A Modular Pipeline for Metabolomic Data Preprocessing

Universitat Oberta de Catalunya

Eduard Pérez Méndez

Statistical Bioinformatics and Machine Learning

Master's degree in Bioinformatics and Biostatistics

Name of the tutor:
Alexandre Sánchez Pla

Name of the SRP:
Carles Ventura Royo

April 15, 2024

UNIVERSITAT DE BARCELONA

Final Work Card

Title of the work:	A Modular Pipeline for Metabolomic Data Preprocessing
Name of the author:	Eduard Pérez Méndez
Name of the tutor:	Alexandre Sánchez Pla
Name of the SRP:	Carles Ventura Royo
Date of delivery:	April 15, 2024
Studies or Prgram:	Master's degree in Bioinformatics and Biostatistics
Area or the Final Work:	Statistical Bioinformatics and Machine Learning
Language of the work:	English
Keywords:	targeted metabolomics, preprocessing, pipeline

Abstract

A maximum of 250 words, detailing the purpose, contexto of application, methodology, results and conclusiones of the work.

"BIG MOTIVATIONAL QUOTE "

AUTHOR NAME

Contents

List of Figures

List of Tables

1. Introduction

1.1 General description

Metabolomics, a powerful and evolving field within the realm of systems biology, plays a pivotal role in unraveling the intricate web of biochemical processes occurring within living organisms. As we delve into the molecular intricacies of biological systems, the generation of vast and complex datasets poses a significant challenge. Challenges in standardizing nutritional metabolomics include experimental design, sample preparation, and data analysis, which impact result validity and reproducibility. Efforts by the international community aim to establish standard procedures and infrastructure for advancing nutritional metabolomics research. This master thesis project aims for the creation of a modular pipeline designed to streamline the processing of targeted metabolomics data to a usable and meaningful dataset for further analysis and biological interpretation.

1.2 Context and justification

Metabolomics is a rapidly evolving field within biology that focuses on the comprehensive study of the metabolite composition of cell types, tissues, organs, or organisms [1–3]. It aims to measure, identify and (semi-)quantify those metabolites. Metabolites are chemical compounds that undergo analysis through conventional chemical assessment methods like mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectrometry. MS approaches are commonly integrated with gas chromatography (GS) and liquid chromatography (LC), leading to the development of two advanced techniques known as gas chromatography-MS (GC-MS) and liquid chromatography-MS (LC-MS). All of these analytical platforms and methodologies generate large amounts of high-dimensional and complex experimental raw data.

However, the statistical analysis of metabolomics data presents significant challenges, attributable not only to the inherent complexity of metabolomics as a research discipline but also to the intricate nature of the data itself. Notwithstanding that numerous studies have explored various methodologies for metabolomic data management, the field still lacks an accepted standard for preprocessing and pretreatment of such data.

One of the obstacles the field encounters is the lack of well defined terminology, as the terms “data preprocessing” and “data pretreatment” have not been used consistently in metabolomics literature [4].

The objectives of data preprocessing/pretreatment encompass two primary aims: firstly, to rectify or mitigate instrumental artifacts and extraneous biological variance, thereby amplifying the signal-to-noise ratio (SNR); and secondly, to effectively transform the data into interpretable spectral profiles through processes such as centering, scaling, and dimensionality reduction [4,5]. The choice of preprocessing and pretreatment methods can signifi-

cantly impact the downstream analysis and interpretation of metabolomic data [6] so the steps should be carefully selected based on the specific characteristics of the data and the research.

By establishing a standardized approach to preprocessing and pretreatment of metabolomic data, the field can improve the quality, comparability, and reproducibility of metabolomic studies. This would facilitate data integration, enable the development of robust statistical models, and enhance our understanding of the complex metabolic processes underlying health and disease.

1.2.1 Preprocessing of data

Given the inherent dissimilarities in data acquisition techniques, unique preprocessing procedures are imperative before embarking on statistical analyses in metabolomics investigations. NMR spectra, for instance, often exhibit signal shifts along the axis due to factors like pH fluctuations [7]. Thus, meticulous preprocessing is indispensable to ensure robust statistical analyses and facilitate inter-spectral signal comparisons. This involves techniques such as binning, peak fitting with spectral databases, and exclusion of unstable or non-informative spectral regions (e.g., water peaks) [3,4,8]. By refining the dataset to a subset of relevant metabolites, statistical methods can effectively discern variations in signal intensity among sample groups [9].

The preprocessing workflows diverge between MS-based and NMR-based metabolomic analyses. In MS-based profiling, data are presented as three-dimensional (3D) tables, in contrast to the two-dimensional (2D) tables derived from GC-MS data preprocessing [4,8]. GC-MS preprocessing entails deconvolution and peak integration to generate intensity profiles for each sample feature corresponding to RT/m/z pairs. Notably, metabolite identification strategies differ between GC-MS and LC-MS methodologies. While GC-MS relies on reproducible mass spectra and extensive databases for metabolite identification based on characteristic fragment ions, MS-based methods prioritize automation, accuracy, peak identification, integration, and annotation [10,11].

While the primary objective of preprocessing is to render data comparable across samples despite instrumental discrepancies, the strategies employed in MS-based methodologies differ from those in NMR-based approaches. Moreover, variations exist between preprocessing methodologies utilized in GC-MS and LC-MS metabolomic analyses, underscoring the intricate nature of metabolomics data preprocessing.

MS-based data preprocessing

MS-based analysis involves the measurement of mass-to-charge ratios (m/z). When combined with either LC or GC, the resulting raw GC/LC-MS data encompass three measured variables: m/z, chromatographic retention time (RT), and intensity count, thereby constituting a three-dimensional (3D) data structure. To streamline the data and eliminate spectral noise and irrelevant biological variability, a two-dimensional (2D) features table is generated through peak picking. This table encompasses all quantified metabolic features from the analyzed samples, with rows corresponding to samples and columns representing variables such as peak areas or intensities, characterized by m/z and retention time in minutes

or scan number (m/z-RT pairs). The preprocessing of MS data involves several steps: 1) denoising and baseline correction; 2) alignment across all samples; 3) peak picking; 4) merging the peaks; and 5) creating a data matrix [3,4,10,12–17].

NMR-based data preprocessing

Similar to MS-based analysis, NMR-based analysis generates a 2D structure of feature data matrix with the samples in the rows and the spectral data points in the columns. Also similar to MS-based analysis, the NMR-based analysis (e.g., 1H NMR analysis) requires data preprocessing to mitigate non-biologically relevant effects. The following data preprocessing steps could be performed: 1) baseline correction; 2) peak binning; 3) peak alignment; 4) quality control; 5) create a data matrix [4,5,15–20]. Preprocessing by either MS or NMR constructs a data matrix containing the relative abundances of a set of mass spectra for a group of samples or subjects under different conditions. The metabolomics data matrix are typically constructed in such a way that each row of the data matrix represents a subject and each column represents the mass spectra (metabolite intensities or metabolite relative abundances, peak or peak intensities).

1.2.2 Pretreatment of Data

Handling Missing Values

Within datasets, missing values or zeros can arise due to a variety of factors, both biological and technical in nature. Categorizations by Sun Xia delineate these zeros into four distinct categories: 1) Structural zeros, 2) Sampling zeros, 3) Values below the limit of detection (LOD), and 4) Zeros derived from negative values that are automatically transformed.

1. **Structural zeros** pertain to peaks absent from a sample or chromatogram due to genuine biological absence rather than technical errors. For instance, if a compound is not present in a biological sample, the corresponding peak for that compound is deemed a structural zero.

2. **Sampling zeros** refer to peaks present in samples but missed during peak picking.

3. **Values below LOD** represent intensities or abundances falling below the detection limit of the mass spectrometer.

4. **Negative value zeros** result from negative intensity or abundance values, considered spectral artifacts or noise, and subsequently transformed to zero.

Identifying the origins of these zeros poses a challenge, and their prevalence presents a significant obstacle for statistical analyses [4,21]. Hence, practical approaches for managing zeros include:

1. **Filtering** based on a threshold, such as the 80% rule.

2. **Imputation** techniques, which can involve substituting zeros with the mean, minimum (or half of the minimum) of non-missing values, or simply zero.

3. Utilizing **missing data estimation algorithms** to employ various methods for handling missing values.

1.3 State of the art

Punto de partida del trabajo (¿Cuál es la necesidad a cubrir? ¿Por qué es un tema relevante? ¿Cómo se resuelve el problema de momento?) y aportación realizada (¿Qué resultado se quiere obtener?).

Es importante tener en cuenta que el trabajo final tiene que ser comprensible para cualquier persona que conozca el área de conocimiento, pero no tiene porque ser experta en el tema del que versa el trabajo.

2. Objectives

Listado de los objetivos del trabajo.

2.1 Main objective

2.2 Specific objectives

3. Sustainable development goals

Esta sección tendría que identificar los impactos positivos y/o negativos del trabajo final en las tres dimensiones de la competencia transversal UOC "Compromiso ético y global".

La Guía transversal sobre la Competencia Ética y Global os ayudará a redactar estos apartados.

4. Approach and methodology

Mención de cuáles son las posibles estrategias para llevar a cabo el trabajo y cuál es la estrategia elegida (desarrollar un producto nuevo, adaptar un producto existente...). Hay que incluir una valoración de por qué esta es la estrategia más apropiada para conseguir los objetivos.

4.1 Methodology

4.2 Planning and calendar

Descripción de los recursos and necesarios para hacer el trabajo, las tareas a realizar y una planificación temporal de cada tarea mediante un diagrama de Gantt o similar. Esta planificación tendría que marcar cuáles son los hitos parciales de cada una de las PEC.

Identificación de los posibles riesgos que pueden hacer que esta planificación no se cumpla y descripción de los planes de mitigación o alternativos en caso de que estos riesgos sean un problema.

4.3 Risk analysis

4.4 Final products

No hay que entrar en detalle: la descripción detallada se hará en el resto de capítulos.

4.5 Chapters structure

Breve explicación de los contenidos de cada capítulo y su relación con el proyecto global.

5. Materials and methods

En estos apartados, hay que describir:

- Los aspectos más relevante del diseño y desarrollo del trabajo.
- La metodología elegida para hacer este desarrollo, describiendo las alternativas posibles, las decisiones tomadas, y los criterios utilizados para tomar estas decisiones.
- Los productos obtenidos.

La estructuración de los capítulos puede variar según el tipo de trabajo.
En caso de que se proceda, se incluirá un apartado de "Valoración económica del trabajo". Este apartado indicará los gastos asociados al desarrollo y mantenimiento del trabajo, así como los beneficios económicos obtenidos y un análisis final sobre la viabilidad del producto.

6. Results

Detallad en este apartado los resultados obtenidos utilizando la metodología descrita en el apartado anterior.

Las figuras tienen que estar explicadas y citadas en el texto, como la 6.1, en la cual se muestra el error en función de la distancia, en unidades arbitrarias. En todas las gráficas tiene que haber el título de los ejes.

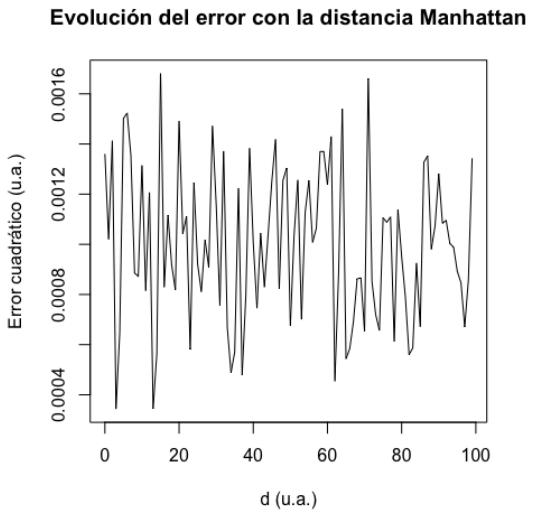

Figure 6.1: Error en función de la distancia en unidades arbitrarias.

7. Conclusion and future vision

Este capítulo tiene que incluir:

- Una descripción de las conclusiones del trabajo:
  - Una vez se han obtenido los resultados, ¿qué conclusiones se extraen?
  - ¿Estos resultados son los esperados? ¿O han sido sorprendentes? ¿Por qué?
- Una reflexión crítica sobre el logro de los objetivos planteados inicialmente:
  - ¿Hemos logrado todos los objetivos? Si la respuesta es negativa, ¿por qué motivo?
- Un análisis crítico del seguimiento de la planificación y metodología a lo largo del producto:
  - ¿Se ha seguido la planificación?
  - ¿La metodología prevista ha sido suficientemente adecuada?
  - ¿Ha habido que introducir cambios para garantizar el éxito del trabajo? ¿Por qué?
- De los impactos previstos en 3, ético-sociales, de sostenibilidad y de diversidad, evaluar/mencionar si se han mitigado (si eran negativos) o si se han conseguido (si eran positivos).
- Si han aparecido impactos no previstos a 3, evaluar/mencionar cómo se han mitigado (si eran negativos) o que han aportado (si eran positivos).
- Las líneas de trabajo futuro que no se han podido explorar en este trabajo y han quedado pendientes.

8. Glossary

Definición de los [4] términos y acrónimos más relevantes utilizados dentro de la Memoria.

Bibliography

1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: The Apogee of the Omics Trilogy. *Nature Reviews Molecular Cell Biology* **13,** 263–269. ISSN: 1471-0080 (Apr. 2012).
2. Zhang, A., Sun, H. & Wang, X. Serum Metabolomics as a Novel Diagnostic Approach for Disease: A Systematic Review. *Analytical and Bioanalytical Chemistry* **404,** 1239–1245. ISSN: 1618-2650 (Sept. 1, 2012).
3. Chen, Y., Li, E.-M. & Xu, L.-Y. Guide to Metabolomics Analysis: A Bioinformatics Workflow. *Metabolites* **12,** 357. ISSN: 2218-1989 (4 Apr. 2022).
4. Ulaszewska, M. M. *et al.* Nutrimetabolomics: An Integrative Action for Metabolomic Analyses in Human Nutritional Studies. *Molecular Nutrition & Food Research* **63,** 1800384. ISSN: 1613-4133 (2019).

Lista numerada de las referencias bibliográficas utilizadas dentro de la memoria. En cada lugar donde se utilice una referencia dentro del texto, se tiene que indicar citando el número de la referencia, por ejemplo: [7].

Es muy importante incluir todas las referencias utilizadas y citarlas apropiadamente, es decir, incluyendo toda la información necesaria para identificar la referencia. La información mínima que se tiene que incluir según el tipo de referencia es:

- Libro: Autores, Título, Edición (si procede) Editorial, Ciudad, Año.
- Artículo de revista: Autores, Título, Nombre de la Revista, Número de Página inicial y final, Número de la revista / Volumen, Año.
- Web: URL y fecha en que se ha visitado.

Información de como citar documentos: http://biblioteca.uoc.edu/es/recursos/citacion-bibliografic

Listado de apartados que son demasiado extensos para incluir dentro de la memoria y tienen un carácter autocontenido (por ejemplo, manuales de usuario, manuales de instalación, etc.)

Dependiendo del tipo de trabajo, es posible que no haya que añadir algún anexo.