DIGE, proteins from different samples are labeled using fluorescent labels (Cy2, Cy3, and Cy5) and then combined before separation and quantification through two-dimensional gel electrophoresis.

This approach has the advantage of minimizing spot pattern variability and the number of gels in an experiment.

The analysis of peptides by mass spectrometry is not, essentially, a quantitative analysis. This fact is due to multiple factors including: different ionization efficiencies of the peptides, the suppression effect when using soft ionization techniques, and a limited dynamic range of some analyzers. Thus, different approaches have been developed to quantify proteins using mass spectrometry. These can be classified into relative and absolute quantitation approaches (Figure 4.6). Relative quantitation approaches compare the relative abundance of individual peptides in a sample to those in different samples, while absolute quantitation approaches determine the exact amount or mass concentration of a protein.

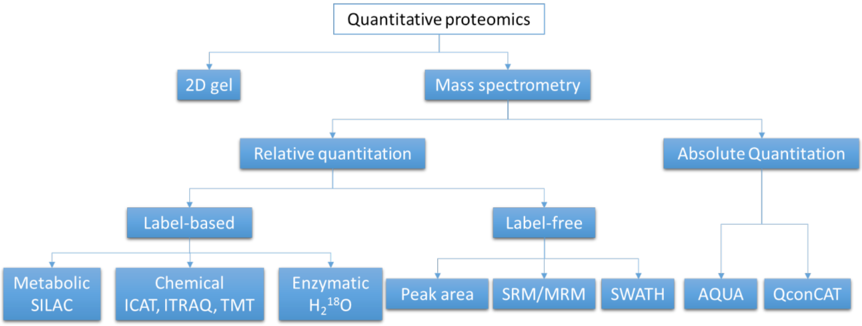


Figure 4.6. Methods for quantitative analysis in proteomics.

Relative quantitation approaches rely on LC-MS and LC-MS/MS experiments and can be carried out using both label and label-free approaches. Label-free approaches depend on spectral counting or peptide peak intensity measurement. The spectral counting approach is based on the number of peptides identified from a given protein to determine its relative abundance, by using different algorithms. These algorithms, for example, emPAI, calculate the predicted abundance of the proteins taking into account the sequence and length of the protein. Relative quantitation using peptide peak intensity measurements is performed by comparing the relative intensity of the MS peptide ions from a protein in different samples. Usually, the peak areas of the several peptides in an LC-MS experiment, typically more than three, are integrated, and the values for each peptide are compared.

Label-based approaches rely on the introduction of different mass tags which alter the mass of the protein or peptide. After this, the samples are combined into a single sample, and the relative quantitation is performed by comparing the relative abundance of either peptide ions (from MS experiments) or peptide product ions (from MS/MS experiments) of the differentially tagged ions. These labels can be introduced metabolically, chemically, or enzymatically at the protein or peptide level during sample preparation. Metabolic labeling involves stable isotope labeling with amino acids in cell culture (SILAC), by culturing cells or organisms in media with tagged analogs of biomolecule monomers (^{2}H, ^{15}N, ^{13}C and ^{18}O). These samples from different conditions can then be pulled together and analyzed simultaneously, avoiding the most common quantitation problems.

In chemical labeling, the isotope label is introduced into proteins or tryptic peptides by a chemical reaction. For example, when using isotope-coded affinity tags (ICAT), a label that binds to cysteine residues with “heavy” (^{13}C) or “light” (^{12}C) reagent is used. The two samples are modified before pooling and then combined for digestion, peptide purification by avidin chromatography and mass spectrometry analysis. Isobaric mass tags, such as tandem mass tag (TMT) and isobaric tags for relative and absolute quantification (ITRAQ) are mass-balanced labels that are used to synthesize tagged tryptic peptides that have the same mass and chromatographic properties. The different mass tags are then identified by tandem mass spectrometry, allowing for quantitative determination of the relative abundance of proteins from up to eight samples.

Enzymatic labeling relies on the use proteases to incorporate a mass label ^{18}O, into the carboxy terminus of peptides. This modification at the C-terminal carboxyl group of proteolytic fragments involves the replacement of two ^{16}O atoms by two ^{18}O atoms when digestion occurs in the presence of $H_2{}^{18}O$. These peptides can then be easily identified in the MS spectra of labeled samples by the presence of a 4Da mass shift from $H_2{}^{16}O$ C-terminal carboxyl group proteolytic peptides.

Absolute mass spectrometry quantitation proteomics usually relies on the use of stable isotope–labeled trypsin-like peptides as internal standards. Absolute quantification is achieved by adding a known amount of these standards to the sample, before the LC-MS experiment. These peptides, known as absolute quantification (AQUA) peptides, have the same retention time as the peptides from the tryptic digestion and absolute quantitation is achieved by using a standard curve to yield the absolute quantitation of the target peptide. Alternative approaches use artificial proteins made of concatenated peptides (concatemer (QconCAT) approach) or make use of protein standards for absolute quantification (PSAQ). In the former approach, each peptide is an internal standard (surrogate) that represents a specific protein, while in the latter a protein that is isotopically labeled and corresponds to the recombinant expressed analog of the protein to be quantified is used. However, these methods are costly and low-throughput approaches and their use have been restricted to a small number of targeted experiments.

To know more:

J. Lovric. Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis. ISBN: 978-0-470-03524-5, Chapters 4.4 and 5.3, 2011.

2.5. Identification of PTMs

Post-translational modifications (PTMs) are chemical modifications of a protein chain after translation, which are responsible for the change of size, composition, function, and location of proteins. More than 900 different PTMs have been identified and included in UNIMOD database (www.**unimod**.org/). The most common post-translational modifications (Table 4.1) include phosphorylation, acetylation, oxidation, alkylation, methylation, and the formation of disulfide bridges. Mass spectrometry is currently the best approach for sequencing peptides, and identifying PTMs, enabling a cost-effective analysis of a large number of samples.

Table 4.1. List of the most common post-translational modifications.

PTM Type	Substrate
N-linked Glycosylation	Asparagine and lysine
O-linked Glycosylation	Lysine, proline, serine, threonine, and tyrosine
C-linked Glycosylation	Tryptophane
Phosphorylation	Serine, threonine, tyrosine, aspartate, histidine or cysteine
Acetylation	N-terminal of some residues and side chain of lysine or cysteine
Amidation	Generally at the C-terminal of a mature active peptide after oxidative cleavage of the last glycine
Hydroxylation	Generally the side chain of asparagine, aspartate, proline or lysine
Methylation	Generally at the N-terminal phenylalanine, the side chain of lysine, arginine, histidine, asparagine or glutamate, and C-terminal cysteine

The PTMs are usually identified by mass spectrometry, by observing a mass shift in mass spectra (MS) of the tryptic peptides (Table 4.2). This mass shift results from a chemical modification that occurs in the side chain of amino acid residues. However, for the unambiguous identification of the location of the PTM, tandem mass spectrometry (MS/MS) experiments are necessary. In these MS/MS data, the mass shift detected in the precursor ion (peptide obtained by tryptic digestion of the modified protein) is identified in the fragment ions having a modified amino acid residue. However, the modified peptides may have other fragmentation pathways that hinder the identification of the site of modification. These pathways are either the loss of the modification as a neutral molecule or as a charged residue.

Most of the PTMs are low-abundance modifications and are labile in MS/MS analyses. Also, many of these modifications are hydrophilic, which might complicate protein sample handling and purification before MS. Also, the presence of protein modifications may affect the cleavage efficiency of proteases, generating unexpected or large peptide products. Certain protein modifications will reduce the ionization and detection efficiency in MS, while multisite protein modification makes the interpretation of the MS/MS data sets complicated and difficult. For these reasons, it is often useful to consider and explore several approaches for mapping PTMs in proteomics. Biochemical purification of cellular compartments, organelles, protein complexes, or of individual proteins is a very useful approach for the analysis of protein modifications, as it reduces the complexity of the protein sample.

It is also very important to use enrichment methods of modified proteins and peptides to decrease the complexity of the sample and therefore increase the number of identified PTMs. Thus, different methods were developed for the selective enrichment of modified proteins for the identification and quantification of PTMs, which include, for example, immunoprecipitation of proteins or PTM enrichment. PTM enrichment can be performed at the protein level, for example, using antibodies, ion metal affinity chromatography (IMAC-phosphoproteins) or immobilized lectins (glycoproteins). PTM enrichment can also be performed at the peptide level, for example, using antibodies (^{Ac}R, ^{Me}K, pTyr), IMAC and TiO_2 (phosphopeptides) or immobilized lectins (glycopeptides).

Table 4.2. Observed mass shift in the MS and MS/MS of modified tryptic peptides.

PTM type	**Δ Monoisotopic Mass (Da)**
Hydroxylation	15.9949
Phosphorylation	
pTyr	79.9663
pSer,	79.9663
pThr	79.9663
Acetylation	42.0105
Methylation	14.0156
Sulfation (sTyr)	79.9568
Deamidation	0.9840
Acylation	
Farnesyl	204.1878
Myristoyl	210.1983
Palmitoyl	238.2296
Glycosylation	
N-linked	>800
O-linked	203, >800
Disulfide bond formation	-2.0157
Nitration of tyrosine	44.9850

To know more:

A.M.N. Silva, R. Vitorino, M.R.M. Domingues, C.M. Spickett, P. Domingues. Post-translational Modifications and Mass Spectrometry Detection. *Free Radic. Biol. Med.*, 2013, 65:925-941.

J. Lovric. Introducing Proteomics: From Concepts to Sample Separation, Mass Spectrometry and Data Analysis. ISBN: 978-0-470-03524-5, Chapter 5.4, 2011.

2.6. Detection and quantification of subcellular protein localization

The field of organelle proteomics has been emerging, aiming to reveal the functions of the proteins in each organelle. However, the elucidation of the subcellular distribution of proteins under different conditions is a major challenge in cell biology, due to the multi-compartmental and dynamic nature of protein localization. Traditionally, subcellular locations of proteins were inferred using more traditional methods, such as immunofluorescence microscopy, green fluorescent protein tagging or biochemical fractionation. Nevertheless, quantitative proteomics workflows have been developed for reliable identification of the protein from whole organelles, as well as for protein assignment to subcellular location. The typical approach is based on subcellular fractionation and enrichment strategies, normally using density gradient centrifugation or commercial kits. Cell organelles may also be enriched by immunoprecipitation with specific antibodies directed against epitopes presented on the surface of the organelle. Mixtures of protein obtained from different organelle fractions are then identified and quantified using the proteomic approaches discussed earlier. Although organelle-based approaches can provide valuable information about specific subcellular compartments in isolation, it is also important to study protein localization in the context of the whole cell to obtain a system-wide view of proteome organization.

To know more:
R. Drissi, M.-L. Dubois, F.-M. Boisvert, Proteomics methods for subcellular proteome analysis. *FEBS J.*, 2013, 280(22):5626-5634.

2.7. Detection and quantification of protein interactions

The role of proteins in the cell depends on the interaction between several different proteins and in the formation of multiprotein complexes. These protein complexes are highly dynamic, and their composition changes over time and with the cell state. Three main approaches are used for the study of protein interaction, which are: affinity pulldown, proximity labeling, and protein correlation profiling. Affinity pulldown uses specific antibodies to isolate the protein of interest and their interacting partners, which are then identified and quantified by mass spectrometry.

Proximity labeling is a technology based on the covalent transfer of biotin labels of a protein to the nearby proteins which are potential interacting proteins. The biotin label in the proteins is then identified using tandem mass spectrometry. Protein correlation profiling is based on the separation using different methods such as density gradient centrifugation or native chromatography and assumes that interacting proteins will co-elute.

Protein-protein interactions in cell signaling are frequently mediated by the interaction of specific amino acid sequences. Despite the central importance of these interactions in cell signaling, the identification of peptide-binding partners is still a very complicated task, and relies mainly on the use of bioinformatics tools, such as STRING (http://string-db.org/).

To know more:
Syafrizayanti, C. Betzen, J.D. Hoheisel, D. Kastelic. Methods for analyzing and quantifying protein-protein interaction. *Expert. Rev. Proteomics*, 2014(2), 11(1):107-20.

3. Data processing and identification of proteins

3.1. Protein sequence databases

As we have discussed previously, protein sequence databases are pivotal in mass spectrometry-based proteomics research. Currently, the principal sources of protein sequence data are translations of nucleotide sequences deposited in the GenBank (https://www.ncbi.nlm.nih.gov/genbank/), EMBL (http://www.ensembl.org/index.html) or DDBJ databases (http://www.ddbj.nig.ac.jp). These databases are integrated into an initiative denominated as International Nucleotide Sequence Database Collaboration (http://www.insdc.org) that gathers nucleotide sequence and annotations of those institutions.

The NCBI protein database (Entrez Proteins,http://www.ncbi.nlm.nih.gov/entrez) has a complete set of deduced proteins. It uses information from the Protein Research Foundation (PRF), Genpet (unreviewed sequences) and RefSeq (manually annotated and reviewed protein sequences) from NCBI (National Center for Biotechnology Information), and from Swiss-Prot (manually annotated and reviewed protein sequences) (Figure 4.7). However, since this database assigns each protein sequence with a unique gene identification (gi) number, it is very large and redundant. NCBI also maintains a non-redundant (nr) protein database, in which sequences and fragment sequences were merged into a single entry.

The UniProt Knowledgebase (UniProtKB) contains extensive curated protein information, including function, classification, and cross-reference and is divided into two different databases: the UniProtKB/Swiss-Prot which is reviewed and manually annotated and UniProtKB/TrEMBL which is the unreviewed section of UniProt.

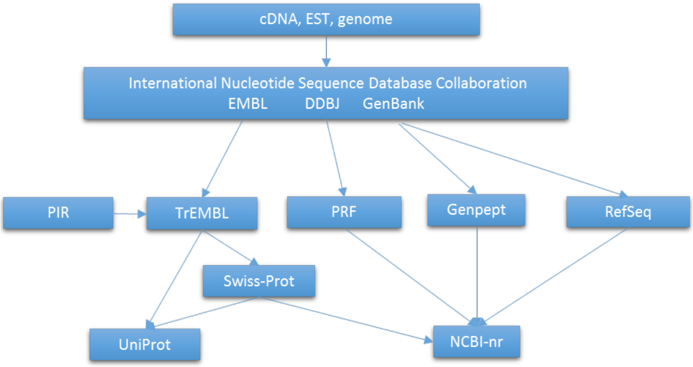


Figure 4.7. Most important protein sequence databases, and their relationships.

To know more:
C. Brooksbank, M.T. Bergman, R. Apweiler, E. Birney, J. Thornton. The European Bioinformatics Institute's data resources 2014. *Nucleic Acids Res.*, 2014, 42(D1):D18-D25.

3.2. Protein database search engines

As discussed previously, for identifying proteins, it is necessary to search the proteins databases for matching peptide masses (PMF approach) or the peptide sequences (PFF approach). Several bioinformatics tools have been designed for this, and they are collectively called protein database search engines. These search engines can be proprietary (Mascot (Matrix Science), Proteome Discoverer (Thermo), ProteinPilot (Sciex), ProteinLynx (Waters), ProteinScape (Bruker) and others) or free (Mascot (limited online search), ProteinProspector, Andromeda, OMSSA, SEQUEST, X!Tandem, Amanda, and many others). Correct MS protein identification depends on many factors, including the algorithm that is used by the search engines and experimental factors that influence the information content in MS data (discussed earlier). Thus, these search engines do not necessarily identify the same proteins in a sample, and usually, some differences are observed. Nevertheless, there are some search engines, for example, SearchGUI, that allow for combining results from multiple search engines. In proteomics, a huge amount of data is acquired and, as such, these searches can take several days per experiment, especially when long LC-MS/MS experiments are used, or when a multiple PTM search is performed.

To know more:
Z.F. Yuan, S. Lin, R.C. Molden, B.A. García. Evaluation of proteomic search engines for the analysis of histone modifications. *Journal of Proteome Research*, 2014, 13:4470-4478.

3.3. Data analysis

The principal challenge of mass spectrometry-based proteomics is related to the analysis and interpretation of the acquired data. This analysis includes integration of taxonomic and functional meta-information from samples containing several hundreds of proteins, in a way that allows interpreting the experiments. To do this, once the identification and quantification of proteins is completed, it is needed to proceed to the functional analysis of the differential proteins relevant for the study. Several free bioinformatics tools can assist in this task, including freeware proteomics pipelines. These proteomic pipelines (OpenMS Proteomics Pipeline, Trans-proteomic pipeline or PeptideShaker) perform different analysis tasks, including assisting in the data analysis. Data analysis usually includes identification of PTMs, comparison and biochemical interpretation of quantitative data, identification of biological processes and protein interactions.

The first step for functional interpretation of the resultant protein list is to connect the protein identifier with its associated Gene Ontology terms. The Gene ontology classification system (http://www.geneontology.org/) defines concepts or classes used to describe gene functions and to classify them from three different perspectives: the molecular function of the gene product, the cellular component where the gene products are active and the biological process where the gene products are active. PANTHER (Protein Analysis Through Evolutionary Relationships) classification system (http://www.pantherdb.org/about.jsp), which is a part of the GO reference Genome Project, is a valuable tool for classifying proteins and their genes to facilitate pathway high-throughput analysis. STRING (http://string-db.org/) and Cytoscape (http://www.cytoscape.org/) are two essential bioinformatics tools that are used for the analysis of protein-protein interaction and designing protein networks. These data can be then categorized to help the scientist in interpreting the results of an experiment. Reactome (http://www.reactome.org/) allows pathway analysis. Finally, Uniprot (http://www.uniprot.org/) is also used as a source of high-quality protein functional information.

To know more:
S.W. Haga, H.F. Wu. Overview of software options for processing, analysis and interpretation of mass spectrometric proteomic data. *Journal of Mass Spectrometry*, 2014, 49:959-969.

4. Practical sessions

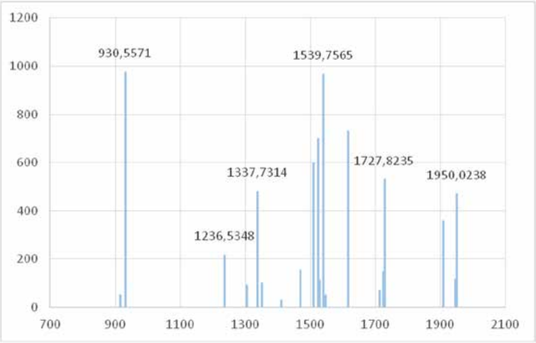
4.1. Protein identification using PMF and PFF approaches with MASCOT

Experiment description

1. 120 µg of protein from human cell line Human Embryo Kidney (HEK 293) was applied onto IPG strips, and 2-DE analysis was performed;
2. After detection and excision with a pipette tip from the gel, the protein spots were digested using trypsin after reduction and alkylation of cysteine residues using dithiothreitol (DTT) and iodoacetamide (IAA);
3. Tryptic peptides were lyophilized and resuspended in 10 µL of a 50% acetonitrile/0.1% formic acid solution. The samples were mixed (1:1) with a saturated matrix solution of a-cyano-4-hydroxycinnamic acid, and aliquots (0.5 µL) were spotted onto the MALDI sample target plate;
4. Peptide mass spectra were obtained on a MALDI-TOF/TOF mass spectrometer (4700 Proteomics Analyzer, Applied Biosystems, Foster City, CA, USA) in the positive ion reflectron mode;
5. For each sample spot, a data-dependent acquisition method was created to select the two most intense peaks, excluding those from the matrix, due to trypsin autolysis or acrylamide peaks, for subsequent MALDI-TOF/TOF MS/MS data acquisition.

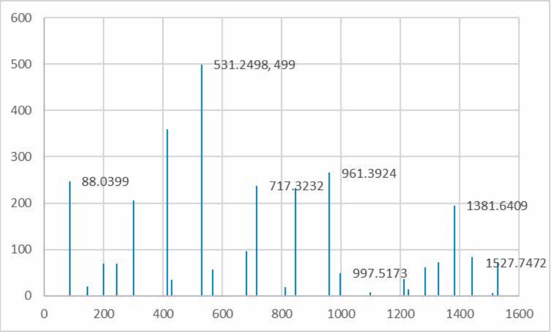
MS spectra of one spot:

m/z
916.4055
930.5571
1236.5348
1303.6696
1337.7314
1349.7409
1408.7048
1467.6749
1507.6632
1507.6686
1522.6397
1523.7924
1527.7532
1538.7780
1539.7565
1544.7386
1615.9085
1712.7815
1723.8684
1727.8235
1907.0442
1944.9352
1950.0238



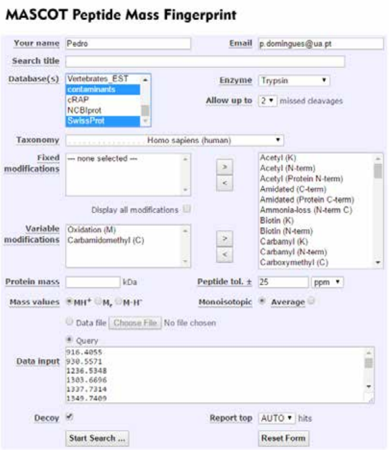
MS/MS of the ion at *m/z* 1527.7532

MS/MS	**AR**
88.0399	246
201.1301	70.3
302.1732	204.9
416.2143	358.6
531.2498	498.7
717.3232	237.9
846.3714	232.3
961.3924	266.1
1098.4522	6.6
1211.5352	35.6
1282.5703	61.6
1381.6409	195.4
1509.7398	5.9
1527.7472	72.3



a) Identification of a protein using the PMF approach

1. Go to the MASCOT online site: (http://www.matrixscience.com/search_form_select.html)
2. Select "perform a search" on the Peptide Mass Fingerprint option (http://www.matrixscience.com/cgi/search_form.pl? FORMVER=2&SEARCH=PMF)
3. Fill in the form. For data input, you can copy and paste the *m/z* of the peptides or choose the data file Mascot MS.txt
4. You should discuss the following:
   a. Why were Swiss Prot and contaminant databases selected?
   b. Why were two missed cleavages allowed?
   c. Why were two variable modifications allowed?
   d. Why is the tolerance 25 ppm?
   e. What is the decoy?



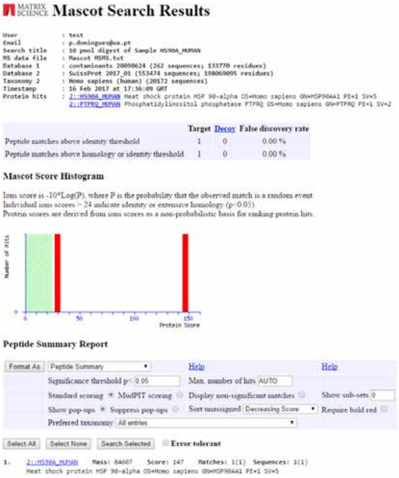
MASCOT Peptide Mass Fingerprint

Your name: Pedro
Email: p.domingues@ua.pt
Search title:
Database(s): Vertebrates_EST, contaminants, cRAP, NCBIprot, SwissProt
Enzyme: Trypsin ▾
Allow up to 2 ▾ missed cleavages
Taxonomy: Homo sapiens (human) ▾
Fixed modifications: --- none selected ---
Display all modifications □
Variable modifications: Oxidation (M), Carbamidomethyl (C)
Acetyl (K)
Acetyl (N-term)
Acetyl (Protein N-term)
Amidated (C-term)
Amidated (Protein C-term)
Ammonia-loss (N-term C)
Biotin (K)
Biotin (N-term)
Carbamyl (K)
Carbamyl (N-term)
Carboxymethyl (C)
Protein mass: kDa
Peptide tol. ±: 25 ppm ▾
Mass values: ● MH^+ ○ M_r ○ $M-H^-$
Monoisotopic ● Average ○
○ Data file Choose File No file chosen
● Query
Data input:

```
916.4055
930.5571
1236.5348
1303.6696
1337.7314
1349.7409
```

Decoy ☑
Report top AUTO ▾ hits

5. After the search, the following Mascot Search Results are obtained:



MATRIX SCIENCE Mascot Search Results

```
User          : test
Email         : p.domingues@ua.pt
Search title  : 10 pmol digest of Sample HS90A_HUMAN
MS data file  : Mascot MSMS.txt
Database 1    : contaminants 20090624 (262 sequences; 133770 residues)
Database 2    : SwissProt 2017_01 (553474 sequences; 198069095 residues)
Taxonomy 2    : Homo sapiens (human) (20172 sequences)
Timestamp     : 16 Feb 2017 at 17:36:09 GMT
Protein hits  : 2::HS90A_HUMAN Heat shock protein HSP 90-alpha OS=Homo sapiens GN=HSP90AA1 PE=1 SV=5
                2::PTPRQ_HUMAN Phosphatidylinositol phosphatase PTPRQ OS=Homo sapiens GN=PTPRQ PE=1 SV=2
```

	Target	Decoy	False discovery rate
Peptide matches above identity threshold	1	0	0.00 %
Peptide matches above homology or identity threshold	1	0	0.00 %

Mascot Score Histogram

Ions score is -10*Log(P), where P is the probability that the observed match is a random event.
Individual ions scores > 24 indicate identity or extensive homology (p<0.05).
Protein scores are derived from ions scores as a non-probabilistic basis for ranking protein hits.

Peptide Summary Report

Format As: Peptide Summary ▾ Help Help
Significance threshold p< 0.05
Max. number of hits AUTO
Standard scoring ● MudPIT scoring ○
Display non-significant matches □
Show sub-sets 0
Show pop-ups ● Suppress pop-ups ○
Sort unassigned Decreasing Score ▾
Require bold red □
Preferred taxonomy All entries ▾

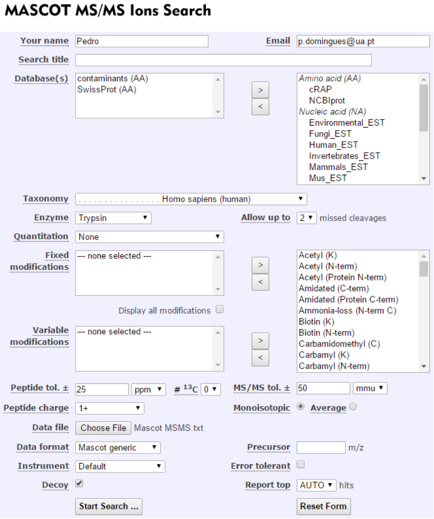
□ Error tolerant

```
1.   2::HS90A_HUMAN   Mass: 84607   Score: 147   Matches: 1(1)  Sequences: 1(1)
     Heat shock protein HSP 90-alpha OS=Homo sapiens GN=HSP90AA1 PE=1 SV=5
```

6. Look closely at the results. What do they mean? Select the protein with the highest score and look at the Protein View: HS90A_HUMAN data.

b) Identification of a protein using the PFF approach

1. Select "perform a search" on the Peptide Mass Fingerprint option (http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS)
2. Fill in the form. For data input, you can copy choose the data file Mascot MSMS.txt. Press "start search…"



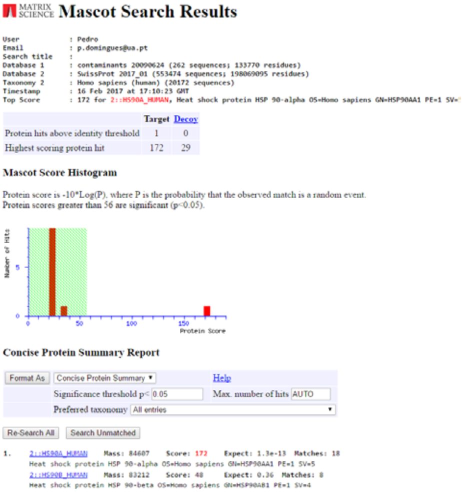
MASCOT MS/MS Ions Search

Your name: Pedro
Email: p.domingues@ua.pt
Search title:
Database(s): contaminants (AA), SwissProt (AA)

Taxonomy: . . . . Homo sapiens (human)
Enzyme: Trypsin
Allow up to 2 missed cleavages
Quantitation: None
Fixed modifications: --- none selected ---
Display all modifications ☐
Variable modifications: --- none selected ---

Peptide tol. ± 25 ppm # ^{13}C 0
MS/MS tol. ± 50 mmu
Peptide charge: 1+
Monoisotopic ● Average ○
Data file: Choose File Mascot MSMS.txt
Data format: Mascot generic
Precursor: m/z
Instrument: Default
Error tolerant ☐
Decoy ☑
Report top AUTO hits

3. Look closely at the Mascot Search Results. What do they mean? Select the protein with the highest score and look at the Protein View: HS90A_HUMAN data



MATRIX SCIENCE Mascot Search Results

```
User          : Pedro
Email         : p.domingues@ua.pt
Search title  :
Database 1    : contaminants 20090624 (262 sequences; 133770 residues)
Database 2    : SwissProt 2017_01 (553474 sequences; 198069095 residues)
Taxonomy 2    : Homo sapiens (human) (20172 sequences)
Timestamp     : 16 Feb 2017 at 17:10:23 GMT
Top Score     : 172 for 2::HS90A_HUMAN, Heat shock protein HSP 90-alpha OS=Homo sapiens GN=HSP90AA1 PE=1 SV=
```

	Target	Decoy
Protein hits above identity threshold	1	0
Highest scoring protein hit	172	29

Mascot Score Histogram

Protein score is -10*Log(P), where P is the probability that the observed match is a random event.
Protein scores greater than 56 are significant (p<0.05).

Concise Protein Summary Report

Concise Protein Summary
Significance threshold p< 0.05
Max. number of hits AUTO
Preferred taxonomy All entries

```
1.    2::HS90A_HUMAN    Mass: 84607    Score: 172    Expect: 1.3e-13  Matches: 18
      Heat shock protein HSP 90-alpha OS=Homo sapiens GN=HSP90AA1 PE=1 SV=5
      2::HS90B_HUMAN    Mass: 83212    Score: 48     Expect: 0.36  Matches: 8
      Heat shock protein HSP 90-beta OS=Homo sapiens GN=HSP90AB1 PE=1 SV=4
```

4. Select number 1 in the query and look at the Peptide View. Look at the peptide fragmentation pattern. Why are b ions more intense than y ions?

4.2. Protein identification from an LC-MS experiment

For this tutorial, you will need to download the following files:

a. Ask your instructor for the raw data of the experiment (1,2GB) and the transformed file (MGF, 100MB)

b. ProteoWizard - This is an open-source interface for converting MS format software. You only need this if you download the RAW file. Additionally, SearchGUI (see below) already includes this application. http://proteowizard.sourceforge.net/

c. mMass (optional) - This is an open-source mass spectrometry interface that can be used for spectra visualisation. It also has several other tools including proteomics tools http://www.mmass.org/

d. Dbtoolkit - This is an open source FASTA data editor for manipulation of fasta sequence database https://github.com/compomics/dbtoolkit

e. SearchGUI - this is an open-source interface for configuring and running proteomics identification search engines. http://compomics.github.io/projects/searchgui.html

f. Peptide shaker - this is a search engine independent platform for interpretation of proteomics identification results. http://compomics.github.io/projects/peptide-shaker.html

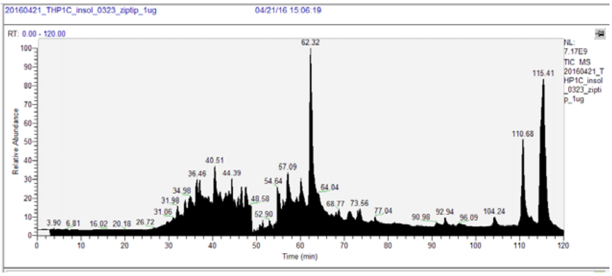
Also, it is very important that you carefully read the peptide shaker full tutorial at https://compomics.com/bioinformatics-for-proteomics/

Experiment description

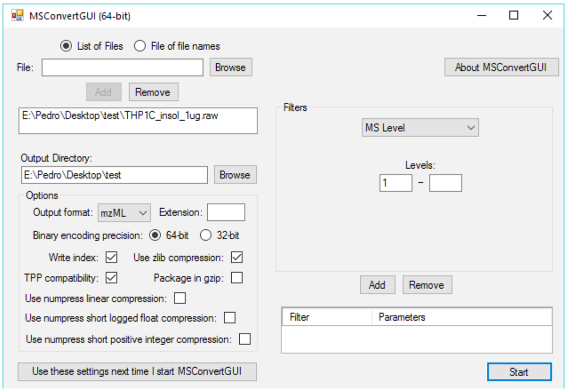
1. For rapid identification of the most abundant intracellular proteins of THP 1 Cell Line (human monocytic cell line), cells were scraped off from the plate using 500 µL PBS/well and sonicated at 53 kHz at 37°C for 30 min. Proteins were precipitated by adding two volumes of acetone to the cell lysate suspensions. Protein suspensions were incubated at -20°C and then centrifuged. Dried protein pellets were resuspended in 30 µL of sample storage buffer. The quantities of the extracted proteins were measured using the Lowry method.
2. In-solution tryptic digestion was performed with trypsin. The protein's cysteine residues were reduced with DTT and alkylated with iodoacetamide. Tryptic peptides were lyophilized and resuspended in 5% ACN/0.1% FA solution.
3. 250 ng of the sample protein extract were analyzed with a QExactive Orbitrap that was coupled to an Ultimate 3000 nano HPLC system. Trap (5 mm × 300 µm I.D.) and an analytical (150 mm × 75 µm I.D.) C18 columns were used.
4. The mass spectrometer was operated in the data-dependent acquisition mode. The ten most intense peaks were subjected to HCD.

4.2.1. Transformation of data to mzML with MSConverter in ProteoWizard

1. Use MSConverter in ProteoWizard to convert the HPC-MS data acquired in the Orbitrap (RAW data file) to a format that can be read by the SearchGUI (MGF data file).
2. If you wish to look at the data in the mMass application, although it can read the mgf file format that is used in SearchGUI, it is best if you convert the RAW (or mgf) data to mzML format.



RAW data chromatogram



● List of Files ○ File of file names

File:

E:\Pedro\Desktop\test\THP1C_insol_1ug.raw

Output Directory: E:\Pedro\Desktop\test

Options

Output format: mzML

Extension:

Binary encoding precision: ● 64-bit ○ 32-bit

Write index: ☑ Use zlib compression: ☑

TPP compatibility: ☑ Package in gzip: □

Use numpress linear compression: □

Use numpress short logged float compression: □

Use numpress short positive integer compression: □

Use these settings next time I start MSConvertGUI

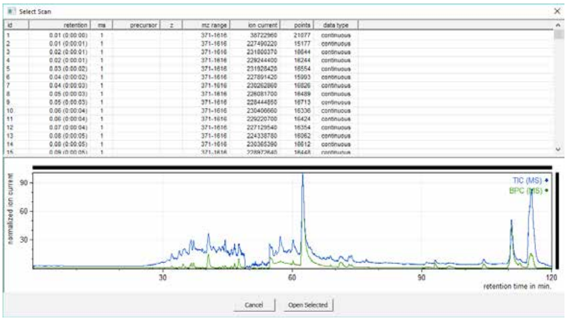
Filters

MS Level

Levels: 1 –

Filter	Parameters

MSConverter application

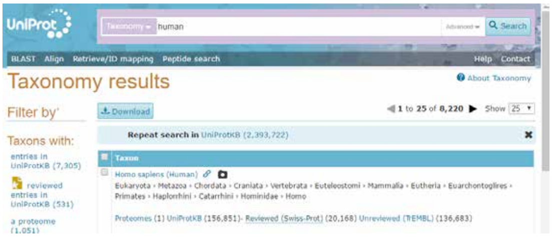


id	retention	ms	precursor	z	mz range	ion current	points	data type
1	0.01 (0:00:00)	1			371-1616	38722960	21077	continuous
2	0.01 (0:00:01)	1			371-1616	227490220	15177	continuous
3	0.02 (0:00:01)	1			371-1616	231800370	16644	continuous
4	0.02 (0:00:01)	1			371-1616	229244400	16244	continuous
5	0.03 (0:00:02)	1			371-1616	231928420	16554	continuous
6	0.04 (0:00:02)	1			371-1616	227891420	15993	continuous
7	0.04 (0:00:03)	1			371-1616	230262860	16826	continuous
8	0.05 (0:00:03)	1			371-1616	226081700	16489	continuous
9	0.05 (0:00:03)	1			371-1616	228444850	16713	continuous
10	0.06 (0:00:04)	1			371-1616	230406660	16336	continuous
11	0.06 (0:00:04)	1			371-1616	229220700	16424	continuous
12	0.07 (0:00:04)	1			371-1616	227129540	16354	continuous
13	0.08 (0:00:05)	1			371-1616	224338780	16062	continuous
14	0.08 (0:00:05)	1			371-1616	230365390	16612	continuous
15	0.09 (0:00:05)	1			371-1616	228972640	16448	continuous

Data observed in the mmass application

4.2.2. Generation of FASTA database from Uniprot (SwissProt)

1. Now, you should download the FASTA file of the proteome of Homo sapiens. This should be done on the taxonomy page of Uniprot site and search Homo Sapiens.



Taxonomy results

Filter by

Taxons with:
entries in UniProtKB (7,305)
reviewed entries in UniProtKB (531)
a proteome (1,051)

1 to 25 of 8,220

Repeat search in UniProtKB (2,393,722)

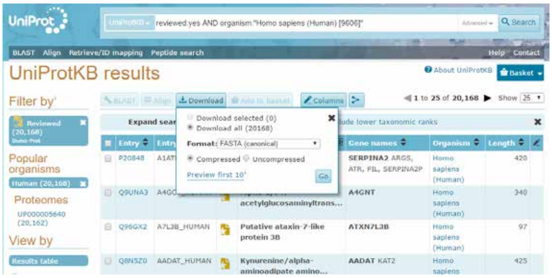
Taxon

Homo sapiens (Human)

Eukaryota › Metazoa › Chordata › Craniata › Vertebrata › Euteleostomi › Mammalia › Eutheria › Euarchontoglires › Primates › Haplorrhini › Catarrhini › Hominidae › Homo

Proteomes (1) UniProtKB (156,851)- Reviewed (Swiss-Prot) (20,168) Unreviewed (TrEMBL) (136,683)

2. Now you should download the Reviewed (Swiss-Prot) FASTA file.



reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

UniProtKB results

Filter by
Reviewed (20,168) Swiss-Prot
Popular organisms
Human (20,168)
Proteomes
UP000005640 (20,162)
View by

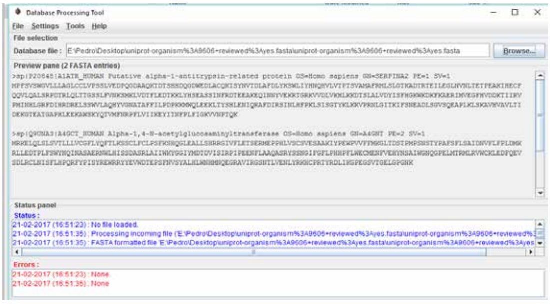
1 to 25 of 20,168

Expand sear... ...lude lower taxonomic ranks

○ Download selected (0)
● Download all (20168)
Format: FASTA (canonical)
● Compressed ○ Uncompressed
Preview first 10

Entry	Entry name		Protein names	Gene names	Organism	Length
P20848	A1AT...			SERPINA2 ARGS, ATR, PIL, SERPINA2P	Homo sapiens (Human)	420
Q9UNA3	A4GC...		...acetylglucosaminyltrans...	A4GNT	Homo sapiens (Human)	340
Q96GX2	A7L3B_HUMAN		Putative ataxin-7-like protein 3B	ATXN7L3B	Homo sapiens (Human)	97
Q8N5Z0	AADAT_HUMAN		Kynurenine/alpha-aminoadipate amino...	AADAT KAT2	Homo sapiens	425

3. You can look at and edit the FASTA file information by using the Dbtoolkit dataBase Processing Tool.



File selection

Database file : E:\Pedro\Desktop\uniprot-organism%3A9606+reviewed%3Ayes.fasta\uniprot-organism%3A9606+reviewed%3Ayes.fasta

Preview pane (2 FASTA entries)

```
>sp|P20848|A1ATR_HUMAN Putative alpha-1-antitrypsin-related protein OS=Homo sapiens GN=SERPINA2 PE=1 SV=1
MPFSVSWGVLLLAGLCCLVPSSLVEDPQGDAAQKTDTSHHDQGDWEDLACQKISYNVTDLAFDLYKSWLIYHNQHVLVTPTSVAMAFRMLSLGTKADTRTEILEGLNVNLTETPEAKIHECF
QQVLQALSRPDTRLQLTTGSSLFVNKSMKLVDTFLEDTKKLYHSEASSINFRDTEEAKEQINNYVEKRTGRKVVDLVKHLKKDTSLALVDYISFHGKWKDKFKAERIMVEGFHVDDKTIIRV
PMINHLGRFDIHRDRELSSWVLAQHYVGNATAFFILPDPKKMWQLEEKLTYSHLENIQRAFDIRSINLHFPKLSISGTYKLKKVPRNLGITKIFSNEADLSGVSQEAPLKLSKAVHVAVLTI
DEKGTEATGAPHLEEKAWSKYQTVMFNRPFLVIIKEYITNFPLFIGKVVNPTQK

>sp|Q9UNA3|A4GCT_HUMAN Alpha-1,4-N-acetylglucosaminyltransferase OS=Homo sapiens GN=A4GNT PE=2 SV=1
MRRELQLSLSVTLLLVCGFLYQFTLKSSCLFCLPSFKSNQGLEALLSHRRGIVFLETSERMEPPHLVSCSVESAAKIYPEWPVVFFMKGLTDSTPMPSNSTYPAFSFLSAIDNVFLFPLDMK
RLLEDTPLFSWYNQINASAERNWLHISSDASRLAIIWKYGGIYMDTDVISIRPIPEENFLAAQASRYSSNGIFGFLPHHPFLWECMENFVEHYNSAIWGNQGPELMTRMLRVWCKLEDFQEV
SDLRCLNISFLHPQRFYPISYREWRRYYEVWDTEPSFNVSYALHLWNHMNQEGRAVIRGSNTLVENLYRKHCPRTYRDLIKGPEGSVTGELGPGNK
```

Status panel

Status :

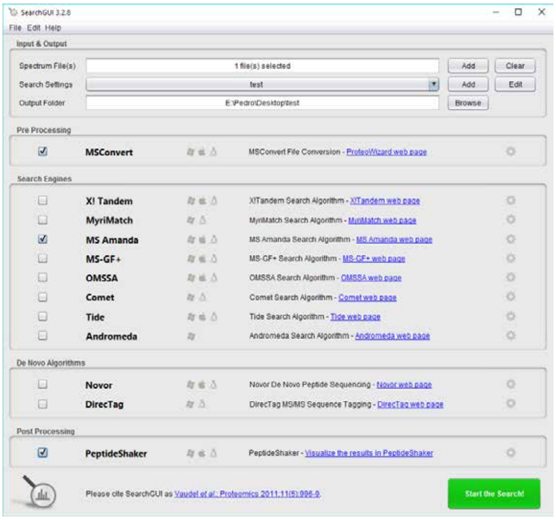
```
21-02-2017 (16:51:23) : No file loaded.
21-02-2017 (16:51:35) : Processing incoming file ('E:\Pedro\Desktop\uniprot-organism%3A9606+reviewed%3Ayes.fasta\uniprot-organism%3A9606+reviewed%3
21-02-2017 (16:51:35) : FASTA formatted file 'E:\Pedro\Desktop\uniprot-organism%3A9606+reviewed%3Ayes.fasta\uniprot-organism%3A9606+reviewed%3Ayes
```

Errors :

```
21-02-2017 (16:51:23) : None
21-02-2017 (16:51:35) : None
```

4.2.3. Search Engines: SearchGUI

1. Open the searchGUI. In the above example, the selected file was the raw file, so MS convert was also selected.
2. In the search settings select "edit" and fill in the form as shown. Variable modifications were chosen (why?).
3. You can also configure the peptide shaker to open the results file, as shown below.
4. The searchGUI will ask if you want to create a concatenated_target_decoy fasta file. Say yes (Why is this important?)



SearchGUI 3.2.8

File Edit Help

Input & Output

Spectrum File(s)	1 file(s) selected	Add	Clear
Search Settings	test	Add	Edit
Output Folder	E:\Pedro\Desktop\test	Browse	

Pre Processing

☑	MSConvert	MSConvert File Conversion - ProteoWizard web page

Search Engines

□	X! Tandem	X!Tandem Search Algorithm - X!Tandem web page
□	MyriMatch	MyriMatch Search Algorithm - MyriMatch web page
☑	MS Amanda	MS Amanda Search Algorithm - MS Amanda web page
□	MS-GF+	MS-GF+ Search Algorithm - MS-GF+ web page
□	OMSSA	OMSSA Search Algorithm - OMSSA web page
□	Comet	Comet Search Algorithm - Comet web page
□	Tide	Tide Search Algorithm - Tide web page
□	Andromeda	Andromeda Search Algorithm - Andromeda web page

De Novo Algorithms

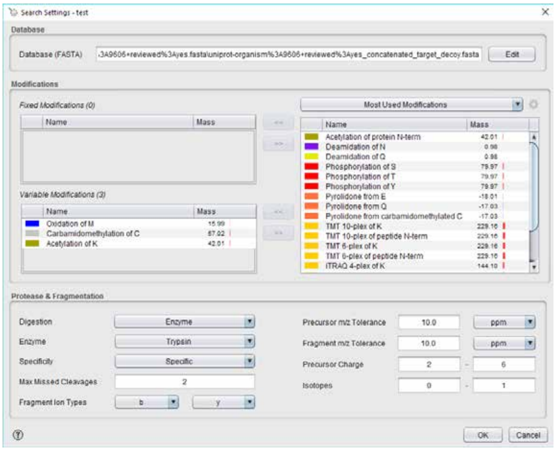
□	Novor	Novor De Novo Peptide Sequencing - Novor web page
□	DirecTag	DirecTag MS/MS Sequence Tagging - DirecTag web page

Post Processing

☑	PeptideShaker	PeptideShaker - Visualize the results in PeptideShaker

Please cite SearchGUI as Vaudel et al.: Proteomics 2011;11(5):996-9.

5. Select the search options and the FASTA database.



Search Settings - test

Database

Database (FASTA) .3A9606+reviewed%3Ayes.fasta\uniprot-organism%3A9606+reviewed%3Ayes_concatenated_target_decoy.fasta Edit

Modifications

Fixed Modifications (0)

	Name	Mass

Variable Modifications (3)

	Name	Mass
🟦	Oxidation of M	15.99
⬜	Carbamidomethylation of C	57.02
🟩	Acetylation of K	42.01

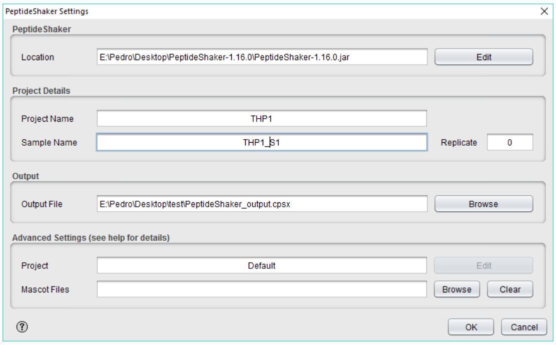
Most Used Modifications

	Name	Mass
🟩	Acetylation of protein N-term	42.01
🟪	Deamidation of N	0.98
🟨	Deamidation of Q	0.98
🟥	Phosphorylation of S	79.97
🟥	Phosphorylation of T	79.97
🟥	Phosphorylation of Y	79.97
🟧	Pyrolidone from E	-18.01
🟧	Pyrolidone from Q	-17.03
🟧	Pyrolidone from carbamidomethylated C	-17.03
🟨	TMT 10-plex of K	229.16
🟨	TMT 10-plex of peptide N-term	229.16
🟨	TMT 6-plex of K	229.16
🟨	TMT 6-plex of peptide N-term	229.16
🟨	iTRAQ 4-plex of K	144.10

Protease & Fragmentation

Digestion	Enzyme	Precursor m/z Tolerance	10.0	ppm
Enzyme	Trypsin	Fragment m/z Tolerance	10.0	ppm
Specificity	Specific	Precursor Charge	2	6
Max Missed Cleavages	2	Isotopes	0	1
Fragment Ion Types	b y			

6. Select the peptideShaker options.



PeptideShaker Settings

PeptideShaker

Location E:\Pedro\Desktop\PeptideShaker-1.16.0\PeptideShaker-1.16.0.jar Edit

Project Details

Project Name THP1

Sample Name THP1_S1 Replicate 0

Output

Output File E:\Pedro\Desktop\test\PeptideShaker_output.cpsx Browse

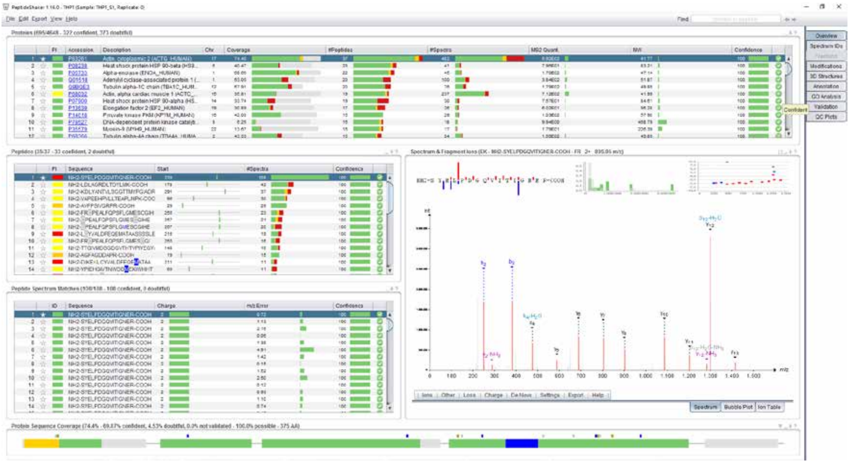
Advanced Settings (see help for details)

Project Default Edit

Mascot Files Browse Clear

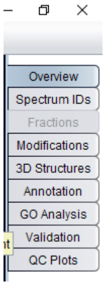
4.2.4. Generation and evaluation of results: PeptideShaker for peptide and protein visualization, and validation. PTM analysis

1. After the searchGUI has performed the search (~3 minutes with an Intel I7-6700K with 16MB of RAM), it will open the results in the peptideshaker platform.
2. Here you will be able to see that ~700 proteins were identified, although 373 have been classified as doubtful (why?).
3. Also, you will be able to see information about the peptides identified for each protein and the mass spectra with the annotated fragmentation pattern.

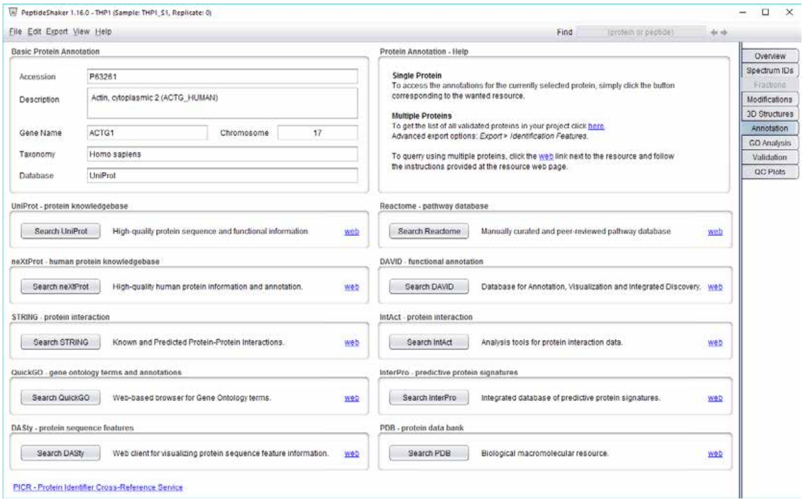

PeptideShaker 1.16.0 - THP1 (Sample: THP1_S1, Replicate: 0)
File Edit Export View Help
Proteins (695/4648 - 322 confident, 373 doubtful)
PI Accession Description Chr Coverage #Peptides #Spectra MS2 Quant. MW Confidence
Peptides (35/37 - 33 confident, 2 doubtful)
PI Sequence Start #Spectra Confidence
Peptide Spectrum Matches
ID Sequence Charge m/z Error Confidence
Spectrum & Fragment Ions
Ions Other Loss Charge DeNovo Settings Export Help
Spectrum Bubble Plot Ion Table
Protein Sequence Coverage
Overview
Spectrum IDs
Fractions
Modifications
3D Structures
Annotation
GO Analysis
Validation
QC Plots

4.2.5. Data analysis: protein information, pathway analysis, and gene ontology

1. Explore the advanced data analysis options of the peptideshaker by opening the modifications tab, and the GO analysis tab.



2. In the annotation tab, you will be able to annotate information for each protein.



Basic Protein Annotation

Accession: P63261

Description: Actin, cytoplasmic 2 (ACTG_HUMAN)

Gene Name: ACTG1 Chromosome: 17

Taxonomy: Homo sapiens

Database: UniProt

Protein Annotation - Help

Single Protein
To access the annotations for the currently selected protein, simply click the button corresponding to the wanted resource.

Multiple Proteins
To get the list of all validated proteins in your project click here.
Advanced export options: *Export > Identification Features.*

To query using multiple proteins, click the web link next to the resource and follow the instructions provided at the resource web page.

UniProt - protein knowledgebase: High-quality protein sequence and functional information

Reactome - pathway database: Manually curated and peer-reviewed pathway database

neXtProt - human protein knowledgebase: High-quality human protein information and annotation.

DAVID - functional annotation: Database for Annotation, Visualization and Integrated Discovery.

STRING - protein interaction: Known and Predicted Protein-Protein Interactions.

IntAct - protein interaction: Analysis tools for protein interaction data.

QuickGO - gene ontology terms and annotations: Web-based browser for Gene Ontology terms.

InterPro - predictive protein signatures: Integrated database of predictive protein signatures.

DASty - protein sequence features: Web client for visualizing protein sequence feature information.

PDB - protein data bank: Biological macromolecular resource.

PICR - Protein Identifier Cross-Reference Service

3. However, if you wish to annotate multiple proteins, you will need to export your results (default protein report) and click the web link next to the resource and follow the instructions provided at the resource web page.

Acknowledgments:

Thanks are due to Fundação para a Ciência e a Tecnologia (FCT, Portugal), the European Union, the Quadro de Referência Estratégica Nacional (QREN), the Programa Operacional Potencial Humano (POPH), the Fundo Europeu de Desenvolvimento Regional (FEDER) and Programa Operacional Factores de Competitividade (COMPETE) for funding the QOPNA (Química Orgânica de Produtos Naturais e Agroalimentares; UID/QUI/00062/2013) research unit and to RNEM (LISBOA-01-0145-FEDER-022125) for the Portuguese Mass Spectrometry Network.