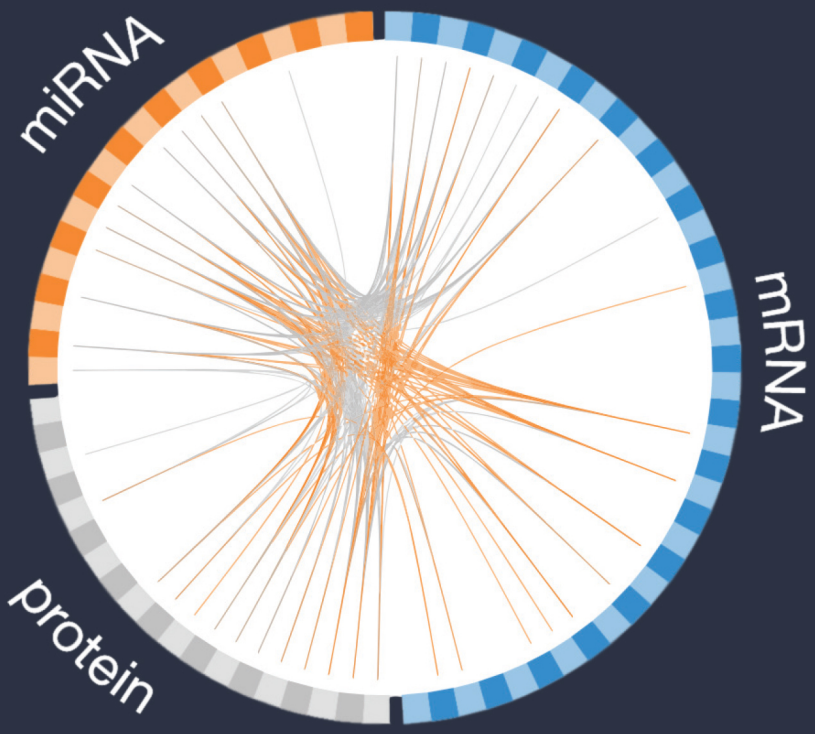
COMPUTATIONAL BIOLOGY SERIES



MULTIVARIATE DATA INTEGRATION USING R

Methods and Applications with the mixOmics Package

Kim-Anh Lê Cao
Zoe Welham

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Multivariate Data Integration Using R

Computational Biology Series

About the Series:

This series aims to capture new developments in computational biology, as well as high-quality work summarizing or contributing to more established topics. Publishing a broad range of reference works, textbooks, and handbooks, the series is designed to appeal to students, researchers, and professionals in all areas of computational biology, including genomics, proteomics, and cancer computational biology, as well as interdisciplinary researchers involved in associated fields, such as bioinformatics and systems biology.

Introduction to Bioinformatics with R: A Practical Guide for Biologists
Edward Curry

Analyzing High-Dimensional Gene Expression and DNA Methylation Data with R
Hongmei Zhang

Introduction to Computational Proteomics
Golan Yona

Glycome Informatics: Methods and Applications
Kiyoko F. Aoki-Kinoshita

Computational Biology: A Statistical Mechanics Perspective
Ralf Blossey

Computational Hydrodynamics of Capsules and Biological Cells
Constantine Pozrikidis

Computational Systems Biology Approaches in Cancer Research
Inna Kuperstein, Emmanuel Barillot

Clustering in Bioinformatics and Drug Discovery
John David MacCuish, Norah E. MacCuish

Metabolomics: Practical Guide to Design and Analysis
Ron Wehrens, Reza Salek

An Introduction to Systems Biology: Design Principles of Biological Circuits
2nd Edition
Uri Alon

Computational Biology: A Statistical Mechanics Perspective
Second Edition
Ralf Blossey

Stochastic Modelling for Systems Biology
Third Edition
Darren J. Wilkinson

Computational Genomics with R
Altuna Akalin, Bora Uyar, Vedran Franke, Jonathan Ronen

An Introduction to Computational Systems Biology: Systems-level Modelling of Cellular Networks
Karthik Raman

Virus Bioinformatics
Dmitrij Frishman, Manuela Marz

Multivariate Data Integration Using R: Methods and Applications with the mixOmics Package
Kim-Anh LeCao, Zoe Marie Welham

Bioinformatics
A Practical Guide to NCBI Databases and Sequence Alignments
Hamid D. Ismail

For more information about this series please visit:
https://www.routledge.com/Chapman--HallCRC-Computational-Biology-Series/book-series/CRCCBS

Multivariate Data Integration Using R

Methods and Applications with the mixOmics Package

Kim-Anh Lê Cao
Zoe Welham


CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

First edition published 2022
by CRC Press
6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742

and by CRC Press
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

ISBN: 9780367460945 (hbk)
ISBN: 9781032128078 (pbk)
ISBN: 9781003026860 (ebk)

DOI: 10.1201/9781003026860

This book has been prepared from camera-ready copy provided by the authors.

From Kim-Anh Lê Cao:
To my parents, Betty and Huy Lê Cao
To the mixOmics team and our mixOmics users,
And to my co-author Zoe Welham without whom this book would not have existed.

For Zoe Welham:
To Joy and Tamara Welham, for their continual support

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Contents

Preface

Context and scope

Modern high-throughput technologies generate information about thousands of biological molecules at different cellular levels in a biological system, leading to several types of omics studies (e.g. transcriptomics as the study of messenger RNA molecules expressed from the genes of an organism, or proteomics as the study of proteins expressed by a cell, tissue, or organism). However, a reductionist approach that considers each of these molecules individually does not fully describe an organism in its environment. Rather, we use multivariate analysis to investigate the simultaneous and complex relationships that occur in molecular pathways. In addition, to obtain a holistic picture of a complete biological system, we propose to integrate multiple layers of information using recent computational tools we have developed through the mixOmics project.

mixOmics is an international endeavour that encompasses methodological developments, software implementation, and applications to biological and biomedical problems to address some of the challenges of omics data integration. We have trained students and researchers in essential statistical and data analysis skills via our numerous multi-day workshops to build capacity in best practice statistical analysis and advance the field of computational statistics for biology. The goal of this book is to provide guidance in applying multivariate dimension reduction techniques for the integration of high-throughput biological data, allowing readers to obtain new and deeper insights into biological mechanisms and biomedical problems.

Who is this book for?

This book is suitable for biologists, computational biologists, and bioinformaticians who generate and work with high-throughput omics data. Such data include – but are not restricted to – transcriptomics, epigenomics, proteomics, metabolomics, the microbiome, and clinical data. Our book is dedicated to research postgraduate students and scientists at any career stage, and can be used for teaching specialised multi-disciplinary undergraduate and Masters's courses. Data analysts with a basic level of `R` programming will benefit most from this resource. The book is organised into three distinct parts, where each part can be skimmed according to the level and interest of the reader. Each chapter contains different levels of information, and the most technical chapters can be skipped during a first read.

Overview of methods in `mixOmics`

The `mixOmics` package focuses on multivariate analysis which examines more than two variables simultaneously to integrate different types of variables (e.g. genes, proteins, metabolites). We use dimension reduction techniques applicable to a wide range of data analysis types. Our analyses can be descriptive, exploratory, or focus on modeling or prediction. Our

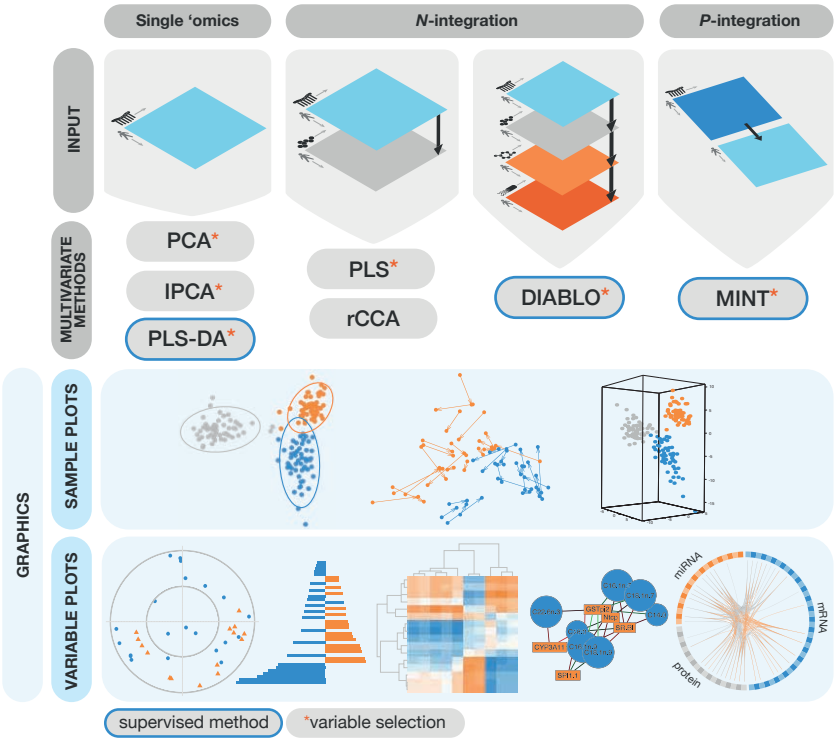


FIGURE 1: *Overview of the methods implemented in the* `mixOmics` *package for the exploration and integration of multiple data sets.* This book aims to guide the data analyst in constructing the research question, applying the appropriate multivariate techniques, and interpreting the resulting graphics.

aim is to summarise these large biological data sets to elucidate similarities between samples, between variables, and the relationship between samples and variables. The `mixOmics` package provides a range of methods to answer different kinds of biological questions, for example to:

- Highlight patterns pertaining to the major sources of variation in the data (e.g. Principal Component Analysis),
- Segregate samples according to their known group and predict group membership of new samples (e.g. Partial Least Squares Discriminant Analysis),
- Identify agreement between multiple data sets (e.g. Canonical Correlation Analysis, Partial Least Squares regression, and other variants),
- Identify molecular signatures across multiple data sets with sparse methods that achieve variable selection.

Key methodological concepts in `mixOmics`

Methods in `mixOmics` are based on *matrix factorisation* techniques, which offer great flexibility in analysing and integrating multiple data sets in a holistic manner. We use *dimension reduction* combined with *feature selection* to summarise the main characteristics of the data and posit novel biological hypotheses.

Dimension reduction is achieved by combining all original variables into a smaller number of artificial *components* that summarise patterns in the original data.

The `mixOmics` package is unique in providing novel multivariate techniques that enable feature selection to identify *molecular signatures*. Feature selection refers to identifying variables that best explain, or predict, the outcome variable (e.g. group membership, or disease status) of interest. Variables deemed irrelevant according to the specific statistical criterion we use in the methods are not taken into account when calculating the components.

Data integration methods use *data projection* techniques to maximise the covariance, or the correlation between, omics data sets. We propose two types of data integration, whether on the same N samples, or on the same P variables (Figure 1).

Finally, our methods can provide either *unsupervised* or *supervised* analyses. Unsupervised analyses are exploratory: any information about sample group membership, or outcome, is disregarded, and data are explored based on their *variance* or *correlation* structure. Supervised analyses aim to segregate sample groups known *a priori* (e.g. disease status, treatments) and identify variables (i.e. biomarker candidates, or molecular signatures) that either explain or separate sample groups.

These concepts will be explained further in Part I.

To aid in interpreting analysis results, `mixOmics` provides insightful graphical plots designed to highlight patterns in both the sample and variable dimensions uncovered by each method (Figure 1).

Concepts not covered

Each `mixOmics` method corresponds to an underlying statistical model. However, the methods we present radically differ from univariate formulations as they do not test one variable at a time, or produce p-values. In that sense, multivariate methods can be considered *exploratory* as they do not enable statistical inference. Our wide range of methods come in many different flavours and can be applied also for predictive purposes, as we detail in this book. 'Classical' univariate statistical inference methods can still be used in our analysis framework after the identification of molecular signatures, as our methods aim to generate novel biological hypotheses.

Who is 'mixOmics'?

The mixOmics project has been developed between France, Australia and Canada since 2009, when the first version of the package was submitted to the CRAN[1]. Our team is composed of core members from the University of Melbourne, Australia, and the Université de Toulouse, France. The team also includes several key contributors and collaborators.

The package implements more than nineteen multivariate and sparse methodologies for omics data exploration, integration, and biomarker discovery for different biological settings, amongst which thirteen were developed by our team (see our list of publications in Section 14.8). Originally, all methods were designed for omics data, however, their application is not limited to biological data only. Other applications where integration is required can be considered, but mostly for cases where the predictor variables are continuous.

[1]The Comprehensive R Architecture Network `https://www.cran.r-project.org`

The package is currently available from Bioconductor[2], with a development version available on GitHub[3]. We continue to maintain and improve the package via new methods, code optimisation and efficient memory storage of R objects.

About this book

Part I: Modern biology and multivariate analysis introduces fundamental concepts in multivariate analysis. *Multi-omics and biological systems* (Chapter 1) compares and contrasts multivariate and univariate analysis, and outlines the advantages and challenges of multivariate analyses. *The Cycle of Analysis* (Chapter 2) details the necessary steps in planning, designing and conducting multivariate analyses. *Key multivariate concepts and dimension reduction in `mixOmics`* (Chapter 3) describes measures of dispersion and association, and introduces key methods in `mixOmics` to manage large data, such as dimension reduction using matrix factorisation and feature selection. *Choose the right method for the right question in `mixOmics`* (Chapter 4) provides an overview of the methods available in `mixOmics` and the types of biological questions these methods can answer.

Part II: mixOmics under the hood provides a deeper understanding of the statistical concepts underlying the methods presented in Part III. *Projection to Latent Structures (PLS)* (Chapter 5) illustrates the different types of algorithms used to solve Principal Component Analysis. We detail in particular the iterative PLS algorithm that projects data onto latent structures (components) for matrix decomposition and dimension reduction, as this algorithm forms the basis of most of our methods. *Visualisation for data integration* (Chapter 6) showcases the variety of graphical outputs offered in `mixOmics` to complement each method. *Performance assessment in supervised analyses* (Chapter 7) describes the techniques employed to evaluate the results of the analyses.

Part III: mixOmics in action provides detailed case studies that apply each method in `mixOmics` to answer pertinent biological questions, complete with example `R` code and insightful plots. We begin with *mixOmics: get started* (Chapter 8) to guide the novice analyst in using the `R` platform for data analysis. Each subsequent chapter is dedicated to one method implemented in `mixOmics`. In *Principal Component Analysis (PCA)* (Chapter 9) and *PLS - Discriminant Analysis (PLS-DA)* (Chapter 12), we introduce different multivariate methods for single omics analysis. The N-integration framework is introduced in *PLS* (Chapter 10) and *Canonical Correlation Analysis* (Chapter 11) for two omics, and $N-$*data integration* (DIABLO, Chapter 13) for multi-omics integration. $P-$*data integration* (MINT, Chapter 14) introduces our latest developments for P-integration to combine independent omics studies. Each of these chapters is organised as follows:

- Aim of the method,
- Research question framed biologically and statistically,
- Principles of the method,
- Input arguments and key outputs,
- Introduction of the case study,
- Quick start R command lines,
- Further options to go deeper into the analysis,
- Frequently Asked Questions,
- Technical methodological details in each Appendix.

[2] `https://www.bioconductor.org/packages/release/bioc/html/mixOmics.html`
[3] `https://github.com/mixOmicsTeam/`

Additional resources related to this book

In addition to the `R` package, the mixOmics project includes a website with extensive tutorials in http://www.mixOmics.org. The `R` code of each chapter is also available on the website. Our readers can also register for our newsletter mailing list, and be part of the mixOmics community on GitHub and via our discussion forum https://mixomics-users.discourse.group/.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Authors

Dr Kim-Anh Lê Cao develops novel methods, software and tools to interpret big biological data and answer research questions efficiently. She is committed to statistical education to instill best analytical practice and has taught numerous statistical workshops for biologists and leads collaborative projects in medicine, fundamental biology or microbiology disciplines. Dr Kim-Anh Lê Cao has a mathematical engineering background and graduated with a PhD in Statistics from the Université de Toulouse, France. She is currently an Associate Professor in Statistical Genomics at the University of Melbourne. In 2019, Kim-Anh received the Australian Academy of Science's Moran Medal for her contributions to Applied Statistics in multidisciplinary collaborations. She has contributed to a leadership program for women in STEMM, including the international Homeward Bound which culminated in a trip to Antarctica, and Superstars of STEM from Science Technology Australia.

Zoe Welham completed a BSc in molecular biology and during this time developed an interest in the analysis of big data. She completed a Master of Bioinformatics with a focus on the statistical integration of different omics data in bowel cancer. She is currently a PhD candidate at the Kolling Institute in Sydney where she is furthering her research into bowel cancer with a focus on integrating microbiome data with other omics to characterise early bowel polyps. Her research interests include bioinformatics and biostatistics for many areas of biology and making that information accessible to the general public.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Part I

Modern biology and multivariate analysis

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

1

Multi-omics and biological systems

Technological advances such as next-generation sequencing and mass spectrometry generate a wealth of diverse biological information, allowing for the monitoring of thousands of variables, or dimensions, that describe a given sample or individual, hence the term 'high dimensional data'. Multi-omics variables represent molecules from different functional levels: for example, transcriptomics for the study of transcripts, proteomics for proteins, and metabolomics for metabolites. However, their complex nature requires an integrative, multidisciplinary approach to analysis that is not yet fully established.

Historically, the scientific community has adopted a *reductionist* approach to data analysis by characterising a very small number of genes or proteins in one experiment to assess specific hypotheses. A holistic approach allows for a deeper understanding of biological systems by adding two new facets to analysis (Figure 1.1): Firstly, by integrating data from different omic functional levels, we move from clarifying a linear process (e.g. the dysregulation of one or two genes) towards understanding the development, health, and disease of an ever-changing, dynamic, hierarchical *system*. Secondly, by adopting a hypotheses-free, data-driven approach, we can build integrated and coherent models to address novel, systems-level hypotheses that can be further validated through more traditional hypotheses.

1.1 Statistical approaches for reductionist or holistic analyses

Compared to a traditional reductionist analysis, multivariate multi-omics analysis drastically differs in its viewpoint and aims. We briefly introduce three types of analysis to illustrate this point:

A **univariate analysis** is a fundamentally reductionist, *hypothesis-driven* approach that is related to inferential statistics (introduced in Section 2.4). A hypothesis test is conducted on one variable (e.g. gene expression or protein abundance) independently from the other variables. Univariate methods make inferences about the population and measure the certainty of this inference through test statistics and p-values. Linear models, t-tests, F-tests, or non-parametric tests fit into a univariate analysis framework. Although interactions between variables are not considered in univariate analyses, when one variable is manipulated in a controlled experiment, we can often attribute the result to that particular variable. In omics studies, where multiple variables are monitored simultaneously, it is difficult to determine which variables influence the biology of interest.

A **bivariate analysis** considers two variables simultaneously, for example, to assess the association between the expression levels of two genes via correlation or linear regression. Such an analysis is often supported by visualisation through scatterplots but can quickly

DOI: 10.1201/9781003026860-1

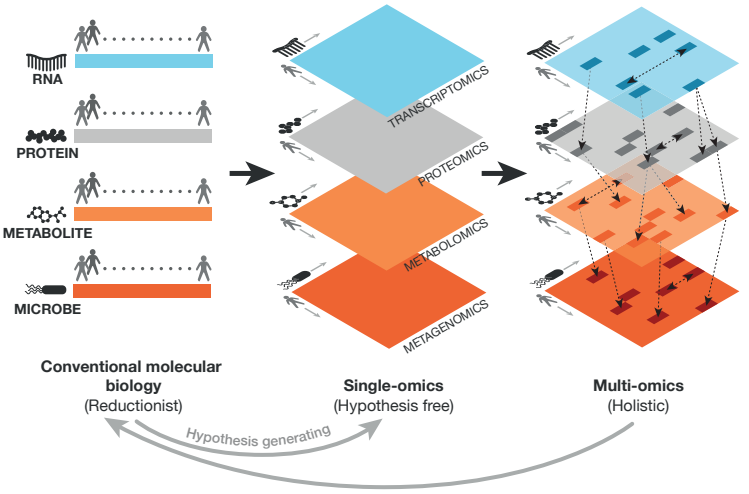


FIGURE 1.1: From reductionism to holism. Until recently, only a few molecules of a given omics type were analysed and related to other omics. The advent of high-throughput biology has ushered in an era of hypothesis-free approaches within a single type of omics data, and across multiple omics from the same set of samples. A holistic approach is now required to understand the different omic functional layers in a biological system and posit novel hypotheses that can be further validated with a traditional reductionist approach. We have omitted DNA as this data type needs to be handled differently in `mixOmics`, see Section 4.2.3.

become cumbersome when dealing with thousands of variables that are considered in a pairwise manner.

A **multivariate analysis** examines more than two variables simultaneously and potentially thousands at a time. In omics studies, this approach can lead to computational issues and inaccurate results, especially when the number of samples is much smaller than the number of variables. Several computational and statistical techniques have been revisited or developed for high-dimensional data. This book focuses on multivariate analyses and extends this to include the integration of multi-omics data sets.

1.2 Multi-omics and multivariate analyses

The aim of omics data integration is to identify associations and patterns amongst different variables and across different omics collected from a sample. Provided appropriate data analysis is conducted, the integration of multiple data sources may also consolidate our confidence in the results when consensus is observed from different experiments.

1.2.1 More than a 'scale up' of univariate analyses

The fundamental difference between multivariate and univariate analysis lies in the scope of the results obtained. Multivariate analysis can unravel groups of variables that share similar patterns in expression across different phenotypes, thus complementing each other to describe an outcome. A univariate analysis may declare the same variables as non-significant, as a variable's ability to explain the phenotype may be subtle and can be masked by individual variation, or confounders (Saccenti et al., 2014). However, with sufficiently powered data, univariate and multivariate methods are complementary and can help make sense of the data. For example, several multivariate and exploratory methods presented in this book can suggest promising candidate variables that can be further validated through experiments, reductionist approaches, and inferential statistics.

1.2.2 More than a fishing expedition

Multivariate analyses, which examine up to thousands of variables simultaneously, are often considered to be 'fishing expeditions'. This somewhat pejorative term refers to either conducting analyses without first specifying a testable hypothesis based on prior research, or, conducting several different analyses on the same data to 'fish' for a significant result regardless of its domain relevance. Indeed, examining a large number of variables can lead to statistically significant results purely by chance.

However, the integration of multi-omics data, with an appropriate experimental design set in an exploratory, rather than predictive approach, offers a tremendous opportunity for discovering associations between omics molecules (whether genes, transcripts, proteins, or metabolites), in normal, temporal or spatial changes, or in disease states. For example, one of our studies identified pathways that were never previously identified as relevant to ontogeny during the first week of human life (Lee et al., 2019). Multi-omics data integration has deepened our understanding of gene regulatory networks by including information from related molecules prior to validation of gene associations with a functional approach (Gligorijević and Pržulj, 2015) and has also efficiently improved functional annotations to proteins instead of using expensive and time-consuming experimental techniques (Ma et al., 2013). Multi-omics can also more easily characterise the relatively small number of genes associated with a particular disease by integrating multiple sources of information (Žitnik et al., 2013). Finally, it has further developed precision medicine by integrating patient- and disease-specific information with the aim to improve prognosis and clinical outcomes (Ritchie et al., 2015).

1.3 Shifting the analysis paradigm

Despite the potential advantages of high-dimensional data, we should keep in mind that quantity does not equal quality. Multivariate data integration is not straightforward: the analyses cannot be reduced to a mere concatenation of variables of different types, or by overlapping information between single data sets, as we illustrate in Figure 1.2. As such, we must shift our traditional view of analysing data.

Biological experimentation often employs univariate statistics to answer clear hypotheses

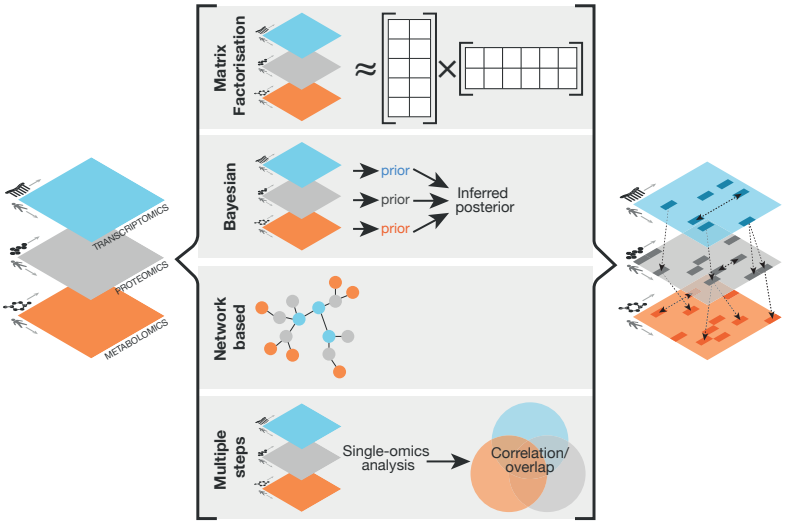


FIGURE 1.2: Types of methods for data integration. Methods for multi-omics data integration are still in active development, and can be broadly categorised into matrix factorisation techniques (the focus of this book), Bayesian, network-based, and multiple-step approaches. The latter deviates from data integration as it considers each data set individually before combining the results.

about the potential causal effect of a given molecule of interest. In high-dimensional data sets, this reductionist approach may not hold due to the sheer amount of molecules that are monitored, and their interactions that might be of biological interest. Therefore, exploratory, *data-driven* approaches are needed to extract information from noise and generate new hypotheses and knowledge. However, the lack of a clear, causal-driven hypothesis presents a challenging new paradigm in statistical analyses.

In univariate hypothesis testing, we report p-values to determine the significance of a statistical test conducted on a *single* variable. In a multivariate setting, however, a p-value assesses the statistical significance of a result while taking into account *all* variables simultaneously. In such analyses, permutation-based tests are common to assess how far from random a result is when the data are reshuffled, but other inference-based methods are currently being developed in the field of multivariate analysis (Wang and Xu, 2021). In `mixOmics` we do not offer such tests, but related methods propose permutation approaches to choose the parameters in the method (see Section 10.7.5).

1.4 Challenges with high-throughput data

There are multiple challenges associated with managing large amounts of biological data, pertaining to specific types of data as well as statistical analysis. To make reliable, valid, and meaningful interpretations, these challenges must be considered, ideally before data collection.

1.4.1 Overfitting

Multivariate omics analysis assesses many molecules that individually, or in combination, can explain the biological outcome of interest. However, these associations may be spurious, as the large number of features can often be combined in different ways to explain the outcome well, despite having no biological relevance. Overfitting occurs when a statistical model captures the noise along with the underlying pattern in the data: if we apply the same statistical model fitted on a high-dimensional data set to a similar but external study, we might obtain different results.[1] The problem of overfitting is a well-known issue in high-throughput biology (Hawkins, 2004). We can assess the amount of overfit using cross-validation or subsampling of the data, as described in Chapter 7.

1.4.2 Multi-collinearity and ill-posed problems

As the number of variables increase, the number of pairwise correlations also increases. *Multi-collinearity* poses a problem in most statistical analyses as these variables bring redundant and noisy information that decreases the precision of the statistical model. Correlations in high-throughput data sets are often spurious, especially when the number of biological samples, or individuals N, is small compared to the number of variables P[2]. The 'small N large P' problem is defined as *ill-posed*, as standard statistical inference methods assume N is much greater than P to generalise the results to the population the sample was drawn from. Ill-posed problems also lead to inaccurate computations.

1.4.3 Zero values and missing values

Data sets may contain a large number of zeros, depending on the type of omics studied and the platform that is used. This is particularly the case for microbiome, proteomics, and metabolomics data: a large number of zeros results in zero-inflated (skewed) data, which can impair methods that assume a normal distribution of the data. *Structural zeros*, or true zeros, reflect a true absence of the variable in the biological environment while *sampling zeros*, or false zeros, may not reflect reality due to experimental error, technological reasons, or an insufficient sample size (Blasco-Moreno et al., 2019). The challenge is whether to consider these zeros as a true zero or missing (coded as `NA` in `R`).

Methods that can handle missing values often assume they are 'missing at random', i.e. missingness is not related to a specific sample, individual, molecule, or type of omics platform. Some methods can estimate missing values, as we present in Appendix 9.A.

[1] Statistical models that overfit have low bias and high variance, meaning that they tend to be complex to fit the training data well, but do not predict well on test data (more details about the bias-variance tradeoff can be found in Friedman et al. (2001) Chapter 2).

[2] In our context, N can also refer to the number of cells in single cell assays, as we briefly mention in Section 14.6.

1.5 Challenges with multi-omics integration

Examining data holistically may lead to better biological understanding, but integrating multiple omics data sets is not a trivial task and raises another series of challenges.

1.5.1 Data heterogeneity

Different omics rely on different laboratory techniques and data extraction platforms, resulting in data sets of different formats, complexity, dimensionalities, information content, and scale, and may be processed using different bioinformatics tools. Therefore, data heterogeneity arises from biological *and* technical reasons and is the main analytical challenge to overcome.

1.5.2 Data size

Integrating multiple omics results in a drastic increase in the number of variables. A filtering step is often applied to remove irrelevant and noisy variables (see Section 8.1). However, the number of variables P still remains extremely large compared to the number of samples N, which raises computational as well as analytical issues.

1.5.3 Platforms

The data integration field is constantly evolving due to ever-advancing technologies with new platforms and protocols, each containing inherent technical biases and analytical challenges. It is crucial that data analysts swiftly adapt their analysis framework to keep apace with these omics-era demands. For example, single cell techniques are rapidly advancing, as are new protocols for their multi-omics analysis.

1.5.4 Expectations for analysis

The field of data integration has no set definition. Data integration can be managed biologically, bioinformatically, statistically, or at the interpretation steps (i.e. by overlapping biological interpretation once the statistical results are obtained). Therefore, the expectations for data integration are diverse; from exploration, and from a low to high-level understanding of the different omics data types. Despite recent advances in single cell sequencing, current technologies are still limited in their ability to parse omics interactions at precise functional levels. Thus, our expectations for data integration are limited, not only by the statistical methods but also by the technologies available to us.

1.5.5 Variety of analytical frameworks

Integrative techniques fully suited to multi-omics biological data are still in development and continue to expand[3]. Different types of techniques can be considered and broadly categorised into (Huang et al. (2017), Figure 1.2):

- Matrix factorisation techniques, where large data sets are decomposed into smaller sub-matrices to summarise information. These techniques use algebra and analysis to optimise specific statistical criteria and integrate different levels of information. Methods in `mixOmics` fit into this category and will be detailed in Chapter 3 and subsequent chapters,
- Bayesian methods, which use assumptions of prior distributions for each omics type to find correlations between data layers and infer posterior distributions,
- Network-based approaches, which use visual and symbolic representations of biological systems, with nodes representing molecules and edges as correlations between molecules, if they exist. Network-based methods are mostly applied for detecting significant genes within pathways, discovering sub-clusters, or finding co-expression network modules,
- Multiple-step approaches that first analyse each single omics data set individually before combining the results based on their overlap (e.g. at the gene level of a molecular signature) or correlation. This type of approach technically deviates from data integration but is commonly used.

1.6 Summary

Modern biological data are high dimensional; they include up to thousands of molecular entities (e.g. genes, proteins, or epigenetic markers) per sample. Integrating these rich data sets can potentially uncover the hierarchical and holistic mechanisms that govern biological pathways. While classical, reductionist, univariate methods ignore these molecular interactions, multivariate, integrative methods offer a promising alternative to obtain a more complete picture of a biological system. Thus, univariate and multivariate methods are different approaches with very little overlap in results but have the advantage of complementarity.

The advent of high-throughput technology has revealed a complex world of multi-omics molecular systems that can be unraveled with appropriate integration methods. However, multivariate methods able to manage high-dimensional and multi-omics data are yet to be fully developed. The methods presented in this book mitigate some of these challenges and will help to reveal patterns in omics data, thus forging new insights and directions for understanding biological systems as a whole.

[3] A comprehensive list of multi-omics methods and software is available at https://github.com/mikelove/awesome-multi-omics.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

2

The cycle of analysis

The *Problem, Plan, Data, Analysis, Conclusion* (PPDAC) cycle is a useful framework for answering an experimental question effectively (Figure 2.1). The mixOmics project emphasises crafting a well-defined biological question (Chapter 4), as this guides data acquisition and preparation (Chapter 8), as well as choosing appropriate multivariate techniques for analysis (Chapter 4). Although this book is focused on analysis and interpretation, careful consideration of each step will maximise a successful analytical outcome.

FIGURE 2.1: PPDAC. The *Problem, Plan, Data, Analysis, Conclusion* cycle proposed by MacKay and Oldford (2000) will guide our multivariate analysis process.

2.1 The *Problem* guides the analysis

Multivariate analysis is appropriate for large data sets where the biological question encompasses a broad domain, rather than parsing the action of a single or small number of variables. Thus, we often require a hypothesis-free investigation based on a *data-driven* approach. However, this does not imply that multivariate analysis is a fishing expedition with no underlying biological question. The experimental design, driven by a well-formulated biological question and the choice of statistical method, will ensure a successful analysis (Shmueli, 2010). Chapter 4 lists several types of biological questions that can be answered with multivariate and integrative methods.

DOI: 10.1201/9781003026860-2