Computational Biology

Cover image: A framework for integrative biology. See page 21, chapter 2 for details.

Computational Biology

Edited by

HOLGER HUSI, DR SC NAT

Codon Publications
Brisbane, Australia

Computational Biology
ISBN: 978-0-9944381-9-5
DOI: http://dx.doi.org/10.15586/computationalbiology.2019

Edited by
Holger Husi, Dr sc nat, Division of Biomedical Science, University of the Highlands and Islands, UK

Published by
Codon Publications
Brisbane, Australia

First Published in October 2019
Printed in Australia

CONTENTS

Doi: http://dx.doi.org/10.15586/computationalbiology.2019

FOREWORD

Our understanding of biology has undergone a revolution in the past 20 years, driven by our ability to capture, store, and interrogate ever-increasing volumes of data. The monumental strides are best illustrated and most visible in the world of genetics and molecular biology in which the power of the discovery of the structure of DNA – for which the Nobel Prize was awarded to Watson, Crick, and Wilkins in 1962 – was only fully unleashed in the late 1990s, when high-performance computers were made available to unlock the secrets of the entire human genome. However, this only heralded the beginning: genomics is only one of a growing number of enormous data sets in biology that requires substantial computing power to realize their full potential. New disciplines have evolved: transcriptomics, proteomics, lipidomics, metabolomics, systems biology, epigenomics, and data analytics are all exponents of this brave new, biological world of "Computational Biology".

This book draws together many of the latest cutting-edge developments in the field of Computational Biology. Each chapter draws on the expertise of world leaders in the field to highlight the utility and potential importance of specific technologies. The breadth of the text is impressive: from Integrative Biology in human diseases through the various branches of metabolomics and proteomics to sequencing and deep learning are all covered. In addition, the key role of statistics in large data set analysis is discussed in a dedicated chapter.

This book would have broad appeal to anybody with an interest in cutting-edge biology. It is important that the computational power that is now available to us to help unravel the seemingly impenetrable complexity of biological systems is fully utilized. The benefits of these technologies are boundless in biology and have yet to be fully realized; precision medicine represents an excellent example of an aspiration in medical development that would be simply unachievable without computational biology at its core.

Prof. Ian Megson BSc, PhD, FHEA, FRBS, FBPhS
Head of Health Research & Innovation
University of the Highlands & Islands, UK
October 2019
Doi: http://dx.doi.org/10.15586/computationalbiology.2019.fr

PREFACE

Computational biology is nowadays one of the cornerstones in biological and medical data analysis and has a long and proud history originating in the 1960s from the fields of biophysics and protein biochemistry, notably the modeling of enzymatic reactions and other kinetic parameters. With the advent of improved and easier to access computing systems came the possibility of exploring biological systems to a much greater depth, especially linked to large-scale analytics platforms of biological samples, such as whole-genome sequencing tools, arrays, mass spectrometry, and many more. Such a considerable volume of data procured in a fast-paced technology-dependent manner required new ways to handle, manage, and analyze the information through improved data analytics streams, which was accomplished by borrowing and applying know-how from other sciences, such as mathematics, statistics, and computer sciences to biology, medicine, and disease analysis. This led to a vast expansion of data repositories and available computational tools feeding into reference databases and constantly improving our understanding of complex biological mechanisms. Ultimately, our ability to handle vast amounts of complex data enables us to integrate the various data streams into a contextualized system through systems approaches, network analysis, and modeling methodologies. Although it is evident that many gaps in our understanding of how any given biological system works still remain, more powerful systems, platforms, and procedures have started to emerge, such as automated decision machines, artificial intelligence, pattern matching approaches, and integrated and integrative data handling protocols, which will help us to continue uncovering new insights.

This book is aimed at both novices and specialists in the field of computational biology and brings together a selection of approaches at the cutting edge of technology and shows both how data and analytics procedures aid us in expanding our understanding of biology and how aberrant or modulated processes can lead to diseases. The first section (chapters 1–5) provides a more general overview. Chapter 1 gives an introduction to image-based systems biology of multicellular spheroids for experimentalists and theoreticians. Here, mathematical models of spheroids are used to investigate cellular interactions since tissues, cells, and even smaller components can be abstracted to spheroids to give a three-dimensional representation of the structural organization within a system. Chapter 2 discusses integrative biology approaches applied to human diseases, in particular, to multifactorial and complex interactions encountered in studying aberrant etiology. Concepts and analytics techniques are introduced for single-layer omics methods as well as procedures to integrate multi-omics data, leading to meaningful and relevant biological insights. Following this is the approach of using machine learning or deep learning in omics data analysis and precision medicine, as described in Chapter 3: deep learning allows us to identify complex patterns and create predictive models from omics data, as well as medical image analysis. Chapter 4 addresses the use of computational biology and bioinformatics in biological sequence analysis, where not only sequence alignments are an important step but also feature detection and selection are of significance. Supervised and

unsupervised learning, neural networks, and hidden Markov models are discussed, as well as deep sequencing or next-generation sequencing data analysis procedures using artificial intelligence and machine learning methods. Statistical procedures to analyze multi-omics data are presented in Chapter 5, using multivariate statistical methods for high-dimensional multiset omics data analysis. Application of canonical correlation analysis, redundancy analysis, and penalized versions are commonly used in omics dataflows, and this chapter gives an overview of how these methods came to match the statistical challenges that come with high-dimensional multiset omics data analysis.

A more specific overview of approaches in computational biology is given in chapters 6–9. Statistical methods for RNA sequencing data analysis are presented in Chapter 6. It covers the statistical models, model assumptions, and challenges encountered in RNA sequencing data analysis, including differential analysis, clustering approaches, and pathway analysis. Here, data analytics packages and embedded statistical methods and how they perform using real-world data are described. Chapter 7 addresses computational epigenomics, ranging from fundamental research to disease prediction and risk assessment. The epigenome encompasses several chemical properties of DNA and DNA-associated proteins that are tissue-specific, distinctive for a disease state, and sensitive to environmental conditions. Mining of genomic data sets and their associated epigenomic features, as well as the computational approaches used to assess statistical significance in comparative analyses, are discussed in this chapter. Chapter 8 discusses computational approaches in proteomics, where an overview of proteomic approaches, biological sample considerations, and data acquisition methods is given. Additionally, data processing software solutions for the various steps and further functional analyses of biological data are presented, which enable the comparison of various data sets as a summation of individual experiments, to cross-compare sample types and other metadata. Chapter 9 reviews cheminformatics and computational approaches in metabolomics using data mining methods and bioinformatics tools, including machine learning approaches. In this chapter, the main technical procedures used in metabolomics data acquisition, data processing, and pipelines, and the ways in which metabolomics data can aid in elucidating aberrant pathways and metabolic dysfunctions in disease, are discussed.

The last two chapters cover more specialized topics. Chapter 10 discusses the nature of feature selection in high-dimensional data using entropy information through statistical inference concepts of entropy in microarray data clustering in order to reduce the multi-dimensionality inherent in the source data to allow data summarization and the specific selection of gene sets associated with modulated conditions such as those found in diseases. The last review, Chapter 11, addresses structural pattern mining approaches applied to cryo-electron tomography using template-based and template-free procedures, where the observation of cellular organelles and macromolecular complexes at nanometer resolution with native conformations requires supervised deep learning-based pattern mining approaches in order to identify and reconstruct biological structures on the cellular as well as molecular level.

A full comprehensive summary of work carried out in the field of computational biology would span many volumes as this discipline is now deeply embedded in practically all large-scale, multi-subject, or integrative data analytics investigations.

Not only is this a very active field in terms of applications but also in the development of novel algorithms, resources, and pipelines. Condensing a torrent of data into contextualized, coherent, and understandable information to explain biology, disease, or to be used in fields such as personalized medicine is no longer possible without the aid of computational tools. Over time, it is expected that new and exciting developments in computational biology will allow us to gain an as yet unprecedented ability to make sense out of seemingly random data in a timely and precise manner. We believe the readers would enjoy the work presented in this book and will be both enlightened and encouraged to further the understanding of the biological world and how we work as a complex assembly of cells and an organism as a whole.

Holger Husi, Dr sc nat
Division of Biomedical Science
University of the Highlands and Islands, UK
October 2019
Doi: http://dx.doi.org/10.15586/computationalbiology.2019.pr

LIST OF CONTRIBUTORS

AEILKO H. ZWINDERMAN, PHD

Department of Clinical Epidemiology
Biostatistics and Bioinformatics
Academic Medical Center
Amsterdam, AZ, The Netherlands

ALI REZA SOLTANIAN, PHD

Modeling of Noncommunicable Diseases Research Center
School of Public Health
Hamadan University of Medical Sciences
Hamadan, Iran

ALYSSON H. URBANSKI, MSC

Department of Clinical and Toxicological Analyses
School of Pharmaceutical Sciences
University of Sao Paulo
Sao Paulo, Brazil

ATTILA CSALA, MSC

Department of Clinical Epidemiology
Biostatistics and Bioinformatics
Academic Medical Center
Amsterdam, AZ, The Netherlands

BELA SANCHES, MPHARM, PHD

Strathclyde Institute of Pharmacy & Biomedical Sciences (SIPBS)
University of Strathclyde, Glasgow
UK

CORAL DEL VAL, PHD

Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain

DONGMEI LI, PHD

Clinical and Translational Science Institute
University of Rochester School of Medicine and Dentistry
Rochester, NY, USA

FATEMEH BAHREINI, PHD

Department of Molecular Medicine and Genetics
Faculty of Medicine
Hamadan University of Medical Sciences
Hamadan, Iran

FLORIAN MARTIN, PHD

PMI R&D, Philip Morris Products S.A.
Neuchâtel, Switzerland

FRANCISCO HERRERA, PHD

Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain

HELDER I. NAKAYA, PHD

Department of Clinical and Toxicological Analyses
School of Pharmaceutical Sciences
University of Sao Paulo
Sao Paulo, Brazil;
Scientific Platform Pasteur/USP
University of Sao Paulo
Sao Paulo, Brazil

HOLGER HUSI, DR SC NAT

Institute of Cardiovascular and Medical Sciences
BHF Glasgow Cardiovascular Research Centre
University of Glasgow
Glasgow, UK;
Division of Biomedical Sciences
Centre for Health Science
University of Highlands and Islands
Inverness, UK

IGOR ZWIR, PHD

Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain

JORDI MARTORELL-MARUGAN, MSC

GENYO, Centre for Genomics and Oncological Research: Pfizer
University of Granada
Andalusian Regional Government
PTS Granada
Granada, Spain

JOSÉ D. ARAUJO, MSC

Department of Clinical and Toxicological Analyses
School of Pharmaceutical Sciences
University of Sao Paulo
Sao Paulo, Brazil

JULIA HOENG, PHD

PMI R&D, Philip Morris Products S.A.
Neuchâtel, Switzerland

KARLA CERVANTES GRACIA, MSC

Basic Sciences Division
Universidad de Monterrey
San Pedro Garza García, N.L. Mexico

MANUEL C. PEITSCH, PHD

PMI R&D, Philip Morris Products S.A.
Neuchâtel, Switzerland

MARCO FERNANDES, MSC, PHD

Department of Psychiatry
Warneford Hospital
Translational Neuroscience and Dementia Research
Oxford University
Oxford, UK;
Institute of Cardiovascular and Medical Sciences
BHF Glasgow Cardiovascular Research Centre
University of Glasgow
Inverness, UK

MIN XU, PHD

Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA

MOHAMED-AMIN CHOUKRALLAH, PHD

PMI R&D, Philip Morris Products S.A.
Neuchâtel, Switzerland

NICOLAS SIERRO, PHD

PMI R&D, Philip Morris Products S.A.
Neuchâtel, Switzerland

NIKOLAI V. IVANOV, PHD

PMI R&D, Philip Morris Products S.A.
Neuchâtel, Switzerland

NILOOFAR RABIEI, MSC

Department of Biostatistics
School of Public Health
Hamadan University of Medical Sciences
Hamadan, Iran

PEDRO CARMONA-SÁEZ, PHD

GENYO, Centre for Genomics and Oncological Research: Pfizer
University of Granada
Andalusian Regional Government
PTS Granada
Granada, Spain

RACHEL CREIGHTON, BS

Department of Bioengineering
University of Washington
Seattle, WA, USA

SABINE C. FISCHER, PHD

Center for Computational and Theoretical Biology
Department of Biology
Universität Würzburg
Würzburg, Germany

SIHAM TABIK, PHD

Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain

USMAN SAEED, MPHIL

Dennemeyer Octimine GmbH
München, Germany;
Department of Bioinformatics
Technical University Munich
Wissenschaftzentrum Weihenstephan
Freising, Germany

XIANGRUI ZENG, BSC

Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA

XIN GAO, PHD

Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division
Computational Bioscience Research Center (CBRC)
King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia

XINDI WU, BSC

Computational Biology Department
Carnegie Mellon University
Pittsburgh, PA, USA

YASSIR BENHAMMOU, MSC

Department of Computer Science and Artificial Intelligence
University of Granada
Granada, Spain

ZAINAB USMAN, MPHIL

Department of Bioinformatics
Technical University Munich
Wissenschaftzentrum Weihenstephan
Freising, Germany

ZHENXI ZHU, BSC

Beijing University of Posts and Telecommunications
Beijing, China

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.cont

1

An Introduction to Image-Based Systems Biology of Multicellular Spheroids for Experimentalists and Theoreticians

Sabine C. Fischer

Center for Computational and Theoretical Biology, Department of Biology, Universität Würzburg, Würzburg, Germany

Author for correspondence: Sabine C. Fischer, Center for Computational and Theoretical Biology, Department of Biology, Universität Würzburg, Emil-Fischer-Str. 32, 97074 Würzburg, Germany. Email: sabine.fischer@uni-wuerzburg.de

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch1

Abstract: Multicellular organisms are inherently three-dimensional. This leads to complex intercellular interactions that cannot be reproduced in two-dimensional cell culture. Instead, three-dimensional spheroids, ball-shaped cell aggregates, arise as model systems. Spheroids provide an accurate in vitro representation of the three-dimensional organization of cells in tissues, and compared to a real tissue, they excel with well-defined experimental conditions, easy handling, and suitability for high-quality imaging. Therefore, spheroids are an experimental system that can be readily combined with mathematical modeling. This chapter shows how image-based systems biology is implemented for multicellular spheroids to study three-dimensional cell–cell interactions. The chapter is intended for experimentalists and theoreticians who plan to extend their research by linkage with other disciplines. The relevant concepts for experimental approaches and quantitative imaging are introduced and linked to mathematical models of spheroids. This results in a list of potential systems biology workflows for typical spheroid research areas in cell biology, cancer biology, and bioprinting. In all three areas, there is a large gap between the details of the mathematical models and the

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

available imaging data. The aim of this chapter is to encourage more interactions of experimentalists and theoreticians to fill this gap in spheroid research.

Keywords: agent-based model; continuous model; microscopy; spheroid formation; spheroid fusion

INTRODUCTION

Biomedical research in areas like cancer, infection, and developmental biology relies more and more on three-dimensional in vitro models including organoids, tissue explants, embryoid bodies, and spheroids (1). A major question in these research areas is how cellular interactions affect the overall behavior of the system. In this regard, multicellular spheroids are the best-studied system. They are three-dimensional, ball-shaped solid cellular aggregates that can be formed from various cell types (2). Originally, they have been formed by cancer cells and implemented as a model system for avascular tumors. Spheroids excel by well-defined experimental conditions, easy handling, as well as the suitability for high-quality imaging and generation of large sample sizes. Hence, they are an ideal model system to address the three-dimensional cellular arrangement, the behavior of individual cells within a tissue-like construct and the contribution of individual cells to the growth of the whole aggregate. Image-based systems biology provides an appropriate conceptual framework to tackle these questions (3). It combines experimental approaches with quantitative imaging data and spatial mathematical modeling (Figure 1). The collaboration of experimentalists and theoreticians

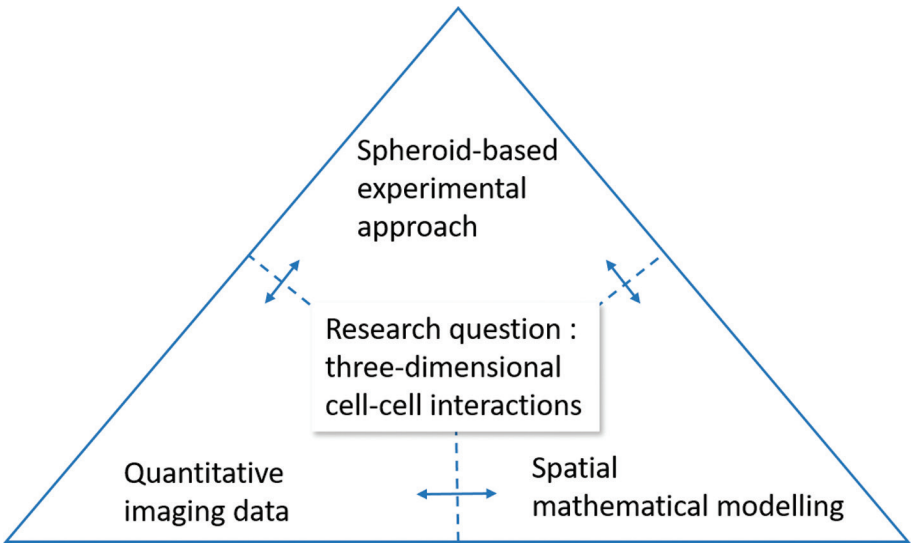


Figure 1 Image-based systems biology in spheroid research. The question of the mechanisms driving three-dimensional cell–cell interactions is tackled by a combination of experiments, quantitative imaging, and mathematical modeling. The three approaches have to be tightly linked.

ideally starts with formulating a common research question. Subsequently, in preliminary studies, the experimental, imaging, and mathematical methods are developed, implemented, and matched. In the main study, all parts interact closely to test the existing hypotheses and generate new ones. This approach requires experts from three different fields and is therefore difficult to implement. In the field of spheroid research, only a few studies have conducted the whole cycle. By introducing the concepts for experimental setups, image, and data analysis as well as mathematical modeling of spheroids and suggesting workflows of how to combine them, this chapter aims at fostering more systems biology approaches in spheroid research.

METHODS

For an image-based systems biology approach, methods from four different categories have to be chosen: experimental approach, imaging, image analysis, and mathematical modeling. The methods have to be inter-linkable and most importantly appropriate for the research question. For example, the quantitative evaluation of the images should provide results that are readily comparable to the statistical readout from the mathematical model. Furthermore, these numbers should provide new insight regarding the question of interest.

Experimental approaches

There are three typical approaches for experimental setups involving spheroids (Table 1). The analysis of spheroid formation mainly focuses on how cells aggregate in three spatial dimensions. This provides insights into cellular aggregation, rearrangement, and adhesion. Fully formed spheroids are typically used to evaluate spheroid growth as well as the viability of the individual cells. Further questions investigated with spheroids could include detailed analyses of the behavior of individual cells like differentiation or potential rearrangement in the spheroid. The fusion of two or more spheroids helps to address questions regarding rearrangement of individual cells or whole cell aggregates as well as cell sorting.

Imaging techniques

Investigating cellular properties within the three-dimensional spheroid context requires the spatial information of each cell and the geometry of the entire spheroid. Classical methods for imaging spheroids have relied on

TABLE 1 Experimental approaches

Research question	Experimental approach
Aggregation, rearrangement, adhesion	Spheroid formation
Spheroid growth, viability, rearrangement, adhesion, differentiation	Fully formed spheroid
Rearrangement, sorting	Spheroid fusion

physical sectioning (4). More recent approaches rely on light microscopy of the intact spheroids. Each approach has its strengths and weaknesses. Hence, choosing the best approach for spheroid imaging depends heavily on the scientific question. Conventional light microscopy such as wide-field microscopy provides a global but two-dimensional picture of spheroids. Due to the high imaging speed and the possibility of imaging multi-well plates, it has widely been used for high-content assays to characterize spheroid viability (5).

Confocal fluorescence microscopy enables imaging of spheroids at the single-cell level in three spatial dimensions. Standard confocal microscopes allow high-throughput imaging of multi-well plates, but the imaging speed is reduced compared to wide-field microscopy. The penetration depth of confocal microscopy is limited by the signal to background or the signal to noise ratio (6), and only small spheroids can be imaged in toto. A further drawback of confocal microscopy is the high risk of phototoxicity and photobleaching (7). Hence, confocal microscopy is only useful for short time-lapse imaging of living spheroids. Light sheet-based fluorescence microscopy (LSFM) allows the imaging of large three-dimensional specimens over periods of several days (7, 8). It provides high acquisition speed and good penetration depth. Photobleaching and phototoxicity are minimized. High-quality imaging with LSFM requires an optimal sample preparation, and standard multi-well plates cannot be used. To achieve full penetration into huge spheroids, multiple views are recorded and subsequently fused (9).

Imaging the sub-micron features of cells requires electron microscopes which have a much higher resolution than light microscopes (10, 11). The physical properties of electron microscopes (e.g., high vacuum) demand specific preparation and staining techniques to reveal the ultrastructure of cells and tissues. Sample preparation has to be optimized such that the number of introduced artifacts is minimal (12).

In summary, imaging techniques differ in the amount of detail they provide, which is typically correlated with the amount of effort for sample preparation and imaging (Table 2). Hence, with increasing complexity of the imaging method, the number of samples that can be measured in a given time decreases. Conducting a power analysis based on the expected variability in the measurements provides information on the number of samples required for a statistically sound analysis (13) and hence further restricts the choice of microscope.

Image analysis

Modern microscopy techniques generate large amounts of data that need to be processed and analyzed. The image analysis has to match the imaging technique and also provide the resulting data in a format that can be readily used as an input for mathematical modeling. Fiji/ImageJ and Icy are two main open-source image analysis platforms that combine a range of standard image analysis tools and advanced plugins in a user-friendly environment (14, 15). In addition, more specialized tools with varying degrees of user-friendliness exist. These are described below in the sections on the specific workflows.

TABLE 2 Imaging and modeling techniques

Imaging techniques

Research question	Imaging technique	Readout	Experience	Exemplary reference
Aggregation, spheroid growth	Wide field	2D projected area	Multi-sample imaging: ✓ Live imaging: ✓ (long-term) No staining necessary: ✓ Sample preparation: easy	(24)
Viability, differentiation	2D imaging of fluorescent or histological section	2D single-cell data of single slice	Multi-sample imaging: ✓ Live imaging: - No staining necessary: - Sample preparation: standard	(29, 30)
Aggregation, rearrangement, sorting, adhesion, viability, and differentiation	Confocal 3D imaging	3D single-cell data	Multi-sample imaging: ✓ Live imaging: ✓ (short-term) No staining necessary: - Sample preparation: advanced	(22)
Aggregation, rearrangement, sorting, adhesion, viability, and differentiation	Light-sheet 3D imaging	3D single-cell data	Multi-sample imaging: - Live imaging: ✓ (long-term) No staining necessary: - Sample preparation: advanced	(17)
Adhesion	Electron microscopy	Subcellular structures	Multi-sample imaging: - Live imaging: - No staining necessary: - Sample preparation: advanced	(10)

Modeling techniques

Research question	Modeling technique	Readout	Experience	Exemplary reference
Spheroid growth	Continuous	Whole spheroid measures including shape	+ Fastest – No single-cell information	(56)
Shape and neighbor changes during aggregation, cellular rearrangement, and sorting; viability, adhesion, differentiation	Cellular Potts model	Typically 2D single-cell data, including shape	+ Allows for complex shapes – Computationally slow, cells restricted to grid	(71)
Positional changes during aggregation, rearrangement, and sorting; viability, adhesion, differentiation	Centroid model	3D single-cell data excluding shape	+ Flexible, convenient for 3D – No shape information, neighborhood has to be approximated	(62)

Modeling techniques

The modeling techniques that have been used in spheroid research so far are distinguished by the details they provide (Table 2). Continuous models regard the spheroid as a whole, neglecting cellular details. The relevant parameters that are typically modeled are shape and size of the whole spheroid over time. Agent-based models, or individual cell-based models as they are sometimes referred to, consider single cells and their respective properties. They are more difficult to implement and require more computational power than continuous models. Agent-based models can be divided into two categories: lattice-free (or off-lattice models) and lattice-based models (16). A common variant of lattice-free models is the centroid model (Figures 2–4). Each cell is assumed to be spherical and defined by its position in space. The cells can have different attributes like radius and protein expression levels. Cell division and cell death can be implemented but the cell shape is neglected. Therefore, the cell neighborhood relations have to be approximated by cell graphs (17).

In lattice-based models, the cells are restricted to a grid. In some variants, a cell is represented by a single grid point on a lattice, and cell movement is implemented as changing from the current grid point to an adjacent grid point. Other lattice-based models depict cells as a cluster of grid points that share the same identifier (Figures 3 and 4). Hence, these models also provide the shape of a cell in addition to its position. A popular representative of this class of models is the Cellular Potts Model.

Several of the modeling approaches have formed the basis for simulation software. Lattice-free software packages like CellSys (18), lattice-based packages such as CompuCell3D (19), and software packages that combine both lattice-based and lattice-free methods, for example, Chaste (20) are promising tools for

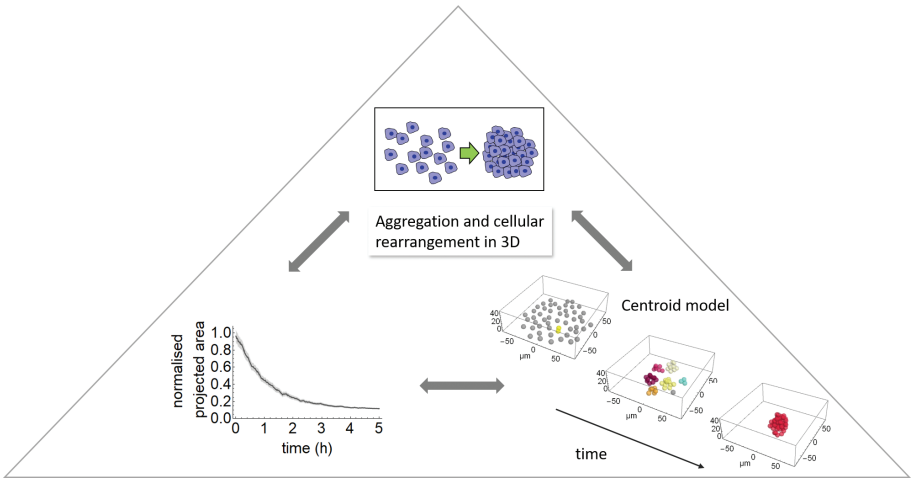


Figure 2 Illustration of the cell biology workflow. Experiments on spheroid formation are combined with quantification of the projected area of the spheroid over time and an agent-based centroid model. The results provide insight into aggregation and cellular rearrangement in three spatial dimensions.

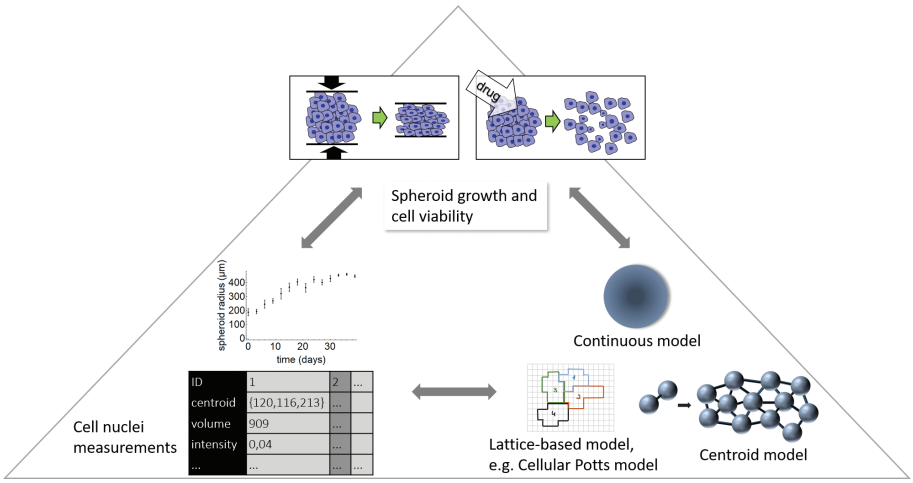


Figure 3 Illustration of the cancer biology workflow. In cancer biology, spheroid growth dynamics, and cell viability are of particular interest. The experiments involve mechanical or chemical perturbations of fully formed spheroids. The quantitative imaging provides global measurements of the spheroid like the radius over time or more detailed single-cell nuclei measurements, depending on the microscopy method employed (see Table 2 for more details). Modeling approaches include continuous models as well as the two types of agent-based models: lattice-based models and centroid models (see Table 2 for more details).

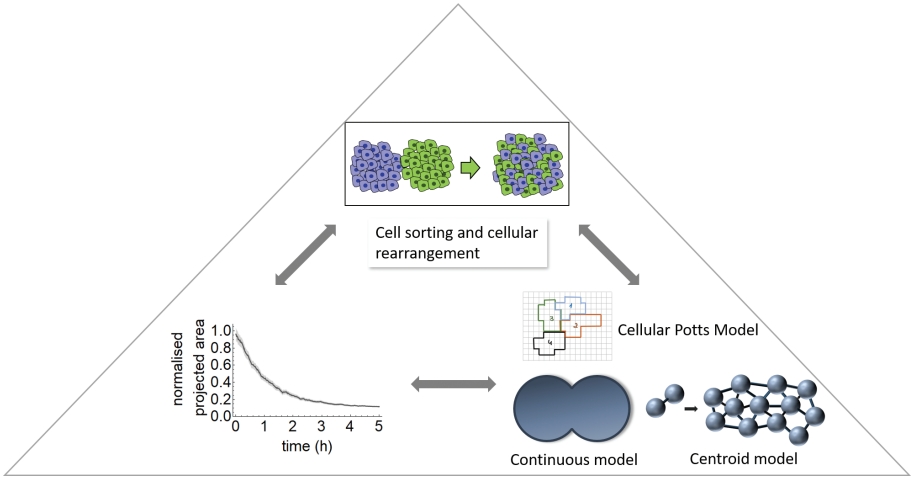


Figure 4 Illustration of the bioprinting workflow. The main spheroid-related question in bioprinting is how cell sorting and cellular rearrangement result in efficient fusion of multiple spheroids. Imaging of spheroid fusion provides the projected area over time. These results have been linked to mathematical models ranging from a continuous approach to agent-based models such as the Cellular Potts model and the centroid model.

implementing individual-based spheroid models. For a given biological question, these packages can form the basis of a model, which then needs to be adapted and extended depending on the required detail.

FIRST WORKFLOW: CELL BIOLOGY

Adhesion-based intercellular interactions and cellular rearrangement play an essential role during tissue development and maintenance. Even though both have been extensively studied in two-dimensional cell cultures, there are still a number of open questions regarding the mechanisms active in the three-dimensional context of a tissue. A systems biology approach based on spheroid formation provides the means to study adhesion and cellular rearrangement in a three-dimensional context. Even though these processes certainly vary between cell types, adhesion molecules, such as cadherins and integrins, as well as cytoskeletal components, like actin and microtubules, play a central role. For the experiments, spheroid formation of different cell types (21, 22) or the same cell type under different conditions is monitored. Potential variations for the same cell type are the expression of different binding proteins (23, 24) or the application of adhesion molecule functional blocking antibodies (21). Studying spheroid formation of cells transfected with wild type or mutant forms of N-cadherin revealed that different cadherin binding sites are responsible for different cell adhesion mechanisms such as the initial binding and the stabilization of an adherence junction. The integrity of spheroids of breast cell lines with different metastatic potential relies on the differential contribution of cadherins, actin, microtubules, and focal adhesion kinase (FAK). In particular, E- or N-cadherin, actin, and microtubules drive the spontaneous aggregation and compaction of the spheroids (21, 22). Breast tumor cell lines that require addition of reconstituted basement membrane (Matrigel) for spheroid formation rely on integrin for correct aggregation (21). The activity of FAK correlates with the metastatic potential of the breast cells (22).

Quantitative imaging data

Visualization of spheroid formation is typically achieved by time-lapse imaging with a wide field or fluorescence microscope. The image acquisition is fast and multi-sample imaging is readily available. This approach provides images of the projected area of the spheroid over time. For the analysis of time-lapse transmission wide field images of spheroid formation, several approaches have been implemented. Saias and colleagues have developed a high-throughput method to monitor and quantify cell aggregation dynamics of colon cancer cells (25). It is based on the segmentation of the projected area of a spheroid. Spheroid edges are identified within the z-projection of the fluorescence image and within a single plane of the transmission image by detecting discontinuities in brightness. The detected spheroid boundary is used to track the spheroid over time. An alternative approach is based on applying a filter with a large kernel to time-lapse spheroid images with fluorescently labeled nuclei and a subsequent binarization (22). This approach does not require a transmission image. If only a transmission image is available, the machine-learning based software ilastik (26) provides a user-friendly environment that is readily applicable (24). Based on manually labeled ground truth data, the algorithm is able to distinguish the spheroid from the background at each time point. Application of ilastik does not require machine-learning expertise. All three segmentation approaches result in time series of the projected area during spheroid formation (Figure 2).

Spatial mathematical modeling

For the modeling, spheroid formation has been represented as three-dimensional cell aggregation (Figure 2) (22, 24). A centroid model has been used in which the cells are assumed to be spheres with a given radius and are defined by their position in space. Cells accumulate to form clusters. Cluster formation occurs through cell–cell binding, cell–cluster binding, and cluster–cluster binding. Separation of cells from clusters is also possible. All parameter values are obtained from experimental measures, except for the density difference between cell and medium, which determines the sinking behavior of the cells in the medium as well as the binding and unbinding probabilities. These parameter values are established by fitting the projected area obtained from the model to the experimental data. Relating the resulting binding and unbinding probabilities to the different experimental conditions (cell type, adhesion molecules present, perturbation with antibodies) reveals the effect of the different perturbations on the cellular binding capacities during spheroid formation.

SECOND WORKFLOW: CANCER BIOLOGY

The most common application of spheroids is in cancer research. Spheroids formed of tumor cells provide a useful model for avascular tumors. Spheroids with diameters above 400–500 μm or more than 30,000 cells establish a concentric cell layering, in which an outer rim of proliferating cells and a layer of quiescent cells surround a necrotic core (17, 27). The applications of spheroids have evolved from drug testing to studying fundamental questions underlying cancer biology (28). The governing question is how different kinds of perturbations affect tumor growth.

Of particular interest are treatments with radiation or drugs, or a combination of the two (29). Different oxygen and glucose concentrations in the medium have also provided effects on tumor spheroid growth dynamics (30). Equally important, but not as straightforward, are studies of cellular responses to mechanical perturbations. Since a tumor is subjected to pressure from the surrounding tissue, this consequently confines the tumor, which is thought to affect the regulation of tumor growth (31). Cells are able to sense mechanical forces either directly by deformation or altered organization of intracellular compartments, such as the cytoskeleton (32) or the cell nucleus (33, 34), or by mechanoreceptors, which transduce the physical into biochemical or electrical signals (35).

Working with spheroids as a tumor model allows to investigate the influence of forces, which mainly depend on the physical properties of the cells and the extracellular matrix. Several studies have explored the role of mechanical stress on spheroid morphology, cell proliferation, and apoptosis, predominantly in the context of cancer research. Different methods exist to apply pressure on spheroids and to quantify the degree of pressure, time, and the treatment with anti-cancer drugs (36, 37).

One option to apply compressive stress on spheroids is an embedment in hydrogel of varying stiffness. These are, for example, composed of protein, agarose, polyacrylamide or polyethylene glycol (36, 38–41). Another option is

incubating a spheroid inside a dialysis bag and applying osmotic pressure with exteriorly added dextran. Long-term compressive stress leads to reduced or inhibited cell proliferation and induction of apoptosis in colon and breast carcinoma cells (37, 42, 43). Tube-like silicone device also provides a means to confine spheroid growth. Spheroids generated from colorectal cancer cells, which grow inside the device, adapt a rod-like shape. The number of mitotic cells increases, but they exhibit spindle defects and enter mitotic arrest upon confinement (44). Finally, physical confinement on growing spheroids has also been applied by encapsulation in alginate shells. Spheroids generated from mouse colon carcinoma cells show increased cell density and altered cellular organization, and cell proliferation is restricted to the outer rim of the spheroid when compressive stress is applied. An increased number of dead cells occurs in the center of the spheroid (31).

Quantitative imaging data

Growth curves of fully formed spheroids are usually obtained by wide field time-lapse imaging, followed by a segmentation of the projected spheroid area or measuring the spheroid radius over time. These time series measurements are complemented by measurements for each individual cell nucleus including its position, size, and intensity of markers, for example, for cell viability. Images of histological sections (29) or fluorescently stained sections from the spheroid center (30) evaluated by standard nuclei segmentation methods provide this information. To obtain the complete three-dimensional information of the cellular distribution, small spheroids can be imaged in toto with a confocal microscope (22). Larger samples require the combination of optical clearing methods with light sheet microscopy (17, 45). For the three-dimensional segmentation of nuclei images, intensity-based methods as well as shape-based methods have been proposed (17, 46–50). From these measurements, cell density measurements of the concentric layering can be extracted (17, 51).

Spatial mathematical modeling

The growth dynamics of fully formed spheroids under different conditions have been a major focus of mathematical modeling since the early 70s (52). Over the years, different strategies for spheroid modeling have emerged. Continuum models consider a spheroid as one entity, while agent-based models focus on single cells and their interactions. They are either implemented as lattice-based models, in which the positions of the cells are restricted by a lattice, or as lattice-free approaches (Figure 3).

A lattice-based model that reproduces spheroid growth dynamics was introduced by Radszuweit and colleagues (53). Recently, the model has been extended for a detailed analysis of the behavior of individual cells in a tumor spheroid of non-small cell lung cancer (NSCLC) cell line (30). Based on the images of two-dimensional spheroid sections, the distributions of dividing cells, necrotic cells, and the extracellular matrix along the radial direction into the spheroid have been quantified for different glucose levels and spheroids of different ages. An iterative refinement of the model to fit the experimental data has revealed a detailed

picture of the effect of growth promoters, growth inhibitors, viability promoters, and inhibitors on the growth dynamics of NSCLC spheroids.

Two different centroid models have been developed simultaneously to study the spatio-temporal growth dynamics (54, 55). Drasdo and Höhme (54) have shown that nutrient limitation has only a small effect on the expansion velocity. It mainly affects the size of the necrotic core. The relation of this agent-based model to a continuous model has been discussed (56). Schaller and Meyer-Hermann (55) have fitted the growth curve of their model spheroids to experimental growth curves to determine the ratios of oxygen and glucose uptake rates. Subsequently, the model has been adapted and expanded to study the effect of radiotherapy on tumor spheroids. The cells surviving the treatment exhibit a synchronization of their cell cycle, resulting in time windows of increased radiation sensitivity of the spheroid. Furthermore, reoxygenation occurs with specific timings upon radiotherapy, creating windows of drug treatment opportunities. Respecting the timings of both processes can increase therapy effectiveness (57–59).

The role of mechanical stress in spheroid growth has mostly been studied using continuum models. The models describe volumetric growth behavior of confined, avascular tumor spheroids (60) or the reorganization of cell aggregates following the release of a homogeneous compression (61). Loessner and colleagues have simulated the effect of both mechanical stimulus and different culturing conditions on spheroid growth (36). They have considered different matrix stiffness, culture timings, and drug treatments. Comparison of the modeling results with experimental data has shown a good agreement. The results on cell proliferation in a mechanically perturbed spheroid and spheroid growth influenced by external pressure mentioned above have been obtained by a systems biology approach including agent-based models (42, 43, 62).

THIRD WORKFLOW: BIOPRINTING

Bioprinting is emerging as an alternative to scaffold-based tissue engineering. One upcoming method is spheroid printing, which relies on pre-formed spheroids that are used as building blocks for tissue generation. The spheroids are dispensed in regular structures, and the engineered tissue emerges through spheroid fusion and maturation. Arranging the spheroids in a circle and subsequent spheroid fusion produces tissue rings (63) that have been proposed as building blocks for vascular trees (63, 64). A similar approach has further led to the formation of tubular structures (65, 66). To generate sheet-like structures of engineered adipose tissue, which one day could be used to regenerate the subcutaneous layer of the skin during reconstructive surgery, spheroids formed from adipose cells were placed in a melt electrowritten scaffold (67). After culturing the constructs for 14 days, a continuous tissue layer arose.

The success of these applications, and spheroid printing in general, relies on perfect spheroid fusion. One factor that influences the fusion is the pre-culture time of the single spheroids (68). Increased pre-culture time of spheroids inversely correlates with the fusion rate, suggesting the influence of cell–cell and cell–ECM contact maturation on spheroid-based tissue fusion. Furthermore, the pre-culture

time of spheroids influences cell-sorting processes that occur during tissue maturation. However, it remains elusive, which further factors control the positioning and collective migration and adhesion of spheroids to form intact and functional tissues.

Quantitative imaging data

Imaging techniques that have been used in the context of spheroid fusion are electron microscopy (67) as well as two-dimensional bright field and fluorescence images (68). Quantitative analysis of these images with ImageJ resulted in the temporal evolution of the size of the microtissue, which is generated by the fusion of the spheroids.

Spatial mathematical modeling

Mathematical modeling of tissue fusion started with the work on cell sorting by Glazier and Graner (69) that was based on the differential adhesion hypothesis by Steinberg (70). Implementing a Cellular Potts model, they determined the effect of different adhesion strengths on the sorting of cells. Sego et al. (71) combined such a Cellular Potts model with continuous diffusion modeling. This provides a representation of the behavior of individual cells as well as global characteristics of molecular-level phenomena. The model can reproduce cell sorting, spheroid fusion, and hole closure dynamics. Combining the cell-level dynamics, in particular cell survival with oxygen diffusion through a spheroid, reveals a sensitivity of the spheroid to externally applied oxygen. Applying external oxygen increases cell viability in the spheroid.

Yang et al. (72) employed a continuous model based on phase field theory to model the fusion of cellular aggregates into larger scale structures such as rings, Y-shapes or T-shapes. Spheroids that are located closely together fuse faster than the less densely packed arrangements. Gaps or errors in aggregate deposition can be directly linked to defects in the final biofabricated tissue construct.

To investigate spheroid fusion in three spatial dimensions, Flenner et al. (73) implemented two agent-based centroid models, a kinetic Monte Carlo method and a cellular particle dynamics method. The outline of two fusing aggregates is well represented by both methods. However, they find that the two simulations show clear differences with respect to the speed of cellular rearrangement. For the kinetic Monte Carlo method, fast movement of individual cells, and hence fast rearrangement, results in a complete cell mixing upon tissue fusion. For the cellular particle dynamics model, the fused tissue still exhibits distinct clusters of the different initial cell types. Experimental data to distinguish between the two results are not available so far. Kinetic Monte Carlo simulations have been extended to tube formation as well as T- and Y-shaped arrangements and the development of vascular tree structures (74, 75). The authors considered uniluminal spheroids as well as heterogeneous spheroids formed from a mixture of cells. They show that geometrically this can work; the question is how the cells survive in such structures. Also, they found that timesaving due to tight packing of the initial configuration of the spheroids is negligible compared to the time the system takes for fusion to steady state.

OUTLOOK: EXPERIMENTS INSPIRED BY THE MATHEMATICAL MODELS

The three exemplary research areas for spheroids have shown that these cell aggregates are a widely applicable in vitro system. Furthermore, major drawbacks of the current state of systems biology approaches for spheroids have become apparent. For all three research areas, agent-based models exist. In the centroid models, the properties of each cell including its position, marker expression, and its connection to neighboring cells are known at each time point. However, in most cases, the quantitative imaging data for testing these detailed predictions from the models are missing. A step further would be to also consider the three-dimensional shape of a cell in a spheroid. Three-dimensional Cellular Potts model exist to tackle this question, but again the experimental data are missing.

Confocal or light-sheet imaging in combination with innovative sample preparation methods (45, 76) and staining protocols (77) can provide the necessary images. Evaluating these images with single-cell-based segmentation and a subsequent analysis with neighborhood graphs (17) can provide the necessary data to refine the complex models. This will provide detailed insight into the spatial interactions of cells in spheroids.

So far, all experimental approaches focus on the spheroid as a closed unit. However, the modeling of spheroid fusion by Fenner et al. raised the question of the cellular dynamics within the spheroid. Do cells in spheroids move and how fast do they rearrange? Three-dimensional live imaging of individual cells in spheroids can provide insight, and the application of the existing cell tracking approaches (78) can yield the quantitative data to validate the mathematical models.

Most applications of spheroids still focus on spheroid growth or intercellular adhesion mainly in the context of cancer. However, these are not the only processes that are different between two-dimensional and three-dimensional cell cultures. Various proteins including keratin, vimentin, heat shock proteins, chaperons, and proteins involved in glucose metabolism have been shown to be differently expressed in two-dimensional cell culture versus three-dimensional cell culture (79). Therefore, the next steps are to address other cellular processes like cell polarization and cell differentiation and to address further diseases apart from cancer.

CONCLUSION

Systems biology approaches are slowly evolving towards spheroid research. In many studies, the different parts (Figure 1) are still well separated. Further efforts have to be made that integrate experiments, quantitative imaging, and mathematical modeling into a whole. This requires close interactions between experts from different disciplines including biology, medicine, physics, mathematics, and computer science. The most integral part of these interactions is a good communication, a common language or the interest to learn and understand the other's language. The collaboration has to start when developing the scientific question.

This ensures that the experimental and theoretical methods that are applied match well and are adequate to address the question of interest.

For some research questions, it might be necessary to adapt the three-dimensional cell culture system. Apart from spheroids, cysts are useful to study cell polarization in epithelia. Embryonic stem cell aggregates like ICM organoids (80), blastoids (81) or gastruloids (82) allow the investigation of cell differentiation during early mammalian embryogenesis. Three-dimensional multicellular structures grown from more specialized stem cells are typically called organoids. They consist of organ-specific cell types and are employed to mimic a variety of human tissues including brain, lung, liver, intestine, kidney, and pancreas. Applications range from studying fundamental questions of organ development and diseases to toxicity testing and personalized medicine (83). The main concepts introduced in this review are readily extendable to these other types of three-dimensional cell culture systems. In all cases, a major effort has to be put on combining information from different sources with the spatial distribution of the cells within the multicellular system.

Acknowledgements: The author thanks Isabell Smyrek, Biena Mathew, Katharina Hötte, Ernst H.K. Stelzer and Silvia Muñoz-Descalzo for their fruitful discussions and valuable inputs, and Ezgi Eylül Bankoglu and Simon Schardt for critical reading of the manuscript.

Conflict of interest: The author declares no potential conflicts of interest with respect to research, authorship, and/or publication of this chapter.

REFERENCES

1. Duval K, Grover H, Han L-H, Mou Y, Pegoraro AF, Fredberg J, et al. Modeling physiological events in 2D vs. 3D cell culture. Physiology. 2017 Jul;32(4):266–77. http://dx.doi.org/10.1152/physiol.00036.2016
2. Sutherland R. Cell and environment interactions in tumor microregions: the multicell spheroid model. Science. 1988 Apr 8;240(4849):177–84. http://dx.doi.org/10.1126/science.2451290
3. Figge MT, Murphy RF. Image-based systems biology. Cytometry A. 2015;87(6):459–61. http://dx.doi.org/10.1002/cyto.a.22663
4. Hirschhaeuser F, Menne H, Dittfeld C, West J, Mueller-Klieser W, Kunz-Schughart LA. Multicellular tumor spheroids: An underestimated tool is catching up again. J Biotechnol. 2010 Jul 1;148(1):3–15. http://dx.doi.org/10.1016/j.jbiotec.2010.01.012
5. Sirenko O, Mitlo T, Hesley J, Luke S, Owens W, Cromwell EF. High-content assays for characterizing the viability and morphology of 3D cancer spheroid cultures. ASSAY Drug Dev Technol. 2015 Sep;13(7):402–14. http://dx.doi.org/10.1089/adt.2015.655
6. Smithpeter CL, Dunn AK, Welch AJ, Richards-Kortum R. Penetration depth limits of in vivo confocal reflectance imaging. Appl Opt. 1998;37(13):2749–54. http://dx.doi.org/10.1364/AO.37.002749
7. Stelzer EHK. Light-sheet fluorescence microscopy for quantitative biology. Nat Methods. 2015 Jan;12(1):23–6. http://dx.doi.org/10.1038/nmeth.3219

8. Pampaloni F, Chang B-J, Stelzer EHK. Light sheet-based fluorescence microscopy (LSFM) for the quantitative imaging of cells and tissues. Cell Tissue Res. 2015 Apr;360(1):129–41. http://dx.doi.org/10.1007/s00441-015-2144-5
9. Swoger J, Verveer P, Greger K, Huisken J, Stelzer EHK. Multi-view image fusion improves resolution in three-dimensional microscopy. Optic Express. 2007 Jun 25;15(13):8029–42. http://dx.doi.org/10.1364/OE.15.008029
10. Almarshad HA, Madhavan M, Hoshino K. Focused ion beam-based milling, imaging and analysis of 3D tumor spheroids. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). 2018 Jul;2018:4480–4483. http://dx.doi.org/10.1109/EMBC.2018.8513165
11. Zhang J, Whitehead J, Liu Y, Yang Q, Leach JK, Liu G. Direct observation of tunneling nanotubes within human mesenchymal stem cell spheroids. J Phys Chem B. 2018 Nov 1;122(43):9920–6. http://dx.doi.org/10.1021/acs.jpcb.8b07305
12. Böttcher B. Transmission electron microscopy: Preparation of specimens. In: John Wiley & Sons, Ltd, editor. eLS. Chichester, UK: John Wiley & Sons, Ltd, 2012; p. a0002998.pub2.
13. Quinn GP, Keough MJ. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press, 2002; p. 537.
14. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: An open-source platform for biological-image analysis. Nat Methods. 2012 Jul;9(7):676–82. http://dx.doi.org/10.1038/nmeth.2019
15. de Chaumont F, Dallongeville S, Chenouard N, Hervé N, Pop S, Provoost T, et al. Icy: an open bioimage informatics platform for extended reproducible research. Nat Methods. 2012 Jul;9(7):690–6. http://dx.doi.org/10.1038/nmeth.2075
16. Glen CM, Kemp ML, Voit EO. Agent-based modeling of morphogenetic systems: Advantages and challenges. PLoS Comput Biol. 2019 Mar 28;15(3):e1006577. http://dx.doi.org/10.1371/journal.pcbi.1006577
17. Schmitz A, Fischer SC, Mattheyer C, Pampaloni F, Stelzer EHK. Multiscale image analysis reveals structural heterogeneity of the cell microenvironment in homotypic spheroids. Sci Rep. 2017 Dec;7(1):43693. http://dx.doi.org/10.1038/srep43693
18. Hoehme S, Drasdo D. A cell-based simulation software for multi-cellular systems. Bioinformatics. 2010 Oct 15;26(20):2641–2. http://dx.doi.org/10.1093/bioinformatics/btq437
19. Swat MH, Thomas GL, Belmonte JM, Shirinifard A, Hmeljak D, Glazier JA. Multi-scale modeling of tissues using CompuCell3D. In: Methods in cell biology. Elsevier; 2012;110:325–366 https://doi.org/10.1016/B978-0-12-388403-9.00013-8
20. Mirams GR, Arthurs CJ, Bernabeu MO, Bordas R, Cooper J, Corrias A, et al. Chaste: An open source C++ library for computational physiology and biology. PLoS Comput Biol. 2013 Mar 14;9(3):e1002970. http://dx.doi.org/10.1371/journal.pcbi.1002970
21. Ivascu A, Kubbies M. Diversity of cell-mediated adhesions in breast cancer spheroids. Int J Oncol. 2007 Dec 1;31:1403–13. http://dx.doi.org/10.3892/ijo.31.6.1403
22. Smyrek I, Mathew B, Fischer SC, Lissek SM, Becker S, Stelzer EHK. E-cadherin, actin, microtubules and FAK dominate different spheroid formation phases and important elements of tissue integrity. Biol Open. 2018 Dec 21;8:10.1242/bio.037051. http://dx.doi.org/10.1242/bio.037051
23. Bunse S, Garg S, Junek S, Vogel D, Ansari N, Stelzer EHK, et al. Role of N-cadherin *cis* and *trans* interfaces in the dynamics of adherens junctions in living cells. Hotchin NA, editor. PLoS ONE. 2013 Dec 2;8(12):e81517. http://dx.doi.org/10.1371/journal.pone.0081517
24. Garg S, Fischer SC, Schuman EM, Stelzer EHK. Lateral assembly of N-cadherin drives tissue integrity by stabilizing adherens junctions. J R Soc Interface. 2015 Mar 6;12(104):20141055. http://dx.doi.org/10.1098/rsif.2014.1055
25. Saias L, Gomes A, Cazales M, Ducommun B, Lobjois V. Cell-cell adhesion and cytoskeleton tension oppose each other in regulating tumor cell aggregation. Cancer Res. 2015 Jun 15;75(12):2426–33. http://dx.doi.org/10.1158/0008-5472.CAN-14-3534
26. Sommer C, Straehle C, Kothe U, Hamprecht FA. Ilastik: Interactive learning and segmentation toolkit. In: 2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. Chicago, IL: IEEE, March 30–April 2, 2011. p. 230–3.

27. Kunz-Schughart LA, Freyer JP, Hofstaedter F, Ebner R. The use of 3-D cultures for high-throughput screening: The multicellular spheroid model. J Biomol Screen. 2004 Jun;9(4):273–85. http://dx.doi.org/10.1177/1087057104265040
28. Ravi M, Paramesh V, Kaviya SR, Anuradha E, Solomon FDP. 3D cell culture systems: Advantages and applications: 3D cell culture systems. J Cell Physiol. 2015 Jan;230(1):16–26. http://dx.doi.org/10.1002/jcp.24683
29. Mao X, McManaway S, Jaiswal JK, Patel PB, Wilson WR, Hicks KO, et al. An agent-based model for drug-radiation interactions in the tumour microenvironment: Hypoxia-activated prodrug SN30000 in multicellular tumour spheroids. PLOS Comput Biol. 2018 Oct 24;14(10):e1006469. http://dx.doi.org/10.1371/journal.pcbi.1006469
30. Jagiella N, Müller B, Müller M, Vignon-Clementel IE, Drasdo D. Inferring growth control mechanisms in growing multi-cellular spheroids of NSCLC cells from spatial-temporal image data. Byrne H, editor. PLoS Comput Biol. 2016 Feb 11;12(2):e1004412. http://dx.doi.org/10.1371/journal.pcbi.1004412
31. Alessandri K, Sarangi BR, Gurchenkov VV, Sinha B, Kiessling TR, Fetler L, et al. Cellular capsules as a tool for multicellular spheroid production and for investigating the mechanics of tumor progression in vitro. Proc Natl Acad Sci. 2013 Sep 10;110(37):14843–8. http://dx.doi.org/10.1073/pnas.1309482110
32. Luo T, Mohan K, Iglesias PA, Robinson DN. Molecular mechanisms of cellular mechanosensing. Nat Mater. 2013 Nov;12(11):1064–71. http://dx.doi.org/10.1038/nmat3772
33. Philip JT, Dahl KN. Nuclear mechanotransduction: Response of the lamina to extracellular stress with implications in aging. J Biomech. 2008 Nov;41(15):3164–70. http://dx.doi.org/10.1016/j.jbiomech.2008.08.024
34. Guilak F, Tedrow JR, Burgkart R. Viscoelastic properties of the cell nucleus. Biochem Biophys Res Commun. 2000;269(3):781–6. http://dx.doi.org/10.1006/bbrc.2000.2360
35. DuFort CC, Paszek MJ, Weaver VM. Balancing forces: Architectural control of mechanotransduction. Nat Rev Mol Cell Biol. 2011 May;12(5):308–19. http://dx.doi.org/10.1038/nrm3112
36. Loessner D, Flegg JA, Byrne HM, Clements JA, Hutmacher DW. Growth of confined cancer spheroids: A combined experimental and mathematical modelling approach. Integr Biol. 2013 Feb 25;5(3):597–605. http://dx.doi.org/10.1039/c3ib20252f
37. Delarue M, Montel F, Vignjevic D, Prost J, Joanny J-F, Cappello G. Compressive stress inhibits proliferation in tumor spheroids through a volume limitation. Biophys J. 2014 Oct;107(8):1821–8. http://dx.doi.org/10.1016/j.bpj.2014.08.031
38. Sieminski AL, Was AS, Kim G, Gong H, Kamm RD. The stiffness of three-dimensional ionic self-assembling peptide gels affects the extent of capillary-like network formation. Cell Biochem Biophys. 2007 Oct 1;49(2):73–83. http://dx.doi.org/10.1007/s12013-007-0046-1
39. Helmlinger G, Netti PA, Lichtenbeld HC, Melder RJ, Jain RK. Solid stress inhibits the growth of multicellular tumor spheroids. Nat Biotechnol. 1997 Aug;15(8):778. http://dx.doi.org/10.1038/nbt0897-778
40. Cheng G, Tse J, Jain RK, Munn LL. Micro-environmental mechanical stress controls tumor spheroid size and morphology by suppressing proliferation and inducing apoptosis in cancer cells. Blagosklonny MV, editor. PLoS One. 2009 Feb 27;4(2):e4632. http://dx.doi.org/10.1371/journal.pone.0004632
41. Paszek MJ, Zahir N, Johnson KR, Lakins JN, Rozenberg GI, Gefen A, et al. Tensional homeostasis and the malignant phenotype. Cancer Cell. 2005 Sep 1;8(3):241–54. http://dx.doi.org/10.1016/j.ccr.2005.08.010
42. Montel F, Delarue M, Elgeti J, Malaquin L, Basan M, Risler T, et al. Stress clamp experiments on multicellular tumor spheroids. Phys Rev Lett. 2011 Oct 24;107(18):188102. http://dx.doi.org/10.1103/PhysRevLett.107.188102
43. Montel F, Delarue M, Elgeti J, Vignjevic D, Cappello G, Prost J. Isotropic stress reduces cell proliferation in tumor spheroids. New J Phys. 2012 May 9;14(5):055008. http://dx.doi.org/10.1088/1367-2630/14/5/055008
44. Desmaison A, Frongia C, Grenier K, Ducommun B, Lobjois V. Mechanical stress impairs mitosis progression in multi-cellular tumor spheroids. Engler AJ, editor. PLoS One. 2013 Dec 3;8(12):e80447. http://dx.doi.org/10.1371/journal.pone.0080447

45. Hoette K, Koch M, Hof L, Tuppi M, Moreth T, Stelzer EHK, et al. Ultra-thin fluorocarbon foils optimize multiscale imaging of three-dimensional native and optically cleared specimens. bioRxiv. 2019 Feb 12. http://dx.doi.org/10.1101/533844
46. Mathew B, Schmitz A, Muñoz-Descalzo S, Ansari N, Pampaloni F, Stelzer EHK, et al. Robust and automated three-dimensional segmentation of densely packed cell nuclei in different biological specimens with Lines-of-Sight decomposition. BMC Bioinformatics. 2015 Dec;16(1):187. http://dx.doi.org/10.1186/s12859-015-0617-x
47. Friebel A, Neitsch J, Johann T, Hammad S, Hengstler JG, Drasdo D, et al. TiQuant: Software for tissue analysis, quantification and surface reconstruction. Bioinformatics. 2015 Oct 1;31(19):3234–6. http://dx.doi.org/10.1093/bioinformatics/btv346
48. Morales-Navarrete H, Segovia-Miranda F, Klukowski P, Meyer K, Nonaka H, Marsico G, et al. A versatile pipeline for the multi-scale digital reconstruction and quantitative analysis of 3D tissue architecture. eLife. 2015 Dec 27;4:e11214. http://dx.doi.org/10.7554/eLife.11214
49. McQuin C, Goodman A, Chernyshev V, Kamentsky L, Cimini BA, Karhohs KW, et al. CellProfiler 3.0: Next-generation image processing for biology. Misteli T, editor. PLoS Biol. 2018 Jul 3;16(7):e2005970. http://dx.doi.org/10.1371/journal.pbio.2005970
50. Lou X, Kang M, Xenopoulos P, Muñoz-Descalzo S, Hadjantonakis A-K. A rapid and efficient 2D/3D nuclear segmentation method for analysis of early mouse embryo and stem cell image data. Stem Cell Rep. 2014 Mar 11;2(3):382–97. http://dx.doi.org/10.1016/j.stemcr.2014.01.010
51. Dini S, Binder BJ, Fischer SC, Mattheyer C, Schmitz A, Stelzer EHK, et al. Identifying the necrotic zone boundary in tumour spheroids with pair-correlation functions. J R Soc Interface. 2016 Oct 1; 13(123):20160649. http://dx.doi.org/10.1098/rsif.2016.0649
52. Byrne HM. Dissecting cancer through mathematics: from the cell to the animal model. Nat Rev Cancer. 2010 Mar;10(3):221–30. http://dx.doi.org/10.1038/nrc2808
53. Radszuweit M, Block M, Hengstler JG, Schöll E, Drasdo D. Comparing the growth kinetics of cell populations in two and three dimensions. Phys Rev E. 2009 May 12;79(5):051907. http://dx.doi.org/10.1103/PhysRevE.79.051907
54. Drasdo D, Hōhme S. A single-cell-based model of tumor growthin vitro: monolayers and spheroids. Phys Biol. 2005 Jul;2(3):133–147. http://dx.doi.org/10.1088/1478-3975/2/3/001
55. Schaller G, Meyer-Hermann M. Multicellular tumor spheroid in an off-lattice Voronoi-Delaunay cell model. Phys Rev E. 2005 May 27;71(5):051910. http://dx.doi.org/10.1103/PhysRevE.71.051910
56. Byrne H, Drasdo D. Individual-based and continuum models of growing cell populations: A comparison. J Math Biol. 2008 Oct 8;58(4):657. http://dx.doi.org/10.1007/s00285-008-0212-0
57. Kempf H, Bleicher M, Meyer-Hermann M. Spatio-temporal cell dynamics in tumour spheroid irradiation. Eur Phys J D. 2010 Oct;60(1):177–93. http://dx.doi.org/10.1140/epjd/e2010-00178-4
58. Kempf H, Hatzikirou H, Bleicher M, Meyer-Hermann M. In silico analysis of cell cycle synchronisation effects in radiotherapy of tumour spheroids. Alber MS, editor. PLoS Comput Biol. 2013 Nov 14; 9(11):e1003295. http://dx.doi.org/10.1371/journal.pcbi.1003295
59. Kempf H, Bleicher M, Meyer-Hermann M. Spatio-temporal dynamics of hypoxia during radiotherapy. Rocha S, editor. PLoS One. 2015 Aug 14;10(8):e0133357. http://dx.doi.org/10.1371/journal.pone.0133357
60. Ciarletta P, Ambrosi D, Maugin GA, Preziosi L. Mechano-transduction in tumour growth modelling. Eur Phys J E. 2013 Mar;36(3):23. http://dx.doi.org/10.1140/epje/i2013-13023-2
61. Giverso C, Preziosi L. Modelling the compression and reorganization of cell aggregates. Math Med Biol. 2012 Jun 1;29(2):181–204. http://dx.doi.org/10.1093/imammb/dqr008
62. Liedekerke PV, Neitsch J, Johann T, Alessandri K, Nassoy P, Drasdo D. Quantitative agent-based modeling reveals mechanical stress response of growing tumor spheroids is predictable over various growth conditions and cell lines. PLoS Comput Biol. 2019 Mar 8;15(3):e1006273. http://dx.doi.org/10.1371/journal.pcbi.1006273
63. Mironov V, Visconti RP, Kasyanov V, Forgacs G, Drake CJ, Markwald RR. Organ printing: Tissue spheroids as building blocks. Biomaterials. 2009 Apr;30(12):2164–74. http://dx.doi.org/10.1016/j.biomaterials.2008.12.084

64. Patra S, Young V. A review of 3D printing techniques and the future in biofabrication of bioprinted tissue. Cell Biochem Biophys. 2016 Jun;74(2):93–8. http://dx.doi.org/10.1007/s12013-016-0730-0
65. Yurie H, Ikeguchi R, Aoyama T, Kaizawa Y, Tajino J, Ito A, et al. The efficacy of a scaffold-free Bio 3D conduit developed from human fibroblasts on peripheral nerve regeneration in a rat sciatic nerve model. Nógrádi A, editor. PLoS One. 2017 Feb 13;12(2):e0171448. http://dx.doi.org/10.1371/journal.pone.0171448
66. Taniguchi D, Matsumoto K, Tsuchiya T, Machino R, Takeoka Y, Elgalad A, et al. Scaffold-free trachea regeneration by tissue engineering with bio-3D printing. Interact Cardiovasc Thorac Surg. 2018 May 1; 26(5):745–52. http://dx.doi.org/10.1093/icvts/ivx444
67. McMaster R, Hoefner C, Hrynevich A, Blum C, Wiesner M, Wittmann K, et al. Tailored melt electrowritten scaffolds for the generation of sheet-like tissue constructs from multicellular spheroids. Adv Healthc Mater. 2019 Apr 1;8(7):1801326. http://dx.doi.org/10.1002/adhm.201801326
68. Rago AP, Dean DM, Morgan JR. Controlling cell position in complex heterotypic 3D microtissues by tissue fusion. Biotechnol Bioeng. 2009;102(4):1231–41. http://dx.doi.org/10.1002/bit.22162
69. Glazier JA, Graner F. Simulation of the differential adhesion driven rearrangement of biological cells. Phys Rev E. 1993 Mar 1;47(3):2128–54. http://dx.doi.org/10.1103/PhysRevE.47.2128
70. Steinberg MS. Reconstruction of tissues by dissociated cells. Science. 1963 Aug 2;141(3579):401–8. http://dx.doi.org/10.1126/science.141.3579.401
71. Sego TJ, Kasacheuski U, Hauersperger D, Tovar A, Moldovan NI. A heuristic computational model of basic cellular processes and oxygenation during spheroid-dependent biofabrication. Biofabrication. 2017 Jun;9(2):024104. http://dx.doi.org/10.1088/1758-5090/aa6ed4
72. Yang X, Mironov V, Wang Q. Modeling fusion of cellular aggregates in biofabrication using phase field theories. J Theor Biol. 2012 Jun;303:110–8. http://dx.doi.org/10.1016/j.jtbi.2012.03.003
73. Flenner E, Janosi L, Barz B, Neagu A, Forgacs G, Kosztin I. Kinetic Monte Carlo and cellular particle dynamics simulations of multicellular systems. Phys Rev E. 2012 Mar 8;85(3):031907. http://dx.doi.org/10.1103/PhysRevE.85.031907
74. Sun Y, Wang Q. Modeling and simulations of multicellular aggregate self-assembly in biofabrication using kinetic Monte Carlo methods. Soft Matter. 2013 Jul 31;9(7):2172–86. http://dx.doi.org/10.1039/c2sm27090k
75. Sun Y, Yang X, Wang Q. In-silicoanalysis on biofabricating vascular networks using kinetic Monte Carlo simulations. Biofabrication. 2014 Jan;6(1):015008. http://dx.doi.org/10.1088/1758-5082/6/1/015008
76. Jeandupeux E, Lobjois V, Ducommun B. 3D print customized sample holders for live light sheet microscopy. Biochem Biophys Res Commun. 2015 Aug;463(4):1141–3. http://dx.doi.org/10.1016/j.bbrc.2015.06.072
77. Smyrek I, Stelzer EHK. Quantitative three-dimensional evaluation of immunofluorescence staining for large whole mount spheroids with light sheet microscopy. Biomed Opt Express. 2017 Feb 1; 8(2):484–99. http://dx.doi.org/10.1364/BOE.8.000484
78. Tinevez J-Y, Perry N, Schindelin J, Hoopes GM, Reynolds GD, Laplantine E, et al. TrackMate: An open and extensible platform for single-particle tracking. Methods. 2017 Feb 15;115:80–90. http://dx.doi.org/10.1016/j.ymeth.2016.09.016
79. Pampaloni F, Stelzer EHK, Leicht S, Marcello M. Madin-Darby canine kidney cells are increased in aerobic glycolysis when cultured on flat and stiff collagen-coated surfaces rather than in physiological 3-D cultures. Proteomics. 2010 Aug 17;10(19):3394–413. http://dx.doi.org/10.1002/pmic.201000236
80. Mathew B, Muñoz-Descalzo S, Corujo-Simon E, Schröter C, Stelzer EHK, Fischer SC. Mouse ICM organoids reveal three-dimensional cell fate clustering. Biophys J. 2019 Jan 8;116:127–41. http://dx.doi.org/10.1016/j.bpj.2018.11.011
81. Rivron NC, Frias-Aldeguer J, Vrij EJ, Boisset J-C, Korving J, Vivié J, et al. Blastocyst-like structures generated solely from stem cells. Nature. 2018 May;557(7703):106. http://dx.doi.org/10.1038/s41586-018-0051-0
82. Beccari L, Moris N, Girgin M, Turner DA, Baillie-Johnson P, Cossy A-C, et al. Multi-axial self-organization properties of mouse embryonic stem cells into gastruloids. Nature. 2018 Oct;562(7726):272. http://dx.doi.org/10.1038/s41586-018-0578-0
83. Clevers H. Modeling development and disease with organoids. Cell. 2016 Jun;165(7):1586–97. http://dx.doi.org/10.1016/j.cell.2016.05.082

2

Integrative Biology Approaches Applied to Human Diseases

Alysson H. Urbanski[1] • José D. Araujo[1] • Rachel Creighton[2] • Helder I. Nakaya[1,3]

[1]Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, Brazil; [2]Department of Bioengineering, University of Washington, Seattle, WA, USA; [3]Scientific Platform Pasteur/USP, University of Sao Paulo, Sao Paulo, Brazil

Author for correspondence: Helder I. Nakaya, Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, SP, 05508, Brazil. Email: hnakaya@usp.br

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch2

Abstract: The study of multifactorial and complex interactions in human diseases has been transformed by the omics revolution. The speed and scale of omics analysis have increased exponentially in the past decades, and it is now easier and faster to generate large amounts of biological data. However, extracting meaningful information from this "sea of data" remains a major challenge. The field of integrative biology utilizes a holistic approach to integrate multilayer biological data. In this chapter, we introduce concepts and techniques for the analysis of single-layer omics data and for integrating multilayer omics datasets to extract meaningful and relevant biological insights. Integrative biology is a promising approach for the study of a wide range of human diseases. We also highlight some current challenges in the field, such as the need for more specialized and interpretable methods, while increasing the accessibility of integrative analysis for the scientific community.

Keywords: integrative biology; multi-omics; proteogenomics; single-layer high-throughput data; systems biology

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

Human diseases involve complex interactions between genes, environment and lifestyle (1). For example, in type 2 diabetes mellitus, there are many behavioral, lifestyle, and genetic risk factors and other pathophysiological abnormalities contributing to hyperglycemia. Major mechanisms of the disease are impaired insulin secretion and insulin resistance in muscle and liver; however, other genes and signaling pathways in different tissues are also involved, such as increased kidney malfunction, inflammation, and neurotransmitter dysfunction (2). Other well-known examples of complex, multigenic, or multifactorial diseases are tumors (3), infectious diseases (4), and cardiovascular diseases (5).

Life sciences research has been revolutionized in past decades by a series of genome-wide technologies, starting with the Human Genome Project in 1990. The speed and scale of genomics analysis increased exponentially after this, facilitated by technologies such as microarrays and high-throughput sequencing (6). Genomics is classified as discovery science, along with other omics such as transcriptomics, miRNAomics, epigenomics, cistromics, proteomics, metabolomics, and microbiomics. The goal of discovery science is to collect and store data describing all the elements of a system (6, 7). As it has become easier and faster to generate large amounts of biological data, new challenges in data analysis and interpretation are emerging (8).

High-throughput data allow us to visualize processes in a certain layer of biological information in an organism or at the single-cell level. A recent example is the association of CD177+ neutrophils to Kawasaki disease through genome-wide transcriptome analysis (9). Additionally, analyzing the metabolome of coronary atherosclerosis patients enabled discovery of several biomarkers of lipid metabolism dysfunctions (10). At a proteomic level, researchers have identified proteins in the brain which are associated with the cognitive trajectory in the elderly (11). Finally, the evolution of single-cell sequencing has allowed the evaluation of these different layers in greater detail (12). The analysis of omics data has advanced the understanding of human diseases, but it is important to remember that these studies represent only one layer of a more complex system.

Network science analyzes the interactions between biomolecules (proteins, RNA, gene sequences), pathways, cells, organs, and even individuals using graph theory methods, and it is an efficient way of extracting information from omics data. Through network analysis, it is possible to identify complex patterns among different components to generate scientific hypotheses regarding the interactions present in health and disease events (13). For example, a recent gene expression network analysis study identified a membrane receptor as a potential therapeutic target for an antiepileptic drug (14). Although the integration of genes into networks gives us a lot of information, it describes only one omics level. Therefore, there is a growing interest in the integration of different omics data (15). In this chapter, we introduce concepts and tools for the analysis of single-layer biological data and integration of multilayer biological data to extract meaningful and relevant biological insights of various human diseases (Figure 1).

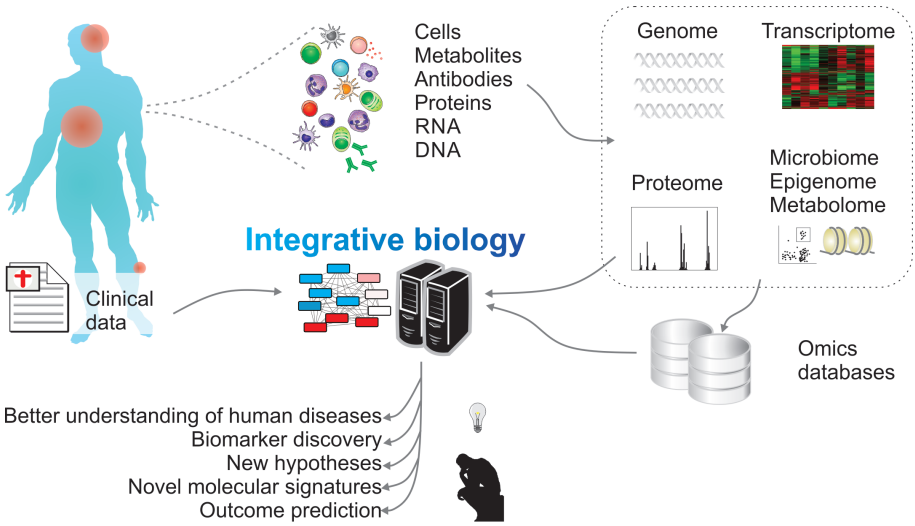


Figure 1 A framework for integrative biology. High-throughput techniques such as transcriptomics, proteomics and metabolomics, in addition to clinical data and other databases, can be used to investigate human diseases through an integrative approach.

APPLICATIONS OF SINGLE-LAYER HIGH-THROUGHPUT DATA

Since the popularization of next-generation sequencing (NGS) and high-throughput mass spectrometry methods, there has been an exponential increase in the generation of biological data, and it is likely that the amount of biological data available will continue to increase. The evolution of high-throughput mass spectrometry has enabled high-resolution visualization of the proteome and metabolome of cells, tissues, and fluids. These data are useful to understand the pathogenic mechanisms, contributing to diagnoses, prognoses, and potential therapeutic interventions.

DNA genomes and exomes can be elucidated using NGS. NGS-based techniques have already overcome the use of microarrays for RNA transcriptome sequencing by enabling the identification of virtually any transcript present in the sample, including unknown transcripts. NGS techniques can also identify differentially expressed genes (DEGs) by applying statistical methods to the expression data (16). Recently, long noncoding RNA (lnc-RNA) (17) and circular RNA (18) molecules have been implicated in the regulation of the innate immune response and can potentially elucidate infectious, autoimmune, and inflammatory disease mechanisms. Despite this, it is important to remember the limitations of studying a heterogeneous mixture of cells. Although the cells may be similar in morphology, localization or other classificatory factors, it is impossible to understand individual cellular features such as metabolic states, transcriptional levels, and metabolic activation using traditional bulk transcriptome sequencing (19).

Thus, RNA sequencing at single-cell level (scRNA-seq) allows a more accurate reconstruction of intracellular and intercellular network interactions (20). Since the first scRNA-seq a decade ago (21), the technology has improved and several protocols and platforms have been developed to respond to the most diverse biological problems, including those related to immune system in health and disease (22, 23). Recently, ultra-high-throughput scRNA-seq techniques based on the droplets strategy, such as Drop-Seq (24), InDrop (25), and 10X Genomics Chromium (26), have gained popularity. These techniques can reduce the cost of sequencing while increasing the throughput by allowing a parallel mRNA profiling of thousands of individual cells by encapsulating them in droplets (27). Raw and processed high-throughput data are stored in several online repositories, making them valuable resources for discovery science approaches (7). The content of the data repositories ranges from genomics and transcriptomics to epigenetics, protein–protein interaction, metabolomics, and microbiome data (Table 1).

Examples of big data generation in specific human disease applications are numerous. Although we do not focus on any specific disease in this chapter, we provide several relevant examples. Zhao et al. performed the transcriptomic profiling of glioma, generating 30 billion reads, from 325 samples in different stages of malignant progression (28). There have also been efforts to investigate in vitro and in vivo response to viral infections, such as influenza and severe acute respiratory syndrome coronavirus, generating dozens of transcriptome and proteome datasets (29). More specific events have also been investigated, such as the methylome of brain metastases that may help to predict individual responses to therapies (30) or the profiling of long non-coding RNA in human hypertrophic cardiomyopathy (31). Data generated from a large-scale multi-omic study, including genome and transcriptome sequencing and proteomic profiling of a large cohort of Alzheimer's disease patients, could improve our knowledge about this pathology (32). In another study, the characterization of *post-mortem* microbial diversity in 188 individuals allowed a better understanding of the *ante-mortem* health condition of some individuals, suggesting that it is possible to estimate the health conditions in living populations from these data (33).

TOOLS FOR THE ANALYSIS OF SINGLE-LAYER HIGH-THROUGHPUT DATA

Ensuring data quality is an essential step in the analysis and integration of omics data. When artifacts and noise are not handled correctly, they can influence the results of the analysis (34). The term "garbage in, garbage out," a common concept in computer science and mathematics, is also applicable in bioinformatics. This phrase means that the output data quality is determined by the input data quality. Several methods can be used to evaluate and control input data quality. One strategy is to determine the statistical significance to avoid false positives, known as the false discovery rate (FDR). Despite a recent debate about the appropriate use of statistical significance, an FDR value of 0.05 or smaller has been generally accepted in academia (35). In addition to the statistical analysis of individual layers, it is important to ensure that the data are biologically meaningful. In this case, the fold-change cut-off is used. The fold-change describes how

TABLE 1 Biological repositories

Database	Description	Reference website
ArrayExpress	Functional genomics data from microarray or NGS. Data types include transcription profiling (mRNA and miRNA), SNP genotyping, chromatin immunoprecipitation (ChIP), and comparative genomic hybridization	https://www.ebi.ac.uk/arrayexpress/
BioGRID	Curated database. Data types include protein–protein, genetic and chemical interactions, and post-translational modifications	https://thebiogrid.org/
dbGAP	Data and results from the interaction of genotype and phenotype	https://www.ncbi.nlm.nih.gov/gap/
ENCODE	Whole-genome database	https://encodeproject.org/
GDC	Genomic, epigenomic, transcriptomic, and proteomic data from cancer samples	https://portal.gdc.cancer.gov/
GEO	Gene expression, hybridization arrays, chips, and microarrays database	https://www.ncbi.nlm.nih.gov/geo/
GTEx	The genotype–tissue expression includes data of tissue-specific gene expression and regulation	https://gtexportal.org/home/
HMDB	Human metabolome database	http://www.hmdb.ca/
ICGC	Cancer genomics database	https://dcc.icgc.org/
IMGT	Immune-related genes sequence database	http://www.imgt.org/
InnateDB	Genes, proteins, interactions, and pathways involved in the innate immune response	https://www.innatedb.com/
MethylomeDB	DNA methylation profiles	http://habanero.mssm.edu/methylomedb/index.html
MGnify	Microbiome database	https://www.ebi.ac.uk/metagenomics/
miRbase	miRNA sequences and annotation	http://www.mirbase.org/
PHISTO	Pathogen–human protein–protein interaction database	http://www.phisto.org/
Reactome	Curated pathway database	https://reactome.org/
SRA	Sequencing and alignment data	https://www.ncbi.nlm.nih.gov/sra
STRING	Protein–protein interaction networks	https://string-db.org/

These databases store raw or processed, and sometimes curated, data derived from different studies and omics technologies.

much a gene or pathway is up- or down-regulated, for example, 2 or 0.5, respectively (36). This kind of analysis allows further downstream integration of the data, since it is possible to associate, for example, a group of DEGs and the metabolic pathways that they belong to (37).

Numerous tools are used to analyze different types of data. Although it is not the focus of this chapter to describe these tools, the concepts of some techniques are described here. Bioconductor is a robust software platform used in the analysis

of omics data (https://www.bioconductor.org/). In bioconductor, there are several packages, mainly in the R scripting language, that provide metrics and methods to evaluate reproducibility, identify outliers and noise. For example, the EdgeR package for gene expression analysis calculates the difference in gene expression for different samples and conditions, considering both the FDR and fold-change of each gene (38). Bioconductor can also be used to analyze high-dimensional mass cytometry (CyTOF) datasets. CyTOF is a platform for collecting high-dimensional phenotypic and functional data for single cells (39). For example, CyTOF can be used to uncover tissue- and disease-associated immune cell subsets (40). A review by Nowicka et al. presents a detailed workflow for CyTOF analyses using the bioconductor platform (41).

Metabolomics provides quantification of metabolites in cells, tissues or biological fluids (42). Several tools are available for the analysis of metabolomics data, including the web tool MetaboAnalyst (43) and the R package MetaboAnalystR (44). Both carry out analyses with the same workflow: (i) Exploratory data analysis; (ii) Metabolic enrichment analysis and metabolic pathway activity prediction; and (iii) Data integration, such as biomarker meta-analysis, joint path analysis, and network explorer. The data input for these tools can be a list of genes or KEGG orthologs.

Single-cell RNA-seq (scRNA-seq) methods are also widely used in studies involving human health (23). To ensure a biologically significant analysis, it is necessary to consider the intrinsic variations of the technique, called batch effects (45). There are several tools that assist in the batch correction process, most of which are based on linear regression, including limma (46), RUVseq (47, 48), and svaseq (49). Other promising approaches for batch correction are based on the detection of mutual nearest neighbors in the high-dimensional gene expression space (50).

The high-dimensional gene expression space is a matter of concern when analyzing scRNA-seq gene expression data. The problem with this high-dimensional space is that it is hard to differentiate the variability between cell populations from the variability between cells within a population, as the distances between cells become more homogenous. High-dimensional data are handled through dimensionality reduction and feature selection. Dimensionality reduction is a process to project data in a smaller dimensional space, preserving some key characteristics of the sample enough to distinguish differences between populations (51). While principal component analysis (PCA) is the recommended tool for RNA-seq, T-distributed stochastic neighbor embedding (tSNE) is the most popular method for dimensionality reduction of scRNA-seq data. PCA is not recommended for scRNA-seq datasets because it is a linear dimensionality reduction algorithm and assumes approximately normally distributed data, while tSNE uses different probability distributions that are more suitable to scRNA-seq data (51). Nonetheless, a recently developed nonlinear dimensionality-reduction technique named uniform manifold approximation and projection (UMAP) outperformed other dimensionality-reduction methods for cell clustering (52). Feature selection reduces the number of dimensions by excluding uninformative genes and identifying the most relevant features for analysis (53). Feature selection in scRNA-seq can be based on correlated expression, highly variable genes (HVG), Michaelis–Menten modeling of dropouts (M3Drop) or spike-in methods (51).

As already mentioned, scRNA-seq enables the identification of transcriptionally distinct cell subpopulations in an otherwise homogeneous cell population. Identification of these groups is typically accomplished through clustering analysis. Clustering approaches can be supervised or unsupervised. If the method uses a known set of gene markers for clustering, it is supervised. Alternatively, unsupervised clustering methods can identify groups without prior information (53). There are many algorithms designed for unsupervised clustering, but the main classes of them are k-means, hierarchical, density-based, and graph clustering (51). For example, through transcriptional clustering analysis of $CD127^{+}$ innate lymphoid cells (ILCs), Björklund et al. uncovered four different cell subpopulations: three different ILCs and natural killer (NK) cells. The group further subdivided the ILC3 group into three new transcriptionally and functionally distinct populations, contributing to the knowledge of ILC biology, and associated inflammatory processes (54).

Clustering analyses in scRNA-seq data can be very useful and informative, but they are not always able to describe dynamic biological processes involved in transitions between different states, such as cellular proliferation and maturation (12). Such events can be computationally modeled through the reconstruction of the cell trajectory and pseudotime estimation (53). Because the cells in a scRNA-seq experiment are unsynchronized, there are different instantaneous timepoints captured that together may represent an entire cell trajectory (55). The term pseudotime refers to an ordering of the cells according to some dynamic process of interest, such as development processes occurring over time. Through pseudotime estimation, cells in different states of a trajectory can be identified, permitting identification of transcriptional changes, branching points in trajectories, and reconstruction of gene regulatory networks (56). Recent efforts have used trajectory and pseudotime methods to better understand human diseases, including hepatitis B (57), osteoarthritis (58), muscular dystrophy (59), and Parkinson's disease (60). As bulk tissue RNA-seq data is more accessible than scRNA-seq data, there is a great interest in the development of deconvolution tools capable of describing the cellular composition of tissue samples, especially in the study of tumors (61).

RNA-seq techniques are also useful for studying the high variability of the immune system and how this may influence disease progression. The immune repertoire is defined as the set of B-cell receptors (BCR) and T-cell receptors (TCR) of an organism. The former directly binds antigen to initiate differentiation of B cells into plasma cells, which then secrete antibodies. The latter recognizes antigens bound to major histocompatibility complex (MHC) molecules displayed on antigen-presenting cells. A robust adaptive immune system relies on the generation of a wide variety of BCRs and TCRs to recognize a varied range of antigens. A highly diverse immune repertoire is generated through V(D)J recombination. Additionally, the BCRs undergo somatic hypermutation, which increases the antigen binding specificity and affinity. Several bioinformatics tools have been developed to accurately determine the immune repertoires from genomic or RNA sequencing data, with a focus on the hypervariable complementarity-determining region 3 (CDR3) sequences. Some of these tools are specific to BCR or TCR, such as TRUST (62) and V'Djer (63), while others can work with both receptor types, such as MiXCR (64). There are also specific tools for scRNA-seq data, such as BASIC (65).

APPLICATIONS OF INTEGRATIVE BIOLOGY TO HUMAN DISEASES

Diseases are accompanied by many simultaneous changes in cell and molecular dynamics, such as gene and protein expression, metabolic pathways, and tissue cell population composition, that can be the cause or consequence of the disease state. An integrative approach to investigate these complex changes and interactions can enable a more holistic understanding of immunology, including inhibition of viral replication, generation of protective immune responses, pathogen evasion of innate and adaptive immunity, and differences in susceptibility between individuals and populations (66).

The central dogma of molecular biology states that the information is transferred sequentially from mRNA to proteins (67). However, this does not always mean there is a perfect correlation between mRNA and protein expression, highlighting the importance of analyzing multiple layers of biological data (68). In fact, now it is clear that the correlation between mRNA and protein expression depends on the cell state. In steady-state conditions, mRNA and protein levels have a strong positive correlation, but during dynamic conditions, including stress responses that are cause or consequence of disease, post-transcriptional processes cause deviations from an ideal positive correlation (69).

MicroRNAs (miRNAs) are short and endogenous RNAs that play important regulatory roles by suppressing mRNA translation by directing mRNA degradation. Again, we might expect a negative correlation between miRNA levels and target protein expression, but the correlation patterns are more complex than expected (70). Nunez et al. observed positively correlated miRNA and mRNA in a mouse model during early stages of alcohol dependence, suggesting that early miRNA activation may play an important role to limit the effect of alcohol-induced genes (71). Recently, an extensive investigation revealed the miRNA–mRNA correlation profile in human peripheral blood mononuclear cells (PBMC) in a rheumatoid arthritis cohort (70), leading to a better understanding of this and other autoimmune diseases (72). Similar efforts are being applied to profile the miRNA-mRNA correlation in tumorigenesis (73).

As personalized and precision medicine evolves, integration of metabolomics data with other layers of information becomes increasingly important. Nakaya et al. (74) used a systems analysis approach to uncover shared molecular signatures that predict influenza antibody response after vaccination. Briefly, they were able to identify transcriptomic signatures of innate immunity that could predict influenza vaccine-induced antibody titers. In addition, they uncovered many miRNA regulators of the response after vaccination. Another example study showing metabolomics integration with proteomics data uncovered signatures of innate immunity, T-cell signaling, and platelet activation related to clinical tolerance to *Plasmodium vivax* (75). Another study showed the association between metabolic pathways and chronic obstructive pulmonary disease (COPD) phenotypes, applying an unbiased metabolomics and transcriptomics approach, enabling the determination of phenotypic and outcome differences (76).

The study of genetic variability is important in the context of human health, since it may be related to differential disease risk in a population. Genome-wide association studies showed that approximately 80% of single-nucleotide polymorphisms (SNPs) associated with human phenotypes are located within non-coding regions, showing the potential association between these regions and the regulation of differential gene expression in health and disease (77) or in pharmacologic susceptibility (78). These non-coding regions may explain part of the variation and tissue-specificity in mRNA expression levels (79). By integrating genomic and transcriptomic data, scientists can find other expression quantitative trait loci (eQTLs) responsible for partial or complete alteration of gene expression (80).

Proteogenomics is an integrative approach between genomic and transcriptomic data, which has greatly advanced the study of several pathologies, especially cancer (81). This approach includes two methods of extracting information. In one method, data from transcriptomics and genomics are used to create protein databases with new peptides that are not present in reference databases. Alternatively, transcriptomics data can be used to validate genomics data and refine gene models (82). For example, Mun et al. performed an extensive proteogenomic characterization of patients with gastric cancer by integrating transcriptional, protein, phosphorylation, and N-glycosylation data (83). The group identified markers that predict a patient's prognosis and how they would respond to treatment. Similarly, this integration of proteogenic data has allowed a better understanding of colon cancer pathology and identification of potential therapeutic targets (84). Integration of metabolome, proteome, and clinical data has also been a powerful approach in fields other than oncology. For example, potential biomarkers for sepsis prognosis have been identified, which may aid in the development of new therapies for patients at higher risk of death (85).

To understand the response to herpes zoster vaccine, Li et al. (86) conducted a multi-layered study combining different datasets including transcriptomics, blood cell population flow cytometry, and plasma cytokine analysis to identify molecular networks correlated with adaptive immunity responses. The analysis revealed high correlations between distinct molecular signatures and biological convergence between the pathways identified by the metabolomic and transcriptomic data. These convergences suggested that the transcription program of blood cells is potentially regulatory in response to metabolic stimuli. For example, the same gene network, consisting of heme biosynthesis, BCR signaling, and inositol phosphate metabolism, was highly expressed in subjects with higher viral load. There were also significant differences between young and old adults, including NK cells frequency and expression of inflammatory genes. This contextualization of immune responses related to vaccination provides a good example of how these new integrative biology techniques may aid in research involving complex molecular responses such as biomarker identification and development of new immunization protocols.

The integration of omics data in health and disease has enabled a more detailed understanding of molecular interactions. This approach has improved the ability to study highly complex diseases including psychiatric diseases (87), pulmonary diseases (88), cardiovascular diseases (89), and the role of the microbiota in inflammatory bowel diseases (90).

TOOLS FOR INTEGRATIVE ANALYSIS

The molecular complexity of many diseases and advances in data integration have popularized studies that integrate different levels of biological data. However, integrative data analysis depends on the data types available and the aims of the study. Consequently, with the emergence of multi-omic data, new challenges have appeared for the development of appropriate statistical computational methods to integrate these data. Methods are required for the integration of the same type of data collected from different studies and the integration of different types of data collected from the same sample, termed horizontal and vertical data integration, respectively (Figure 2) (91). Although not discussed in detail, we briefly review some concepts of omics data integration.

In addition to horizontal and vertical data integration, multiple layers of data can be integrated using top–down and bottom–up approaches. Bottom–up integration consists of associating genomics and/or transcriptomics data with proteomics, metabolomics and/or clinical data in order to predict global changes in a cell or organism, such as phenotypic responses and key pathways. In contrast, a top–down approach consists of parallel clustering of different categories of data for automated and unified integration (92).

One bottom–up method used frequently in the integration of multiple omic layers is the search for correlations (93). This approach is based on regression methods and seeks to find elements that vary simultaneously in different layers, such as the search for SNPs and eQTLs that influence gene expression and are responsible for disease phenotypes (94). Co-expression network analysis is an informative bottom–up approach that can improve our knowledge in functional annotation and disease gene prediction (95). Recently, an integrative tool, CEMiTool, for the identification of co-expression modules was developed (95). In addition to unsupervised identification of co-expression modules, this tool allows automated integration with gene set enrichment analysis (96), which can identify whether the co-expression gene module is enriched for some relevant biological pathway and associated with a phenotype. This tool can also integrate co-expression modules with protein–protein interaction data, which is useful to identify the key regulators of a network (95). Other bottom–up approaches include clustering of DNA, mRNA, miRNA, protein, metabolite, epigenetic, network, and manual annotation data for later integration. These approaches are concisely described in a review by Yu and Zeng (92).

MixOmics is a multi-omic integrative computational tool based on the R language that is useful in a wide variety of omic studies. It is dedicated to the multivariate analysis of biological datasets with a specific focus on data exploration, dimensionality reduction, and data visualization (97). It offers a wide range of supervised statistical analysis methods that integrate multiple omic data to analyze relationships between these data. The methods include canonical correlation analysis, partial least squares regression, and PCA to perform discriminant analysis, horizontal or vertical integration, and the identification of molecular signatures (98, 99). Assuming the data have been normalized by specific methods (depending on its nature), mixOmics can explore and integrate different types of biological data. The input can be based on both discrete and continuous data such

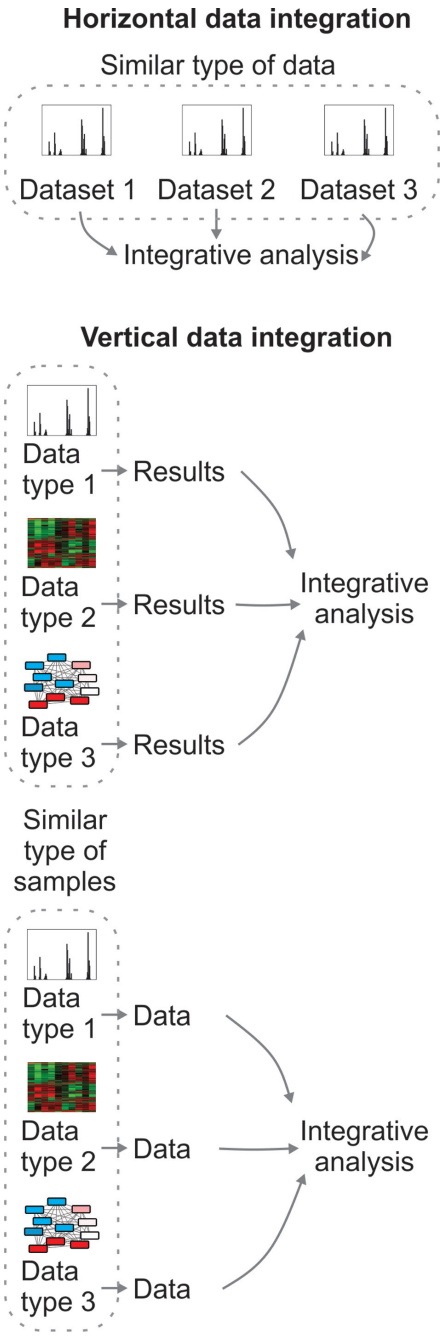


Figure 2 Horizontal and vertical data integration. Horizontal integration joins the similar data type of *n* datasets for analysis, while vertical integration combines different data types from the similar types of samples. Vertical analysis can integrate individually generated results (middle panel) or extract complex patterns directly from the data in parallel (bottom).

as mass spectrometry, microarray, proteomics, and metabolomics, or data generated by sequencing, such as RNA-seq, 16S, and metagenomic shotgun.

In contrast, a top–down approach consists of the parallel clustering of different categories of data for automated and unified integration (92). Top–down methods consist of statistical and machine learning tools such as joint models (100), Bayesian analysis (101), factor analysis (102), multiple kernel learning (103), deep learning (104), and simultaneous clustering (105). There are many useful data integration methods, and the method selection depends on the nature of the data to be analyzed. With the increasing availability of data on public databases and the development of new methods, the tendency is for greater use of omic data integration.

CHALLENGES

With the continued advancement of NGS technologies, omics science is expected to move towards an increasingly integrative approach. With this shift, managing the vast amount of data generated and integrating these data in a significant way remains a challenge (106, 107). There are concerns about the data reproducibility and accessibility (108) and efforts to overcome this, such as the FAIR principles (109). The FAIR guideline suggests ways to data become Findable, Accessible, Interoperable, and Reusable. Additionally, curated databases and improved software-database interoperability would facilitate data integration (110). Another part of the solution is the popularization of open source sharing platforms, such as GitHub, enabling developers and users to share and review their codes and scripts, as well as develop tools in collaboration with other researchers (111). A particular issue is to go beyond finding correlations to infer causality between two or more elements, such as concentration of metabolites and levels of gene expression (112). This remains a great challenge for integrative biology, which relies on molecular studies, both in vitro and in vivo, to attest the causation (93). It is important to develop new analytical methods to produce results that are easy to interpret, since the interpretation of the results can be another challenge as great as the creation of new tools (110). Finally, the evolution of integrative biology also depends on massive computational resources, both for data storage and analysis (113).

CONCLUSION

Although a huge amount of biological data is being generated at incredible pace, this is not being translated to knowledge. A large fraction of the data has the potential to be applied in clinical practice, but they are idle in repositories or waiting for the development of proper methods for data integration and interpretation. Traditionally, these data are generated by conventional hypothesis-driven methodologies. In this approach, the hypothesis is stated, tested and then accepted or refuted, based on the outcome. Alternatively, the popularization of high-throughput technologies spreads the data-driven hypothesis, or hypothesis-free, approach. In data-driven hypothesis definition, models are created after data

analysis and only then a hypothesis is formulated and tested. This integrative and systems approach can reproduce complex disease states and, therefore, has higher chances of clinical implementation. Hypothesis-driven generation and data-driven hypothesis generation are non-exclusive, since the latter can use the data produced by the former to create useful models for new hypothesis-driven studies. In this context, collaboration between bioinformatics and wet lab experts is essential for integrating multiple layers of information, which is, and will continue to be, very useful for elucidating how disease processes occur and for the development of new therapeutic interventions.

Acknowledgement: This work was supported by the São Paulo Research Foundation (FAPESP; grants 2018/14933-2, 2018/21934-5 and 2013/08216-2) and a grant from the Innovative Medicines Initiative 2 Joint Undertaking (IMI2 JU) under the VSV-EBOPLUS (grant number 116068) project.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship, and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Hunter DJ. Gene–environment interactions in human diseases. Nat Rev Genet. 2005 Apr;6(4): 287–98. http://dx.doi.org/10.1038/nrg1578
2. DeFronzo RA, Ferrannini E, Groop L, Henry RR, Herman WH, Holst JJ, et al. Type 2 diabetes mellitus. Nat Rev Dis Prim. 2015 Dec 23;1(1):15019. http://dx.doi.org/10.1038/nrdp.2015.19
3. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell. 2011 Mar 4;144(5):646–74. http://dx.doi.org/10.1016/j.cell.2011.02.013
4. Khor CC, Hibberd ML. Shared pathways to infectious disease susceptibility? Genome Med. 2010 Aug 10;2(8):52. http://dx.doi.org/10.1186/gm173
5. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nat Rev Cardiol. 2015 Nov 13;12(11):627–42. http://dx.doi.org/10.1038/nrcardio.2015.152
6. Weaver JMJ, Ross-Innes CS, Fitzgerald RC. The "–omics" revolution and oesophageal adenocarcinoma. Nat Rev Gastroenterol Hepatol. 2014 Jan 27;11(1):19–27. http://dx.doi.org/10.1038/nrgastro.2013.150
7. Ideker T, Galitski T, Hood L. A new approach to decoding life: Systems Biology. Annu Rev Genomics Hum Genet. 2001 Sep 28;2(1):343–72. http://dx.doi.org/10.1146/annurev.genom.2.1.343
8. Nakaya HI, Li S, Pulendran B. Systems vaccinology: Learning to compute the behavior of vaccine induced immunity. Wiley Interdiscip Rev Syst Biol Med. 2012;4(2):193–205. http://dx.doi.org/10.1002/wsbm.163
9. Ko T-M, Chang J-S, Chen S-P, Liu Y-M, Chang C-J, Tsai F-J, et al. Genome-wide transcriptome analysis to further understand neutrophil activation and lncRNA transcript profiles in Kawasaki disease. Sci Rep. 2019 Dec 23;9(1):328. http://dx.doi.org/10.1038/s41598-018-36520-y
10. Gao X, Ke C, Liu H, Liu W, Li K, Yu B, et al. Large-scale metabolomic analysis reveals potential biomarkers for early stage coronary atherosclerosis. Sci Rep. 2017 Dec 18;7(1):11817. http://dx.doi.org/10.1038/s41598-017-12254-1

11. Wingo AP, Dammer EB, Breen MS, Logsdon BA, Duong DM, Troncosco JC, et al. Large-scale proteomic analysis of human brain identifies proteins associated with cognitive trajectory in advanced age. Nat Commun. 2019 Dec 8;10(1):1619. http://dx.doi.org/10.1038/s41467-019-09613-z
12. Chen H, Albergante L, Hsu JY, Lareau CA, Lo Bosco G, Guan J, et al. Single-cell trajectories reconstruction, exploration and mapping of omics data with STREAM. Nat Commun. 2019 Dec 23;10(1):1903. http://dx.doi.org/10.1038/s41467-019-09670-4
13. Gosak M, Markovič R, Dolenšek J, Slak Rupnik M, Marhl M, Stožer A, et al. Network science of biological systems at different scales: A review. Phys Life Rev. 2018 Mar;24:118–35. http://dx.doi.org/10.1016/j.plrev.2017.11.003
14. Srivastava A, George J, Karuturi RKM. Transcriptome analysis. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology, vol. 3. 1st ed., Cambridge, MA: Elsevier, 2019. p. 729–805.
15. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. Brief Bioinform. 2017 Jun 30;19(6):1370–81. http://dx.doi.org/10.1093/bib/bbx066
16. Costa-Silva J, Domingues D, Lopes FM. RNA-Seq differential expression analysis: An extended review and a software tool. Wei Z, editor. PLoS One. 2017 Dec 21;12(12):e0190152. http://dx.doi.org/10.1371/journal.pone.0190152
17. Jiang M, Zhang S, Yang Z, Lin H, Zhu J, Liu L, et al. Self-recognition of an inducible host lncRNA by RIG-I feedback restricts innate immune response. Cell. 2018 May;173(4):906–19.e13. http://dx.doi.org/10.1016/j.cell.2018.03.064
18. Liu C-X, Li X, Nan F, Jiang S, Gao X, Guo S-K, et al. Structure and degradation of circular RNAs regulate PKR activation in innate immunity. Cell. 2019 May;177(4):865–80.e21. http://dx.doi.org/10.1016/j.cell.2019.03.046
19. Cristinelli S, Ciuffi A. The use of single-cell RNA-Seq to understand virus–host interactions. Curr Opin Virol. 2018 Apr;29:39–50. http://dx.doi.org/10.1016/j.coviro.2018.03.001
20. Wu AR, Wang J, Streets AM, Huang Y. Single-cell transcriptional analysis. Annu Rev Anal Chem. 2017 Jun;10(1):439–62. http://dx.doi.org/10.1146/annurev-anchem-061516-045228
21. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-seq whole-transcriptome analysis of a single cell. Nat Methods. 2009 May 6;6(5):377–82. http://dx.doi.org/10.1038/nmeth.1315
22. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods. 2016 Apr 7;13(4):329–32. http://dx.doi.org/10.1038/nmeth.3800
23. See P, Lum J, Chen J, Ginhoux F. A Single-cell sequencing guide for immunologists. Front Immunol. 2018 Oct 23;9:2425. http://dx.doi.org/10.3389/fimmu.2018.02425
24. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015 May;161(5):1202–14. http://dx.doi.org/10.1016/j.cell.2015.05.002
25. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015 May;161(5):1187–201. http://dx.doi.org/10.1016/j.cell.2015.04.044
26. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017 Apr 16;8(1):14049. http://dx.doi.org/10.1038/ncomms14049
27. Zhang X, Li T, Liu F, Chen Y, Yao J, Li Z, et al. Comparative analysis of droplet-based ultra-high-throughput single-cell RNA-seq systems. Mol Cell. 2019 Jan;73(1):130–42.e5. http://dx.doi.org/10.1016/j.molcel.2018.10.020
28. Zhao Z, Meng F, Wang W, Wang Z, Zhang C, Jiang T. Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. Sci Data. 2017 Apr 14;4(1):170024. http://dx.doi.org/10.1038/sdata.2017.24
29. Aevermann BD, Pickett BE, Kumar S, Klem EB, Agnihothram S, Askovich PS, et al. A comprehensive collection of systems biology data characterizing the host response to viral infection. Sci Data. 2014 Dec 14;1(1):140033. http://dx.doi.org/10.1038/sdata.2014.33
30. Salomon MP, Orozco JIJ, Wilmott JS, Hothi P, Manughian-Peter AO, Cobbs CS, et al. Brain metastasis DNA methylomes, a novel resource for the identification of biological and clinical features. Sci Data. 2018 Nov 6;5:180245. http://dx.doi.org/10.1038/sdata.2018.245

31. Liu X, Ma Y, Yin K, Li W, Chen W, Zhang Y, et al. Long non-coding and coding RNA profiling using strand-specific RNA-seq in human hypertrophic cardiomyopathy. Sci Data. 2019 Dec 13;6(1):90. http://dx.doi.org/10.1038/sdata.2018.245
32. Wang M, Beckmann ND, Roussos P, Wang E, Zhou X, Wang Q, et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. Sci Data. 2018 Sep 11;5:180185. http://dx.doi.org/10.1038/sdata.2018.185
33. Pechal JL, Schmidt CJ, Jordan HR, Benbow ME. A large-scale survey of the postmortem human microbiome, and its potential to provide insight into the living health condition. Sci Rep. 2018 Dec 10;8(1):5724. http://dx.doi.org/10.1038/s41598-018-23989-w
34. Haile S, Corbett RD, Bilobram S, Bye MH, Kirk H, Pandoh P, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. Nucleic Acids Res. 2019 Jan 25;47(2):e12–e12. http://dx.doi.org/10.1093/nar/gky1142
35. Baker M. Statisticians issue warning over misuse of P values. Nature. 2016 Mar 7;531(7593):151. http://dx.doi.org/10.1038/nature.2016.19503
36. Wang Y, Sun M. Transcriptome data analysis: Methods and protocols. New York: Humana Press, Springer; 2018. 238 p.
37. Kuleshov M V., Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016 Jul 8;44(W1):W90–7.
38. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan 1;26(1):139–40. http://dx.doi.org/10.1093/bioinformatics/btp616
39. Olsen LR, Leipold MD, Pedersen CB, Maecker HT. The anatomy of single cell mass cytometry data. Cytometry A. 2019 Feb;95(2):156–72. http://dx.doi.org/10.1002/cyto.a.23621
40. van Unen V, Li N, Molendijk I, Temurhan M, Höllt T, van der Meulen-de Jong AE, et al. Mass cytometry of the human mucosal immune system identifies tissue- and disease-associated immune subsets. Immunity. 2016 May;44(5):1227–39. http://dx.doi.org/10.1016/j.immuni.2016.04.014
41. Nowicka M, Krieg C, Crowell HL, Weber LM, Hartmann FJ, Guglietta S, et al. CyTOF workflow: Differential discovery in high-throughput high-dimensional cytometry datasets. F1000Research. 2019 May 24;6:748. http://dx.doi.org/10.12688/f1000research.11622.3
42. Azad RK, Shulaev V. Metabolomics technology and bioinformatics for precision medicine. Brief Bioinform. 2018 Jan 3;bbx170:1–15. http://dx.doi.org/10.1093/bib/bbx170
43. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018 Jul 2;46(W1):W486–94. http://dx.doi.org/10.1093/nar/gky310
44. Chong J, Yamamoto M, Xia J. MetaboAnalystR 2.0: From raw spectra to biological insights. Metabolites. 2019 Mar 22;9(3):57. http://dx.doi.org/10.3390/metabo9030057
45. Hicks SC, Townes FW, Teng M, Irizarry RA. Missing data and technical variability in single-cell RNA-sequencing experiments. Biostatistics. 2018 Oct;19(4):562–78. http://dx.doi.org/10.1093/biostatistics/kxx053
46. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr;43(7):e47–e47. http://dx.doi.org/10.1093/nar/gkv007
47. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan 1;8(1):118–27. http://dx.doi.org/10.1093/biostatistics/kxj037
48. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014 Sep;32(9):896–902. http://dx.doi.org/10.1038/nbt.2931
49. Leek JT. svaseq: Removing batch effects and other unwanted noise from sequencing data. Nucleic Acids Res. 2014 Dec;42(21):e161. http://dx.doi.org/10.1093/nar/gku864
50. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018 May 2;36(5):421–7. http://dx.doi.org/10.1038/nbt.4091
51. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Aspects Med. 2018 Feb;59:114–22. http://dx.doi.org/10.1016/j.mam.2017.07.002
52. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019 Jan;37(1):38–44. http://dx.doi.org/10.1038/nbt.4314

53. Chen G, Ning B, Shi T. Single-Cell RNA-Seq Technologies and related computational data analysis. Front Genet. 2019 Apr 5;10:317. http://dx.doi.org/10.3389/fgene.2019.00317
54. Björklund ÅK, Forkel M, Picelli S, Konya V, Theorell J, Friberg D, et al. The heterogeneity of human CD127+ innate lymphoid cells revealed by single-cell RNA sequencing. Nat Immunol. 2016 Apr;17(4):451–60. http://dx.doi.org/10.1038/ni.3368
55. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018 Aug 7;50(8):96. http://dx.doi.org/10.1038/s12276-018-0071-8
56. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol. 2018 Apr 16;14(4):e8046. http://dx.doi.org/10.15252/msb.20178046
57. Cheng Y, Zhu YO, Becht E, Aw P, Chen J, Poidinger M, et al. Multifactorial heterogeneity of virus-specific T cells and association with the progression of human chronic hepatitis B infection. Sci Immunol. 2019 Feb 8;4(32):eaau6905. http://dx.doi.org/10.1126/sciimmunol.aau6905
58. Ji Q, Zheng Y, Zhang G, Hu Y, Fan X, Hou Y, et al. Single-cell RNA-seq analysis reveals the progression of human osteoarthritis. Ann Rheum Dis. 2019 Jan;78(1):100–10. http://dx.doi.org/10.1136/annrheumdis-2017-212863
59. van den Heuvel A, Mahfouz A, Kloet SL, Balog J, van Engelen BGM, Tawil R, et al. Single-cell RNA sequencing in facioscapulohumeral muscular dystrophy disease etiology and development. Hum Mol Genet. 2019 Apr 1;28(7):1064–75. http://dx.doi.org/10.1093/hmg/ddy400
60. Lang C, Campbell KR, Ryan BJ, Carling P, Attar M, Vowles J, et al. Single-cell sequencing of iPSC-dopamine neurons reconstructs disease progression and identifies HDAC4 as a regulator of Parkinson cell phenotypes. Cell Stem Cell. 2019 Jan;24(1):93–106.e6. http://dx.doi.org/10.1016/j.stem.2018.10.023
61. Finotello F, Trajanoski Z. Quantifying tumor-infiltrating immune cells from transcriptomics data. Cancer Immunol Immunother. 2018 Jul 14;67(7):1031–40. http://dx.doi.org/10.1007/s00262-018-2150-z
62. Li B, Li T, Wang B, Dou R, Zhang J, Liu JS, et al. Ultrasensitive detection of TCR hypervariable-region sequences in solid-tissue RNA–seq data. Nat Genet. 2017 Apr 1;49(4):482–3. http://dx.doi.org/10.1038/ng.3820
63. Mose LE, Selitsky SR, Bixby LM, Marron DL, Iglesia MD, Serody JS, et al. Assembly-based inference of B-cell receptor repertoires from short read RNA sequencing data with V'DJer. Bioinformatics. 2016 Dec 15;32(24):3729–34. http://dx.doi.org/10.1093/bioinformatics/btw526
64. Bolotin DA, Poslavsky S, Davydov AN, Frenkel FE, Fanchi L, Zolotareva OI, et al. Antigen receptor repertoire profiling from RNA-seq data. Nat Biotechnol. 2017 Oct 11;35(10):908–11. http://dx.doi.org/10.1038/nbt.3979
65. Canzar S, Neu KE, Tang Q, Wilson PC, Khan AA. BASIC: BCR assembly from single cells. Bioinformatics. 2016 Oct 2;btw631. http://dx.doi.org/10.1093/bioinformatics/btw631
66. Nakaya HI. Systems biology of infectious diseases and vaccines. In: Eils R, Kriete A, editors. Computational Systems Biology. 2nd ed., San Diego, CA: Academic Press, 2014. p. 331–58.
67. CRICK F. Central dogma of molecular biology. Nature. 1970 Aug;227(5258):561–3. http://dx.doi.org/10.1038/227561a0
68. Anderson L, Seilhamer J. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis. 1997;18(3–4):533–7. http://dx.doi.org/10.1002/elps.1150180333
69. Liu Y, Beyer A, Aebersold R. On the dependency of cellular protein levels on mRNA abundance. Cell. 2016 Apr;165(3):535–50. http://dx.doi.org/10.1016/j.cell.2016.03.014
70. Wang L, Zhu J, Deng F-Y, Wu L-F, Mo X-B, Zhu X-W, et al. Correlation analyses revealed global microRNA–mRNA expression associations in human peripheral blood mononuclear cells. Mol Genet Genomics. 2018 Feb 6;293(1):95–105. http://dx.doi.org/10.1007/s00438-017-1367-4
71. Nunez YO, Truitt JM, Gorini G, Ponomareva ON, Blednov YA, Harris R, et al. Positively correlated miRNA-mRNA regulatory networks in mouse frontal cortex during early stages of alcohol dependence. BMC Genomics. 2013;14(1):725. http://dx.doi.org/10.1186/1471-2164-14-725
72. Olsen NJ, Moore JH, Aune TM. Gene expression signatures for autoimmune disease in peripheral blood mononuclear cells. Arthritis Res Ther. 2004;6(3):120–8. http://dx.doi.org/10.1186/ar1190
73. Kumar V, Kumar V, Chaudhary AK, Coulter DW, McGuire T, Mahato RI. Impact of miRNA-mRNA profiling and their correlation on medulloblastoma tumorigenesis. Mol Ther Nucleic Acids. 2018 Sep;12:490–503. http://dx.doi.org/10.1016/j.omtn.2018.06.004

74. Nakaya HI, Hagan T, Duraisingham SS, Lee EK, Kwissa M, Rouphael N, et al. Systems analysis of immunity to influenza vaccination across multiple years and in diverse populations reveals shared molecular signatures. Immunity. 2015 Dec;43(6):1186–98. http://dx.doi.org/10.1016/j.immuni.2015.11.012
75. Gardinassi LG, Arévalo-Herrera M, Herrera S, Cordy RJ, Tran V, Smith MR, et al. Integrative metabolomics and transcriptomics signatures of clinical tolerance to Plasmodium vivax reveal activation of innate cell immunity and T cell signaling. Redox Biol. 2018 Jul;17:158–70. http://dx.doi.org/10.1016/j.redox.2018.04.011
76. Cruickshank-Quinn CI, Jacobson S, Hughes G, Powell RL, Petrache I, Kechris K, et al. Metabolomics and transcriptomics pathway approach reveals outcome-specific perturbations in COPD. Sci Rep. 2018 Dec 20;8(1):17132. http://dx.doi.org/10.1016/j.redox.2018.04.011
77. Manolio TA. Genomewide association studies and assessment of the risk of disease. N Engl J Med. 2010 Jul;363(2):166–76. http://dx.doi.org/10.1056/NEJMra0905980
78. Gamazon ER, Huang RS, Cox NJ, Dolan ME. Chemotherapeutic drug susceptibility associated SNPs are enriched in expression quantitative trait loci. Proc Natl Acad Sci. 2010 May 18;107(20):9287–92. http://dx.doi.org/10.1073/pnas.1001827107
79. Gerrits A, Li Y, Tesson BM, Bystrykh L V., Weersing E, Ausema A, et al. Expression quantitative trait loci are highly sensitive to cellular differentiation state. PLoS Genet. 2009 Oct;5(10):e1000692. http://dx.doi.org/10.1073/pnas.1001827107
80. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet. 2009 Mar;10(3):184–94. http://dx.doi.org/10.1038/nrg2537
81. Kumar D, Bansal G, Narang A, Basak T, Abbas T, Dash D. Integrating transcriptome and proteome profiling: Strategies and applications. Proteomics. 2016 Oct;16(19):2533–44. http://dx.doi.org/10.1002/pmic.201600140
82. Nesvizhskii AI. Proteogenomics: Concepts, applications and computational strategies. Nat Methods. 2014 Nov 30;11(11):1114–25. http://dx.doi.org/10.1038/nmeth.3144
83. Mun D-G, Bhin J, Kim S, Kim H, Jung JH, Jung Y, et al. Proteogenomic characterization of human early-onset gastric cancer. Cancer Cell. 2019 Jan;35(1):111–24.e10. http://dx.doi.org/10.1016/j.ccell.2018.12.003
84. Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell. 2019 May;177(4):1035–49.e19.
85. Langley RJ, Tsalik EL, Velkinburgh JC, Glickman SW, Rice BJ, Wang C, et al. An integrated clinico-metabolomic model improves prediction of death in sepsis. Sci Transl Med. 2013 Jul 24;5(195):195ra95. http://dx.doi.org/10.1126/scitranslmed.3005893
86. Li S, Sullivan NL, Rouphael N, Yu T, Banton S, Maddur MS, et al. Metabolic Phenotypes of Response to Vaccination in Humans. Cell. 2017 May;169(5):862–77.e17. http://dx.doi.org/10.1126/scitranslmed.3005893
87. Johnson E, Bierut L, Cox N. Integrative omics in psychiatric diseases: Tools for discovery and understanding biology. Eur Neuropsychopharmacol. 2019;29:S741–2. http://dx.doi.org/10.1016/j.euroneuro.2017.06.073
88. Hobbs BD, Chimakurthi L, Morrow JD, Wang X, Liu Y-Y, Celli BR, et al. Integrative omics to discover novel subtypes in a chronic obstructive pulmonary disease lung tissue cohort. Am J Respir Crit Care Med. 2019;199:A6092. http://dx.doi.org/10.1164/ajrccm-conference.2019.199.1_MeetingAbstracts.A6092
89. Leon-Mimila P, Wang J, Huertas-Vazquez A. Relevance of multi-omics studies in cardiovascular diseases. Front Cardiovasc Med. 2019 Jul 17;6:91. http://dx.doi.org/10.3389/fcvm.2019.00091
90. Segal JP, Mullish BH, Quraishi MN, Acharjee A, Williams HRT, Iqbal T, et al. The application of omics techniques to understand the role of the gut microbiota in inflammatory bowel disease. Therap Adv Gastroenterol. 2019 Jan 24;12:175628481882225. http://dx.doi.org/10.1177/1756284818822250
91. Wu C, Huang BE, Chen G, Lovenberg TW, Pocalyko DJ, Yao X. Integrative analysis of diseaseland omics database for disease signatures and treatments: A bipolar case study. Front Genet. 2019 Apr 30;10:396. http://dx.doi.org/10.3389/fgene.2019.00396
92. Yu XT, Zeng T. Integrative analysis of omics big data. In: Huang T, editor. Computational Systems Biology. Methods Mol Biol. New York, NY: Humana Press; 2018;1754:109–35.
93. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 Dec;18(1):83. http://dx.doi.org/10.1186/s13059-017-1215-1

94. Sun YV, Hu Y-J. Integrative analysis of multi-omics data for discovery and functional studies of complex human diseases. Adv Genet. 2016;93:147–90.
95. Russo PST, Ferreira GR, Cardozo LE, Bürger MC, Arias-Carrasco R, Maruyama SR, et al. CEMiTool: A bioconductor package for performing comprehensive modular co-expression analyses. BMC Bioinformatics. 2018 Dec 20;19(1):56. http://dx.doi.org/10.1186/s12859-018-2053-1
96. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct;102(43):15545–50. http://dx.doi.org/10.1073/pnas.0506580102
97. Rohart F, Gautier B, Singh A, Lê Cao K-A. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov 3;13(11):e1005752. http://dx.doi.org/10.1371/journal.pcbi.1005752
98. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics. 2019 Sep 1; 35(17):3055–62. http://dx.doi.org/10.1093/bioinformatics/bty1054
99. Rohart F, Eslami A, Matigian N, Bougeard S, Lê Cao K-A. MINT: A multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics. 2017 Dec 27;18(1):128. http://dx.doi.org/10.1186/s12859-017-1553-8
100. Geng P, Tong X, Lu Q. An integrative U method for joint analysis of multi-level omic data. BMC Genet. 2019 Dec 10;20(1):40. http://dx.doi.org/10.1186/s12863-019-0742-z
101. Ickstadt K, Schäfer M, Zucknick M. Toward integrative Bayesian analysis in molecular biology. Annu Rev Stat Its Appl. 2018 Mar 7;5(1):141–67. http://dx.doi.org/10.1146/annurev-statistics-031017-100438
102. Argelaguet R, Velten B, Arnol D, Dietrich S, Zenz T, Marioni JC, et al. Multi-Omics factor analysis—A framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018 Jun 20;14(6): e8124. http://dx.doi.org/10.15252/msb.20178124
103. Speicher NK, Pfeifer N. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. Bioinformatics. 2015 Jun 15;31(12):i268–75. http://dx.doi.org/10.1093/bioinformatics/btv244
104. Zhang L, Lv C, Jin Y, Cheng G, Fu Y, Yuan D, et al. Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet. 2018 Oct 18;9:477. http://dx.doi.org/10.3389/fgene.2018.00477
105. Narayanan M, Vetta A, Schadt EE, Zhu J. Simultaneous clustering of multiple gene expression and physical interaction datasets. PLoS Comput Biol. 2010 Apr 15;6(4):e1000742. http://dx.doi.org/10.1371/journal.pcbi.1000742
106. D'Argenio V. The high-throughput analyses era: Are we ready for the data struggle? High Throughput. 2018 Mar 2;7(1):8. http://dx.doi.org/10.1371/journal.pcbi.1000742
107. Palsson B, Zengler K. The challenges of integrating multi-omic data sets. Nat Chem Biol. 2010 Nov 18;6(11):787–9. http://dx.doi.org/10.1038/nchembio.462
108. Nekrutenko A, Taylor J. Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. Nat Rev Genet. 2012 Sep 17;13(9):667–72. http://dx.doi.org/10.1038/nrg3305
109. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016 Dec 15;3(1):160018.
110. Pinu FR, Beale DJ, Paten AM, Kouremenos K, Swarup S, Schirra HJ, et al. Systems biology and multi-omics integration: Viewpoints from the metabolomics research community. Metabolites. 2019 Apr;9(4):76. http://dx.doi.org/10.3390/metabo9040076
111. Misra BB, Langefeld CD, Olivier M, Cox LA. Integrated omics: Tools, advances, and future approaches. J Mol Endocrinol. 2019 Jan 13;62:R21-R45. https://doi.org/10.1530/JME-18-0055.
112. Altman N, Krzywinski M. Association, correlation and causation. Nat Methods. 2015 Oct 29; 12(10):899–900. http://dx.doi.org/10.1038/nmeth.3587
113. Yin Z, Lan H, Tan G, Lu M, Vasilakos AV, Liu W. Computing platforms for big biological data analytics: Perspectives and challenges. Comput Struct Biotechnol J. 2017;15:403–11. http://dx.doi.org/10.1016/j.csbj.2017.07.004

3

Deep Learning in Omics Data Analysis and Precision Medicine

Jordi Martorell-Marugán[1] • Siham Tabik[2] • Yassir Benhammou[2] • Coral del Val[2] • Igor Zwir[2] • Francisco Herrera[2] • Pedro Carmona-Sáez[1]

[1]GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Granada, Spain; [2]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

Authors for correspondence: Francisco Herrera, Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain. Email: herrera@decsai.ugr.es; Pedro Carmona-Sáez, GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, PTS Granada, Avenida de la Ilustración 114 – 18016, Granada, Spain. Email: pedro.carmona@genyo.es

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch3

Abstract: The rise of omics techniques has resulted in an explosion of molecular data in modern biomedical research. Together with information from medical images and clinical data, the field of omics has driven the implementation of personalized medicine. Biomedical and omics datasets are complex and heterogeneous, and extracting meaningful knowledge from this vast amount of information is by far the most important challenge for bioinformatics and machine learning researchers. In this context, there is an increasing interest in the potential of deep learning (DL) methods to create predictive models and to identify complex patterns from these large datasets. This chapter provides an overview of the main applications of DL methods in biomedical research, with focus on omics data analysis and precision medicine applications. DL algorithms and the most popular architectures are introduced first. This is followed by a review of some of the main applications and problems approached by DL in omics data and medical image analysis. Finally, implementations for improving the diagnosis, treatment, and classification of complex diseases are discussed.

In: *Computational Biology*. Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

Keywords: artificial neural networks; biomedical informatics; deep learning; omics data analysis; precision medicine

INTRODUCTION

The amount of available biological data has increased exponentially since the emergence of high-throughput technologies such as microarrays and next-generation sequencing (1), introducing biology to the big data era. These methods initiated the so-called omics revolution, where large amounts of omics data providing global information about different properties of genes, proteins or biomolecules can be generated within a short period of time in a cost-effective way. These methods have revolutionized biomedical research by providing a more comprehensive understanding of the biological system under study and the molecular mechanisms underlying disease development. The generation of such a large amount of data in biomedicine requires the application of advanced informatics techniques in order to extract new insights and expand our knowledge about diseases, improve diagnosis, and design personalized treatments. In this context, DL algorithms have become one of the most promising methods in the area (2).

DL is a subset of machine learning (ML) algorithms characterized by the use of artificial neural networks (ANN). ANNs are inspired by biological neural networks in a sense that they are formed by interconnected artificial neurons, which receive an input, apply a transformation to the data, and return an output (which can be an input for another neuron). DL is gaining popularity as a powerful approach that can encode and learn from heterogeneous and complex data, in both supervised and unsupervised settings. DL methods have achieved considerable improvements in classical artificial intelligence challenges like language processing, speech recognition, and image recognition (3). In the context of biomedical research, DL methods have drawn the attention of many researchers, and there is an increasing number of applications in omics data analysis. Omics data analysis is frequently impeded by low signal to noise ratios, datasets with large number of variables and relatively small number of samples or large analytical variance. In this context, DL techniques have already over-performed previous methods in terms of sensitivity, specificity and efficiency (4). In addition, DL algorithms not only have the challenge of analyzing each kind of data separately but also have the challenge of integrating different omics layers or even other sources of information such as medical images or clinical health records. This big data analysis and integration is fueling the implementation of personalized medicine approaches allowing early detection and classification of diseases or personalized therapies for each patient depending on their biochemical background. This chapter reviews the main applications of DL methods to omics data analysis with a focus on the types of analysis, challenges, and opportunities in precision medicine.

DEEP LEARNING METHODS

DL networks are a class of ML algorithms whose aim is to determine a mathematical function f that maps a number of inputs, x, to their corresponding outputs, y, such as $y = f(x)$. A simple feedforward network $y = f(x;w) = LN(LN\text{-}1(\ldots L1(x))$ is

defined as a composition of N nonlinear transformations $Li(1<=i<=N)$ where each function Li corresponds to a hidden layer activation, and w is the learnable weight contained in all filter bank layers that are updated during the training.

Under the supervised learning approach, the training of these networks is often done iteratively in which a set of training data, also called batch, with their ground truth labels are provided to the network as input. After a feedforward of this batch through the network's layers, the output layer computes the loss function as the difference between the calculated prediction and the correct response. After computing the loss function, all layers' weights are updated so that the loss error of the next iteration is minimized. This weight-tuning operation is performed using a back-propagation algorithm (5) where the error function gradient is propagated in the opposite direction through the network after a batch of feedforwards to adjust filter banks, thereby learning the value of the parameter w that results in the best function approximation.

Deep feedforward neural network (DFF)

DFFs, also called multilayer perceptrons, constitute the simplest DL architecture. In these models, the input information x flows to its corresponding output y through an intermediate function f being evaluated and learned inside the neural network layers. These models are called feedforward since there are no feedback connections in which outputs of the model are fed back into themselves.

Convolutional neural network (CNN)

CNNs are the most adequate DNNs to deal with high multi-dimensional data like medical images. In medical imaging applications, CNNs act like a long dimensionality reduction process, binding input images to their classification scores outputs (e.g., disease or healthy patient). The building block layers of a CNN are convolutional layer, pooling layer, and fully connected layer. Generally, DL CNNs are applied with a transfer learning strategy to enhance their performance in dealing with relatively small datasets. Transfer learning consists of transferring prior learned knowledge from a source domain into a target domain. This approach is carried out by using one of the well-known CNNs pre-trained on a large dataset such as ImageNet (6), either for further training on the new data or to reuse it as a features extractor (7). Rawat and Wang (8) wrote a more comprehensive review on CNNs history and their architectures. Some of the most influential CNNs are summarized in Table 1.

Recurrent neural network (RNN)

RNNs are neural networks used especially for sequential data in a way that the reached output decision at time step $t - 1$ affects the decision which will be reached one moment later at time step t. These networks have two input sources, the present and the recent past, which are combined to determine how they respond to new data.

TABLE 1 **Summary of some of the most influential CNNs**

CNN	Layers	Parameters	Comments
LeNet	5	60 000	First CNN to be trained on a large dataset (5, 87)
AlexNet	7	60 million	Variation of LeNet. First CNN model to win the prestigious ILSVRC[a] in 2012 (88).
GoogLeNet	22	4 million	Winner of ILSVRC[a] in 2014 (89). The main contribution is the inception module which is composed of different parallel small convolutions.
VGGNet	16	-	Initially the runner-up in ILSVRC 2014 behind GoogleNet (90)
ResNet	18, 34, 50, 101 or 152	11.7 million – 60.2 million	To overcome the gradient vanishing issue, ResNet authors (91) proposed using a residual function $F(x) = H(x) - x$, where $H(x)$ is the standard mapping function that we want to learn with an input x through few stacked non-linear layers. By reformulating it as $H(x) = F(x) + x$, where $F(x)$ and x represent the stacked non-linear layers and the identity function, respectively. Based on their hypothesis, it is better to optimize the reformulated residual mapping function $F(x)$ than optimizing the original mapping $H(x)$.
DenseNet	121, 161, 169 or 201	8 million – 20 million	Presented in (92) to take advantage from previous findings regarding CNN's depth increasing and identity shortcut connections. The specificity of this new network architecture is that each layer is connected to all its previous and next layers.

[a]Large Scale Visual Recognition Challenge.

Long-/short-term memory (LSTM)

The main drawback of RNNs is the vanishing gradient problem. To address this issue, a variant of RNN called LSTM was proposed. LSTMs aim to preserve the error that can be back-propagated through time and layers. In fact, they allow recurrent nets to continue to learn over many steps by maintaining a more constant error. LSTMs contain information outside the normal flow of the recurrent network in a gated cell. Information can be stored in, written to or read from a cell, much like data in a computer's memory.

Deep belief network (DBN)

To learn deep features representation, a DBN (9) is built with a concatenation of several restricted Boltzmann machine (RBM) stacked on each other. RBM is the core component of DBN models (10), being a generative stochastic model that can

be used either for unsupervised or supervised learning. It is composed of two layers, an input visible layer and an adjacent hidden layer trained with the aim to learn a probability distribution in the input set. Nevertheless, unlike original Boltzmann machine (11), intra-connections between hidden–hidden or visible–visible layers in an RBM are disjointed forming a bipartite graph.

Autoencoder (AE)

Generally, AEs act in an unsupervised manner trying to learn a distribution of a given dataset (12) and are often used as a dimensionality reduction network (13). AEs try to learn a mapping function $M\ w,\ b\ (x) = x' \approx x$ throughout stacked hidden layers mapping an input data x to its similar identity x' Generally, an AE is composed of an encoder and a decoder. The first one is trying to learn a set of low-dimensional representation features z, while the second is trying to reconstruct a similar copy of x using only learned features z. A special case of AEs is sparse autoencoder (SAE) (14), where sparsity is introduced into the hidden units by making the number of nodes in the hidden layer z bigger than in the input layer x. When several SAEs with only their encoding parts are stacked on each other, we obtain a stacked sparse autoencoder (SSAE) which is often trained in a bottom–up greedy fashion to learn deep feature representation from the data (14).

DEEP LEARNING APPLICATIONS IN OMICS DATA ANALYSIS

DL algorithms are specially suitable to analyze complex, heterogeneous, and high-dimensional data such as omics datasets (15). This section reviews some cases of omics data analyses in which DL methods have provided significant insights, and the next section provides an overview of some of the main applications in the context of precision medicine, such as biomarker discovery for disease classification. A summary of the main applications is provided in Figure 1.

Genomics and sequence analysis

Genomics uses a set of techniques to analyze DNA sequences for studying the structure and function of genomes, gene regulation, and genetic alterations that can be associated with several diseases. During the last years, DL methods have been applied to genomics data to address several questions. For instance, Poplin et al. developed a method to detect single-nucleotide polymorphisms (SNP) and indels by applying CNNs, which outperformed previous tools (16). In this context, other approaches have applied ResNets (17), DFF (18) or CNN (19) to predict the pathogenic consequences of genetic variants. In addition, Xie et al. applied DFFs and SAEs to predict the effect of genetic variants on gene expression (20). In the field of functional genomics, DL algorithms have been applied to predict enhancers' sequences and regulatory motifs in the genome (21–25) from heterogeneous sources of data (histone modifications, chromatin accessibility and so on). Wang et al. applied CNNs to quantify transcription factor (TF)-DNA binding affinities (26). Oubounyt et al. combined a CNN and an

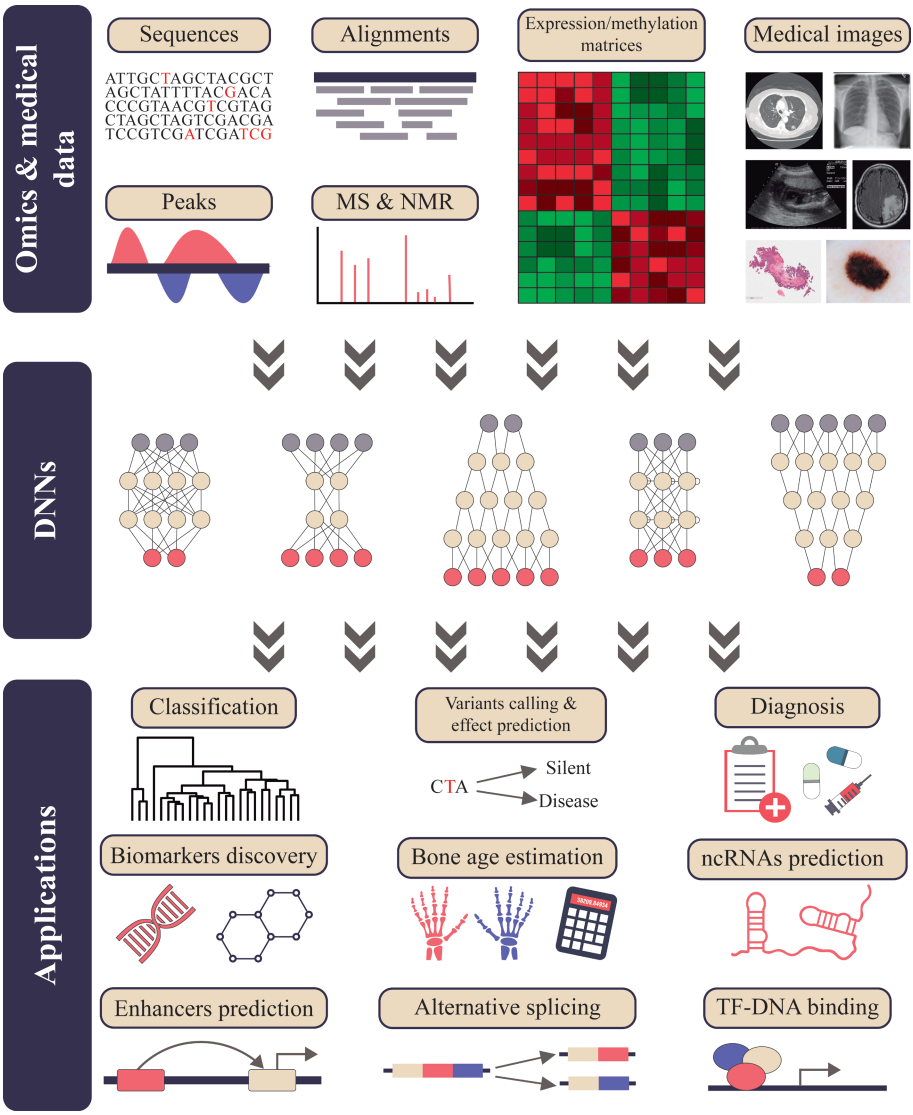


Figure 1 DNNs have been applied to several biological data types. At the top, there are the different types of data. At the middle, there are some examples of DNNs structures. At the bottom, there are some of the main applications achieved with these methodologies. Source of medical images: TCIA (93) for MRI and CT; Chest X-Ray database (94) for X-Ray; MedPix® (https://medpix.nlm.nih.gov) for US; TCGA (58) for the histopathological image and ISIC (https://www.isic-archive.com) for the skin lesion. Some graphical elements were downloaded from Stockio (https://www.stockio.com/) and Freepik (https://www.freepik.com/).

LSTM to predict promoter sequences in genes (27). DL algorithms have also helped to identify splice junctions through CNN (28).

Genome-wide association studies

Another important field of application of genomics techniques is the screening of genetic regions (loci) that associate with diseases/phenotypes, what is termed genome-wide association studies (GWAS). In this context, GWAS analyses identify SNPs in genomic locations that are incorporated into risk prediction models traditionally analyzed by polygenic risk scores (29). However, this method presents certain limitations such as the inability to reduce the missing heritability, dealing with epistasis, assumption of a global linear association model or the replication of results in different samples (30).

As an alternative, supervised learning algorithm, especially DL models, is gaining relevance in this field. Promising results have been shown by Montaez et al. (31) that developed a DL framework for the classification of obesity as a binary phenotype. However, the predictive capacity of these genetic markers is weak because it is based on single locus. More recently, Fergus et al. (32) modeled the epistatic effects of SNPs using SSAEs to classify term and preterm births observations in African-American women. Although it shows a good performance in classification and the capture of loci interactions, it suffers from the common black-box problem. The selected SNPs loose the GWAS context making it very difficult to evaluate their contribution to the phenotype. A different approach is the one proposed by PGMRA (33), a deep unsupervised and data-driven ML method designed for fusing genotypic–phenotypic analysis in a semi-supervised fashion including unsupervised non-negative matrix factorization (NMF) method as an AE (13), multiobjective optimization and pooling, interpretable association of types of knowledge, and labeling the associations. Each layer has its own learning process and constitutes the input of the next layer. The results from PGMRA are interpretable and have been able to decrease the missing heritability and identify the epistatic sets of markers that are composed of the genotypic–phenotypic architecture of a disease or trait (34).

Transcriptomics

Transcriptomics quantifies the expression level of all RNA transcripts that are produced in a cell. Transcriptomics raw data are usually processed to generate expression matrices containing an estimate of expression level of each gene or transcript across several samples and conditions, which are typically the input of DL methods. There is a broad range of transcriptomics applications in which DL has been successfully applied. For example, one of the main goals of gene expression data is the analysis of alternative splicing (i.e., the synthesis of different transcripts isoforms from the same gene). In this context, Zhang et al. notably achieved to analyze differential splicing between different samples using RNA-seq data and combining a DNN and a Bayesian statistical model (35). On the other hand, CNNs have been applied to identify actual splice junctions from false positives generated during RNA-seq reads alignment (36). In addition, Jha et al. proposed a model to integrate RNA-seq and CLIP-seq data in order to improve the study of alternative splicing (37).

Another major research focus in transcriptomics is the prediction of other types of RNAs, such as non-coding RNAs (ncRNAs), and the characterization of their expression. In this context, Hill et al. proposed an RNN to differentiate between coding and non-coding RNAs (38), demonstrating the capability of their algorithm to identify ncRNAs without providing their model with previous knowledge. Tripathi et al. developed a method to detect long ncRNAs (lncRNAs) (39). They reached a remarkable 99% accuracy rate applying a DFF to reference databases. Long intergenic ncRNAs (lincRNAs), a type of lncRNAs which are transcribed in intergenic regions, have been also successfully predicted feeding an AE with previous knowledge about lincRNAs (40).

Epigenomics

Epigenomic studies identify modifications in DNA that comprise markers that can potentially alter gene expression without modifying the DNA sequence itself. There are several epigenetic markers such as DNA methylation, histone modification, and specifically positioned nucleosomes. DNA methylation perhaps is the most studied epigenetic modification. DNA methylation studies generate methylation matrices that, like gene expression matrices, can be used for biomarker discovery or disease classification problems. In this context, DL methods have been used to accurately predict the sequences recognized by DNA- and RNA-binding proteins using CNNs (41). A key advantage of this method is the capability to integrate data from different technologies used in epigenomics studies, like chromatin immunoprecipitation (ChIP)-seq or cross-linking immunoprecipitation (CLIP)-seq. DNase I sequencing data have been also used for predicting the three-dimensional chromatin state in a cell using CNN (42). On the other hand, Wang et al. accurately predicted DNA methylation state feeding SAEs with sequence and Hi-C data (43). Histone modifications, similar to DNA methylation, do not affect DNA sequence but can modify its availability to the transcriptional machinery. Using CNNs, Yin et al. designed an algorithm to predict these histone modifications by integrating sequence and DNase data (44). In addition, Singh et al. used a CNN to infer gene expression from histone modifications data (45), while Sekhon et al. used a LSTM to predict differential gene expression, also from histone modifications data (46).

Proteomics and metabolomics

Proteomics comprises a set of techniques that can be used to quantify expression levels, post-translational modifications or localization of proteins in a cell or a biological sample. Metabolomics is the study of a complete metabolome, which are small molecules that participate in general metabolic reactions. The technologies used by these omics-streams are, among others, mass spectrometry (MS) or nuclear magnetic resonance (NMR), and the first challenge for researchers in this field is to assign raw instrumental signals to proteins or metabolites.

In proteomics, the most common experimental strategy is to split proteins into short amino acid chains (peptides) and to analyze these peptides in an MS. The MS output signals are compared to peptide profiles stored in public or proprietary databases to identify them. However, these databases are still incomplete

and inaccurate. In this context, Zhou et al. developed a software that uses a LSTM network to predict peptide MS/MS spectra (47). Knowing peptide spectra a priori facilitates the task of assigning MS/MS spectra to peptides comparing them to the theoretical spectra. Another proteomics application is de novo peptide sequencing, which is essential for proteins characterization. In this field, Tran et al. surpassed previous software combining CNN and LSTMs networks to effectively accomplish such a difficult task (48). Once the collection of peptides has been sequenced in a proteomics sample, the next challenge is to identify the proteins of origin of such peptides. Kim et al. addressed this problem through a CNN (49), getting better results than other dedicated libraries for this task. DL has been also applied to predict protein secondary structures from their amino acid sequences (50).

NMR technology is essential for both proteomics and metabolomics data generation. However, it has the technical limitation to return many noise signals that should be filtered in order to improve accuracy. Kobayashi et al. automated this necessary step by applying CNNs to remove noise peaks from NMR spectra (51), thereby improving the performance.

Applying DL methods to metabolomics data is especially challenging because they are unable to identify specific factors that contribute to individual samples, which is essential in these types of experiments (52). Despite this fact, some DL applications have been developed in this field providing interesting results. For instance, Date and Kikuchi combined DNN and mean decrease accuracy metric to analyze NMR-based metabolomics data (52). Asakura et al. also applied DNNs to metabolomics data, overperforming other ML applications (53).

CLINICAL APPLICATIONS AND PRECISION MEDICINE

Precision medicine basically aims to move away from general therapies for a broad population to individualized targeted therapies and treatment protocols depending on each patient's molecular background (54), or establish preventive medicine strategies based on disease susceptibility estimation (55). Omics data have a key role in this transition as they enable studying diseases from several simultaneous levels (e.g., DNA sequence, gene expression, and medical images) and identifying which parts of the complex biological functions are altered. In this new scenario, several ML-based approaches have been applied to medicine (56). However, although ML has been demonstrated to be useful in several precision medicine applications, it has some disadvantages that can be overcome by DL architectures. For instance, ML performance has a strong dependence on the data preprocessing to extract features, while DL models include this feature extraction (57).

Biomarker discovery and patient classification

One of the most common applications of omics technologies in biomedical research is the identification of new biomarkers for early diseases diagnosis, treatment response, and classification. The availability of large amounts of public omics data, especially in cancer, such as The Cancer Genome Atlas (TCGA) (58), has permitted the identification of new biomarkers with both DL and

non-DL strategies. A promising study applied an SDAE to classify breast cancer samples from the TCGA database into healthy or diseased using gene expression data (59). In addition, this method identified a set of highly interactive genes which could be good cancer biomarkers. Gene expression data from TCGA have also been exploited to accurately differentiate samples into different cancer types (60). On the other hand, Si et al. used an AE to classify healthy and breast cancer patients using methylation data (61), while Chatterjee et al. used CNN to classify different cancer types by their methylation patterns, achieving very promising results (62). Multiple omics (RNA-seq, miRNA-seq, and methylation data) have been combined by Chaudhary et al. to classify liver cancer patients into different survival groups (63). Authors used TCGA data to train their AE model, but they expect to improve their method using more clinical data in the future. In a similar work, Olivier et al. integrated the same kinds of omics data from TCGA to stratify bladder cancer patients by their survival chances (64). They used an AE approach to split patients into two survival groups. They also used these clusters to identify biomarkers linked to survival rates. Biomarkers for Alzheimer's disease have also been proposed using DFFs (65). Another precision oncology application is a tool developed by Yuan et al. to classify cancer types based on somatic mutations (66). The authors combined a DFF with other statistical techniques. They trained and tested their method with TCGA data for 12 cancer types.

Medical imaging

Medical imaging is one of the main tools for the transition from traditional medicine to precision medicine. This section reviews some DL-based imaging applications in the context of disease classification and diagnosis.

In skin cancer, the first step for diagnosing is based on visual inspections by dermatologists. Consequently, skin cancer diagnosis is a classical image recognition problem where researchers have applied ML methods and image recognition approaches. In a recent work, Estava et al. trained a CNN with thousands of clinical images to automatically identify whether a skin lesion is a skin cancer symptom (67). With their method, they obtained results as good as a panel of expert dermatologists. Some other studies addressed this problem with CNNs (68), all of them with promising results, and it is expected that this research will be translated in a few years into mobile applications able to accurately diagnose skin cancer lesions.

In the context of brain cancer, tumor segmentation is essential to define the shape and size of the tumor and apply diagnoses and therapies accordingly. This tumor segmentation is usually made manually by doctors using magnetic resonance imaging (MRI) images. However, this crucial task is very time-consuming and subjective. Therefore, there has been a lot of interest in automating tumor segmentation from MRI data. This task is very challenging because MRI data consist of 3D images where tumors are very different between patients, and in addition, they are very heterogeneous images depending on the device and experimental procedures employed (69). Several researchers addressed this challenge using CNNs (70–72) or SAEs (73).

Analysis of histopathological images is one of the most common tests for cancer diagnosis. As with brain tumor segmentation, the analysis of images is manually performed by pathologists, which is a time-consuming task. In this context,

several attempts have been made in order to automate this process. Litjens et al. reported a CNNs-based strategy for prostate and breast cancer diagnosis (74), although their results are very preliminary and much more research is necessary in this field. In addition, Xie et al. recently combined different DL algorithms to classify breast cancer subtypes from histopathological images (75). Colorectal polyps have been also classified applying a ResNet (76).

Computed tomography (CT) is used for the diagnosis of several diseases due to its capacity to generate three-dimensional anatomic images. Some DL approaches will likely enable the use of CT images in precision medicine. Roth et al., for instance, proposed the application of CNNs to automatically classify CT images into the different human anatomical parts (77). Such classification is the first step in many CT-based diagnostic strategies. There are also some specific applications in this field, for instance, for pancreas segmentation (78) or coronary artery calcium scoring (79).

Ultrasound (US) imaging is another imaging technique with many medical applications, for instance, in heart dysfunctions diagnosis. Carneiro and Nascimento innovated this field using DBNs to left ventricle endocardium tracking, allowing the automatic detection of different cardiopathies (80). On the other hand, Lekadir et al. applied a CNN to characterize carotid plaque composition (81). In addition, Biswas et al. developed a DL method to characterize liver US images, allowing the diagnosis and stratification of liver pathologies (82).

Some DL methods have been also applied to X-ray images. For instance, Nasr-Esfahani et al. used a CNN to detect vessel regions, a necessary step for coronary artery disease diagnosis (83). Bone age assessment is a common technique to detect growth abnormalities, and currently, it is done manually by comparing the X-ray images from databases. However, some authors applied DL algorithms to automate this process (84, 85).

Finally, facial images are being used with very promising results for automatic disease diagnosis. In a very recent work, Gurovich et al. have presented a facial analysis framework for genetic syndrome classification (86). They used patient facial images and CNNs to quantify similarities of facial features to hundreds of syndromes outperforming clinicians in diagnosis tasks.

CONCLUSION

Omics technologies are not only changing the way we study biomedicine but also introducing novel analytical challenges to bioinformatics analysts. DL is a promising approach to analyze these complex and heterogeneous datasets to drive precision medicine. This chapter reviewed some of the most common DL applications in omics data analysis and precision medicine. Although these methods have been used with very promising results, there are important considerations to take into account. The most successful application of DL in biomedical research to date has been in supervised learning; therefore, a crucial step is to avoid biases in training sets as quality of learning depends on the quality of the input data. No single method is universally applicable, and the choice of whether and how to use DL approaches will be problem-specific. Conventional analytical approaches will remain valid and have advantages when data are scarce or if the aim is to assess

statistical significance, which is currently difficult using DL methods. Another limitation of DL is the increased complexity, which applies both to model design and to the required computing environment. The application of DL methods to omics and precision medicine is a very new field. Although there are still some limitations, there is an increasing interest and research efforts that is resolving the major shortcomings and providing with very promising applications. The increasing availability of a larger number of omics datasets, medical images and clinical health records is fuelling the promising applications of DL technology that in the near future will play an increasingly important role in this field.

Acknowledgement: JMM was partially funded by Ministerio de Economía, Industria y Competitividad. This work was partially supported by Junta de Andalucía (PI-0173-2017).

Conflict of Interest: The authors declare no potential conflicts of interest with respect to research, authorship and/or publication of this chapter.

Copyright and Permission Statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D20–6. http://dx.doi.org/10.1093/nar/gkv1352
2. Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. OMICS J Integr Biol. 2018 Oct 1;22(10):630–6. http://dx.doi.org/10.1089/omi.2018.0097
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44. http://dx.doi.org/10.1038/nature14539
4. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851–69.
5. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989 Dec;1(4):541–51. http://dx.doi.org/10.1162/neco.1989.1.4.541
6. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE (Institute of Electrical and Electronics Engineers); 2009. p. 248–55. http://dx.doi.org/10.1109/CVPR.2009.5206848
7. Orenstein EC, Beijbom O. Transfer learning and deep feature extraction for planktonic image data sets. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, CA: IEEE (Institute of Electrical and Electronics Engineers); 2017. p. 1082–8.
8. Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 2017;29(9):2352–449. http://dx.doi.org/10.1162/neco_a_00990
9. Hinton GE. Deep belief networks. Scholarpedia. 2009 May 31;4(5):5947. http://dx.doi.org/10.4249/scholarpedia.5947
10. Fischer A, Igel C. An introduction to restricted Boltzmann machines. In: Alvarez L, Mejail M, Gomez L, Jacobo J, editors. Progress in pattern recognition, image analysis, computer vision, and applications. Berlin: Springer; 2012. p. 14–36. (Lecture Notes in Computer Science).

11. Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for Boltzmann Machines*. Cogn Sci. 1985;9(1):147–69. http://dx.doi.org/10.1207/s15516709cog0901_7
12. Rumelhart DE, McClelland JL. Learning internal representations by error propagation. In: Parallel distributed processing: Explorations in the microstructure of cognition: Foundations [homepage on the Internet]. MITP; 1987 [cited 2019 May 16]. Available from: https://ieeexplore.ieee.org/document/6302929
13. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006 Jul 28;313(5786):504–7. http://dx.doi.org/10.1126/science.1127647
14. Ng A, others. Sparse autoencoder. CS294A Lect Notes. 2011;72(2011):1–19.
15. Zhang Z, Zhao Y, Liao X, Shi W, Li K, Zou Q, et al. Deep learning in omics: A survey and guideline. Brief Funct Genomics. 2018 Sep 26; http://dx.doi.org/10.1093/bfgp/ely030
16. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7. http://dx.doi.org/10.1038/nbt.4235
17. Qi H, Chen C, Zhang H, Long JJ, Chung WK, Guan Y, et al. MVP: Predicting pathogenicity of missense variants by deep neural networks. bioRxiv. 2018 Feb 2;259390. http://dx.doi.org/10.1101/259390
18. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinforma Oxf Engl. 2015 Mar 1;31(5):761–3. http://dx.doi.org/10.1093/bioinformatics/btu703
19. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015 Oct;12(10):931–4. http://dx.doi.org/10.1038/nmeth.3547
20. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017 Nov 17;18(Suppl 9):845. http://dx.doi.org/10.1038/nmeth.3547
21. Liu F, Li H, Ren C, Bo X, Shu W. PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. Sci Rep. 2016 22;6:28517. http://dx.doi.org/10.1038/srep28517
22. Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics. 2017 Dec 1;18(Suppl 13):478. http://dx.doi.org/10.1186/s12859-017-1878-3
23. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: A general computational framework for predicting enhancers. Nucleic Acids Res. 2015 Jan;43(1):e6. http://dx.doi.org/10.1093/nar/gku1058
24. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics. 2018 31;19(1):202. http://dx.doi.org/10.1186/s12859-018-2187-1
25. Eser U, Churchman LS. FIDDLE: An integrative deep learning framework for functional genomic data inference. bioRxiv. 2016 Oct 17;081380. http://dx.doi.org/10.1101/081380
26. Wang M, Tai C, E W, Wei L. DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic Acids Res. 2018 Jun 20;46(11):e69. http://dx.doi.org/10.1093/nar/gky215
27. Oubounyt M, Louadi Z, Tayara H, Chong KT. DeePromoter: Robust promoter predictor using deep learning. Front Genet. 2019;10:286. http://dx.doi.org/10.3389/fgene.2019.00286
28. Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. SpliceRover: Interpretable convolutional neural networks for improved splice site prediction. Bioinforma Oxf Engl. 2018 Dec 15;34(24):4180–8. http://dx.doi.org/10.1093/bioinformatics/bty497
29. Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 2009 Oct;5(10):e1000678. http://dx.doi.org/10.1371/journal.pgen.1000678
30. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009 Aug;460(7256):748–52. http://dx.doi.org/10.1038/nature08185
31. Montaez CAC, Fergus P, Montaez AC, Hussain A, Al-Jumeily D, Chalmers C. Deep learning classification of polygenic obesity using genome wide association study SNPs. In: 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE, 2018. p. 1–8.

32. Fergus P, Montanez A, Abdulaimma B, Lisboa P, Chalmers C, Pineles B. Utilising deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American Women. IEEE/ACM Trans Comput Biol Bioinform. IEEE, 2018:1–1. http://dx.doi.org/10.1109/TCBB.2018.2868667
33. Arnedo J, del Val C, de Erausquin GA, Romero-Zaliz R, Svrakic D, Cloninger CR, et al. PGMRA: A web server for (phenotype x genotype) many-to-many relation analysis in GWAS. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W142–149. http://dx.doi.org/10.1093/nar/gkt496
34. Arnedo J, Svrakic DM, Del Val C, Romero-Zaliz R, Hernández-Cuervo H, Molecular genetics of Schizophrenia Consortium, et al. Uncovering the hidden risk architecture of the schizophrenias: Confirmation in three independent genome-wide association studies. Am J Psychiatry. 2015 Feb 1; 172(2):139–53. http://dx.doi.org/10.1176/appi.ajp.2014.14040435
35. Zhang Z, Pan Z, Ying Y, Xie Z, Adhikari S, Phillips J, et al. Deep-learning augmented RNA-seq analysis of transcript splicing. Nat Methods. 2019;16(4):307–10. http://dx.doi.org/10.1038/s41592-019-0351-9
36. Zhang Y, Liu X, MacLeod J, Liu J. Discerning novel splice junctions derived from RNA-seq alignment: A deep learning approach. BMC Genomics. 2018 Dec 27;19(1):971. http://dx.doi.org/10.1186/s12864-018-5350-1
37. Jha A, Gazzara MR, Barash Y. Integrative deep models for alternative splicing. Bioinforma Oxf Engl. 2017 Jul 15;33(14):i274–82. http://dx.doi.org/10.1093/bioinformatics/btx268
38. Hill ST, Kuintzle R, Teegarden A, Merrill E, Danaee P, Hendrix DA. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. Nucleic Acids Res. 2018 Sep 19;46(16):8105–13. http://dx.doi.org/10.1093/nar/gky567
39. Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK. DeepLNC, a long non-coding RNA prediction tool using deep neural network. Netw Model Anal Health Inform Bioinforma. 2016 Jun 10; 5(1):21. http://dx.doi.org/10.1007/s13721-016-0129-2
40. Yu N, Yu Z, Pan Y. A deep learning method for lincRNA detection using auto-encoder algorithm. BMC Bioinformatics. 2017 Dec 6;18(Suppl 15):511. http://dx.doi.org/10.1186/s12859-017-1922-3
41. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015 Aug;33(8):831–8. http://dx.doi.org/10.1038/nbt.3300
42. Schreiber J, Libbrecht M, Bilmes J, Noble WS. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv. 2018 Jul 15;103614. http://dx.doi.org/10.1101/103614
43. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo Y-Y, et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep. 2016 Jan 22;6:19598. http://dx.doi.org/10.1038/srep19598
44. Yin Q, Wu M, Liu Q, Lv H, Jiang R. DeepHistone: A deep learning approach to predicting histone modifications. BMC Genomics. 2019 Apr 4;20(Suppl 2):193. http://dx.doi.org/10.1186/s12864-019-5489-4
45. Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: Deep-learning for predicting gene expression from histone modifications. Bioinforma Oxf Engl. 2016 01;32(17):i639–48. http://dx.doi.org/10.1093/bioinformatics/btw427
46. Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. Bioinforma Oxf Engl. 2018 01;34(17):i891–900. http://dx.doi.org/10.1093/bioinformatics/bty612
47. Zhou X-X, Zeng W-F, Chi H, Luo C, Liu C, Zhan J, et al. pDeep: Predicting MS/MS spectra of peptides with deep learning. Anal Chem. 2017 05;89(23):12690–7. http://dx.doi.org/10.1021/acs.analchem.7b02566
48. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. Proc Natl Acad Sci U S A. 2017 Jul 18; http://dx.doi.org/10.1073/pnas.1705691114
49. Kim M, Eetemadi A, Tagkopoulos I. DeepPep: Deep proteome inference from peptide profiles. PLoS Comput Biol. 2017 Sep;13(9):e1005661. http://dx.doi.org/10.1371/journal.pcbi.1005661
50. Spencer M, Eickholt J, Jianlin Cheng null. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinform. 2015 Feb;12(1):103–12. http://dx.doi.org/10.1109/TCBB.2014.2343960

51. Kobayashi N, Hattori Y, Nagata T, Shinya S, Güntert P, Kojima C, et al. Noise peak filtering in multi-dimensional NMR spectra using convolutional neural networks. Bioinforma Oxf Engl. 2018 Dec 15; 34(24):4300–1. http://dx.doi.org/10.1093/bioinformatics/bty581
52. Date Y, Kikuchi J. Application of a deep neural network to metabolomics studies and its performance in determining important variables. Anal Chem. 2018 06;90(3):1805–10. http://dx.doi.org/10.1021/acs.analchem.7b03795
53. Asakura T, Date Y, Kikuchi J. Application of ensemble deep neural network to metabolomics studies. Anal Chim Acta. 2018 Dec 11;1037:230–6. http://dx.doi.org/10.1016/j.aca.2018.02.045
54. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016 16;17(9):507–22. http://dx.doi.org/10.1038/nrg.2016.86
55. Chen R, Snyder M. Promise of personalized omics to precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2013 Feb;5(1):73–82. http://dx.doi.org/10.1002/wsbm.1198
56. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019 04;380(14):1347–58. http://dx.doi.org/10.1056/NEJMra1814259
57. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: New computational modelling techniques for genomics. Nat Rev Genet. 2019 Apr 10; http://dx.doi.org/10.1038/s41576-019-0122-6
58. Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20. http://dx.doi.org/10.1038/ng.2764
59. Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. Pac Symp Biocomput Pac Symp Biocomput. 2017;22:219–29. http://dx.doi.org/10.1142/9789813207813_0022
60. Lyu B, Haque A. Deep learning based tumor type classification using gene expression data. bioRxiv. 2018 Jul 11;364323. http://dx.doi.org/10.1101/364323
61. Si Z, Yu H, Ma Z. Learning deep features for DNA methylation data analysis. IEEE Access. 2016;4:2732–7. http://dx.doi.org/10.1109/ACCESS.2016.2576598
62. Chatterjee S, Iyer A, Avva S, Kollara A, Sankarasubbu M. Convolutional neural networks in classifying cancer through DNA methylation. ArXiv180709617 Cs Q-Bio Stat [Internet]. 2018 Jul 24 [cited 2019 May 6]; Available from: http://arxiv.org/abs/1807.09617
63. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-Omics integration robustly predicts survival in liver cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2018 Mar 15;24(6):1248–59. http://dx.doi.org/10.1158/1078-0432.CCR-17-0853
64. Poirion OB, Chaudhary K, Garmire LX. Deep learning data integration for better risk stratification models of bladder cancer. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:197–206.
65. Zafeiris D, Rutella S, Ball GR. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. Comput Struct Biotechnol J. 2018;16:77–87. http://dx.doi.org/10.1016/j.csbj.2018.02.001
66. Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, et al. DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics. 2016 Dec 23;17(Suppl 17):476. http://dx.doi.org/10.1186/s12859-016-1334-9
67. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 02;542(7639):115–8. http://dx.doi.org/10.1038/nature21056
68. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: Systematic review. J Med Internet Res. 2018 Oct 17;20(10):e11936. http://dx.doi.org/10.2196/11936
69. Işın A, Direkoğlu C, Şah M. Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Comput Sci. 2016 Jan 1;102:317–24. http://dx.doi.org/10.1016/j.procs.2016.09.407
70. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging. 2016;35(5):1240–51. http://dx.doi.org/10.1109/TMI.2016.2538465

71. Naceur MB, Saouli R, Akil M, Kachouri R. Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images. Comput Methods Programs Biomed. 2018 Nov;166:39–49. http://dx.doi.org/10.1016/j.cmpb.2018.09.007
72. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. Med Image Anal. 2017;35:18–31. http://dx.doi.org/10.1016/j.media.2016.05.004
73. Xiao Z, Huang R, Ding Y, Lan T, Dong R, Qin Z, et al. A deep learning-based segmentation method for brain tumor in MR images. In: 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). Atlanta, GA: IEEE, 2016. p. 1–6.
74. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016 23;6:26286. http://dx.doi.org/10.1038/srep26286
75. Xie J, Liu R, Luttrell J, Zhang C. Deep learning based analysis of histopathological images of breast cancer. Front Genet. 2019;10:80. http://dx.doi.org/10.3389/fgene.2019.00080
76. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inform. 2017;8:30. http://dx.doi.org/10.4103/jpi.jpi_34_17
77. Roth HR, Lee CT, Shin H-C, Seff A, Kim L, Yao J, et al. Anatomy-specific classification of medical images using deep convolutional nets. 2015 IEEE 12th Int Symp Biomed Imaging ISBI. 2015 Apr. p. 101–4. http://dx.doi.org/10.1109/ISBI.2015.7163826
78. Roth HR, Farag A, Lu L, Turkbey EB, Summers RM. Deep convolutional networks for pancreas segmentation in CT imaging. ArXiv150403967 Cs. 2015 Mar 20;94131G. http://dx.doi.org/10.1117/12.2081420
79. Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Išgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med Image Anal. 2016;34:123–36. http://dx.doi.org/10.1016/j.media.2016.04.004
80. Carneiro G, Nascimento JC. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Trans Pattern Anal Mach Intell. 2013 Nov;35(11):2592–607. http://dx.doi.org/10.1109/TPAMI.2013.96
81. Lekadir K, Galimzianova A, Betriu A, Del Mar Vila M, Igual L, Rubin DL, et al. A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound. IEEE J Biomed Health Inform. 2017;21(1):48–55. http://dx.doi.org/10.1109/JBHI.2016.2631401
82. Biswas M, Kuppili V, Edla DR, Suri HS, Saba L, Marinhoe RT, et al. Symtosis: A liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm. Comput Methods Programs Biomed. 2018;155:165–77. http://dx.doi.org/10.1016/j.cmpb.2017.12.016
83. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Ward K, Jafari MH, et al. Vessel extraction in X-ray angiograms using deep learning. Conf Proc IEEE Eng Med Biol Soc. 2016;2016:643–6. http://dx.doi.org/10.1109/EMBC.2016.7590784
84. Lee JH, Kim KG. Applying deep learning in medical images: The case of bone age estimation. Healthc Inform Res. 2018 Jan;24(1):86–92. http://dx.doi.org/10.4258/hir.2018.24.1.86
85. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal. 2017;36:41–51. http://dx.doi.org/10.1016/j.media.2016.10.010
86. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60–4. http://dx.doi.org/10.1038/s41591-018-0279-0
87. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, et al. Handwritten digit recognition with a back-propagation network. In: Touretzky DS, editor. Advances in neural information processing systems 2 [homepage on the Internet]. Morgan-Kaufmann; 1990 [cited 2019 May 16]. p. 396–404. Available from: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf
88. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25 [homepage on the Internet]. Curran Associates, Inc.; 2012 [cited 2019 May 16].

p. 1097–1105. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf
89. Szegedy C, Wei Liu, Yangqing Jia, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [homepage on the Internet]. Boston, MA, USA: IEEE; 2015 [cited 2019 May 16]. p. 1–9. Available from: http://ieeexplore.ieee.org/document/7298594/
90. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv14091556 Cs [homepage on the Internet]. 2014 Sep 4 [cited 2019 May 16]; Available from: http://arxiv.org/abs/1409.1556
91. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE, 2016. p. 770–8.
92. Huang G, Liu Z, Maaten L v d, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017. p. 2261–9.
93. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging. 2013 Dec;26(6): 1045–57. http://dx.doi.org/10.1007/s10278-013-9622-7
94. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017. p. 3462–71.

4

Biological Sequence Analysis

Usman Saeed[1,2] • Zainab Usman[2]

[1]Dennemeyer Octimine GmbH, München, Germany; [2]Department of Bioinformatics, Technical University Munich, Wissenschaftzentrum Weihenstephan, Freising, Germany

Author for correspondence: Usman Saeed, Dennemeyer Octimine GmbH, Landaubogen 1-3, 81373 München, Germany. Email: usman.saeed08@gmail.com

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch4

Abstract: This chapter focuses on several biological sequence analysis techniques used in computational biology and bioinformatics. The first section provides an overview of biological sequences (nucleic acids and proteins). Bioinformatics helps us understand complex biological problems by investigating similarities and differences that exist at sequence levels in poly-nucleic acids or proteins. Alignment algorithms such as dynamic programming, basic local alignment search tool and HHblits are discussed. Artificial intelligence and machine learning methods have been used successfully in analyzing sequence data and have played an important role in elucidating many biological functions, such as protein functional classification, active site recognition, protein structural features identification, and disease prediction outcomes. This chapter discusses both supervised and unsupervised learning, neural networks, and hidden Markov models. Sequence analysis is incomplete without discussing next-generation sequencing (NGS) data. Deep sequencing is highly important due to its ability to address an increasingly diverse range of biological problems such as the ones encountered in therapeutics. A complete NGS workflow to generate a consensus sequence and haplotypes is discussed.

Keywords: dynamic programming; machine learning; next-generation sequencing; pairwise alignment; sequence analysis.

In: *Computational Biology*. Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

It has been estimated that over 12 million different species exist on the planet (1). The biodiversity across all life forms including plants, animals, and microbes can be attributed to their unique genomic and proteomic composition. Like an instruction manual that guides about all the sequential tasks to be done in the right order to accomplish a process, the biological organisms have all the details in their genes, creating combinations of nucleotides resulting in the diversity that we see in the biological world. There are two types of nucleic acids, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). In 1953, Watson and Crick proposed that the DNA is made up of two long poly-nucleotide chains comprising of four nucleotides, namely adenine (A), guanine (G), cytosine (C), and thymine (T) (2). In RNA, however, thymine is replaced by the nucleotide uracil (U) as a complementary nucleotide to adenine. The strands in both DNA and RNA have a polyphosphate backbone with adjacent nucleotides forming polyphosphate di-ester bonds. DNA is a double-stranded structure; the two chains are twisted around each other with hydrogen bonds between the base portions of nucleotides holding the two chains together. The sequence of bases in DNA is of crucial importance as it contains the code to the formation of diverse proteins and hence the complexity and diversity of life. The unique order of bases in DNA results in the creation of basic hereditary units called genes. In 2003, the human genome project initially estimated 20,000 genes in the human genome (3, 4), and these estimates were later revised to 25,000–30,000 genes (5). Based on the sequence of DNA, enzymes like RNA polymerase create single-stranded messenger RNA (mRNA) that later translate into proteins. This whole process of decoding the DNA sequence into a protein is referred to as the "central dogma of life" (6). Depending on different organisms, all genes may not code for proteins. Composed of amino acids, proteins are much more complicated than nucleic acids. There are 20 major amino acids which make up proteins, and each protein can have them assembled in different numbers and order. Amino acid sequence of proteins is also of crucial importance as it not only determines the physiochemical properties of proteins but also determines the different conformations they can create in a three-dimensional space (7). These conformational changes result in complicated protein structures that in turn allows them to serve unique biological functions, for example, transport, functional regulation, and homeostasis. Therefore, the importance of nucleotide sequence in DNA/RNA and of amino acids in proteins cannot be overstated.

Sequence comparison of DNA can allow us to compare the differences at gene level across different organisms and species. Comparative genomics is a branch of science that uses bioinformatics techniques extensively to trace the genes across multiple species and study their similarities and differences. Such studies help us infer the functional and structural characteristics of newly found or existing proteins. Programmatically, biological sequence analysis is not much different than comparing strings and text, and thus, developing the concept of alignment is important. Sequences evolving over species and clades through mutations include insertions, deletions (indels), and mismatches. When comparing two biological sequences, an alignment is generated to view differences between the sequences at each position.

PAIRWISE ALIGNMENT AND DYNAMIC PROGRAMMING

Pairwise alignment involves comparing two sequences against each other and finding the best possible alignment between them. The process involves scoring at each position for match, mismatch, and indels. Since matches are preferred over deletions, matches are normally assigned the highest scores, and lowest for insertions. Similarity between two sequences is inversely proportional to the number of mismatches and indels in their alignment. Although the scoring for alignment can be as simple as +1 for match, 0 for mismatch, and –2 for insertion, different scoring models have been developed based on the statistically relevant frequencies of one amino acid changing into another.

Needleman–Wunsch algorithm

Initially developed by Needleman and Wunsch in 1970, the algorithm is based on dynamic programming and allows for global or end-to-end alignment of two sequences (8). The algorithm involves three main steps, namely initialization, calculation, and trace back. A matrix of dimensions *i, j* is initialized, where *i* and *j* are the length of two sequences under comparison. In the second step, *F(i, j)* highest score for each comparison at each position is calculated,

$$F(i, j) = max\ \{F(i–1, j–1) + s(x_i, y_i), F(i–1, j) – d, F(i, j–1) – d\}$$

where "$s(x_i, y_i)$" is the match/mismatch score and "d" is the penalty for deletion.

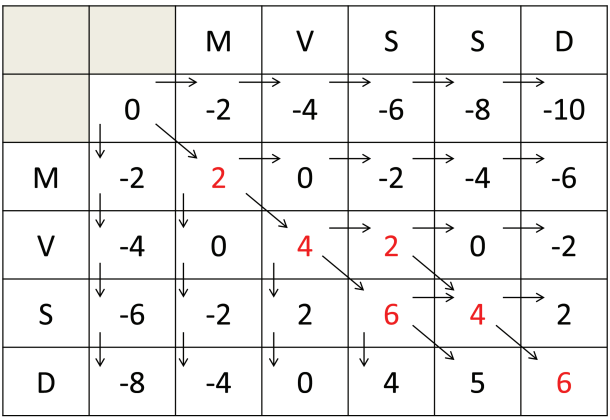
After the maximum score for each position in the matrix is calculated (Figure 1), trace back starts from the last cell (bottom right) in the matrix. Each step involves moving from the current cell to the one from which the value of the current cell was derived. A match or mismatch is assigned if the maximum score was derived from a diagonal cell. Insertion/deletion is assigned if the score was derived from the top or left cell. After the trace back is completed, we have two sequences aligned end to end with each other with an optimal alignment score (9).

Smith–Waterman algorithm

Initially proposed by Smith and Waterman in 1981, the algorithm allows for local sequence alignment and is like the Needleman–Wunsch algorithm (10). Local sequence alignment can be used in situations where it is required to align smaller subsequences of two sequences. In the biological context, such a situation may arise while searching for a domain or motif within larger sequences. The algorithm comprises of the same steps as Needleman–Wunsch; however, there are two main differences. Computation of max score also includes an option of 0:

$$F(i, j) = max\ \{0, F(i–1, j–1) + s(x_i, y_i), F(i–1, j) – d, F(i, j–1) – d\}$$

Assignment of "0" as max score corresponds to starting a new alignment. It allows for alignments to end anywhere within the matrix. The trace back therefore starts from the highest value of *F(i, j)* in the matrix and ends where it encounters 0.



Alignment 1:

M	V	S	S	D
M	V	S	-	D

Alignment 2:

M	V	S	S	D
M	V	-	S	D

Figure 1 **Needleman–Wunsch matrix.** The calculation uses scores for match +2, mismatch –1, and gap –2. The arrows show the matrix cell from where the value is generated. Red-coloured cell values show the trace back that creates alignment.

HEURISTIC LOCAL ALIGNMENT

One main challenge in bioinformatics sequence analysis is decoding the vast number and length of sequences. These big data of protein and DNA sequence databases (over 100 million sequences) come from species across the tree of life. Although the local alignment methods based on dynamic programming are quite accurate and guarantee to find an optimally scored alignment, they are slow and not practical for sequence alignments against databases with millions of sequences. The time complexity of dynamic programming algorithms is $O(mn)$, that is, the product of sequence lengths. In the initial attempts to improve the speed for sequence comparisons, heuristic algorithms like BLAST (11), BLAT (12), and FASTA (13, 14) were created. Further advancements in the efficiency of similarity search algorithms came with algorithms like LSCluster (15), Usearch (16), Vsearch (17), Diamond (18) and Ghostx (19). In general, these algorithms search for exact matches and extend the alignment from those matches trying to estimate the optimal scoring alignment.

Basic Local Alignment Search Tool, initially developed by Altschul and colleagues (11), is based on the idea that the best scoring sequence alignment would contain the highest number of identical matches or highly scoring sub-alignments. The algorithm carries out the following steps: (i) reduce the query sequence into small subsequences called seeds, (ii) search these seeds across the database for exact matches, and (iii) extend the exact matches into an un-gapped alignment until a maximal scoring extension is reached. The use of seeds to first search for

exact matches greatly increases the whole searching process and the un-gapped alignment misses only a small set of significant matches. The accuracy and sensitivity of BLAST made it amongst the most widely used search algorithm in the biological world. A variant of BLAST named Position-Specific-Iterative BLAST (PSI-BLAST) extends the basic BLAST algorithm (20). PSI-BLAST performs multiple iterations of BLAST and uses the hits found in one iteration as a query for the next iteration. Although slower due to sheer amount of calculations required, PSI-BLAST is considered a reliable tool to find distant homology relationships.

Although BLAST and PSI-BLAST are extensively used, recently developed methods offer results with higher accuracy and sensitivity. Hidden Markov models (HMM) have been used efficiently for numerous applications to understand and explore biological data. One such example is HMM–HMM-based lightning fast sequence search (HHblits) introduced in 2012 (21). The tool can be used as an alternative for BLAST and PSI-BLAST and is 50 to 100 times more sensitive. The high sensitivity of the tool can be attributed to the algorithm which relies on comparing the HMM representations of the sequences. Although profile–profile or HMM–HMM alignments are very slow to compute, the prefilter in HHblits reduces the required alignments from millions to thousands, thus giving it a considerable speed advantage. HHblits represents each sequence in the database as a profile HMM. Prefiltering reduces the number of HMM comparisons for similarity search by selecting only those target sequences where the largest un-gapped alignment exists, and a Smith–Waterman based alignment reveals a significant E-value.

MACHINE LEARNING AND SEQUENCE ANALYSIS

Biological data provide amongst the perfect use cases of machine learning and artificial intelligence algorithms. This is the reason that researchers in the field of bioinformatics and computational biology have used statistical analysis and inference since the very beginning. Techniques like maximum likelihood (22) and neighbor joining (23) have been used for comparative genomics. Naïve Bayes and Markov chains have been extensively used for sequence analysis. Logistic regressions, support vector machines, and random forests have been used in numerous applications ranging from prediction of protein sequence or structural elements to classification of proteins into different structural and functional classes. With the development of deep neural networks, we observe an increase in the use of the algorithms like long short-term memory (LSTM) (24) and convolutional neural networks (CNN or ConvNet) (25) to predict the different features and behavior of proteins, for example, protein contact prediction and prediction of post-translational modifications.

Machine learning methods are broadly divided into two types, supervised and un-supervised learning. Based on the inherent features of the data, if it is not labeled and cannot be assigned to any type, then classification is done using unsupervised learning. For instance, the classification of proteins into different groups is done based on their sequence similarity to each other. K-means clustering algorithm (26) and Markov clustering (27) can be used in unsupervised classification. On the other hand, if the data are labeled into different sets, this information can be used to train the computer by showing it positive and negative examples.

Once the training is complete, the accuracy of training can be tested using similar data not used in the training dataset. Any classification technique following training and testing procedures using labeled data is termed supervised machine learning. Examples for this type of learning include SVM, HMM, random forest, and CNN.

Hidden Markov Models

HMM is a statistical method that can be used to predict the probability of occurrence for a future event. HMMs provide the foundations for a range of complex models that can be used for multiple sequence alignment, profile searches or detection of sequence elements. In order to understand the HMMs and their use in biological data, consider the example of binding site recognition on a DNA sequence. There is an observable sequence of nucleotides which in the right order hides underneath a binding site. We can observe the nucleotide sequence, but the presence or absence of a binding site remains hidden to us. HMMs are particularly suited for such problems because they use observed frequencies to calculate emission and transition probabilities to decipher the hidden states. An HMM involves two types of probabilities, transition and emission probabilities. The probability of moving from one state to another is called the transition probability. The probability to observe a variable within a state is called emission or output probability.

Figure 2 shows a schematic HMM with basic architecture and elements. HMMs have been used not only to create sequence profiles but also to create probabilistic model representation of protein clusters. Pfam is an example database that clusters proteins based on their functional elements and represents them with HMM. The downside to HMMs is that they assume a future event depends only on the event that happened immediately before and not in the distant past. This creates a limitation to use standard HMMs in complex cases where sequence elements influence each other that may be close in the three-dimensional space but sequentially lie far from each other. Outside of the biological world, one such example is autocomplete or word suggestions. The words appearing in suggestion are directly dependent on the word that appeared immediately before the present suggestion.

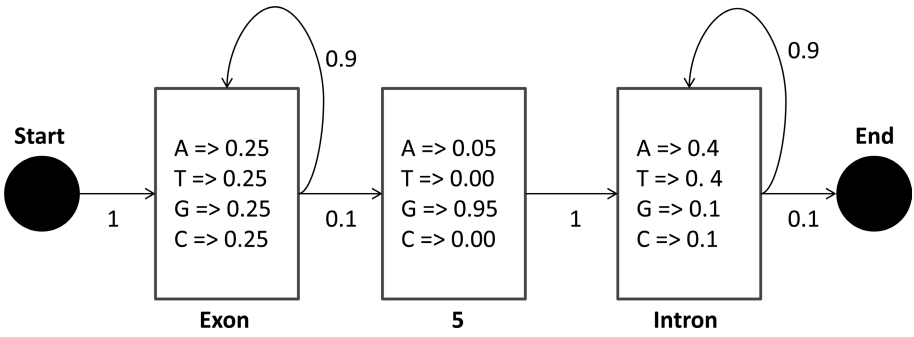


Figure 2 Hidden Markov model. The HMM is designed to predict the G rich splice site. The value inside the boxes show emission probabilities, that is, the probability for each nucleotide to appear while the values outside show transition probabilities to move from one state to the next. HMM representation adapted from (9).

Neural networks

Artificial neural networks is another classification technique with numerous applications in computational biology. Neuron is the basic unit of an artificial neural network. Each neuron can have multiple input connections with weights assigned to each of them. The output value from the neurons is calculated according to its activation function. A neural network may consist of multiple layers, with each layer containing multiple neurons. Figure 3 shows a multi-layered neural network with 32 neurons and 192 edges. Neural networks are used in supervised learning and classification. This approach uses labeled data and follows the main steps listed below:

(i) *Dataset*: Divide the data into training sets and testing set (mostly 70–30% split or 60–40% split, respectively).
(ii) *Training*: Use the training data to traverse over the neuron and estimate the output.
(iii) *Iterate*: Based on the difference between the actual and estimated output, calculate the error and adjust the weights accordingly. Repeat step 2.
(iv) *Testing*: After multiple iterations between step 2 and 3, the model is trained and can be tested. Use the test set (unseen data for model) to compute the output. As the actual label is known, the accuracy and sensitivity can be calculated based on the correct (true positives or true negatives) and incorrect classifications (false positives or false negative).
(v) *Validation*: The training- and test-set splits are randomized and new sets are created from the existing dataset. This new test-train split is then used again iterating over steps 2–4. The idea is to create a model independent for generalized datasets. Depending on situations, there can be multiple iterations for this step and hence referred to k-fold cross validation.

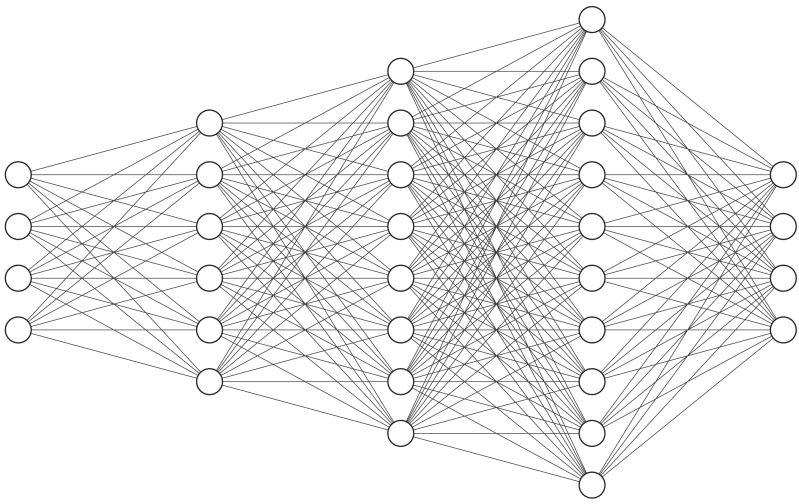


Figure 3 Neural network representation. Each node represents a neuron, and the edges depict weights that connect the neurons between layers. After each iteration, the weights are adjusted to correct for error.

In order to assess the performance of the model, outputs are calculated from different models based on different activation functions or even different neural network architectures. Sensitivity (recall) and accuracy are calculated for each of the models, and the best performing model should have a high recall rate.

The performance of machine learning in general and neural networks in particular depends highly on the quality of the data. A high-quality data would have low noise/junk while having a high homogeneity. Noise in biological data can refer to ambiguous sequence elements or incorrect labels. A high homogeneity results in an equal distribution of diversity in data across different data splits. Assuring the good quality of data before model training is a very important and time-consuming step for data scientists. If the training dataset is not a homogenous representative of the population, it can lead to a biased classification in the models. A bias model can show promising results for the testing dataset but fails in the actual world. This happens because the model is trained to classify only those types of cases that it observed during the training, and a bias sample resulted in a skewed perception of the real-world scenario. The quality of classification from neural networks also depends highly on the training iterations and size of datasets. While the ability for high-powered computation has greatly increased in the last decade, coupled with biological big data, neural networks can be used to train accurate classifiers. Neural networks have now evolved into their more complex form called "Dense Networks" or "Deep learning." These networks (e.g., LSTM) comprise numerous neurons and high number of hidden layers between the input and output layers (hence deep network). Although the depth of a network results in a better-quality model, they are difficult to train due to the requirement of high computing power.

NEXT-GENERATION SEQUENCING

The last three decades have seen a continuous evolution of sequencing technologies. Starting from traditional Sanger sequencing to whole genome shot gun sequencing by Craig Venter and later next-generation sequencing (NGS) (4). The latest amongst these is the "Nanopore," highly compact and efficient sequencing that connects to a computer via USB; it is easily transportable and fits on a small desktop. The technology that initially required thousands of dollars per nucleotide is much cheaper now. An NGS pipeline comprises of two main sections: a wet lab section involves sample preparation, amplification, and sequencing; and the second section involves a bioinformatics workflow that uses the data generated by the wet lab to derive a sequence and other information. It is important to note that the bioinformatics section involves sequence analysis algorithms that are based on statistical and heuristic techniques to analyze and generate sequences. This section focuses on the bioinformatics aspect of NGS since it has evolved an ecosystem of computational algorithms and pipelines around it for accurate and efficient sequencing. NGS is a massively parallel sequencing technology, also referred as high-throughput sequencing, that allows analysis of large fragments of DNA and RNA genomes with high sensitivity, much more quickly and cheaply than the

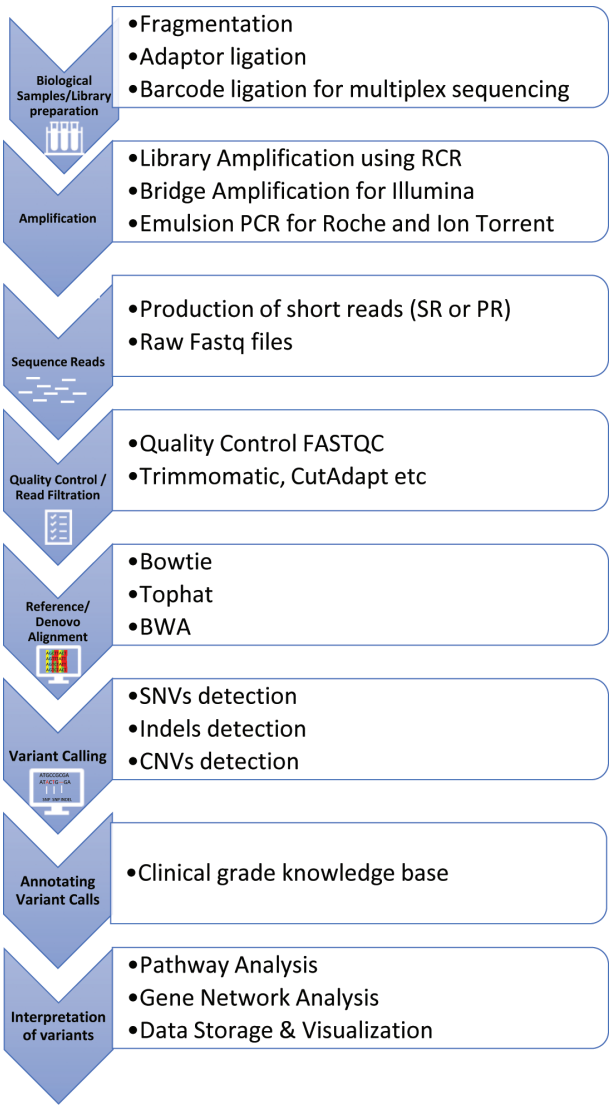


Figure 4 Overview of NGS data analysis workflow. The steps involved in high-throughput sequencing of biological data: (i) biological samples/library preparation, (ii) amplification, (iii) sequence reads, (iv) quality control/read filtration, (v) alignment, (vi) variant calling, (vii) annotating variant calls, and (viii) interpretation of variants.

previously used Sanger sequencing methodology. In NGS, different platform technologies follow the same eight major steps (Figure 4):

(i) *Library preparation*: The first step in NGS workflow involves preparation of high-quality and high-yield sequence library. The isolated genomic DNA or RNA is sheared into smaller fragments ranging from 150–5000 base pairs (bp)

depending on the sequencing platform. The desired library can be created using either of the two fragmentation approaches, mechanical shearing or enzyme-based fragmentation (28, 29). Mechanical shearing methods include acoustic shearing, needle-shear, sonication, and nebulization, whereas enzyme-based methods involve transposons and restriction enzymes (endonucleases) (30). The small fragments known as reads have short overhangs (sticky ends) of 5'-phosphate and 3'-hydroxl groups. These ends are repaired by adenylation at 3' ends resulting in adapter ligation that is important for amplification. During library preparation, unique barcodes can be added to the fragments facilitating multiple sequencing of various samples in the same run (31).

(ii) *Amplification*: The goal of this step is to create thousands of copies for each read. The library is loaded onto the flow-cell, and the fragments are amplified using clonal amplification methods such as emulsion PCR or bridge amplification. In emulsion PCR, the library is amplified within a tiny water droplet floating in an oil solution (32, 33). In bridge amplification, the single-stranded DNA from the library is hybridized to the flow-cell's surface-bound forward and reverse oligos that are complementary to the library adapter sequences. Hybridized at one end, the singe-stranded DNA then folds over to form a bridge and binds to adapter-complementary oligos at the other end. DNA polymerase adds nucleotides to amplify DNA, and a clonal cluster is generated as the original strand is washed away leaving complementary strands of amplified DNA attached to the flow cell. (34).

(iii) *Sequencing*: The amplified individual sequences are sequenced using different platforms and sequencing technologies that include Illumina (Solexa) sequencing, Roche 454 sequencing, and Ion Torrent (Proton/PGM sequencing). Illumina (Solexa) sequencing works by simultaneously identifying DNA bases (A, T, C or G), and each base emits a unique fluorescent signal as it is added to the nucleic acid chain. Illumina sequencing involves 100–150 bp read length. Illumina has some variations that mainly differ in the amount of DNA sequenced in one run (Table 1). Roche 454 sequencing is based on pyrosequencing; a technique that detects pyrophosphate release, again

TABLE 1 Comparison of Illumina sequencing platforms

Sequencing platforms	Run time	Max output (Gb)	Max read number (million)	Max read length (bp)
iSeq Series	9–17.5 hours	1.2	4	2 × 150
MiniSeq Series	4–24 hours	7.5	25	2 × 150
MiSeq Series	4–55 hours	15	25	2 × 300
NextSeq Series	13–20 hours	120	400	2 × 150
HiSeq Series	<1–3.5 days	1500	5000	2 × 150
HiSeq X Series	<3 days	1800	6000	2 × 150

Different attributes and key features of different Illumina platforms include run time, maximum output, maximum read number, and maximum read length.

using fluorescence, after nucleotides are incorporated by polymerase to a new strand of DNA. Roche 454 sequencing produces sequence reads of up to 1000 bp in length. Like Illumina, it does this by sequencing multiple reads at once by reading optical signals as bases are added. Ion Torrent (Proton / PGM sequencing) measures the direct release of H+ (protons) from the incorporation of individual bases by DNA polymerase and therefore differs from the previous two methods as it does not measure light. As in other kinds of NGS, the input DNA or RNA is fragmented, this time ~200 bp. These sequencing technologies result in raw sequencing reads (20 to 1000 bp) stored in the FASTQ format which contains both the nucleotide sequence and its corresponding quality scores. These reads can be either "single-ended" or "paired-ended." Paired-end reads are produced when the fragment size used in the sequencing process is much longer (typically 250–500 bp long).

(iv) *Quality control and read filtration*: After sequencing is complete, the read data are in electronic form and can be processed to generate a whole genome or a specific gene sequence using a bioinformatics NGS pipeline. Although quality control and filtration is the fourth step in generating a full analyzable sequence, it is the first step in a bioinformatics NGS pipeline. Read filtration involves removing low confidence and erroneous reads from the dataset. The amplified raw reads pass through quality control check using FastQC (35) that can produce a detailed report on the number, quality, and coverage of reads. These methods mostly work on sequence analysis techniques like clustering short reads to calculate their frequency and quality scores. It is followed by read filtration, clipping of adapters and low-quality base pairs from 3' and 5' ends using software such as CutAdapt (36), trimmomatic (37) and others.

(v) *Alignment*: Once the read quality is acceptable, millions of raw sequence reads (single-end or paired-end) are mapped and aligned using either a reference based assembly (in which reference sequence is available) or de novo assembly (in the absence of a reference sequence). The sequence reads of variable lengths are aligned using different bioinformatics alignment tools such as BWA (38), Bowtie (39), and TopHat (40). These heuristic-based aligners allow fast sequence alignment and generate a consensus sequence from the alignment by searching the overlapping portions of the reads and merging them into longer reads in order to construct a region of interest, that is, genes or a whole genome. The main aim of this step is to generate a consensus sequence from the millions of reads. A consensus sequence shows the genetic makeup at the time of the sample collection. This step marks the completion of sequence generation for a partial or a whole genome. The following steps are important for an in-depth analysis beyond generation of only a single sequence.

(vi) *Variant identification*: NGS is not only time efficient but also provides the data for an in-depth sequence analysis. Variant analysis uses the reads file to determine the conserved and variable nucleotides at specific positions. As this process involves statistical calculations spanning over millions of reads, it is both a time and computationally intensive process. Bootstrap resampling of reads can be used to assess the quality of variant calling scores. The variations within the genomic sequences such as single-nucleotide polymorphisms (SNPs), single-nucleotide variants (SNV), and indels (insertions

and deletions) are detected using software such as SAMtools (41), Genome Analysis Toolkit (GATK) (42), and VarScan (43, 44). Both SAMtools and GATK use the Bayesian probabilistic approach to identify true variants from alignment errors, whereas VarScan uses a heuristic approach. Most NGS methods for SNV detection are designed to detect germline variations in an individual's genome, whereas the variations that are identified within a population are referred as SNPs.

(vii) *Annotation*: The genetic variants detected are annotated based on the published peer-reviewed literature and public genetic variant databases.

(viii) *Interpretation of variants*: Lastly, medical professionals will interpret these variants and obtain the patient's clinical history in order to establish a most accurate diagnosis. This includes examining different disease pathways and gene network analysis and identifying actual mutations causing a disease.

APPLICATIONS OF NGS IN CLINICAL PRACTICE

The NGS technologies have several applications in research to solve a diverse range of biological problems. Comprehensive analysis of NGS data includes whole-genome sequencing, gene expression determination, transcriptome profiling, and epigenetics. NGS has enabled the researches to sequence large segments of the genome (i.e., whole-genome sequencing) and provides insights into identifying and understanding the genetic variants such as SNPs, insertions, and deletions of DNA, and rearrangements such as translocation and inversions associated with diseases for further targeted studies (45). Researchers use RNA sequencing (RNASeq) to uncover genome-wide transcriptome characterization and profiling (46). Analysis involving genome-wide gene expression (i.e., gene transcription, post-transitional modifications, and translation) and the molecular pathway analysis provide a deeper understanding of gene regulation in neurological, immunological, and other complex diseases. Other applications include studying heritable changes in gene regulation that occur without a change in the DNA sequence. Epigenetics play a significant role in growth, development, and disease progression. The studies on epigenetic changes in cancer provide insight into important tumorigenic pathways (47, 48).

CONCLUSION

Sequence analysis is a broad area of research with sub-domains. Alignment of sequences can reveal important information concerning the structural and functional sites within sequences. It is used to explore the evolutionary path of sequences by identifying the sequence orthologs and homologs. Sequence analysis also involves the use of machine learning techniques for classification and prediction of sequence elements. Statistical methods are used to create sequence profiles and identify other distantly related sequences with a higher precision. Advancement of sequencing technologies has resulted in a next-generation era that opened the doors to personalized medicine and haplotype/quasi-species detection. With correctly organized NGS pipelines, it is possible to analyze the effects of drugs directly at the sequence level.

Acknowledgement: Usman Saeed acknowledges the funding support by Dennemeyer Octimine GmbH, Landaubogen 1-3, 81373 München, Germany.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship, and/or publication of this chapter.

REFERENCES

1. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 2016;113(21):5970–5. http://dx.doi.org/10.1073/pnas.1521291113
2. Watson JD, Crick FH. Molecular structure of nucleic acids; A structure for deoxyribose nucleic acid. Nature. 1953;171(4356):737–8. http://dx.doi.org/10.1038/171737a0
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860–921. http://dx.doi.org/10.1038/35057062
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304–51. http://dx.doi.org/10.1126/science.1058040
5. Pennisi E. Human genome. A low number wins the GeneSweep Pool. Science. 2003;300(5625):1484. http://dx.doi.org/10.1126/science.300.5625.1484b
6. Crick F. Central dogma of molecular biology. Nature. 1970;227(5258):561–3. http://dx.doi.org/10.1038/227561a0
7. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181(4096):223–30. http://dx.doi.org/10.1126/science.181.4096.223
8. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48(3):443–53. http://dx.doi.org/10.1016/0022-2836(70)90057-4
9. Durbin R, Eddy SR, Krogh A, Mitchison GJ. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.
10. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7. http://dx.doi.org/10.1016/0022-2836(81)90087-5
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. http://dx.doi.org/10.1016/S0022-2836(05)80360-2
12. Kent WJ. BLAT—The BLAST-like alignment tool. Genome Res. 2002;12(4):656–64. http://dx.doi.org/10.1101/gr.229202
13. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985;227(4693):1435–41. http://dx.doi.org/10.1126/science.2983426
14. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85(8):2444–8. http://dx.doi.org/10.1073/pnas.85.8.2444
15. Husi H, Skipworth RJ, Fearon KC, Ross JA. LSCluster, a large-scale sequence clustering and aligning software for use in partial identity mapping and splice-variant analysis. J Proteomics. 2013;84:185–9. http://dx.doi.org/10.1016/j.jprot.2013.04.006
16. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–1. http://dx.doi.org/10.1093/bioinformatics/btq461
17. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: A versatile open source tool for metagenomics. PeerJ. 2016;4:e2584. http://dx.doi.org/10.7717/peerj.2584
18. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60. http://dx.doi.org/10.1038/nmeth.3176

19. Suzuki S, Kakuta M, Ishida T, Akiyama Y. GHOSTX: An improved sequence homology search algorithm using a query suffix array and a database suffix array. PLoS One. 2014;9(8):e103833. http://dx.doi.org/10.1371/journal.pone.0103833
20. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402. http://dx.doi.org/10.1093/nar/25.17.3389
21. Remmert M, Biegert A, Hauser A, Soding J. HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9(2):173–5. http://dx.doi.org/10.1038/nmeth.1818
22. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52(5):696–704. http://dx.doi.org/10.1080/10635150390235520
23. Saitou N, Nei M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
24. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80. http://dx.doi.org/10.1162/neco.1997.9.8.1735
25. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. http://dx.doi.org/10.1145/3065386
26. Forgy EW. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. Biometrics. 1965;21(3):768–9.
27. Van Dongen S. Graph clustering by flow simulation. Utrecht: University of Utrecht, 2000.
28. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: Overviews and challenges. Biotechniques. 2014;56(2):61–4, 6, 8, passim. http://dx.doi.org/10.2144/000114133
29. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One. 2011;6(11):e28240. http://dx.doi.org/10.1371/journal.pone.0028240
30. Marine R, Polson SW, Ravel J, Hatfull G, Russell D, Sullivan M, et al. Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA. Appl Environ Microbiol. 2011;77(22):8071–9. http://dx.doi.org/10.1128/AEM.05610-11
31. Dodt M, Roehr JT, Ahmed R, Dieterich C. FLEXBAR—Flexible Barcode and Adapter Processing for Next-Generation Sequencing Platforms. Biology (Basel). 2012;1(3):895–905. http://dx.doi.org/10.3390/biology1030895
32. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc Natl Acad Sci U S A. 2003;100(15):8817–22. http://dx.doi.org/10.1073/pnas.1133470100
33. Nakano M, Komatsu J, Matsuura S, Takashima K, Katsura S, Mizuno A. Single-molecule PCR using water-in-oil emulsion. J Biotechnol. 2003;102(2):117–24. http://dx.doi.org/10.1016/S0168-1656(03)00023-3
34. Pemov A, Modi H, Chandler DP, Bavykin S. DNA analysis with multiplex microarray-enhanced PCR. Nucleic Acids Res. 2005;33(2):e11. http://dx.doi.org/10.1093/nar/gnh184
35. Andrews, S. FastQC: a quality control tool for high throughput sequence data. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc
36. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17:10–2. http://dx.doi.org/10.14806/ej.17.1.200
37. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20. http://dx.doi.org/10.1093/bioinformatics/btu170
38. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Cambridge: Broad Institute of Harvard and MIT; 2013.
39. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25. http://dx.doi.org/10.1186/gb-2009-10-3-r25
40. Trapnell C, Pachter L, Salzberg SL. TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11. http://dx.doi.org/10.1093/bioinformatics/btp120

41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. http://dx.doi.org/10.1093/bioinformatics/btp352
42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. http://dx.doi.org/10.1101/gr.107524.110
43. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: Variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25(17):2283–5. http://dx.doi.org/10.1093/bioinformatics/btp373
44. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22(3):568–76. http://dx.doi.org/10.1101/gr.129684.111
45. Peng L, Bian XW, Li DK, Xu C, Wang GM, Xia QY, et al. Large-scale RNA-Seq Transcriptome Analysis of 4043 Cancers and 548 Normal Tissue Controls across 12 TCGA Cancer Types. Sci Rep. 2015;5:13413. http://dx.doi.org/10.1038/srep13413
46. Koh Y, Park I, Sun CH, Lee S, Yun H, Park CK, et al. Detection of a distinctive genomic signature in Rhabdoid Glioblastoma, A rare disease entity identified by whole exome sequencing and whole transcriptome sequencing. Transl Oncol. 2015;8(4):279–87. http://dx.doi.org/10.1016/j.tranon.2015.05.003
47. Kinsella M, Harismendy O, Nakano M, Frazer KA, Bafna V. Sensitive gene fusion detection using ambiguously mapping RNA-Seq read pairs. Bioinformatics. 2011;27(8):1068–75. http://dx.doi.org/10.1093/bioinformatics/btr085
48. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MG, et al. deFuse: An algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol. 2011;7(5):e1001138. http://dx.doi.org/10.1371/journal.pcbi.1001138

5

Multivariate Statistical Methods for High-Dimensional Multiset Omics Data Analysis

Attila Csala • Aeilko H. Zwinderman

Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam, The Netherlands

Author for correspondence: Attila Csala, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Academic Medical Center, Amsterdam 1105 AZ, The Netherlands. Email: a@csala.me

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch5

Abstract: This chapter covers the state-of-the-art multivariate statistical methods designed for high-dimensional multiset omics data analysis. Recent biotechnological developments have enabled large-scale measurement of various biomolecular data, such as genotypic and phenotypic data, dispersed over various omics domains. An emergent research direction is to analyze these data sources using an integrated approach to better model and understand the underlying biology of complex disease conditions. However, comprehensive analysis techniques that can handle both the size and complexity, and at the same time can account for the hierarchical structure of such data, are lacking. An overview of some of the developments in multivariate techniques for high-dimensional omics data analysis, highlighting two well-known multivariate methods, canonical correlation analysis (CCA) and redundancy analysis (RDA), is provided in this chapter. Penalized versions of CCA are widespread in the omics data analysis field, and there is recent work on multiset penalized RDA that is applicable to multiset omics data. How these methods meet the statistical challenges that come with high-dimensional multiset omics data analysis and help to further our understanding of the human condition in terms of health and disease are presented. Additionally, the current challenges to be resolved in the field of omics data analysis are discussed.

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

Keywords: canonical correlation analysis; high-dimensional data analysis; integrative omics data; multivariate statistics; redundancy analysis

INTRODUCTION

High-throughput sequencing methods such as the Affymetrix GeneChip 1994, Illumina SNP genotyping 2001 and Illumina BeadChip 2005 have provided the possibility of collecting millions of molecular variables (i.e., biomolecular data) from biological samples (1). Simultaneously, developments in knowledge databases including the Kyoto Encyclopedia of Genes and Genomes 1995, Human Genome Project 2003 and 1000 Genomes Project 2015, along with the formation of large biobanks such as the Estonian Genome Project 2000 and the UK Biobank 2006, have provided new means to store and manage biomolecular data. National computing services and leading data science companies have established large-scale computer facilities (e.g., Globus Genomics 2013, Helix Nebula 2013 and European Open Science Could 2019) to enable routine access and analysis of extremely large databases (2, 3). Many biomedical research institutions have established biobanks to store and manage both organic tissue and in silico data of patients on genetic and genomic variations, epigenetic measurements, and gene- and protein-expressions in various tissues, along with disease phenotypes and treatment response (1, 3).

These technological developments in the biomedical field, sometimes collectively referred to as the biotechnological revolution, have created new opportunities to better understand the human condition in terms of health and disease. The development and application of statistical methods that aim to analyze and understand large-scale biomolecular data is referred to as the field of biomolecular big data analysis. The topic of this chapter is omics data analysis, which is a subfield of biomolecular big data analysis. Omics data analysis aims to analyze and understand large-scale biomolecular data from more than one omics data source, where omics is shorthand for a range of -omics domains such as genomics, epigenomics, transcriptomics, proteomics, lipidomics, metabolomics and microbiomics. The field of omics data analysis has two main objectives (4–6):

(i) To understand the underlying biology of disease conditions with emphasis on mechanisms and etiology
(ii) To improve our ability to predict, prevent and treat disease conditions (i.e., translational medicine).

While there has been considerable progress on these objectives for simple monogenic disease conditions (7), such progress has been slow for complex poly- and omnigenic disease conditions (5, 8, 9). The main reason for the relatively low progress in complex conditions is often attributed to the lag between the technologies to collect such vast amounts of biomolecular data and the techniques to analyze and understand such data (10). Current technologies can measure vast amounts of data on simple as well as on complex disease conditions. However, complex conditions presumably have multifaceted underlying biological pathways that the current techniques are unable to model from the available large-scale data sources (9, 11, 12).

Advancements in biotechnology offer the possibility to routinely collect, store and analyze high-dimensional omics data. The high-dimensionality of such biomolecular data refers to the routine practice of collecting biomarkers and disease phenotypes (i.e., biomolecular variables) on a large-scale, often measured in the thousands to millions, while the number of available samples (i.e., patients) is usually measured mostly in the hundreds (i.e., variables >> samples). The collection and analysis of vast numbers of biomolecular variables is hoped to help biomedical scientists to better understand the human condition in terms of health and disease. The main goal of omics data analysis is to model biological pathways in biomolecular data sources in such a way that the biological pathways best model the genetic architecture and the overall underlying biology of disease conditions (8). The resulting biological pathway models then can be used to understand the mechanisms and etiology of disease conditions and ultimately be used to improve our ability to treat such conditions. In light of these possibilities, many scientists believe that personalized medicine at an extremely detailed molecular scale will be possible in the near future (13, 14).

This chapter provides an overview of the development of techniques that are aimed at analyzing and understanding large-scale biomolecular data, with emphasis on multivariate techniques for omics data analysis. Multivariate techniques can: (i) handle the simultaneous analysis of multiple high-dimensional omics data sources, (ii) provide biologically interpretable results, (iii) have well-defined objective functions (no-black box methods) and (iv) preferably have open source software implementations. A perspective on the gap between the technologies that collect, store and manage large-scale biomolecular data and the techniques that analyze and understand such data (i.e., the technology-technique gap) is provided. The four periods in the history of omics data analysis (Table 1) that are well distinguishable in terms of paradigm shifts and the way the biomedical scientific community approaches large-scale biomolecular data are described. Although there are various statistical methods available to analyze omics data, many of them do not meet certain requirements. Thus, the so-called supervised machine learning techniques, which require labeled data for classification (15, 16), are excluded. An excellent review that describes supervised and unsupervised techniques can

TABLE 1 The four periods of development of multivariate techniques and the associated paradigm shifts

Period	Time	Technique	Paradigm Shift
1	Early 2000s	Univariate approach	Associating one or a subset of biomarkers with a single-disease phenotype
2	Late 2000s	Multivariate approach	Associating subset of biomarkers and disease-phenotypes with each other
3	Early 2010s	Multiset multivariate approach	Associating subsets of biomarkers and disease-phenotypes with each other from various data sources
4	Late 2010s	Hierarchical multiset multivariate approach	Associating one or a subset of dependent disease-phenotypes with subsets of independent biomarkers from various data sources

be found in Ref. (17). Also, methods that can be considered multivariate techniques but do not have well-defined objective functions are excluded (12, 18–20). Overall reviews on multivariate techniques for omics data analysis can be found in Refs. (14, 21–25).

EARLY 2000s: THE UNIVARIATE APPROACH

Historically, most techniques focus on analyzing the association between a single disease phenotype and one, or a subset of, biomarker(s) from a particular omics data source. This approach has been widespread since the early 2000s in genome-wide association studies (GWASs) (7). The study published in 2002 by Ozaki et al. on myocardial infarction is widely regarded as the first successful GWAS study (26). Generally, a GWAS aims to analyze the association between a single disease phenotype and one or a subset of biomarkers, which translates to a monothematic model (1). This is often referred to as the univariate approach, since there is only a single dependent variable, namely a disease phenotype, that is associated with one or a subset of independent variables, namely the biomarker(s). Biological pathways modeled by the univariate model are then composed by a single disease phenotype and one or a subset of biomarker(s). This univariate approach, especially in the GWAS framework, has made considerable contributions to biomarker discovery for monogenic and genetically complex conditions (8, 27). However, many biomedical scientists argue that univariate approaches are suboptimal for the pursuit of objectives (i) and (ii) mentioned above, especially when applied to data collected on patients with complex poly- or omnigenic conditions (1, 8, 9, 11).

LATE 2000s: MULTIVARIATE APPROACHES

Complex poly- or omnigenic conditions have complex biological pathways, composed of multiple biomarkers that can be associated with more than one disease phenotype. That is, biological pathways of complex conditions can be best modeled in omics data by associating multiple biomarkers with multiple disease phenotypes. The emergence of this hypothesis resulted in the development of multivariate techniques for omics data analysis, since some multivariate techniques are able to associate multiple disease phenotypes with multiple biological markers.

Penalized canonical correlation analysis

Among the first multivariate statistical methods that were developed for omics data analysis are the modified versions of canonical correlation analysis (CCA). CCA is a well-known multivariate technique that aims to subtract linear combinations of variables (i.e., canonical variates) from two data sources, in a way that the canonical variates maximally correlate with each other (28). The objective function of CCA is:

$$\underset{\boldsymbol{a},\boldsymbol{b}}{arg\,\max}\; cor(\boldsymbol{Xa},\boldsymbol{Yb}), \tag{1}$$

where $\boldsymbol{X}$ denotes the first data source and $\boldsymbol{Xa}$ denotes a linear combination of the variables from $\boldsymbol{X}$, and $\boldsymbol{Y}$ denotes the second data source and $\boldsymbol{Yb}$ denotes a linear combination of the variables from $\boldsymbol{Y}$. $\boldsymbol{Xa}$ and $\boldsymbol{Yb}$ are the canonical variates, and the correlation between the canonical variates is called the canonical correlation. Thus, the objective function of CCA is to maximize the canonical correlation.

CCA applied to omics data results in a set of biomarkers from one omics data source that maximally correlates with a set of biomarkers or disease phenotypes from a second data source. Note that CCA does not distinguish between dependent and independent variables. Also, CCA, in its organic form, is not applicable to omics data, since the high-dimensional nature of omics data (i.e., variables >> samples) causes CCA to fail to subtract canonical variates from the data sources. Modified versions of CCA that solve this issue have started to appear from the late 2000s, among them are penalized canonical correlation analysis (penalizedCCA) (29), regularized canonical correlation analysis (rCCA) (30), sparse canonical correlation analysis (sCCA) (31) and penalized canonical correlation analysis (pCCA) (32). These studies applied a form of penalization to the organic CCA framework, which makes penalized forms of CCA applicable to high-dimensional data and, in most cases, results in a model that includes only a subset of the original variables from the data sources (i.e., variable selection) (33). Variable selection is a desirable property when the original variables are too numerous to be interpretable in the results of the analysis, which is exactly the case with omics data. The exact properties of variable selection depend on the type of penalization applied to CCA, and an overview on penalization methods can be found in Ref. (34). In general, penalized forms of CCA have the same objective function as the generic CCA, that is, it aims to maximize the correlation between linear combinations of two (sub)sets of variables. Applying penalized forms of CCA to omics data results in a model with a (sub)set of biomarkers that maximally correlate with a (sub)set of disease phenotypes or biomarkers penalizedCCA, sCCA and pCCA facilitate variable selection, while sCCA uses a penalization form that makes it applicable to high-dimensional data but does not facilitate variable selection.

Penalized partial least squares regression

Other multivariate statistical methods that were developed in the late 2000s for omics data analysis are modified versions of partial least squares regression (PLS). PLS is a set of general least squares regression techniques applied in an iterative algorithmic framework, and, in fact, CCA is a special case of PLS (35). In general, PLS techniques aim to subtract two sets of linear combinations of variables (i.e., latent variables) from two data sources in a way that the covariance between the latent variables is maximized (36). The objective function of PLS is:

$$arg \max_{\boldsymbol{a},\boldsymbol{b}} cov(\boldsymbol{Xa},\boldsymbol{Yb}), \tag{2}$$

where $\boldsymbol{X}$ denotes the first data source and $\boldsymbol{Xa}$ denotes a linear combination of the variables from $\boldsymbol{X}$, and $\boldsymbol{Y}$ denotes the second data source and $\boldsymbol{Yb}$ denotes a linear combination of the variables from $\boldsymbol{Y}$. $\boldsymbol{Xa}$ and $\boldsymbol{Yb}$ are the latent variables. The objective function of the generic PLS is to maximize the covariance between the latent variables. While this objective function can be modified based on the regression

techniques used in the iterative framework (35), the early applications of PLS to omics data aimed to maximize the covariance between the latent variables.

PLS applied to omics data results in a linear combination of biomarkers between two data sources that have maximum covariance with each other. Similar to CCA, PLS in its organic form is not applicable to omics data, since high-dimensional data (i.e., variables >> samples) cause the general least squares regression techniques in PLS to fail to subtract linear combinations from the data sources. Lê Cao et al. introduced a penalized version of PLS, called the sparse PLS (sPLS), to solve this issue (37). Other PLS-based methods are sparse partial least squares regression (sPLSR) (38), sparse PLS-discriminant analysis (sPLS-DA) (39) and two-way orthogonal PLS (O2PLS) (40). sPLS, sPLSR, sPLS-DA and O2PLS facilitate variable selection, which is a desirable property, as discussed above in the case of penalized CCA.

EARLY 2010S: MULTISET MULTIVARIATE APPROACHES

From the mid-2010s, the need has become apparent for multiset techniques that are able to analyze multiple sets of omics data sources simultaneously (i.e., integrated or multiset techniques). The developments of such methods were motivated by the hypothesis that biological pathways are composed of a collection of biomarkers and disease phenotypes that are not constrained to one or two biological domains. This hypothesis was probably influenced by the relatively new field of systems biology.

Systems biology advocates that properties of biological organisms can be best modeled by assessing its multiple components and the interactions of its various biological domains simultaneously (41). Thus, system biology claims that system properties, such as the function and mechanism of complex conditions, can be better assessed through a system-wide approach (i.e., integrating and analyzing different parts of an organism simultaneously) in contrast to the so-called reductionist approach (i.e., analyzing different parts of an organism separately). Translating this to omics data analysis, one may hypothesize that techniques constrained to one or two omics domains result in a monothematic type of knowledge and possibly miss modeling system-wide properties of complex conditions. In fact, omics domains are not discrete and separable biological entities, as the reductionist approach advocates, but they can rather be better conceptualized as different biomolecular data sources measuring the manifestation of particular biological pathways across different biological sections in the organism. In other words, various omics data sources can be seen as measurements of biomarkers and disease phenotypes of particular conditions present in the patient, dispersed over various biomolecular sections. Therefore, for complex poly- and omnigenic conditions, integrated analysis of multiple omics data sources should be favored (1).

Generalized penalized canonical correlation analysis

The simultaneous analysis of multiple omics domains created the anticipation that multiset techniques will enable better biological pathway models through the

discoveries of biomarkers and disease phenotypes that are dispersed over multiple biomolecular domains (42). One group of such multiset techniques is based on generalized penalized CCA (43), which is the generalization of penalized CCA to multiple data sources. The objective function of generalized penalized CCA is similar to that of CCA in Equation 1, but instead of maximizing the canonical correlation of two canonical variates, it maximizes the canonical correlation of multiple canonical variates

$$arg \max_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_J} \sum_{j,k=1,j\neq k}^{J} c_{ij}\, cor\left(\boldsymbol{X}_j\boldsymbol{a}_j, \boldsymbol{X}_k\boldsymbol{a}_k\right), \quad (3)$$

where $\boldsymbol{X}_j$ denotes the jth data source and $\boldsymbol{X}_j\boldsymbol{a}_j$ denotes a linear combination of the variables from $\boldsymbol{X}_j$. $\boldsymbol{X}_j\boldsymbol{a}_j$ is the jth canonical variate and c_{jk} indicates whether two data sources are connected; $c_{jk} = 1$ if $\boldsymbol{X}_j$ and $\boldsymbol{X}_k$ are connected and 0 otherwise (43).

Generalized penalized CCA applied to omics data results in multiple sets of biological variables that maximally correlate with each other, thereby enabling the simultaneous analysis of multiple biomarkers and disease phenotypes that are dispersed over multiple omics domains. Variations of generalized penalized CCA for omics data analysis started to appear in the mid-2010s, among them are generalized CCA (gCCA) (44), sparse generalized canonical correlation analysis (sGCCA) (45) and data integration analysis for biomarker discovery using latent components (DIABLO) (46). sGCCA and DIABLO facilitate variable selection, while gCCA does not.

Penalized multi-block partial least squares regression

Another group of multiset techniques belong to the extended versions of penalized PLS. These techniques, called multi-block penalized PLS, have a similar objective function to that of penalized PLS in Equation 2 (as generalized penalized CCA relates to penalized CCA). We omit the equation, as it is almost identical to Equation 3, but instead of the correlation, the covariances between the multiple latent variables are maximized. Multi-block penalized PLS applied to omics data results in multiple sets of biomarkers or disease phenotypes that have maximum covariance with each other. Some of the early applications of multi-block penalized PLS to omics data analysis are sparse Multi-Block PLS (sMBPLS) regression (47) and Sparse multi-block PLSR (Sparse MBPLSR) (48). Both sMBPLS and Sparse MBPLSR facilitate variable selection.

A summary of multivariate methods for one-, two-, and multiset omics data analysis can be found in (23). These multiset methods, based on CCA and PLS, are able to detect multiple highly associated biomarkers and disease phenotypes dispersed over multiple biological domains. Note that all the multivariate techniques described so far are aiming to maximize either the correlation or covariance between linear combinations of (sub)sets of biomarkers and disease phenotypes. Therefore, they can at best be used to pursue our understanding of the mechanisms of complex disease. However, in order to understand disease etiology, analyzing the correlation and covariance between linear combinations of subsets of variables is not sufficient (4, 5, 11).

LATE 2010s: HIERARCHICAL MULTISET MULTIVARIATE APPROACHES

Since the mid-2010s, the need for techniques that are not only able to help detect correlated biomarkers and biological pathways of disease phenotypes, but also could aid in detection of causal relationships and understanding disease etiology, has become more apparent (4, 5, 11). This need was motivated by the hypothesis that omics domains have an inherent hierarchical relationship in terms of possible interactions. One of the earliest hypotheses for such a hierarchical relationship model for biomolecular domains, called the Central dogma of molecular biology, was published in the 1970s, sketching plausible interactions between what we call today genomics, transcriptomics and proteomics (49). The Central dogma postulates that genetic information is transferred from genomics to proteomics through transcriptomics. As of today, there are multiple hypotheses on the possible hierarchical structure between the various omics domains, with most implying a genetic information flow from the genome to the phenome (11). In other words, there is a hierarchical structure between genome and phenome in terms of the phenome being dependent on the genome. Thus, in order to better understand disease etiology for complex conditions, multiset multivariate techniques that are able to account for a hierarchical structure between omics domains in terms of dependent and independent data sources should be favored. Redundancy analysis (RDA), the multivariate equivalent of regression analysis, accounts for the genetic information flow in omics domains by distinguishing between dependent and independent omics data sources.

Penalized multi-block redundancy analysis

RDA can be seen as the multivariate extension of univariate regression analysis. RDA aims to subtract linear combinations of independent variables (i.e., latent variables) from a data source in a way that the latent variables explain the most variance in a second dependent data source (50). The objective function of RDA is:

$$arg\max_{\boldsymbol{a}} \sum_{q=1}^{Q} cor\left(\boldsymbol{Xa}, \boldsymbol{y_q}\right)^2, \tag{4}$$

where $\boldsymbol{X}$ denotes the independent data source, $\boldsymbol{Xa}$ denotes a linear combination of the variables from $\boldsymbol{X}$ and $\boldsymbol{y_q}$ denotes the qth variable from the dependent data source (with a total of Q variables). $\boldsymbol{Xa}$ is a latent variable, and the sum of the squared correlations between the latent variable and all the variables of $\boldsymbol{Y}$ is called the redundancy index. Thus, the objective function of RDA is to maximize the redundancy index. Note that RDA maximizes the sum of squared pairwise correlations between a linear combination of variables from an independent data source and between variables of a dependent data source. The aim of RDA is then to find a linear combination of the independent variables that explains the most variance in all the dependent variables. Similarly, we could describe the CCA (or PLS) techniques we presented earlier as techniques aiming to explain maximum variance in their canonical variate (or latent variable) pairs. But the CCA and PLS techniques

do not distinguish between dependent and independent data sources, since in Equation 1, and in Equation 2, the objective function is maximized with respect to the canonical variates, and latent variables, from both data sources, and thus, the variables in both data sources are regarded as independent variables. In Equation 4, the objective function of RDA is maximized with respect to the latent variable of $\boldsymbol{X}$, and the variables from $\boldsymbol{Y}$ are not transformed and are regarded as the dependent variables.

RDA applied to omics data results in a set of independent biomarkers from one data source that explains the most variance in the dependent disease phenotypes from a second data source. RDA accounts for the hierarchical structure between data sources in terms of dependent and independent variables. RDA, in its organic form, is not applicable to omics data, since high-dimensional data cause RDA to fail to subtract latent variables from the independent data source. Similarly, as with CCA and PLS, this can be solved by introducing penalization to RDA. The first penalized RDA, called regularized linear redundancy analysis (regRDA), appeared in the late 2000s (51), and its first application to omics data analysis, called sparse redundancy analysis (sRDA), was in the late 2010s (52). sRDA facilitates variable selection and regRDA does not.

Penalized RDA is able to account for the hierarchical structure between two data sources, and its multiset extension is able to account for the hierarchical structure between multiple data sources. The objective function of multiset penalized RDA is similar to that of RDA in Equation 4, but instead of maximizing the redundancy index between the independent latent variable and all the dependent variables, it maximizes the sum of redundancy indices of multiple latent variable with all the dependent variables (53):

$$arg \max_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_j} \sum_j^J \sum_q^Q cor\left(\boldsymbol{X}_j\boldsymbol{a}_j, \boldsymbol{y}_q\right)^2, \tag{5}$$

where $\boldsymbol{X}_j$ denotes the jth independent data source and $\boldsymbol{y}_q$ denotes the qth variable from the dependent data source (with a total of Q variables). $\boldsymbol{X}_j\,\boldsymbol{a}_j$ denotes the jth linear combination of the variables from $\boldsymbol{X}_j$.

Multiset penalized RDA applied to omics data results in multiple sets linear combinations of independent biomarkers that explain the most variance in the dependent disease phenotypes. Therefore, multiset penalized RDA enables the simultaneous analysis of multiple biomolecular variables that are dispersed over multiple omics domains, while it accounts for the hierarchical structure between the data sources. One application of multiset penalized RDA is multiset sparse redundancy analysis (multi-sRDA) (53), which facilitates variable selection. A summary of the multivariate methods reviewed in this text can be found in Table 2.

CONCLUSION

We examined the state-of-the-art techniques aimed to analyze and understand large-scale biomolecular data. As also reported by others, we likewise identified a technology–technique gap, namely the gap between technologies to collect, store and manage large-scale biomolecular data and the techniques to analyze

TABLE 2 **Multivariate statistical methods for high-dimensional omics data analysis, a chronological overview**

Name	Multiset	Variable selection	Hierarchical	Year	Reference
Penalized CCA (pCCA)	no	yes	no	2007	(28)
Regularized CCA (rCCA)	no	no	no	2008	(29)
Sparse PLS (sPLS)	no	yes	no	2008	(36)
Sparse CCA (sCCA)	no	yes	no	2009	(30)
Penalized CCA (pCCA)	no	yes	no	2009	(31)
Sparse partial least squares regression (sPLSR)	no	yes	no	2009	(37)
Sparse PLS-discriminant analysis (sPLS-DA)	no	yes	no	2011	(38)
Regularized generalized CCA (rGCCA)	yes	no	no	2011	(42)
sparse Multi-Block PLS (sMBPLS) regression	yes	yes	no	2012	(46)
Generalized CCA (gCCA)	yes	no	no	2014	(43)
Sparse generalized canonical correlation analysis (sGCCA)	yes	yes	no	2014	(44)
Sparse multi-block PLSR (Sparse MBPLSR)	yes	yes	no	2015	(47)
Two-Way Orthogonal PLS (O2PLS)	no	yes	no	2016	(39)
Sparse RDA (sRDA)	no	yes	yes	2017	(51)
Multiset sRDA	yes	yes	yes	2018	(52)
Data Integration Analysis for Biomarker discovery using Latent cOmponents (DIABLO)	yes	yes	no	2019	(45)

The first column contains the names, column *Multiset* indicates whether the method is applicable for multiple omics sets, column *Variable selection* indicates whether the method facilitates variable selection and column *Hierarchical* indicates whether the method is able to account for the hierarchical structure between omics data sources. This table is complementary to and based on the tables that can be found in (23).

and understand such data. We described four periods in the history of omics data analysis that are well distinguishable in terms of paradigm shifts in the way the biomedical scientific community approaches large-scale biomolecular data. We highlighted some of the main effects of these major paradigm shifts on the advancement of the omics data analysis field. The main motivation to switch from univariate to multiset multivariate techniques is that analytical techniques constrained to one or two omics domains result in a monothematic type of knowledge and likely miss modeling system-wide properties of complex conditions. Omics domains are not discrete and separable biological entities as reductionist-type approaches. They should be conceptualized as various

biomolecular data sources measuring the manifestations of biological pathways across various biological sections in an organism. Therefore, various omics domains can be seen as sources for biomarkers and disease phenotypes of particular conditions present in patients, dispersed over various biomolecular sections. We described multiset multivariate methods that aim to identify associated biomarkers and disease phenotypes dispersed over various biomolecular sections and therefore provide optimized biological pathway models of complex conditions. Therefore, to pursue objectives (i) and (ii) mentioned in the introduction section for complex poly- and omnigenic conditions, multiset multivariate techniques should be favored over univariate ones. To pursue objective (ii), techniques that aim to identify causal associations should be favored. We describe techniques that aim to identify causal relationships by modeling the hierarchical structure between omics domains in terms of interactions between biomarkers and disease phenotypes from various omics domains. As of today, there are multiple hypotheses on the possible hierarchical structure between the various omics domains, and most of these hierarchical structures aim to model the genetic information flow from the genome to the phenome. We conclude that, in order to pursue objectives (i) and (ii) for complex conditions, a prominent research direction for the omics data analysis field is the development and application of hierarchical multiset multivariate approaches.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

REFERENCES

1 Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: The rise of omics data and their integration in biomedical sciences. Brief Bioinform. 2018 Mar 1;19(2):286–302. http://dx.doi.org/10.1093/bib/bbw114

2. Berger B, Peng J, Singh M. Computational solutions for omics data. Nat Rev Genet. 2013 Apr 18;14(5):333–46. http://dx.doi.org/10.1038/nrg3433

3. Langmead B, Nellore A. Cloud computing for genomic data analysis and collaboration. Nat Rev Genet. 2018 Jan 30;19(4):208–19. http://dx.doi.org/10.1038/nrg.2017.113

4. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol. 2017 Dec 5;18(1):83. http://dx.doi.org/10.1186/s13059-017-1215-1

5. Gallagher MD, Chen-Plotkin AS. The post-GWAS era: From association to function. Am J Hum Genet. 2018 May;102(5):717–30. http://dx.doi.org/10.1016/j.ajhg.2018.04.002

6. Pingault JB, O'Reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nat Rev Genet. 2018;19(9):566–80. http://dx.doi.org/10.1038/s41576-018-0020-3

7. Visscher PM, Goddard ME, Derks EM, Wray NR. Evidence-based psychiatric genetics, AKA the false dichotomy between common and rare variant hypotheses. Mol Psychiatry. 2012;17(5):474–85. http://dx.doi.org/10.1038/mp.2011.65

8. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. Genetic architecture: The shape of the genetic contribution to human traits and disease. Nat Rev Genet. 2017 Dec 11;19(2):110–24. http://dx.doi.org/10.1038/nrg.2017.101
9. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common disease is more complex than implied by the Core Gene Omnigenic Model. Cell. 2018 Jun;173(7):1573–80. http://dx.doi.org/10.1016/j.cell.2018.05.051
10. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: Challenges and opportunities. BMC Med Genomics. 2015;1–12. http://dx.doi.org/10.1186/s12920-015-0108-y
11. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97. http://dx.doi.org/10.1038/nrg3868
12. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: Integrative methods for multi-omics data. Brief Bioinform. 2017 Jun 30;19(June 2017):1370–81. http://dx.doi.org/10.1093/bib/bbx066
13. Karczewski KJ, Snyder MP. Integrative omics for health and disease. Nat Rev Genet. 2018 Feb 26;19(5):299–310. http://dx.doi.org/10.1038/nrg.2018.4
14. Kim M, Tagkopoulos I. Data integration and predictive modeling methods for multi-omics datasets. Mol Omics. 2018 Feb 12;14(1):8–25. http://dx.doi.org/10.1039/C7MO00051K
15. Li Y, Wu FX, Ngom A. A review on machine learning principles for multi-view biological data integration. Brief Bioinform. 2018;19(2):325–40.
16. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. Cell. 2018;173(7):1581–92. http://dx.doi.org/10.1016/j.cell.2018.05.015
17. Huang S, Chaudhary K, Garmire LX. More is better: Recent progress in multi-omics data integration methods. Front Genet. 2017 Jun 16;8(JUN):1–12. http://dx.doi.org/10.3389/fgene.2017.00084
18. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2016 Jul 29;18(5):bbw068. http://dx.doi.org/10.1093/bib/bbw068
19. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018 Apr;15(141):142760. http://dx.doi.org/10.1098/rsif.2017.0387
20. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019 Jan 26;51(1):12–8. http://dx.doi.org/10.1038/s41588-018-0295-5
21. Zierer J, Menni C, Kastenmüller G, Spector TD. Integration of "omics" data in aging research: From biomarkers to systems biology. Aging Cell. 2015 Dec;14(6):933–44. http://dx.doi.org/10.1111/acel.12386
22. Bersanelli M, Mosca E, Remondini D, Giampieri E, Sala C, Castellani G, et al. Methods for the integration of multi-omics data: Mathematical aspects. BMC Bioinformatics. 2016 Dec 20;17(S2):S15. http://dx.doi.org/10.1186/s12859-015-0857-9
23. Meng C, Zeleznik OA, Thallinger GG, Kuster B, Gholami AM, Culhane AC. Dimension reduction techniques for the integrative analysis of multi-omics data. Brief Bioinform. 2016;17(October 2015):628–641, bbv108. http://dx.doi.org/10.1093/bib/bbv108
24. Dihazi H, Asif AR, Beißbarth T, Bohrer R, Feussner K, Feussner I, et al. Integrative omics—From data to biology. Expert Rev Proteomics. 2018 Jun 3;15(6):463–6. http://dx.doi.org/10.1080/14789450.2018.1476143
25. Rappoport N, Shamir R. Multi-omic and multi-view clustering algorithms: Review and cancer benchmark. Nucleic Acids Res. 2019 Jan 25;47(2):1044. http://dx.doi.org/10.1093/nar/gky1226
26. Ozaki K, Yozo O, Aritoshi I, Akihiko S, Ryo Y, Tatsuhiko T, et al. Functional SNPs in the Lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002;32(4):650–4. http://dx.doi.org/10.1038/ng1047
27. Mills MC, Rahal C. A scientometric review of genome-wide association studies. Commun Biol. 2019 Dec 7;2(1):9. http://dx.doi.org/10.1038/s42003-018-0261-x
28. Hotelling H. Relations between two sets of variates. Biometrika. 1936 Dec 1;28(3/4):321. http://dx.doi.org/10.2307/2333955
29. Waaijenborg S, Zwinderman AH. Penalized canonical correlation analysis to quantify the association between gene expression and DNA markers. BMC Proc. 2007;1 Suppl 1:S122. http://dx.doi.org/10.1186/1753-6561-1-S1-S122
30. Gonzalez I, Déjean S, Martin P, Baccini A. CCA : An R Package to extend canonical correlation analysis. J Stat Softw. 2008;23(12):1–14. http://dx.doi.org/10.18637/jss.v023.i12

31. Parkhomenko E, Tritchler D, Beyene J. Sparse canonical correlation analysis with application to genomic data integration. Stat Appl Genet Mol Biol. 2009 Jan 6;8(1):1–34. http://dx.doi.org/10.2202/1544-6115.1406
32. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Stat Appl Genet Mol Biol. 2009 Jan 9;8(1):1–27. http://dx.doi.org/10.2202/1544-6115.1470
33. Tibshirani R. Regression selection and shrinkage via the Lasso. J Roy Stat Soc Ser B. 1996;58:267–88. http://dx.doi.org/10.1111/j.2517-6161.1996.tb02080.x
34. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol.). 2005 Apr;67(2):301–20. http://dx.doi.org/10.1111/j.1467-9868.2005.00503.x
35. Esposito Vinzi V, Russolillo G. Partial least squares algorithms and methods. Wiley Interdiscip Rev Comput Stat. 2013 Jan;5(1):1–19. http://dx.doi.org/10.1002/wics.1239
36. Esposito Vinzi V, Chin WW, Henseler J, Wang H, editors. Handbook of partial least squares. Berlin, Heidelberg: Springer Berlin Heidelberg; 2010.
37. Lê Cao K-A, Rossouw D, Robert-Granié C, Besse P. A sparse PLS for variable selection when integrating omics data. Stat Appl Genet Mol Biol. 2008 Jan 18;7(1):35. http://dx.doi.org/10.2202/1544-6115.1390
38. Chun H, Keleş S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics. 2009 May;182(1):79–90. http://dx.doi.org/10.1534/genetics.109.100362
39. Lê Cao K-A, Boitard S, Besse P. Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011 Dec 22;12(1):253. http://dx.doi.org/10.1186/1471-2105-12-253
40. Bouhaddani S El, Houwing-Duistermaat J, Salo P, Perola M, Jongbloed G, Uh H-W. Evaluation of O2PLS in Omics data integration. BMC Bioinformatics. 2016;17 Suppl 2(2):11. http://dx.doi.org/10.1186/s12859-015-0854-z
41. Tavassoly I, Goldfarb J, Iyengar R. Systems biology primer: The basic methods and approaches. Essays Biochem. 2018 Oct 26;62(4):487–500. http://dx.doi.org/10.1042/EBC20180003
42. Haas R, Zelezniak A, Iacovacci J, Kamrad S, Townsend S, Ralser M. Designing and interpreting "multi-omic" experiments that may change our understanding of biology. Curr Opin Syst Biol. 2017 Dec;6(September):37–45. http://dx.doi.org/10.1016/j.coisb.2017.08.009
43. Tenenhaus A, Tenenhaus M. Regularized generalized canonical correlation analysis. Psychometrika. 2011 Apr 17;76(2):257–84. http://dx.doi.org/10.1007/s11336-011-9206-8
44. Shen C, Sun M, Tang M, Priebe CE. Generalized canonical correlation analysis for classification. J Multivar Anal. 2014 Sep;130:310–22. http://dx.doi.org/10.1016/j.jmva.2014.05.011
45. Tenenhaus A, Philippe C, Guillemot V, Le Cao K-A, Grill J, Frouin V. Variable selection for generalized canonical correlation analysis. Biostatistics. 2014 Jul 1;15(3):569–83. http://dx.doi.org/10.1093/biostatistics/kxu001
46. Singh A, Shannon CP, Gautier B, Rohart F, Vacher M, Tebbutt SJ, et al. DIABLO: An integrative approach for identifying key molecular drivers from multi-omics assays. Birol I, editor. Bioinformatics. 2019 Jan 18;35(January):1–8, 3055–3062.
47. Li W, Zhang S, Liu C-C, Zhou XJ. Identifying multi-layer gene regulatory modules from multi-dimensional genomic data. Bioinformatics. 2012 Oct 1;28(19):2458–66. http://dx.doi.org/10.1093/bioinformatics/bts476
48. Karaman İ, Nørskov NP, Yde CC, Hedemann MS, Bach Knudsen KE, Kohler A. Sparse multi-block PLSR for biomarker discovery when integrating data from LC–MS and NMR metabolomics. Metabolomics. 2015 Apr 14;11(2):367–79. http://dx.doi.org/10.1007/s11306-014-0698-y
49. Crick F. Central dogma of molecular biology. Nature. 1970 Aug 8;227(5258):561–3. http://dx.doi.org/10.1038/227561a0
50. van den Wollenberg AL. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika. 1977 Jun;42(2):207–19. http://dx.doi.org/10.1007/BF02294050
51. Takane Y, Hwang H. Regularized linear and kernel redundancy analysis. Comput Stat Data Anal. 2007 Sep;52(1):394–405. http://dx.doi.org/10.1016/j.csda.2007.02.014
52. Csala A, Voorbraak FPJM, Zwinderman AH, Hof MH. Sparse redundancy analysis of high-dimensional genetic and genomic data. Bioinformatics. 2017 Oct 15;33(20):3228–34. http://dx.doi.org/10.1093/bioinformatics/btx374
53. Csala A, Hof MH, Zwinderman AH. Multiset sparse redundancy analysis for high-dimensional omics data. Biom J. 2018 Nov;61:1–18, 406–423. http://dx.doi.org/10.1002/bimj.201700248

6

Statistical Methods for RNA Sequencing Data Analysis

Dongmei Li

Clinical and Translational Science Institute, University of Rochester School of Medicine and Dentistry, Rochester, NY, USA

Author for correspondence: Dongmei Li, Clinical and Translational Science Institute, University of Rochester School of Medicine and Dentistry, 265 Crittenden Boulevard CU 420708, Rochester, NY, USA. Email: Dongmei_Li@urmc.rochester.edu

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch6

Abstract: This chapter will review the statistical methods used in RNA sequencing data analysis, including bulk RNA sequencing and single-cell RNA sequencing. RNA sequencing data analysis has been widely used in biomedical and biological research to identify genes associated with certain conditions or diseases. Many statistical methods have been proposed to analyze bulk and single-cell RNA sequencing data. Several studies have compared the performance of different statistical methods for RNA sequencing data analysis through simulation studies and real data evaluations. This chapter will summarize the statistical methods and the evaluation results for comparing different statistical analysis methods used for RNA sequencing data analysis. It will cover the statistical models, model assumptions, and challenges encountered in the RNA sequencing data analysis. It is hoped that this chapter will help researchers learn more about the statistical perspective of the RNA sequencing data analysis and enable them to choose appropriate statistical analysis methods for their own RNA sequencing data analysis.

Keywords: bulk RNA sequencing; data analysis; differential analysis; RNA sequencing; single-cell RNA sequencing

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

RNA sequencing, including bulk RNA sequencing and single-cell RNA sequencing, is a popular technology used in biological and biomedical fields (1, 2). Figure 1 shows the analysis flow of RNA sequencing data. In RNA sequencing experiments, RNAs of interest need to be extracted first from the cells and then converted to complementary DNA (cDNA) to be sequenced by high-throughput platforms. Next, the sequenced short cDNA fragments are mapped to a genome or a transcriptome, and the summarized count data are derived to estimate the expression levels for each gene or isoform (3–5). Finally, statistical methods or

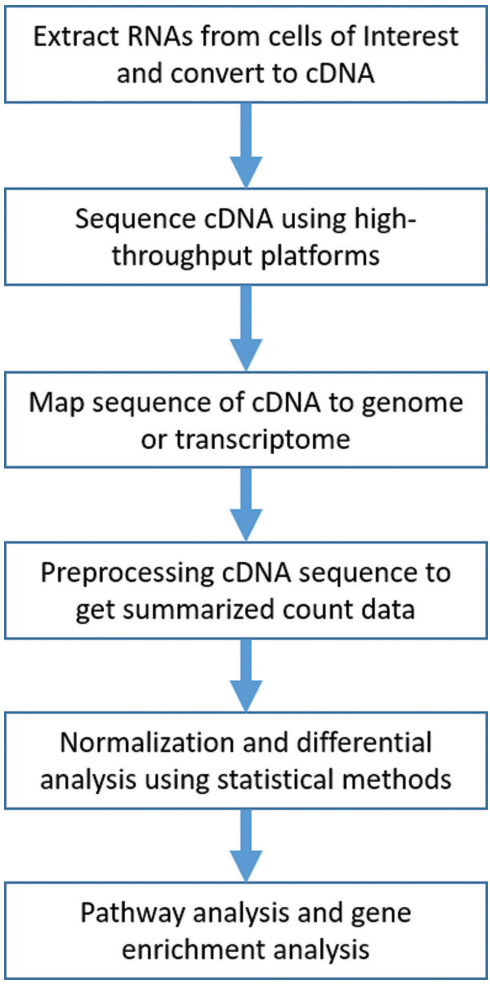


Figure 1 **Analysis flow of RNA sequencing data.**

machine learning methods are applied to the summarized count data after normalization to evaluate transcription levels under different biological and biomedical conditions, to discover novel transcripts and isoforms, and to detect alternative splicing and splice junctions (6). The single-cell RNA sequencing, in addition, allows to understand gene expression pattern within the cell; to identify cell heterogeneity, cell population, and sub-population; and to examine the effects of low copy mRNA distribution and transcriptional regulation (7). Pathway analysis and gene enrichment analysis are usually performed further on selected significant genes after differential analysis (8, 9).

RNA sequencing has been widely used to study the mechanism of complex disease, identify potential biomarkers for clinical indications and infer gene pathways (10–12). Similar to bulk RNA sequencing, single-cell RNA sequencing has been applied to identify cell populations, infer gene regulatory networks, and track different cell lineages (13–15). Single-cell RNA sequencing also has the potential to identify drug-resistant clones, assist non-invasive biopsy diagnosis, and infer stem cell regulatory networks (16–18).

As the sequencing technology advances rapidly, the cost of both bulk RNA sequencing and single-cell RNA sequencing also dramatically decreased (18, 19). With this massive amount of RNA sequencing data now available, it is very challenging to obtain accurate information from the data and further transform this information into useful knowledge (20, 21). Differential gene expression analysis also has its own challenges. The distribution of read coverage might be different along the genome attributed to the variation of genome compositions. Meanwhile, larger genes have more mapped reads than smaller genes although their expression levels might be the same. Furthermore, many biological variations sometimes cannot be accounted for in the data analysis due to relatively small sample sizes for each experimental condition. This chapter focuses on the statistical analysis methods used for differential analysis in both bulk RNA sequencing and single-cell RNA sequencing data. Commonly used statistical methods, their model assumptions, and tests for RNA sequencing differential analysis are discussed (Table 1). The simulation results of comparing different statistical methods and challenges encountered in the data analysis are summarized. Recommendations on the selection of appropriate statistical methods for RNA sequencing differential analysis are also provided.

STATISTICAL METHODS FOR BULK RNA SEQUENCING DIFFERENTIAL ANALYSIS

Current popular methods for bulk RNA-seq differential analysis methods could be classified into four categories based on the type of statistical methods used for differential analysis: (i) *t*-test analogical methods (Cuffdiff and Cuffdiff2) (22, 23), (ii) Poisson or negative binomial model-based methods (edgeR, DESeq, DESeq2, baySeq, EBSeq) (24–29), (iii) non-parametric methods (SAMseq and NOIseq) (30–32), and (iv) linear models (voom and sleuth) (33, 34).

TABLE 1 **Summary of gene differential expression analysis methods for bulk RNA and single-cell RNA sequencing data**

Bulk RNA sequencing data			
Method	**Read count distribution assumption/model**	**Differential analysis test**	**Reference**
Cuffdiff and Cuffdiff2	Similar to *t*-distribution on log-transformed data	*t*-test analogical method	(22, 23)
edgeR	Negative binomial distribution	Exact test analogous to Fisher's exact test or likelihood ratio test	(24, 25)
DESeq	Negative binomial distribution	Exact test analogous to Fisher's exact test	(26)
DESeq2	Negative binomial distribution	Wald test	(27)
baySeq	Negative binomial distribution	Posterior probability through Bayesian approach	(28)
EBSeq	Negative binomial-beta empirical Bayes model	Posterior probability through Bayesian approach	(29)
SAMseq	Non-parametric method	Wilcoxon rank statistics based permutation test	(30)
NOIseq	Non-parametric method	Corresponding logarithm of fold change and absolute expression differences have a high probability than noise values	(31, 32)
voom	Similar to *t*-distribution with empirical Bayes approach	Moderated *t*-test	(33)
Sleuth	Additive response error model	Likelihood ratio test	(34)
Single-cell RNA sequencing data			
Method	**Read count distribution assumption/model**	**Differential analysis test**	**Reference**
SCDE	Two-component mixture model with Poisson and negative binomial distributions	Posterior probability of being differentially expressed through Bayesian approach	(40)
MAST	Hurdle model with indicator variable and logistic regression	Differences in summarized regression coefficients between groups through bootstrap method	(41)
scDD	Bayesian modeling approach	Bayes factor score through permutation method	(42)
DEsingle	zero-inflated negative binomial model	Likelihood ratio test	(43)
SigEMD	Logistic regression and Wald test for selecting genes with zero count and then impute zero counts using the Lasso regression model	Non-parametric test based on Earth Mover's Distance (EMD) through permutation method	(44)

Table continued on following page

TABLE 1 Summary of gene differential expression analysis methods for bulk RNA and single-cell RNA sequencing data (Continued)

Single-cell RNA sequencing data			
Method	**Read count distribution assumption/model**	**Differential analysis test**	**Reference**
SINCERA	Exact or normal distribution	Welch's *t*-test or Wilcoxon rank sum test	(46)
D^3E	Discrete distribution	Cramér-von Mises test, Kolmogorov–Smirnov test or likelihood ratio test	(47)
EMDomics	Distribution functions are different	EMD-based permutation test	(48)
Monocle2	Generalized linear model approach	Likelihood ratio test	(45, 51)
Linnorm	*t*-distribution with empirical Bayes approach	Moderated *t*-test	(49)
Discriminative Learning	Multiple logistic regression model	Likelihood ratio test	(50)

Cuffdiff and Cuffdiff2

Both the Cuffdiff and Cuffdiff2 methods use the *t*-test analogical method to test the changes in gene expression levels between different groups (22, 23). The mean gene expression level for each gene is determined using the maximum likelihood estimating method for different groups. Then, the mean difference of the logarithm-transformed gene expression levels of the estimated gene expression levels is used as the numerator in the *t*-test analogical method, and the estimated variance of the mean differences in logarithm is estimated using the delta method. The power of the *t*-test analogical method in Cuffdiff and Cuffdiff2 depends on the length of the transcripts tested as longer transcripts yield more reads. Thus, the results from Cuffdiff and Cuffdiff2 are biased toward a higher probability of identifying longer transcripts or genes. The major differences between Cuffdiff and Cuffdiff2 are methods used to extrapolate the estimated gene expression levels. Cuffdiff determines the estimated gene expression levels using the maximum likelihood method with the Bayesian approach and Poisson distribution assumption, while the Cuffdiff2 method improves the estimation of gene expression levels through modeling cross-replicate variability in transcript-level counts and adopts the negative binomial distribution assumption for those estimated counts.

edgeR

For each gene in each sample, edgeR assumes that the summarized count follows a negative binomial distribution with mean equal to the multiplication of library size and relative abundance (the gene expression levels), and the variance for each

gene is a function of the mean (24, 25). The genewise dispersion is estimated using a conditional maximum likelihood method through the empirical Bayes approach. For gene differential expression testing, edgeR uses either an exact test analogous to Fisher's exact test with consideration of overdispersion or a likelihood ratio test within a negative binomial generalized log-linear model framework.

DESeq and DESeq2

DESeq uses a modified negative binomial model implemented in edgeR (26). DESeq estimates the variance based on the relative abundance of the gene through a data-driven approach. DESeq tests gene expression differences between groups using an exact test analogous to Fisher's exact test with test statistics as the sum of total count within each group and across groups. DESeq2 takes a generalized linear model approach to model the group differences in relative abundance, which can also accommodate more complex study designs (27). DESeq2 assumes that the dispersion follows a log normal prior distribution with means being a function of normalized counts for each gene. DESeq2 uses an empirical Bayes approach to integrate the dispersion and fold change estimates and tests the gene differential expression using the Wald test.

baySeq

baySeq assumes that the summarized count data follow a negative binomial distribution and use the whole dataset to obtain a prior distribution for the estimated model parameters (28). The data dispersion is approximated using the maximum likelihood method. The baySeq method uses a posterior probability of non-differential expression between groups and a Bayesian FDR estimate to select significantly differentially expressed genes between groups.

EBSeq

EBSeq assumes that within each biological condition, the expected count from each gene follows a negative binomial distribution (29). Within each group, the mean of gene expressions is a function of the variance of gene expressions. The variance of gene expressions follows a beta distribution with the two parameters estimated using the expectation-maximization (EM) algorithm. For the gene expression differential tests between groups, EBSeq obtains a posterior probability of genes being differentially expressed between groups through Bayes' rule using the EM algorithm within the negative binomial-beta empirical Bayes model framework. EBSeq also uses a Bayesian FDR estimate to assist the selection of significantly differentially expressed genes.

SAMseq

SAMseq is a non-parametric method proposed for differential gene expression testing between groups (30). For between-group comparisons, SAMseq uses the two-sample Wilcoxon rank statistics. SAMseq uses a re-sampling procedure to

account for different sequencing depths in the differential data analysis. The null distribution of the Wilcoxon rank statistic and FDR are estimated using the permutation method.

NOIseq

NOIseq is also a non-parametric method for testing differential gene expression between groups through ratio of fold change and absolute expression differences (31, 32). NOIseq uses sequencing-depth corrected and normalized RNA sequencing count data and models the noise distribution by contrasting the logarithm of fold change and absolute expression differences between groups. NOIseq considers a gene to be differentially expressed between groups if the corresponding logarithm of fold change and absolute expression difference values have a high probability to be higher than noise values.

voom

voom takes a linear modeling strategy to model the count data (33). It determines the mean–variance relationship based on the delta rule and Taylor's theorem and obtains the estimate for variance through the piecewise linear function defined by the fitted LOWESS curve. voom also generates a weight for each observation and uses the estimated variance and weight as the input in the limma empirical Bayes analysis pipeline. The gene expression differential analysis between groups is tested using the moderated *t*-statistics.

Sleuth

Sleuth uses an additive response error model with the total between-sample variability being an additive of biological variance and inferential variance (34). The biological variance is composed of between-sample variation and variation during the library preparation process. The inferential variance includes variation due to random sequencing of fragments and variation coming from computational inference procedures. Sleuth tests gene differential expression between groups using the likelihood ratio test.

STATISTICAL METHODS COMPARISONS FOR BULK RNA SEQUENCING DIFFERENTIAL ANALYSIS

In 2013, Soneson conducted an extensive comparison of 11 methods used for bulk RNA sequencing differential analysis through both simulation studies and real RNA sequencing data examples (35). The methods Soneson compared include edgeR, DESeq, baySeq, EBSeq, SAMseq, and voom, described before. The comparison of those methods showed that all methods had low power with small sample sizes, and there was no optimal method applicable for all conditions. voom performed well under many conditions and was robust to outliers and computationally efficient. However, voom performed worse when the variances were

unequal between groups. SAMseq requires larger sample sizes (at least 4–5 samples per group) to detect significantly differentially expressed genes. The comparison also found that DESeq was often overly conservative, and edgeR was too liberal with a larger number of false positives. Both baySeq and EBSeq were computationally less efficient. baySeq showed highly variable results when significant genes were all modulated in one direction, and the results were largely affected by outliers. EBseq had a poor false discovery rate (FDR) control in most situations and was relatively robust to outliers.

Previous experimental validation of selected differentially expressed genes from multiple RNA sequencing differential expression analysis methods (Cuffdiff2, edgeR, DESeq2) found a high FDR of the Cuffdiff2 method and high false negative rates of the DESeq2 method (36). The edgeR method had relatively higher sensitivity and specificity than the Cuffdiff2 and DESeq2 methods. In addition, the experimental validation also showed that pooled samples in the experiments suffered from lower positive predictive values than individual samples.

Using results from qRT-PCR as the gold standard, an extended review of eight RNA sequencing differential analysis methods (baySeq, DESeq, DESeq2, EBSeq, edgeR, voom, NOIseq, and SAMseq) was conducted to determine their precision, accuracy, and sensitivity (37). By comparing the results from qRT-PCR and selected differentially expressed genes from each of the eight methods, it was found that voom, NOIseq, and DESeq2 showed more consistent results than the other methods. In addition, the significantly differentially expressed genes selected by consensus of baySeq, DESeq2, voom, and NOISeq had the best performance indicators on precision, accuracy, and sensitivity. Furthermore, the investigation also found that mapping methods in the pre-processing step of RNA sequencing data analysis had minimal effect on downstream RNA sequencing gene differential analysis, given that a reference genome for the RNA sequencing data was available.

A recent investigation of six RNA sequencing differential analysis methods (DESeq, DESeq2, edgeR, SAMseq, EBSeq, and voom) focused on their stability measured by the area under the correlation curve (38). Among the explored factors that have a potential to affect the stability of RNA sequencing differential analysis methods, fold changes of truly differentially expressed genes and their variability seem largely to affect the stability of those methods. Larger sample size is associated with increased stability, and a sample size of 10 or larger in each group results in a plateau on stability. DESeq2 and edgeR were less likely to be affected by outliers on their stability measurements.

STATISTICAL METHODS FOR SINGLE-CELL RNA SEQUENCING DIFFERENTIAL ANALYSIS

Single-cell RNA sequencing is becoming popular in recent years to better understand the stochastic process and gene regulations in a granular resolution (13, 15, 16, 39). The commonly used gene differential expression analysis in single-cell RNA sequencing can be classified into two categories, with one category modeling excess zeros (SCDE, MAST, scDD, DEsingle, and SigEMD) (40–44) and the other category without modeling the excess zeros in the single-cell RNA sequencing data (DESeq2, SINCERA, D^3E, EMDomics, Monocle2, Linnorm, and Discriminative Learning)

(12, 27, 45–50). DESeq2 is a popular method used for bulk RNA sequencing data analysis, which is also often used for analyzing single-cell RNA sequencing data for testing of differential expression between groups.

Single-cell differential expression (SCDE)

SCDE uses a two-component mixture model for the gene expression data from single-cell RNA sequencing experiments (40). The excess zero part (dropouts) is modeled by a Poisson distribution with user-specified thresholds for the mean (such as 0.1). The expressed genes are modeled by a negative binomial distribution technique. For gene differential expression analysis between groups, SCDE takes a Bayesian approach to obtain the posterior probability of a gene being expressed in one group and then uses a fold expression difference between groups as the test statistics with empirical P-values calculated to select differentially expressed genes.

MAST

MAST uses a hurdle model approach for single-cell RNA sequencing gene differential expression analysis (41). MAST assumes conditional independence between expression rate and expression levels for each gene. MAST uses an indicator variable to denote whether a gene is expressed in a cell and fits a logistic regression for the discrete indicator variable. For genes expressed in a cell, MAST fits a normally distributed linear model. The gene differential expression analysis between groups is tested using the differences in summarized regression coefficients between groups. The null distribution of the test statistics is estimated through a bootstrap method with empirical Bayes approach regularizing model parameters.

scDD

scDD is also based on a Bayesian modeling approach to detect differentially expressed genes between groups (42). ScDD models the excess zeros using a logistic regression and models the non-zero gene expressions using a conjugate Dirichlet process mixture model of normal distributions. For testing gene differential expressions, scDD calculates an approximate Bayes factor score that compares the probability of differential expression with the probability of non-differential expression. The empirical P-values for the differential expression tests are computed using a permutation method.

DEsingle

DEsingle uses a zero-inflated negative binomial (ZINB) model to characterize the read counts and excess zeros in single-cell RNA sequencing data (43). The ZINB model has two components, one modeling the excess zeros through an indicator variable multiplied by the proportion of constant zeros and the other modeling the positive gene expressions through a negative binomial model multiplied by the proportion of non-zeros. The gene differential expression analysis is conducted through likelihood ratio tests within the ZINB model framework.

SigEMD

Different from other excess zero modeling methods for single-cell RNA sequencing differential analysis, SigEMD takes an additional step in modeling the excess zeros (44). SigEMD first uses logistic regression and the Wald test to select genes with zero counts that are affecting gene expression distributions, then SigEMD imputes those zero counts through a Lasso regression model. The gene differential analysis between groups is conducted using a non-parametric test based on Earth Mover's Distance (EMD). The P-values are computed using a permutation method.

SINCERA

SINCERA is a pipeline developed for single-cell RNA sequencing data analysis (46). SINCERA can be used for the pre-processing of single-cell RNA sequencing data, identifying cell types and key gene expression regulators, selecting differentially expressed genes, and predicting gene signatures. For gene differential analysis between groups, SINCERA uses the Welch's *t*-test when the sample size of both groups is >5; otherwise, SINCERA uses the Wilcoxon rank sum test. SINCERA also includes the SAMseq algorithm as an optional method for selecting differentially expressed genes from single-cell RNA sequencing data.

D^3E

D^3E is a discrete distribution method used for single-cell RNA sequencing gene differential expression analysis (47). To identify genes differentially expressed between groups, D^3E uses either the Cramér-von Mises test, the Kolmogorov–Smirnov test or the likelihood ratio test. To test the hypothesis of the driving mechanism in apparent changes, D^3E fits a transcriptional burst model to the expression data for each gene through a method of moments or a Bayesian approach. Following the transcriptional burst model, parameter changes between groups will be calculated.

EMDomics

EMDomics detects significantly differentially expressed genes between groups for single-cell RNA sequencing data by comparing the two distribution functions of gene expressions between groups (48). EMDomics compares the differences between groups based on EMD, a commonly used approach to compare two histograms in imaging analysis. EMDomics measures the differences between two normalized distributions of the two groups through normalized total cost of transforming distributions between groups. Permutation test is used to compute the P-values for the EMD tests.

Monocle2

Using the census algorithm, Monocle2 converts the relative single-cell RNA sequencing expression levels into relative counts for each gene without experimental spike-in controls (45, 51). The census algorithm in Monocle2 estimates

the total number of mRNAs in each cell by calculating the ratio of the total number of single-mRNA genes to the fraction of the library contributed by them and then rescales the transcript per million (TPM) in single cell values into mRNA counts for each gene. Monocle2 tests gene differential expression between groups through a likelihood ratio test for comparing a full generalized linear model with additional effects to a reduced generalized linear model based on negative binomial distributions.

Linnorm

Linnorm proposes a new normalization and transformation method for single-cell RNA sequencing data analysis (49). The normalization and transformation parameters are calculated based on stably expressed genes across different cells. Linnorm uses the moderated *t*-statistics in the limma package for gene differential expression analysis through the empirical Bayes approach to centralize the estimated variances from the data.

Discriminative learning

Discriminative learning uses the multiple logistic regression framework (50). Different from previous single-cell RNA sequencing differential analysis methods, discriminative learning uses the group labels as the outcome variables and uses the gene expression levels and other characteristics of the samples as the predictor variables to identify genes significantly associated with the group labels through likelihood ratio tests.

STATISTICAL METHODS COMPARISON FOR SINGLE-CELL RNA SEQUENCING DIFFERENTIAL ANALYSIS

A previous comparison of six methods (SCDE, MAST, D^3E, Monocle, edgeR, DESeq) for single-cell RNA sequencing differential analysis examined the performance of those methods under different unimodal or bimodal distributions (52). The comparison found significant differences among those methods regarding precision, recall, empirical power, and overall performance. The investigation did not suggest an optimal method that performs better than other methods under all scenarios. A call for new differential analysis methods for single-cell RNA sequencing data was suggested as a result from the comparisons.

Another evaluation of 36 approaches for gene differential expression analysis in single-cell RNA sequencing data found remarkable differences in the performance of those approaches (53). They also found the gene differential expression analysis methods developed specifically for single-cell RNA sequencing data did not perform generally better than the methods developed for bulk RNA sequencing data.

A recent comprehensive evaluation of single-cell RNA sequencing differential analysis methods compared 11 differential analysis methods, including SCDE, MAST, scDD, DEsingle, SigEMD, SINCERA, D^3E, EMDomics, Monocle2, edgeR,

and DESeq2 (54). The gene expression values from real single-cell RNA sequencing experiments are multimodal with excess zeros, which makes the gene expression differential analysis challenging. Currently, there is no method available that can handle both multimodality and excess zeros. The comparison showed that no single method performs uniformly better than other methods under all circumstances. In general, non-parametric methods that could handle multimodality perform better than methods modeling excess zeros, while methods modeling excess zeros resulted in higher true positive rates and lower false positive rates. Gene differential expression analysis methods developed for single-cell RNA sequencing data had similar performance as those methods developed for bulk RNA sequencing data. In addition, low agreement was found among the selected genes from those differential analysis methods for single-cell RNA sequencing data. This recent evaluation also indicates the need of new differential analysis methods for single-cell RNA sequencing data.

CONCLUSION

As RNA sequencing technology is getting increasingly popular and more advanced in the biomedical and biological fields, coupled with a decrease of the cost for RNA sequencing experiments, more RNA sequencing differential analysis methods will be developed to identify differentially expressed genes between groups. For gene differential analysis methods used for both bulk RNA sequencing and single-cell RNA sequencing data, there is no consensus on an optimal method under all circumstances, although DESeq2 is currently very popular for gene expression differential analysis for bulk RNA sequencing data within the bioinformatics community. Remarkable differences were also found among different gene expression differential analysis methods in terms of numbers and characteristics of selected differentially expressed genes. Gene expression differential analysis methods specific for single-cell RNA sequencing data have a similar performance as methods developed for bulk RNA sequencing data, when both were used for single-cell RNA sequencing data. Evaluations of commonly used gene expression differential analysis methods for RNA sequencing data indicate a need for better differential analysis methods, especially for single-cell RNA sequencing data. Taking consensus of the selected differentially expressed genes from multiple methods could improve accuracy and reduce the false discovery rate, but it could also increase the false negative rate. New methods that integrate multiple approaches with both reduced false positives and reduced false negatives might be the direction for future differential analysis method development.

Acknowledgement: This work was supported by the National Cancer Institute of the National Institutes of Health (NIH) and the Food and Drug Administration (FDA) Center for Tobacco Products under Award Number U54CA228110. Dr. Li's time is supported in part by the University of Rochester CTSA award number UL1 TR002001 from the National Center for Advancing Translational Sciences of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or the Food and Drug Administration (FDA).

Conflict of Interest: The author declares no potential conflict of interest with respect to research, authorship, and publication of this chapter.

Copyright and permission statement: To the best of my knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, et al. RNA sequencing for research and diagnostics in clinical oncology. Semin Cancer Biol. 2019. http://dx.doi.org/10.1016/j.semcancer.2019.07.010
2. Wang T, Johnson TS, Shao W, Lu Z, Helm BR, Zhang J, et al. BERMUDA: A novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol. 2019;20(1):165. http://dx.doi.org/10.1186/s13059-019-1764-6
3. Tian L, Dong X, Freytag S, Le Cao KA, Su S, JalalAbadi A, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods. 2019;16(6):479–87. http://dx.doi.org/10.1038/s41592-019-0425-8
4. Ferrall-Fairbanks MC, Ball M, Padron E, Altrock PM. Leveraging single-cell RNA sequencing experiments to model intratumor heterogeneity. JCO Clin Cancer Inform. 2019;3:1–10. http://dx.doi.org/10.1200/CCI.18.00074
5. Wang L, Felts SJ, Van Keulen VP, Pease LR, Zhang Y. Exploring the effect of library preparation on RNA sequencing experiments. Genomics. 2018. http://dx.doi.org/10.1016/j.ygeno.2018.11.030
6. Parker BJ. Statistical methods for transcriptome-wide analysis of RNA methylation by bisulfite sequencing. Methods Mol Biol. 2017;1562:155–67. http://dx.doi.org/10.1007/978-1-4939-6807-7_11
7. Petegrosso R, Li Z, Kuang R. Machine learning and statistical methods for clustering single-cell RNA-sequencing data. Brief Bioinform. 2019. http://dx.doi.org/10.1093/bib/bbz063
8. Reimand J, Isserlin R, Voisin V, Kucera M, Tannus-Lopes C, Rostamianfar A, et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. Nat Protoc. 2019;14(2):482–517. http://dx.doi.org/10.1038/s41596-018-0103-9
9. Siavoshi A, Taghizadeh M, Dookhe E, Piran M. Gene expression profiles and pathway enrichment analysis to identification of differentially expressed gene and signaling pathways in epithelial ovarian cancer based on high-throughput RNA-seq data. bioRxiv. 2019:566331. http://dx.doi.org/10.1101/566331
10. Costa V, Aprile M, Esposito R, Ciccodicola A. RNA-Seq and human complex diseases: Recent accomplishments and future perspectives. Eur J Hum Genet. 2013;21(2):134–42. http://dx.doi.org/10.1038/ejhg.2012.129
11. Akond Z, Alam M, Mollah MNH. Biomarker identification from RNA-seq data using a robust statistical approach. Bioinformation. 2018;14(4):153–63. http://dx.doi.org/10.6026/97320630014153
12. Xiong H, Guo H, Xie Y, Zhao L, Gu J, Zhao S, et al. RNAseq analysis reveals pathways and candidate genes associated with salinity tolerance in a spaceflight-induced wheat mutant. Sci Rep. 2017;7(1):2731. http://dx.doi.org/10.1038/s41598-017-03024-0
13. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):96. http://dx.doi.org/10.1038/s12276-018-0071-8
14. Kotliar D, Veres A, Nagy MA, Tabrizi S, Hodis E, Melton DA, et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. Elife. 2019;8. http://dx.doi.org/10.7554/eLife.43803

15. Haque A, Engel J, Teichmann SA, Lonnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9(1):75. http://dx.doi.org/10.1186/s13073-017-0467-4
16. Hedlund E, Deng Q. Single-cell RNA sequencing: Technical advancements and biological applications. Mol Aspects Med. 2018;59:36–46. http://dx.doi.org/10.1016/j.mam.2017.07.003
17. Pizzolato G, Kaminski H, Tosolini M, Franchini DM, Pont F, Martins F, et al. Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRVdelta1 and TCRVdelta2 gammadelta T lymphocytes. Proc Natl Acad Sci U S A. 2019;116(24):11906–15. http://dx.doi.org/10.1073/pnas.1818488116
18. Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights. 2015;9(Suppl 1):29–46. http://dx.doi.org/10.4137/BBI.S28991
19. Chen G, Ning B, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Front Genet. 2019;10:317. http://dx.doi.org/10.3389/fgene.2019.00317
20. Williams AG, Thomas S, Wyman SK, Holloway AK. RNA-seq data: Challenges in and recommendations for experimental design and analysis. Curr Protoc Hum Genet. 2014;83:11.3.1–20. http://dx.doi.org/10.1002/0471142905.hg1113s83
21. Finotello F, Di Camillo B. Measuring differential gene expression with RNA-seq: Challenges and strategies for data analysis. Brief Funct Genom. 2015;14(2):130–42. http://dx.doi.org/10.1093/bfgp/elu035
22. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–15. http://dx.doi.org/10.1038/nbt.1621
23. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31(1):46–53. http://dx.doi.org/10.1038/nbt.2450
24. Robinson MD, McCarthy DJ, Smyth GK. edgeR: A bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. http://dx.doi.org/10.1093/bioinformatics/btp616
25. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40(10):4288–97. http://dx.doi.org/10.1093/nar/gks042
26. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. http://dx.doi.org/10.1186/gb-2010-11-10-r106
27. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550. http://dx.doi.org/10.1186/s13059-014-0550-8
28. Hardcastle TJ, Kelly KA. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:422. http://dx.doi.org/10.1186/1471-2105-11-422
29. Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BM, et al. EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. Bioinformatics. 2013;29(8):1035–43. http://dx.doi.org/10.1093/bioinformatics/btt087
30. Li J, Tibshirani R. Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. Stat Methods Med Res. 2013;22(5):519–36. http://dx.doi.org/10.1177/0962280211428386
31. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: A matter of depth. Genome Res. 2011;21(12):2213–23. http://dx.doi.org/10.1101/gr.124321.111
32. Tarazona S, Furio-Tari P, Turra D, Pietro AD, Nueda MJ, Ferrer A, et al. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 2015;43(21):e140.
33. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):R29. http://dx.doi.org/10.1186/gb-2014-15-2-r29
34. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods. 2017;14(7):687–90. http://dx.doi.org/10.1038/nmeth.4324
35. Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91. http://dx.doi.org/10.1186/1471-2105-14-91

36. Rajkumar AP, Qvist P, Lazarus R, Lescai F, Ju J, Nyegaard M, et al. Experimental validation of methods for differential gene expression analysis and sample pooling in RNA-seq. BMC Genom. 2015;16. http://dx.doi.org/10.1186/s12864-015-1767-y
37. Costa-Silva J, Domingues D, Lopes FM. RNA-seq differential expression analysis: An extended review and a software tool. PLoS One. 2017;12(12). http://dx.doi.org/10.1371/journal.pone.0190152
38. Lin BQ, Pang Z. Stability of methods for differential expression analysis of RNA-seq data. BMC Genom. 2019;20. http://dx.doi.org/10.1186/s12864-018-5390-6
39. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: A tutorial. Mol Syst Biol. 2019;15(6):e8746. http://dx.doi.org/10.15252/msb.20188746
40. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2. http://dx.doi.org/10.1038/nmeth.2967
41. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16:278. http://dx.doi.org/10.1186/s13059-015-0844-5
42. Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, et al. A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol. 2016;17(1):222. http://dx.doi.org/10.1186/s13059-016-1077-y
43. Miao Z, Deng K, Wang X, Zhang X. DEsingle for detecting three types of differential expression in single-cell RNA-seq data. Bioinformatics. 2018;34(18):3223–4. http://dx.doi.org/10.1093/bioinformatics/bty332
44. Wang T, Nabavi S. SigEMD: A powerful method for differential gene expression analysis in single-cell RNA sequencing data. Methods. 2018;145:25–32. http://dx.doi.org/10.1016/j.ymeth.2018.04.017
45. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6. http://dx.doi.org/10.1038/nbt.2859
46. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: A pipeline for single-cell RNA-seq profiling analysis. PLoS Comput Biol. 2015;11(11):e1004575. http://dx.doi.org/10.1371/journal.pcbi.1004575
47. Delmans M, Hemberg M. Discrete distributional differential expression (D3E) – A tool for gene expression analysis of single-cell RNA-seq data. BMC Bioinformatics. 2016;17:110. http://dx.doi.org/10.1186/s12859-016-0944-6
48. Nabavi S, Schmolze D, Maitituoheti M, Malladi S, Beck AH. EMDomics: A robust and powerful method for the identification of genes differentially expressed between heterogeneous classes. Bioinformatics. 2016;32(4):533–41. http://dx.doi.org/10.1093/bioinformatics/btv634
49. Yip SH, Wang P, Kocher JA, Sham PC, Wang J. Linnorm: Improved statistical analysis for single cell RNA-seq expression data. Nucleic Acids Res. 2017;45(22):e179. http://dx.doi.org/10.1093/nar/gkx828
50. Ntranos V, Yi L, Melsted P, Pachter L. A discriminative learning approach to differential expression analysis for single-cell RNA-seq. Nat Methods. 2019;16(2):163–6. http://dx.doi.org/10.1038/s41592-018-0303-9
51. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with Census. Nat Methods. 2017;14(3):309–15. http://dx.doi.org/10.1038/nmeth.4150
52. Dal Molin A, Baruzzo G, Di Camillo B. Single-cell RNA-sequencing: Assessment of differential expression analysis methods. Front Genet. 2017;8:62. http://dx.doi.org/10.3389/fgene.2017.00062
53. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61. http://dx.doi.org/10.1038/nmeth.4612
54. Wang T, Li B, Nelson CE, Nabavi S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. BMC Bioinformatics. 2019;20(1):40. http://dx.doi.org/10.1186/s12859-019-2599-6

7

Computational Epigenomics: From Fundamental Research to Disease Prediction and Risk Assessment

Mohamed-Amin Choukrallah • Florian Martin • Nicolas Sierro • Julia Hoeng • Nikolai V. Ivanov • Manuel C. Peitsch

PMI R&D, Philip Morris Products S.A., Neuchâtel, Switzerland

Author for correspondence: Mohamed-Amin Choukrallah, PMI R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, CH-2000 Neuchâtel, Switzerland.
E-mail: MohamedAmin.Choukrallah@pmi.com

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch7

Abstract: Over the past two decades, rapid advances in DNA sequencing technologies have allowed genome-wide interrogation of epigenetic features. The epigenome landscape encompasses a growing number of chemical properties of DNA and DNA-associated proteins; these properties are tissue-specific, distinctive for disease state and sensitive to environmental exposures. The epigenetic field has rapidly evolved from basic research investigations, aiming to understand the nature and function of epigenetic marks, to clinical and preclinical applications, where vast epigenetic information is used for risk assessment and disease prediction. The large diversity of epigenetic marks is mirrored by the complex variability of their genomic patterns and distributions. Mining of large-scale genomic datasets relies strongly on computational approaches and statistical models that should be carefully selected and adapted to fit the nature of the signals analyzed and the hypotheses tested. Here, we review recent advances in computational approaches used to analyze epigenetic data, with an emphasis on histone modifications and DNA methylation. We discuss the standard workflows for data acquisition, processing, and transformation, as well as the computational approaches used to assess statistical significance in comparative analyses. We also discuss the

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia.
ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

prediction methods utilized to associate epigenetic modifications with human disorders and environmental factors.

Keywords: data modeling; disease prediction; DNA methylation; epigenetics; histone modifications.

INTRODUCTION

Gene expression is regulated by the interaction between DNA molecules and DNA-binding proteins such as transcription factors (TFs), coactivator, and corepressor complexes. Some of these complexes modify the chromatin structure and its transcription competency. Chemical modifications to DNA and DNA-associated proteins (histones) and non-coding RNAs are considered the main epigenetic mechanisms controlling genome activity. The term epigenetics was first coined by Conrad Waddington to describe a set of causal heritable mechanisms that translate genotypes to phenotypes (1). More recent definitions describe epigenetics as modifications that regulate gene expression without altering the DNA sequence.

Epigenetic mechanisms are generally assessed by measuring their associated chemical tags or marks on target molecules, such as the methylation of cytosine or the acetylation of histone residues. The epigenome of a cell can be defined as the combination of all epigenetic marks at a given time across the genome that synergistically dictates the usage of the underlying DNA sequence. Given the large number of known epigenetic marks, and probably a much larger number of unknown ones, as well as the limitations of current epigenomics methods, the epigenome cannot be assessed as a whole, and current studies capture snapshots of only a small fraction of it. Epigenetic marks are dynamic and reversible and can undergo rapid changes during development and in response to various exposures, including drug treatment. Epigenetic alterations are also associated with a number of diseases and can be used as diagnostic and prognostic biomarkers of disease onset and progression, respectively.

The majority of epigenetic studies seek to identify changes between experimental conditions and further leverage this information to explain other molecular or physiological alterations. The results of such comparative studies mainly rely on the statistical approach and selection criteria used. Here, we review current knowledge of epigenetic mechanisms, with a focus on DNA methylation and histone modifications. We describe the workflows used to process and interpret epigenetic data generated by next-generation sequencing technologies. We also discuss the main computational approaches applied to identify differentially regulated loci and prediction methods used to associate epigenetic changes with human disorders or environmental exposures.

DNA MODIFICATIONS

Cytosine methylation at the 5-carbon position (5mC) is the most frequent DNA modification in eukaryotes. In mammals, 5mC occurs almost exclusively in the

context of CpG dinucleotides (2), with the cytosines in both strands usually being methylated. The majority of CpG sites in mammalian genomes are methylated except at active regulatory elements (REs) (3, 4). 5mC is catalyzed by DNA methyltransferases, DNMT1, DNMT3a, and DNMT3b. DNMT1 is a maintenance enzyme that ensures the inheritance of 5mC patterns during DNA replication, while DNMT3a and DNMT3b catalyze de novo DNA methylation (5). Formation of 5mC is reversible and can be converted by ten–eleven translocation (TET) enzymes to 5-hydroxymethylcytosine (5hmC), 5-formylcytosine (5fC), and 5-carboxylcytosine (5caC) through three consecutive oxidation reactions (6), ultimately leading to the unmethylated cytosine. Initially, 5mC oxidation derivatives were considered as intermediates in the process of DNA demethylation. However, recent investigations indicate that they may represent distinct epigenetic states with regulatory functions (7). Although 5mC was historically associated with gene silencing, genome-wide investigations have shown that 5mC readout depends on the genomic context. While a high level of DNA methylation at REs is indicative of transcriptional silencing, gene bodies show high levels of DNA methylation regardless of their expression status (3, 8). In addition to cytosines, eukaryotic DNA can also be methylated at the nitrogen-6 position of adenosine bases (6mA) (9). In contrast to cytosine modifications, adenosine modifications have received less attention and will not be discussed in this chapter.

ASSESSMENT OF CYTOSINE MODIFICATIONS

Cytosine modifications can be assessed by various methods (10) involving two main technologies, high-throughput sequencing and methylation arrays. While methylation arrays are restricted to annotated loci such as promoters and a fraction of known enhancers, sequencing methods can potentially cover every cytosine in the genome. Whole-genome bisulfite sequencing (WGBS) is currently the gold standard technique for assessing cytosine modifications at single-base resolution across the entire genome. WGBS is based on the bisulfite reaction that converts unmodified cytosines (uCs) into uracils while 5mC and 5hmC bases are protected from the conversion (Figure 1). After DNA amplification and high-throughput sequencing, uCs are read as thymines, whereas 5mC and 5hmC are read as cytosines. Given that WGBS cannot distinguish 5mC from 5hmC (11), the measured signal represents the sum of both modifications. However, the contribution of each modification strongly depends on its relative abundance in the investigated tissue or cell type. 5hmC levels are generally very low in mammalian cells and vary across cell types and tissues. 5hmC is abundant in the brain but extremely low in blood and spleen and almost undetectable in cultured cell lines (12). 5hmC levels can be assessed by a subtractive approach through the combination of oxidative bisulfite sequencing (oxBS-seq) (13, 14) and bisulfite sequencing (BS-seq). oxBS-seq implies an oxidation step that converts 5hmC into 5fC, which is further converted by the bisulfite reaction into uracil and read as thymine after sequencing, similar to uCs. Therefore, oxBS-seq identifies real 5mC, while BS-seq identifies 5mC + 5hmC (Figure 1). Consequently, subtracting the oxBS-seq signal from the BS-seq signal allows the computation of 5hmC levels (15), on the condition that the oxBS- and BS-conversion rates are very close to 100%.

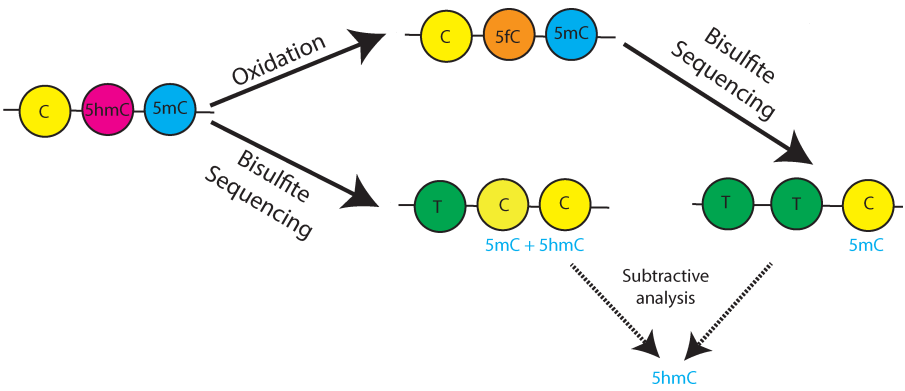


Figure 1 During bisulfite sequencing (BS-seq), unmodified cytosines (C) are read as thymines (T), while methylated (5mC) and hydroxymethylated cytosines (5hmC) are protected from bisulfite conversion and read as C. In this scenario, BS-seq does not discriminate 5mC from 5hmC. Oxidative bisulfite sequencing (oxBS-seq) includes an oxidation step, during which 5hmC is converted to 5fC and read as T after bisulfite sequencing, similar to unmodified C, while only 5mC is read as C. 5hmC proportions are computed by subtracting oxBS-seq signals from BS-seq signals.

The majority of DNA-methylation investigations are based on bisulfite conversion and assume that 5mC is the major cytosine modification and, therefore, neglect the contribution of 5hmC. The current chapter discusses only WGBS analysis, which is applicable to all BS-seq data, and the term DNA methylation will refer to 5mC + 5hmC as measured by BS-seq, unless otherwise stated.

Sequence alignment, read count, and methylation calling

The first step in analyzing DNA methylation data is the alignment (mapping) of sequencing reads to the reference genome. To maximize the rate of read mapping, it is recommended to trim sequencing adapters and low-quality bases at read ends. This process can be performed by using Trim Galore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore) or cutadapt tools (16). C-to-T bisulfite conversion results in reduced complexity of the converted DNA and subsequent loss of complementarity with the reference genome. The bisulfite reaction and subsequent DNA amplification produce four individual strands from a single original fragment. Additionally, bisulfite libraries can be directional or non-directional, with the first approach preserving strand specificity in contrast to the second (17). BS-seq analysis tools such as Bismark (18) and QuasR (19) take into consideration these parameters and try to identify the best unique alignment by running four alignment processes simultaneously. Firstly, reads are C-to-T or G-to-A (reverse strand) converted and aligned to an equivalently converted genome. The alignment process is time-consuming and requires considerable computing resources. For large studies, a high-performance computing cluster is required.

As the assessment of cytosine status strongly depends on C-to-T conversion, the bisulfite conversion efficiency must be controlled for every experiment. Spiking samples with unmethylated lambda phage DNA provide an accurate estimation of bisulfite conversion as all cytosines in this genome should be converted. High-quality experiments produce conversion rates greater than 99%.

In classical WGBS experiments, where 5mC and 5hmC cannot be distinguished, both modifications are reported as methylated cytosines. In this context, the methylation level of cytosines is reported as the ratio of the number of reads with "C" (5mc + 5hmC) over the number of reads with either "C" or "T" (5mC + 5hmC + C). These reads originate from a population of cells with variable methylation states. Therefore, the methylation ratio ranges from 0 to 1, where 0 corresponds to a fully unmethylated state and 1 indicates a fully methylated state. Tools such as Bismark and QuasR produce count matrices containing the number of methylated and unmethylated reads for every cytosine that can be used for further analysis. In studies combining the oxBS-seq and BS-seq approaches, 5mC, 5hmC, and uC proportions can be computed using maximum likelihood estimates (20) or binomial modeling (21). It is important to mention that calculating the simple difference between BS and oxBS signals as an estimate of 5hmC can produce negative proportions and sums (5mC + 5hmC + uC) greater than 1. Such inconsistencies may simply represent sequencing artifacts or low-coverage biases and have no biological significance.

DNA methylation patterns

Cytosine methylation, as measured by WGBS in the human (3) and mouse (4) genomes, has shown that 5mC occurs mostly in the context of CpG dinucleotides, while non-CpG methylation is a rare event. The 5mC frequency of individual CpGs has a bimodal distribution, with the majority of CpGs being highly methylated and a small subset of CpGs showing an unmethylated state. In addition to these two categories, a third population of CpGs shows an intermediate range of methylation ranging from 10 to 50% (Figure 2). At the genome scale, the methylome can be segmented into three distinct classes: fully methylated regions (FMRs),

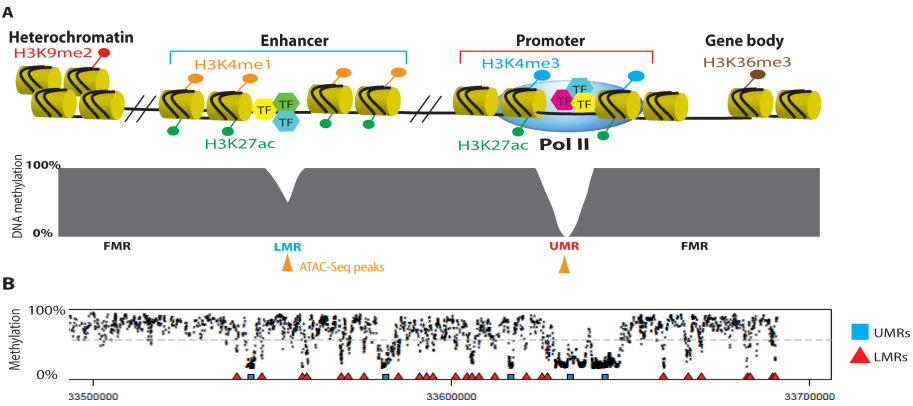


Figure 2 Epigenetic landscape. (A) Genomic distribution of the main epigenetic marks. Histones are depicted as yellow cylinders; black lines represent DNA, and modifications of histone tails are shown as circles. Transcription factors (TF) are indicated by hexagons. Grey blocks represent DNA methylation patterns. UMRs, LMRs and FMRs show the unmethylated regions, low-methylated regions and fully methylated regions, respectively. Assay for transposase-accessible chromatin using sequencing (ATAC-Seq) peaks are depicted in orange. (**B**) Genome browser snapshot illustrating DNA methylation patterns. ac: acetylation; H: histone; K: lysine; me: methylation; Pol II: RNA polymerase II.

unmethylated regions (UMRs), and low-methylated regions (LMRs) (4). FMRs represent 90% of the genome and are enriched at inter- and intragenic regions. UMRs correspond to the majority of CpG islands and active promoters, while LMRs exhibit enhancer features such as specific histone marks and binding of TFs (4) (Figure 2). While the majority of genomic regions fit this classification, some cell types contain contiguous regions showing disordered states of methylation ranging from 0 to 100%, with little similarity between neighboring CpGs (3). These loci were termed "partially methylated domains" (PMDs). It is important to mention that PMDs and LMRs are two distinct methylation profiles, and particular attention should be paid to the behavior of PMDs in comparative investigations. Methylation distribution in PMDs may affect the identification of LMRs (22) and the computing of differential methylation between experimental groups. The presence of PMDs can be evaluated using, for instance, the MethylSeekR package (22), and whether or not to exclude them from the analysis depends on the study purpose.

Computing differential methylation

After methylation calling, the next step is to compare methylation profiles between experimental groups to identify differentially methylated cytosines (DMCs) or differentially methylated regions (DMRs). Selecting the appropriate statistical model is the most important step in computing differential methylation for studies with biological replicates. The different statistical methods used to call DMCs/DMRs were summarized in an excellent review by Wreczycka et al. (23), which addressed additional aspects of DNA methylation analysis. In general, regression models are the best choice for comparing methylation profiles in studies with several replicates per experimental group. The selected method should take into consideration intra-group variability, which is more pronounced in in vivo and clinical studies. Beta-binomial distribution is a natural choice for computing differential methylation when biological replicates are available as it can correct for technical sampling and intra-group variability. A number of tools, such as DSS (24) and RADMeth (25), are based on the beta-binomial distribution. These tools compute differential methylation for single cytosines and require predefined file formatting. For more flexibility, the beta-binomial distribution is implemented in a number of R packages such as TailRank (https://cran.r-project.org/web/packages/TailRank/index.html) and AOD (https://cran.r-project.org/web/packages/aod/index.html).

DMRs are usually called based on the FDR-adjusted P-value from the fitted statistical model. The value of methylation difference can also be used as an additional selection parameter. As the outcome of this approach is based on the cutoffs used, it is important to have a global quantitative view of data distribution by generating volcano or simple scatter plots. Another important parameter in calling DMRs is the read coverage at the investigated position, because accurate evaluation of methylation differences between samples requires decent read coverage. In our own work, we set the minimum number of reads to 15; however, this parameter can be changed depending on the data at hand. Computing differential methylation should also take into consideration a number of covariates such as age, sex, and other potential confounding factors. Finally, genetic variations can also affect methylation status, and particular attention should be paid to C/T single-nucleotide polymorphism (SNPs).

In addition to differential methylation, increased variability in methylation levels can also be observed at some loci in response to exposures (26, 27) or in relation to some diseases (28). These variably methylated regions (VMRs) have been suggested to be regions of stochastic epigenetic variations (29) that may indicate a certain degree of flexibility in the control of local chromatin structure. VMRs have been observed to occur preferentially at enhancers and 3′-untranslated regions (3′UTRs) (30), suggesting a potential role in gene regulation. VMRs can be called simply by calculating the variance; however, this approach is sensitive to intra-individual and technical variations. The multiple hypothesis testing approach has been suggested to call VMRs by distinguishing biological variability from intra-individual variations (31). Although this approach was applied to methylation arrays, it can also be adapted to sequencing data.

DNA methylation in disease research and risk assessment

Genome-wide association studies (GWAS) have been designed to identify risk-associated SNPs that can be used as prediction tools in clinical investigations or for personalized medicine. Similarly, epigenome-wide association studies (EWAS) aim to derive potential associations between epigenetic marks and a particular trait, disease or exposure–response profile. To date, the vast majority, if not all, of EWAS have been based on DNA methylation. Therefore, these investigations should rather be termed "methylome-wide association studies" (MWAS). MWAS have been mainly conducted in the context of tumorigenesis, where the methylome of cancer cells is characterized by global hypomethylation except at some CpG islands that undergo hypermethylation (32). MWAS have been also conducted in relation to various diseases and phenotypes. The EWASdb database records 1319 MWAS associated with 302 diseases and/or phenotypes, including autoimmune, metabolic, and exposure-related disorders, to name a few (33).

To date, most MWAS have been based on methylation arrays that interrogate mainly annotated regions and poorly cover the complex network of distal REs. Given the central role of distal REs in genome regulation, it is crucial to interrogate the association of these loci with the traits of interest. For example, WGBS investigations in the mouse lung showed that cigarette smoke exposure mainly alters DNA methylation at candidate enhancers (identified as LMRs), while promoters are less affected (34). The importance of distal RE is also illustrated by the fact that the majority of GWAS-identified hits are located in non-coding regions with potential regulatory function, arguing for their informative value in both GWAS and EWAS.

Ideally, a comprehensive MWAS would assess cytosine status at a genome-wide level using WGBS and oxBS-seq in parallel to discriminate 5mC from 5hmC. However, this scenario requires high read coverage to accurately evaluate methylation variations. Finally, an adequate sample size is required to assure sufficient power to detect methylation differences (35). These requirements make whole-genome investigations costly for studies involving large cohorts. Alternatively, cytosine methylation can be investigated for a defined set of genomic targets using capture techniques (36). This approach allows the design of custom sets of loci to address specific needs and to increase the read coverage per site, while reducing the cost.

MWAS must also take into consideration inter- and intra-individual variations in DNA methylation levels. Unlike genetic information, where all cell types share the same genome, the epigenome varies between cell types and tissues. Thus, epigenome profiling in peripheral sources such as blood and saliva may not recapitulate the variations occurring in specific target organs. Cell heterogeneity in liquid biopsies may also complicate the use of DNA methylation variations as reliable biomarkers. Additionally, epigenetic marks change over time, are sensitive to environmental factors and health status, and may be affected by genetic variants. The aforementioned confounding factors and many others should be considered during the experimental design and computational framework.

DNA methylation results are generally reported as the difference in mean methylation ratios between experimental groups, with P-values derived from a sound statistical model. In most MWAS, the effect size is modest. It is rare to observe cytosines moving from the unmethylated state to fully methylated state or vice versa except when comparing methylomes from different cellular origins (37). Usually, the association between the response variable (disease, exposure, and the like) and explanatory variable (DNA methylation level) relies on the P-value, while the effect size is neglected, thus reducing the applicability of MWAS in personalized medicine. For example, the cg03636183 CpG site located in the *F2RL3* gene is considered a strong marker of cigarette-smoke exposure in blood samples. The median methylation level of this CpG is 95% in never smokers and 83% in smokers (38). Despite the methylation difference of 12%, this site still belongs to the category of fully methylated CpGs and can hardly be used to distinguish smokers from non-smokers. However, this site and many others are reproducibly found to be differentially methylated in independent cohorts in relation to smoking. Given the binary nature of 5mC at the allele level, these reproducible, but weak, variations can be explained by the cellular heterogeneity of blood samples. Some DNA methylation variation may reflect the cell-type composition of blood samples (39, 40). As mentioned earlier, the measured methylation level represents the average of events occurring in a population of cells. Therefore, cell-type-specific variations may be diluted in the averaged bulk signal. It has been shown that many loci, including cg03636183 in *F2RL3* and cg05575921 in *AHRR*, exhibit distinct patterns of smoking-associated methylation variations across blood-cell types (41). Investigating DNA methylation variations in specific cell types may reduce the bias linked to cell heterogeneity and allow more accurate detection of cell-type-specific DMRs or DMCs.

Leveraging epigenetic associations to causal biologic mechanisms is still challenging. DNA methylation variations can be the cause or consequence of the investigated phenotype. This complex interaction is illustrated by the chronology of promoter hypermethylation in cancer cells. It has been reported that some transcriptionally silenced promoters in healthy cells become aberrantly hypermethylated during tumorigenesis, implying that the hypermethylation of some loci is likely the consequence, rather than the cause, of tumorigenesis (42). Despite the lack of clear causality to cancer etiology, DNA methylation levels of a limited set of loci have been used to develop diagnostic tests for colorectal, prostate, and bladder cancers (43). Cologuard®, a DNA methylation-based diagnostic kit (Exact Sciences Corporation, WI, USA), was the first stool DNA screening test approved by the U.S. Food and Drug Administration for colorectal cancer.

Machine learning in MWAS

Machine-learning (ML) algorithms are promising tools for identifying methylome variations predictive or indicative of certain phenotypes or exposures. These algorithms seek to identify a set of loci (features) whose methylation levels can be used as a signature to categorize samples from different experimental groups (classification methods) or to estimate continuous metrics such as age (regression methods). In the context of MWAS, classification algorithms have been mainly used to classify cancer samples. Random forest (RF)-based supervised learning is one of the most used ML algorithm in MWAS (37, 44, 45). For example, this algorithm has been used to construct a DNA methylation signature based on 20 loci for stratifying different types of brain metastasis. This signature also showed a good performance on samples from a test set that was not used to train the model (37). The good classification power of this signature is probably due to the cell-type-specific DNA methylation patterns of primary tumors. The RF algorithm has been also used to build DNA methylation signatures to classify different tumor types, including breast, kidney, and thyroid carcinomas (44), and to classify central nervous system tumors (45). DNA methylation has been also used to classify subtypes and predict treatment outcome in patients with childhood acute lymphoblastic leukemia using the nearest shrunken centroids (NSC) approach.

Regression algorithms have been also applied to methylome data, mainly in the context of age prediction. DNA methylation of a limited set of CpGs has been used to build age predictors in humans (46, 47) and mice (48, 49). One of the first epigenetic predictors of age, termed the Horvath clock (46), is a multi-tissue predictor based on 353 CpGs and can estimate chronological age in test samples with a median error of 3.6 years. This model has been derived from 8000 methylomes using elastic-net regression. After this pioneering work, a number of other DNA methylation clocks have been developed using other tissues and regression algorithms (50). Regularized linear regressions are the most used algorithms for building age predictors. The regression method selected depends on the data at hand and the questions to be answered. Although the most accurate DNA methylation clocks are derived from elastic-net regression, the beneficial effects of anti-aging interventions are better computed by ridge regression-based clocks (49). ML approaches have mainly been applied to data generated by methylation arrays and are only starting to be used for sequencing datasets. In the context of sequencing datasets, the aforementioned limitations for computing differential methylation are also valid for ML approaches, and particular attention should be paid to poorly covered sites.

CHROMATIN REGULATION

Chromatin is a DNA–protein complex, the primary function of which is to organize the genetic material in a compact form to fit into the nucleus. The fundamental chromatin unit, the nucleosome, consists of 147 DNA base pairs wrapped around histone octamers. Multiple histone residues, mainly at histone tails, can undergo covalent post-translational modifications (PTMs), including methylation, phosphorylation, acetylation, and SUMOylation. PTM regulation involves

three families of epigenetic enzymes: writers that catalyze the addition of various chemical groups, readers that recognize and interpret these modifications, and erasers that remove them to ensure dynamic epigenetic regulation (51).

Genome-wide mapping of PTMs has identified their functional association with chromatin properties, transcriptional competency, DNA-damage repair, and DNA replication. The combination of PTMs at a particular locus shapes the local chromatin structure and modulates transcriptional activity. This combinatorial regulatory code has been termed "histone code." For example, trimethylation of lysine 4 of histone 3 (H3K4me3) marks actively transcribed promoters, while monomethylation of the same residue (H3K4me1) marks active enhancers. Acetylation of any histone residue (e.g., H3K27ac, H3K9ac or H3K14ac) is always associated with active REs (Figure 2). Other PTMs are associated with transcriptional repression. For example, H3K9me2 is a key marker of heterochromatin domains (52), and H3K27me3 indicates polycomb group silenced loci (53).

Profiling histone modifications by ChIP sequencing

Chromatin immunoprecipitation (ChIP) followed by high-throughput sequencing (ChIP-seq) is a powerful technique for profiling the genomic distribution of PTMs and other DNA-binding proteins such as TFs and epigenetic enzymes. ChIP-seq involves an immunoprecipitation (IP) step using antibodies directed against the target protein. The captured DNA is further subjected to next-generation sequencing, and the resulting reads are mapped to the reference genome to identify the binding sites of target proteins (Figure 3). The general assumption is that target protein binding sites will produce more reads than the rest of the genome, which will be covered by the sequencing noise/background captured by unspecific binding of the IP antibody. The sequencing noise is generally assessed by sequencing a fraction of the input chromatin prior to the IP step. This noise is not uniform and reflects local chromatin accessibility, amplification, and mappability biases. The interpretability of ChIP-seq experiments strongly depends on antibody specificity, the amount of starting material, and epitope integrity after cell lysis and chromatin shearing. The impact of these parameters is reflected by the signal-to-noise ratio in the sequencing data.

Once the sequencing reads are aligned to the reference genome, the next step is to identify genomic regions that are enriched for the target protein. This step is usually termed peak-calling, because the first ChIP-seq experiments were mainly designed to map TFs and resulted in very short enriched regions (0.5 kb to 1 kb) with a peak shape and clear summit (maximum read density) when visualized on genome browsers. However, not all ChIP-seq experiments generate narrow peaks. Some PTMs such as heterochromatin marks are uniformly enriched in very large regions with no clear summit, while other PTMs such as active promoter marks (e.g., H3K4me3 and H3K9ac) are enriched in relatively short regions (1 kb to 2 kb) with a clear local maximum read density. More complex patterns include a mixture of narrow peaks and diffused regions such as the H3K27me3 mark. The majority of peak-calling tools (listed in two reviews (54, 55) were designed to detect narrow peaks and may not perform accurately on ChIP-seq experiments with broad and diffuse enriched regions. However, some tools such as the popular MACS (56) and Epic (57) have included new parameters to model mixed enrichment events in recent updates.

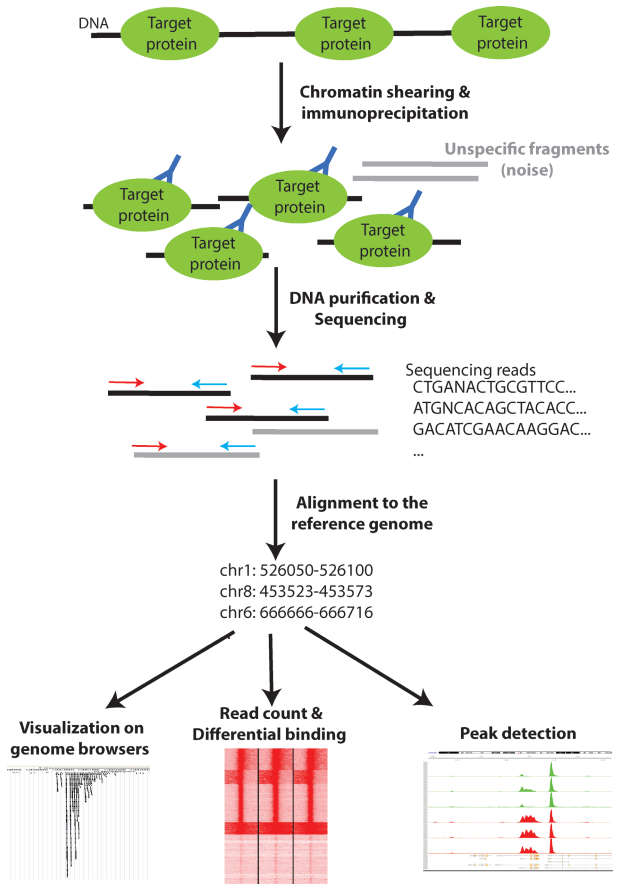


Figure 3 ChIP-seq workflow. Sheared chromatin is incubated with an antibody directed against the target protein. Upon purification, the captured DNA is then subjected to high-throughput sequencing, and the resulting reads are aligned to the reference genome. Aligned reads can be visualized on genome browsers and computed to identify binding sites of target proteins.

Modeling the background distribution of reads is an important step in peak detection and can be performed from the input control, but not all studies include this control. Consequently, the majority of algorithms model the intrinsic background of ChIP samples. While this approach performs well for narrow-peak experiments, it provides poor results for diffuse enriched regions. In our opinion, the input control should be included in all ChIP-seq experiments. A simple scatter plot comparison of read counts over genomic windows from ChIP samples versus the input control (Figure 4) provides a primary evaluation of ChIP-seq quality. Additionally, we believe that the input control is mandatory for investigating heterochromatin marks, given their genomic distribution. Finally, particular attention should be paid to repetitive elements that produce very short peaks with a high number of reads, as these peaks represent sequencing biases rather than binding events.

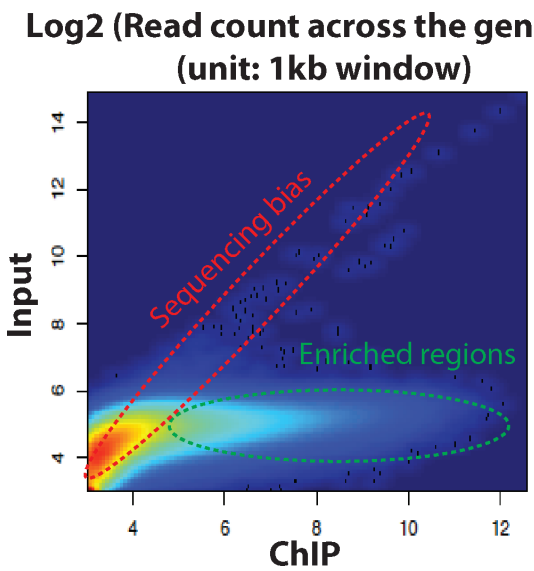


Figure 4 Example of a high-quality ChIP-seq experiment. The genome-wide ChIP signal (number of reads per 1-kb window) is plotted against the corresponding input control. Enriched/bound loci (indicated in green) show higher read numbers in the ChIP sample than in the input control. Unbound loci show low read numbers in the ChIP sample. Reads originated from unspecific binding and/or sequencing biases are present equally in ChIP and input samples (indicated in red).

The initial goal of ChIP-seq experiments was to investigate the genomic distribution of DNA-binding proteins in the context of basic research, and the first ChIP-seq studies rarely included biological replicates. Comparative analyses mainly consisted of detecting differentially enriched regions at defined coordinates, such as annotated promoters based on read count cut-offs. With the continuous decrease in sequencing costs and widespread application of the technique, including in clinical investigations, most recent studies include biological replicates. A number of approaches have been suggested to leverage biological replicates to improve the accuracy of peak detection (58). Most of these methods compare the overlap between peaks detected independently in the different replicates and select confident peaks based on reproducibility. While this approach is convenient for identifying highly confident-enriched regions, it is not suitable for identifying significantly differentially enriched regions based on read counts in comparative analyses (e.g., case vs. control).

To identify differentially enriched regions between experimental groups, we suggest that the analysis should include all potentially bound regions, even those with low confidence. If a peak caller is used, peaks from different replicates and experimental groups can be merged to build a unique set of loci. A more holistic approach consists of assessing differential binding/enrichment at genomic windows along the chromosomes. Once a consensus set of loci is defined, read counts can be generated for all replicates. Differential binding can then be computed based on read counts similar to differential expression in RNA sequencing data.

This step can be performed using the DEseq2 package (59), which uses negative binomial distribution to compute the statistical significance between groups. Other packages such as Diffbind and MMDiff have been developed specifically for differential ChIP-seq analysis. Diffbind uses DEseq2 internally but offers the possibility to integrate input controls, while MMDiff takes in account the distribution of reads within the enriched regions. The choice of which approach to use is dictated by the questions to be answered, number of replicates, and availability of control experiments. Although the majority of available tools perform a normalization step, it is important to ensure the scaling of unequal datasets by library size.

Genome-wide chromatin investigations are rarely conducted in clinical studies because of the complexity of chromatin properties, amount of starting material required, and multiplicity of processing steps. Additionally, histone modification patterns are cell-type-specific and need to be generated from target organs rather than from peripheral sources, which restricts the investigations to postoperative and post-mortem samples. A search for clinical trials involving chromatin among the 308,830 clinical trial records available in the ClinincalTrials.gov database resulted in only 82 and 16 hits for the terms chromatin and ChIP-seq, respectively. The recent adaptation of the ChIP-seq protocol to small cancer biopsies (60) may, however, facilitate the future use of this approach in clinical studies.

Assessing chromatin accessibility by ATAC-seq

The assay for transposase-accessible chromatin using sequencing (61) (ATAC-seq) allows detection of accessible (i.e., open) chromatin regions, which are mainly active REs and TF-binding sites. ATAC-seq is based on a process called tagmentation, which involves simultaneous fragmentation and sequencing-adapter ligation. This reaction is carried out with a hyperactive mutant of Tn5 transposase that inserts sequencing adapters into open chromatin regions. Reads produced from these regions during high-throughput sequencing are used to detect peaks, similar to ChIP-seq data. While ChIP-seq ideally requires a few million cells, a standard ATAC-seq experiment requires only 50,000 cells, making it more suitable for studies with a limited amount of starting material. Although ATAC-seq provides no information about the identity of the binding proteins, ATAC-seq-enriched regions show high overlap with active RE-associated PTMs such as H3K4me3 and H3K27ac. ATAC-seq has been recently used to investigate open chromatin distribution in 23 cancer types (62).

CONCLUSION

Advances in sequencing technologies have enabled scientists to reveal the striking immensity of gene-regulation mechanisms and, particularly, the large repertoire of epigenetic pathways. Although a number of these mechanisms are now well understood, many others remain to be elucidated. For example, the human genome codes for hundreds of TFs, but only a small fraction of them have been studied (63). Similarly, the roles of many histone and DNA modifications remain unclear. Transcriptional alterations play a central role in almost all human

disorders, and these alterations are very likely preceded by changes in epigenetic patterns and TF binding and/or activity.

The diversity of measurable epigenetic marks holds the promise of using epigenetic events as early markers of human disorders and for providing mechanistic clues to disease etiology. However, this initial excitement about epigenetic markers has been tempered by the complexity of their biological outcomes and their interactions with other molecular signals (e.g., gene expression). At the molecular level, most, if not all, epigenetic marks are binary, and their variations in some loci can, in theory, be used to monitor a number of biological processes. However, most of the observed epigenetic changes in EWAS are modest and reflect the average of events occurring in a heterogeneous population of cells. Advances in single-cell investigations may help unveil more reliable epigenetic markers as exemplified by the recent characterization of DNA methylation profiles of circulating tumor cells using single-cell methylomes (64) and single-cell ChIP-seq investigation of breast cancer heterogeneity (65). Another limitation of currently available EWAS is the poor investigation of non-coding regions that contain most of the distal REs and represent the vast majority of disease-associated variations at the genetic level.

In our opinion, improvement of EWAS outcomes should be articulated around three main axes: reduction of cell-type heterogeneity, increase in genome coverage, and combination of a larger panel of epigenetic marks. Overcoming these challenges will require massive computational and technical efforts in both academic and industrial research. Generating interpretable genome-wide data from low cell number or single-cell samples will likely be the next breakthrough in clinical investigations. This new type of data will require the development of new computational approaches prioritizing personalized assessment rather than group comparisons. Computational investigations should also leverage the diversity of epigenetic marks together with other omics data to better understand the flow of events leading to disease onset, possibly identifying combinatorial markers for disease progression and drug response.

Conflict of interest: The authors declare no potential conflict of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced.

REFERENCES

1. Waddington CH. The epigenotype. 1942. Int J Epidemiol. 2012 Feb;41(1):10–13. http://dx.doi.org/10.1093/ije/dyr184
2. Lei H, Oh SP, Okano M, Juttermann R, Goss KA, Jaenisch R, et al. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. Development (Cambridge, England). 1996 Oct;122:3195–205.
3. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009 Nov 19;462:315–22. http://dx.doi.org/10.1038/nature08514

4. Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Scholer A, et al. DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature. 2011 Dec 14;480:490–5. http://dx.doi.org/10.1038/nature10716
5. Lyko F. The DNA methyltransferase family: A versatile toolkit for epigenetic regulation. Nat Rev Genet. 2018 Feb;19(2):81–92. http://dx.doi.org/10.1038/nrg.2017.80
6. Wu H, Zhang Y. Mechanisms and functions of Tet protein-mediated 5-methylcytosine oxidation. Genes Dev. 2011 Dec 1;25(23):2436–52. http://dx.doi.org/10.1101/gad.179184.111
7. Hill PWS, Amouroux R, Hajkova P. DNA demethylation, Tet proteins and 5-hydroxymethylcytosine in epigenetic reprogramming: An emerging complex story. Genomics. 2014 Nov 1;104(5):324–33. http://dx.doi.org/10.1016/j.ygeno.2014.08.012
8. Baubec T, Colombo DF, Wirbelauer C, Schmidt J, Burger L, Krebs AR, et al. Genomic profiling of DNA methyltransferases reveals a role for DNMT3B in genic methylation. Nature. 2015 Apr 9;520(7546):243–7. http://dx.doi.org/10.1038/nature14176
9. Wu TP, Wang T, Seetin MG, Lai Y, Zhu S, Lin K, et al. DNA methylation on N(6)-adenine in mammalian embryonic stem cells. Nature. 2016 Apr 21;532(7599):329–33. http://dx.doi.org/10.1038/nature17640
10. Epigenomics in tobacco risk assessment: Opportunities for integrated new approaches – ScienceDirect [Internet]. [cited 2019 Jun 7]. Available from: https://www.sciencedirect.com/science/article/pii/S2468202018300573
11. Nestor C, Ruzov A, Meehan R, Dunican D. Enzymatic approaches and bisulfite sequencing cannot distinguish between 5-methylcytosine and 5-hydroxymethylcytosine in DNA. BioTechniques. 2010 Apr;48(4):317–19. http://dx.doi.org/10.2144/000113403
12. Nestor CE, Ottaviano R, Reddington J, Sproul D, Reinhardt D, Dunican D, et al. Tissue type is a major modifier of the 5-hydroxymethylcytosine content of human genes. Genome Res. 2012 Mar;22(3):467–77. http://dx.doi.org/10.1101/gr.126417.111
13. Booth MJ, Branco MR, Ficz G, Oxley D, Krueger F, Reik W, et al. Quantitative sequencing of 5-methylcytosine and 5-hydroxymethylcytosine at single-base resolution. Science. 2012 May 18;336(6083):934–7. http://dx.doi.org/10.1126/science.1220671
14. Booth MJ, Ost TWB, Beraldi D, Bell NM, Branco MR, Reik W, et al. Oxidative bisulfite sequencing of 5-methylcytosine and 5-hydroxymethylcytosine. Nat Protoc. 2013 Oct;8(10):1841–51. http://dx.doi.org/10.1038/nprot.2013.115
15. Skvortsova K, Zotenko E, Luu P-L, Gould CM, Nair SS, Clark SJ, et al. Comprehensive evaluation of genome-wide 5-hydroxymethylcytosine profiling approaches in human DNA. Epigenetics Chromatin [Internet]. 2017 Apr 20 [cited 2018 Oct 15];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5397694/
16. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011 May 2;17(1):10–12. http://dx.doi.org/10.14806/ej.17.1.200
17. Chen P-Y, Cokus SJ, Pellegrini M. BS Seeker: Precise mapping for bisulfite sequencing. BMC Bioinformatics. 2010 Apr 23;11:203. http://dx.doi.org/10.1186/1471-2105-11-203
18. Krueger F, Andrews SR. Bismark: A flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011 Jun 1;27(11):1571–2. http://dx.doi.org/10.1093/bioinformatics/btr167
19. Gaidatzis D, Lerch A, Hahne F, Stadler MB. QuasR: Quantification and annotation of short reads in R. Bioinformatics. 2015 Apr 1;31:1130–2. http://dx.doi.org/10.1093/bioinformatics/btu781
20. Kiihl SF, Martinez-Garrido MJ, Domingo-Relloso A, Bermudez J, Tellez-Plaza M. MLML2R: An R package for maximum likelihood estimation of DNA methylation and hydroxymethylation proportions. Stat Appl Genet Mol Biol. 2019 Jan 17;18(1):pii:/j/sagmb.2019. http://dx.doi.org/10.1515/sagmb-2018-0031
21. Xu Z, Taylor JA, Leung Y-K, Ho S-M, Niu L. oxBS-MLE: An efficient method to estimate 5-methylcytosine and 5-hydroxymethylcytosine in paired bisulfite and oxidative bisulfite treated DNA. Bioinformatics. 2016 Jan;32(23):3667–9. http://dx.doi.org/10.1093/bioinformatics/btw527
22. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013 Sep;41(16):e155. http://dx.doi.org/10.1093/nar/gkt599

23. Wreczycka K, Gosdschan A, Yusuf D, Grüning B, Assenov Y, Akalin A. Strategies for analyzing bisulfite sequencing data. J Biotechnol. 2017 Nov 10;261:105–15. http://dx.doi.org/10.1016/j.jbiotec.2017.08.007
24. Feng H, Conneely KN, Wu H. A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res. 2014 Apr;42(8):e69. http://dx.doi.org/10.1093/nar/gku154
25. Dolzhenko E, Smith AD. Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinformatics. 2014 Jun 24;15:215. http://dx.doi.org/10.1186/1471-2105-15-215
26. Jenkins TG, James ER, Alonso DF, Hoidal JR, Murphy PJ, Hotaling JM, et al. Cigarette smoking significantly alters sperm DNA methylation patterns. Andrology. 2017 Nov;5(6):1089–99. http://dx.doi.org/10.1111/andr.12416
27. Vaz M, Hwang SY, Kagiampakis I, Phallen J, Patil A, O'Hagan HM, et al. Chronic cigarette smoke-induced epigenomic changes precede sensitation of bronchial epithelial cells to single-step transformation by KRAS mutations. Cancer Cell. 2017 Sep 11;32(3):360–376.e6. http://dx.doi.org/10.1016/j.ccell.2017.08.006
28. Webster AP, Plant D, Ecker S, Zufferey F, Bell JT, Feber A, et al. Increased DNA methylation variability in rheumatoid arthritis-discordant monozygotic twins. Genome Med. 2018 Sep 4;10(1):64. http://dx.doi.org/10.1186/s13073-018-0575-9
29. Feinberg AP, Irizarry RA. Stochastic epigenetic variation as a driving force of development, evolutionary adaptation, and disease. PNAS. 2010 Jan 26;107(Suppl 1):1757–64. http://dx.doi.org/10.1073/pnas.0906183107
30. Garg P, Joshi RS, Watson C, Sharp AJ. A survey of inter-individual variation in DNA methylation identifies environmentally responsive co-regulated networks of epigenetic variation in the human genome. PLoS Genet. 2018 Oct 1;14(10):e1007707. http://dx.doi.org/10.1371/journal.pgen.1007707
31. Jaffe AE, Feinberg AP, Irizarry RA, Leek JT. Significance analysis and statistical dissection of variably methylated regions. Biostatistics. 2012 Jan;13(1):166–78. http://dx.doi.org/10.1093/biostatistics/kxr013
32. Sproul D, Meehan RR. Genomic insights into cancer-associated aberrant CpG island hypermethylation. Brief Funct Genomics. 2013 May;12(3):174–90. http://dx.doi.org/10.1093/bfgp/els063
33. Liu D, Zhao L, Wang Z, Zhou X, Fan X, Li Y, et al. EWASdb: Epigenome-wide association study database. Nucleic Acids Res. 2019 Jan 8;47(D1):D989–93. http://dx.doi.org/10.1093/nar/gky942
34. Choukrallah M-A, Sierro N, Martin F, Baumer K, Thomas J, Ouadi S, et al. Tobacco Heating System 2.2 has a limited impact on DNA methylation of candidate enhancers in mouse lung compared with cigarette smoke. Food Chem Toxicol. 2019 Jan 1;123:501–10. http://dx.doi.org/10.1016/j.fct.2018.11.020
35. Rakyan VK, Down TA, Balding DJ, Beck S. Epigenome-wide association studies for common human diseases. Nat Rev Genet. 2011 Jul 12;12(8):529–41. http://dx.doi.org/10.1038/nrg3000
36. Li Q, Suzuki M, Wendt J, Patterson N, Eichten SR, Hermanson PJ, et al. Post-conversion targeted capture of modified cytosines in mammalian and plant genomes. Nucleic Acids Res. 2015 Jul 13;43(12):e81. http://dx.doi.org/10.1093/nar/gkv244
37. Orozco JIJ, Knijnenburg TA, Manughian-Peter AO, Salomon MP, Barkhoudarian G, Jalas JR, et al. Epigenetic profiling for the molecular classification of metastatic brain tumors. Nat Commun [Internet]. 2018 Nov 6 [cited 2019 Jun 14];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6219520/
38. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. Am J Hum Genet. 2011 Apr 8;88:450–7. http://dx.doi.org/10.1016/j.ajhg.2011.03.003
39. Bauer M, Linsel G, Fink B, Offenberg K, Hahn AM, Sack U, et al. A varying T cell subtype explains apparent tobacco smoking induced single CpG hypomethylation in whole blood. Clin Epigenetics. 2015;7:81. http://dx.doi.org/10.1186/s13148-015-0113-1
40. Bauer M, Fink B, Thürmann L, Eszlinger M, Herberth G, Lehmann I. Tobacco smoking differently influences cell types of the innate and adaptive immune system-indications from CpG site methylation. Clin Epigenetics. 2015;7:83. http://dx.doi.org/10.1186/s13148-016-0249-7

41. Su D, Wang X, Campbell MR, Porter DK, Pittman GS, Bennett BD, et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. PLoS One. 2016;11(12):e0166486. http://dx.doi.org/10.1371/journal.pone.0166486
42. Sproul D, Nestor C, Culley J, Dickson JH, Dixon JM, Harrison DJ, et al. Transcriptionally repressed genes become aberrantly methylated and distinguish tumors of different lineages in breast cancer. Proc Natl Acad Sci U S A. 2011 Mar 15;108:4364–9. http://dx.doi.org/10.1073/pnas.1013224108
43. Kronfol MM, Dozmorov MG, Huang R, Slattum PW, McClay JL. The role of epigenomics in personalized medicine. Expert Rev Precis Med Drug Dev. 2017;2(1):33–45. http://dx.doi.org/10.1080/23808993.2017.1284557
44. Celli F, Cumbo F, Weitschek E. Classification of large DNA methylation datasets for identifying cancer drivers. Big Data Research. 2018 Sep;13:21–8. http://dx.doi.org/10.1016/j.bdr.2018.02.005
45. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. DNA methylation-based classification of central nervous system tumours. Nature. 2018 Mar;555(7697):469–74.
46. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(10):R115. http://dx.doi.org/10.1186/gb-2013-14-10-r115
47. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013 Jan 24;49(2):359–67. http://dx.doi.org/10.1016/j.molcel.2012.10.016
48. Multi-tissue DNA methylation age predictor in mouse. PubMed – NCBI [Internet]. [cited 2019 Apr 11]. Available from: https://www.ncbi.nlm.nih.gov/pubmed/28399939
49. Thompson MJ, Chwiałkowska K, Rubbi L, Lusis AJ, Davis RC, Srivastava A, et al. A multi-tissue full lifespan epigenetic clock for mice. Aging (Albany NY). 2018 Oct 21;10(10):2832–54. http://dx.doi.org/10.18632/aging.101590
50. Weidner CI, Lin Q, Koch CM, Eisele L, Beier F, Ziegler P, et al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 2014 Feb 3;15(2):R24. http://dx.doi.org/10.1186/gb-2014-15-2-r24
51. Kouzarides T. Chromatin modifications and their function. Cell. 2007 Feb 23;128(4):693–705. http://dx.doi.org/10.1016/j.cell.2007.02.005
52. Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. Nature. 2001 Mar 1;410(6824):116–20. http://dx.doi.org/10.1038/35065132
53. Bernstein E, Duncan EM, Masui O, Gil J, Heard E, Allis CD. Mouse polycomb proteins bind differentially to methylated histone H3 and RNA and are enriched in facultative heterochromatin. Mol Cell Biol. 2006 Apr;26(7):2560–9. http://dx.doi.org/10.1128/MCB.26.7.2560-2569.2006
54. Wilbanks EG, Facciotti MT. Evaluation of algorithm performance in ChIP-Seq peak detection. PLoS One. 2010 Jul 8;5(7):e11471. http://dx.doi.org/10.1371/journal.pone.0011471
55. Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. Nat Methods. 2009 Nov;6(11 Suppl):S22–32. http://dx.doi.org/10.1038/nmeth.1371
56. Zhang Y. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137. http://dx.doi.org/10.1186/gb-2008-9-9-r137
57. Stovner EB, Sætrom P. epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics. 2019 Mar 28. http://dx.doi.org/10.1093/bioinformatics/btz232
58. Yang Y, Fear J, Hu J, Haecker I, Zhou L, Renne R, et al. Leveraging biological replicates to improve analysis in ChIP-seq experiments. Comput Struct Biotechnol J [Internet]. 2014 Jan 31 [cited 2019 Jun 16];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3962196/
59. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014 Dec 5;15(12):550. http://dx.doi.org/10.1186/s13059-014-0550-8
60. Singh AA, Schuurman K, Nevedomskaya E, Stelloo S, Linder S, Droog M, et al. Optimized ChIP-seq method facilitates transcription factor profiling in human tumors. Life Sci Alliance [Internet]. 2018 Dec 28 [cited 2019 Jun 19];2(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6311467/
61. Buenrostro J, Wu B, Chang H, Greenleaf W. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr Protoc Mol Biol. 2015 Jan 5;109:21.29.1–21.29.9. http://dx.doi.org/10.1002/0471142727.mb2129s109

62. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018 Oct 26;362(6413):eaav1898. http://dx.doi.org/10.1126/science.aav1898
63. Li YF, Altman RB. Systematic target function annotation of human transcription factors. BMC Biol. 2018 10;16(1):4. http://dx.doi.org/10.1186/s12915-017-0469-0
64. Gkountela S, Castro-Giner F, Szczerba BM, Vetter M, Landin J, Scherrer R, et al. Circulating tumor cell clustering shapes DNA methylation to enable metastasis seeding. Cell. 2019 Jan 10;176(1–2):98–112.e14. http://dx.doi.org/10.1016/j.cell.2018.11.046
65. Grosselin K, Durand A, Marsolier J, Poitou A, Marangoni E, Nemati F, et al. High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. Nat Genet. 2019 Jun;51(6):1060. http://dx.doi.org/10.1038/s41588-019-0424-9

8

Computational Approaches in Proteomics

Karla Cervantes Gracia[1] • Holger Husi[2,3]

[1]Basic Sciences Division, Universidad de Monterrey, San Pedro Garza García, N.L. Mexico; [2]Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK; [3]Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, UK

Author for correspondence: Holger Husi, Division of Biomedical Sciences, University of the Highlands and Islands, Centre for Health Science, Inverness IV2 3JH, UK.
Email: Holger.Husi@uhi.ac.uk

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch8

Abstract: Understanding of biological processes and aberrations in disease conditions has over the years moved away from the study of single molecules to a more holistic and all-encompassing view to investigate the entire spectrum of proteins. This method, termed proteomics, has been enabled principally by mass spectrometry techniques. The power of mass spectrometry-based proteomics lays in its ability to investigate an entire proteome and associated expression or modification states of a huge amount of proteins in one single experiment. This massive amount of data requires a high level of automation in data processing to render it into a reduced set of information that can be used to answer the initial hypotheses, explore the biology or contextualize molecular changes associated with a physiological attribute. This chapter gives an overview of the most common proteomic approaches, biological sample considerations and data acquisition methods, data processing, software solutions for the various steps and further functional analyses of biological data. This enables the comparison of various datasets as a summation of individual experiments, to cross-compare sample types and other metadata. There are many approach pipelines in existence that cover specialist disciplines and data analytics steps, and it is a certainty that many more data analysis methodologies will be generated over the coming years, but it also emphasizes the

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia.
ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

inherent place of proteomic technologies in research in elucidating the nature of biological processes and understanding of disease etiology.

Keywords: data analysis; mass spectrometry; proteomics; software; workflows

INTRODUCTION

The development and improvement of high-throughput techniques in "omic" science have paved the way not only to a broader view of the molecules involved in a specific condition but also to generate networks of all interacting elements (genes, proteins, and metabolites) to gain a better understanding of how a specific biological system works. Despite the over-abundance of genomics research in this field, there is so much more complexity left out in a system that can be explained by the understanding and integration of proteomic data. The proteome is more complex and is not as stable as the genome, and it is not only based on what is observed in the genome but also influenced by several factors. Protein expression depends on tissue type, environmental stimuli, and post-translational modifications (PTM) that influence its level of activity, structure, function and regulation (1, 2). Moreover, life depends on proteins, as they are responsible for many complex processes within a cell, from replication, gene transcription and translation to cellular senescence and death. Therefore, by having a better understanding of the proteome, a wider comprehension of cellular regulation can be achieved. Proteomics is the high-throughput study of proteins incorporating the identification, quantitation, analysis and comparison of differential expression of proteins from samples under specific biological conditions. The characterization of the proteome involves the identification of structure, function, interactions and modifications (3).

Because of its improved sensitivity and specificity, mass spectrometry (MS) proteomics is the most widely used approach, and it is considered the method of choice to obtain global measurements of proteins (4). The most common and classic applications of proteomics are to characterize large datasets to create an inventory of identified proteins in different tissue or cellular samples, as well as to generate lists of differentially expressed proteins from samples under specific conditions (5). However, these data alone lack a biological meaning, and therefore, it is essential to pursue additional approaches to allow a better interpretation of biological processes (6, 7).

Qualitative and quantitative methods are also of importance in network analyses. Qualitative approaches are much more common. Although quantitative network analysis can generate more specialized results and are better adapted to generate new insights and advancement in biomedical research by unraveling the significant proteins that interplay in a disease, producing new diagnostic hypothesis, the standardization and homogenization of its analysis still need improvement to establish its reliability and reproducibility when analyzing high-throughput data (8, 9).

High-throughput technologies and bioinformatic tools are fundamental for proteomics data interpretation to discover new biological insights on cellular processes, disease etiology and biomarker candidates. Although these tools are under

continuous update and new approaches are implemented, the development of harmonized benchmarks for datasets and analysis, as well as to establish gold-standard workflows, is imperative to produce more reliable and reproducible results, and by doing so, it will help to overcome the challenges of proteomics data interpretation (10, 11).

The acquisition of a vast amount of high-quality raw spectra using MS is nowadays a relatively simple task with the right equipment and involves a high level of automation, which however leads to a fundamental, and crucial, step to mathematically and statistically interrogate the data and ultimately match it to a library of known or hypothetical molecules. This is of particular importance in strategies such as shotgun proteomics and other large-scale MS screens, whereas specific applications such as selective or multiple reaction monitoring (SRM/MRM) have a different requirement for the entire workflow and require appropriate specific software solutions (12). Figure 1 shows a general overview of a typical proteomics workflow, starting from protein and peptide preparation from tissues to the computational procedures to obtain a list of molecules with associated confidence or significance scores that can then be analyzed further.

SAMPLE TYPES AND SAMPLE PROCESSING APPROACHES

Body homeostasis is maintained through specialized systems, which are orchestrated by the interplay between cells, tissues and organs. Each system anomaly can be better described by specific samples; therefore, to characterize diseases, it becomes essential to analyze the proteome of the appropriate samples. Many sample types are suitable for proteomic analysis, including cells, organs, tissues and body fluids. Biomarker discovery helps to identify pathological states, track disease progression and improve diagnostics or disease etiology, which are some of the common applications when using these sample types (13). An important

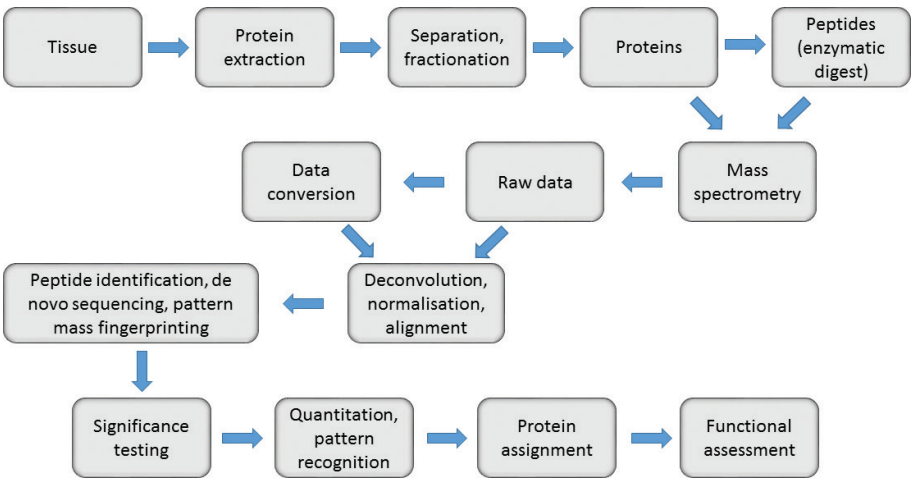


Figure 1 Flowchart and procedures for a generic proteomics pipeline.

factor to be considered for the success of a proteomic approach is the quality and quantity of the sample, due to the challenge that its complexity implies for MS techniques. As the detection rates of proteins using MS are directly related to the absolute quantity of these biomolecules in a sample, high-abundance proteins tolerate losses during processing quite well. However, the detection rates of low-abundance proteins are usually much more sensitive towards loss due to common instrumentation and processes, and therefore, the preferred workflow is microproteomics, which minimizes this loss and increases the sample processing efficiency (14). Samples, such as the ones derived from cancer or tumour cells, exemplify low-abundance protein samples. They should be analyzed by specific microproteomic workflows with special adaptations of the techniques for sample preparation, cleaning, fractionation and separation to ensure minimal losses before analysis and increase sensitivity of nano/microgram-samples that allows maximal identification of low-abundance proteins (15, 16).

Cell lines

A general overview of disadvantages and applications of each sample-source type is presented in Table 1. Although heterogeneous cell populations that compose tissues can be individually isolated and analyzed, cell lines are believed to reflect the protein composition of primary cells and specific tissues. Moreover, the reproducibility in proteomic analysis using cell lines is one of its main advantages over other sample types. It also allows proteomic subcellular analysis (17). Several applications of proteomic analyses using cell lines have been established to investigate molecular pathways of specified cell types, differences between normal and disease phenotypes, and different stages of diseases (18, 19). However, problems in cell line culturing are rather common if no quality control is carried out and can lead to unreliable results if not detected and treated. The most common problems with cell lines are genomic instability, infections by microorganisms that could alter cell turnover and protein expression patterns and cross-contamination leading to the growth of a mixture of cell types, affecting the results of the study even before proteomic analysis can be performed (20).

Tissue culture

Proteomic analysis with tissue culture as a sample is also a very informative approach. It allows the interaction and analysis of the diverse cell types involved in a disease, leading to a broader view of the biological systems of importance in pathology. It is basically based on the growth of tissue outside the organism, under controlled conditions. Tissue samples for this are obtained through surgery from humans or animals. Tissue culture-based proteomic profiling is useful for understanding the biological mechanisms underlying a disease, biomarker and therapeutic target identification as well as effects in a sample due to viral, drug or genetic changes (21). Techniques, such as 3D co-culture systems, fresh tissue proteomics and tumour spheroid models, have improved the analysis and results (22). However, its accessibility remains as its major downside, and no accurate track in disease progression can be performed without re-sampling.

Organ

Organ samples can be maintained under specific culture conditions, and its different cell types can be analyzed. It is the most difficult sample to obtain from humans, and since biofluids are secreted from several organs and make proteins more accessible, they are the sample of choice for biomarker discovery and pathology research (23, 24). However, reliability and reproducibility are still issues to be addressed, before they can be eventually established as a good source of clinical proteomics. Like the tissue samples, animal models serve as a good source of organ samples. They provide a controlled environment and the possibility to follow up the changes in proteomic profiling throughout the course of a disease. The major drawback using animal models is that they cannot accurately predict how a system works in humans (25). However, in order to overcome this issue and to have a better and broader understanding of the interaction of human proteome within a system, new engineered model systems have been created, such as multiorgan lab-on-a-chip platforms, that show a better correlation with human systems than animal models, mimicking the key aspects of responses like drug treatment (26, 27).

Exosomes

Besides the analysis of the proteome in cells, proteins secreted by the cells have gained attention when unraveling the etiology of diseases. All together, these proteins are known as secretome, and a specific component of the secretome that has been studied in relation to pathology is the exosome. Exosomes are membrane vesicles, differentiated from other vesicles by size and expression of the CD81 protein. They have a very low abundance of proteins, which is undetectable using biofluid analysis (28). Among these proteins are some that are specific to the biological fluid or cell, making the exosomes an interesting source for biomarkers to advance the identification and understanding of pathologies (29).

Biological fluids

Depending on the purpose of the research, diverse body fluids can be collected and processed for proteomic analysis. A general overview of disadvantages and applications of each sample-source type is presented in Table 1. Among the commonly analyzed biological fluids in proteomics are blood, serum, plasma, cerebrospinal fluid (CSF), urine, saliva and semen. The fluctuation in their protein levels is expected to reflect pathophysiological conditions; however, some drawbacks such as protein content, high abundance of masking proteins, and sample instability can lead to complex interpretations (30). Blood, serum and plasma are the most common biological fluids in proteomic research due to its non-invasive nature and its high concentration of protein/peptides, as well as the assumption that blood reflects the pathophysiological state of several organs. Biofluids such as urine and CSF are not the most desirable samples for proteomics because they contain a lower protein/peptide concentration (31). In addition, a complicated collection process is a hindrance in obtaining a reasonable amount of CSF sample (31).

TABLE 1 Biological fluids overview: Applications and disadvantages

Biological fluid	Applications	Disadvantages
Serum and plasma	Serum and plasma have been used for multiple proteomics–based biomarker discovery studies.	Dynamic qualitative and quantitative range of proteins; small number of highly abundant proteins can mask potential biomarkers; biomarker of interest can be lost upon the removal of highly abundant proteins.
Cerebrospinal fluid (CSF)	Potential diagnostic utility in neurodegenerative diseases including Alzheimer's, multiple sclerosis and Parkinson's.	Requires a lumbar puncture or a spinal tap, invasive procedures. Traumatic punctures can alter CSF protein expression levels and skew a diagnosis; small volumes of samples obtained; yield a highly dynamic range of protein concentrations; small number of highly abundant proteins can mask potential biomarkers; depletion techniques are neither time nor cost-effective techniques; biomarker of interest can be lost upon the removal of highly abundant proteins.
Urine	Good source of biomarkers for urogenital and systemic diseases.	Definition of disease-specific biomarkers is complicated; significant changes in the proteome throughout the day can be connected with the time of collection, fluid intake, diet, exercise, circadian rhythms and circulatory levels of various hormones; presence of MS hampering salts; lower concentration of proteins/peptides compared to serum and plasma.
Saliva	Most of the biomolecules that are usually detected in urine and blood can also be found in salivary secretions; about 30% of blood proteins are also present in saliva.	Very low concentration of proteins; very rapid protein degradation in whole saliva at room temperature, this may occur during saliva collection and handling.
Semen	Applications in research areas such as reproduction and prostate cancer, and used for many purposes in the diagnosis of male fertility.	Small number of highly abundant proteins can mask potential biomarkers; biomarker of interest can be lost upon the removal of high abundant proteins.
Circulating tumour cells (CTC)	Practical application in diagnosis and disease treatment, determine the prognosis of metastatic progression or relapse, monitor anti-cancer treatments, understand the mechanism of metastatic disease and develop new strategies in disease treatment.	Very low abundance of CTC in blood; cell heterogeneity makes it difficult to isolate the whole CTC population.

Although the technicalities of sample collection, management and storage are known to be of vital importance to keep the composition and quality of the sample to be reproducible and reliable, there is no commonly accepted standardization protocol for bio-sampling procedures. Variables, such as storage times; temperatures and number of freeze-thaw cycles; removal of additives, such as heparin to prevent clotting; as well as the consumables, such as collection and processing tubes, are important parameters to be considered in order to avoid differences in protein composition among samples (31, 32). Bio-sampling optimization and standardization are essential steps to improve reproducibility for accurate correlations among different studies (33).

DATA ACQUISITION

Proteomics has become a feasible and a promising approach with the advancements in MS methodologies. MS/MS innovations and possible combinations are constantly under improvement, and nowadays, it has become the gold standard for any kind of proteomic studies. Furthermore, high-resolution mass spectrometers have been recently adapted for high-throughput proteomics (34). MS/MS has a high impact in lowering sample complexity by the isolation of precursor ions through a mass filter, as well as their fragmentation and further detection by high-resolution mass analyzers (35). Moreover, for each of these steps, technologies have been developed to identify and distinguish peptides more accurately, with a better resolution, coverage and reproducibility. Also, computer tools have been under constant development to improve the analysis of the complex outcome data (Table 2). In order to achieve a more accurate protein identification, three main approaches have been described: bottom–up (BU), top–down (TD) and, more recently, middle–down (MD) (Figure 2).

Bottom–up data analysis

In contrast to TD and MD proteomics analysis, for BU data analysis, a deconvolution step is not required when implementing ESI, due to the rare generation of double- and triple-charged fragment ions (36). Mass spectra raw data are commonly processed by Proteome Discoverer or MaxQuant platforms using several search engines, such as Sequest, Mascot, Andromeda, X!Tandem and COMET, usually against UniProt databases (37–39). MaxQuant software can also determine protein quantitation and estimate the error of PTM false localization. For downstream correlation and clustering analysis, the identified proteins are commonly processed in the Perseus platform (38, 40, 41). To reduce data complexity, principal component analysis (PCA) has been the method of choice, and also to identify the relatedness of the differentially expressed proteins within and among samples (39, 41). Moreover, to interpret the potential function of the datasets obtained, the DAVID platform is commonly used to enrich them with Gene Ontology terms, KEGG pathway information and InterPro protein domains (39, 42). Additionally, constructed networks are commonly visualized in Cytoscape, and in order to identify functional and physical associations among mRNA and protein data, the STRING database is used (43). All MS data are usually deposited

TABLE 2 Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses

Name	Scope	URL
Reference databases		
UniProt/SwissProt	Proteome assemblies	https://www.uniprot.org
RefSEQ	Genome, transcriptome and proteome assemblies	https://www.ncbi.nlm.nih.gov/refseq/
Tandem mass spectra protein/peptide search engines		
SEQUEST, Comet	Cross-correlation-based scoring, commercial (SEQUEST), open source, free (Comet)	https://proteomicsresource.washington.edu/protocols06/sequest.php
Mascot	Probability-based scoring, commercial	http://www.matrixscience.com/
X!Tandem	Statistical confidence (expectation value) scoring, pattern matching, open source	https://www.thegpm.org/tandem/
Andromeda	Probabilistic scoring model, open source	http://coxdocs.org/doku.php?id=maxquant:andromeda:start
ProLuCID	Three-tier scoring system, binomial probability, cross-correlation calculated Z-score, open source	http://fields.scripps.edu/yates/wp/?page_id=17
Platforms, integrated pipelines		
Proteome Discoverer	Commercial, fully integrated solution for Thermo Fisher instruments, outputs protein/peptide lists, also works with other data formats	https://planetorbitrap.com/proteome-discoverer
Progenesis QI	Commercial, integrated solution from Waters, outputs protein/peptide lists, also works with other data formats	http://www.nonlinear.com/progenesis/qi-for-proteomics/
ProteinPilot	Commercial, fully integrated solution for AB Sciex instruments, outputs protein/peptide lists, also works with other data formats	https://sciex.com/products/software/proteinpilot-software
Integrated Proteomics Pipeline (IP2)	Commercial, uses the ProLuCID search engine, outputs protein/peptide lists, offers limited downstream analysis	http://www.integratedproteomics.com/products.html
Trans-Proteomic Pipeline (TPP)	Modular open-source standardized data processing pipeline	http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
PatternLab	Open-source raw MS file data processing, includes the Comet peptide search engine, Microsoft Windows specific	http://patternlabforproteomics.org/index.html
MaxQuant	Open-source data processing workflow, includes Anromeda peptide search engine, Linux and Windows distributions	https://www.maxquant.org/
Skyline	Modular open-source standardized data processing pipeline, workflow editor, Microsoft Windows specific	https://skyline.ms/project/home/software/Skyline/begin.view
OpenMS	Modular open-source standardized data processing pipeline, workflow editor, OS system independent	https://www.openms.de/
Taverna	Framework for biocomputational tools, workflow editor, open source, re-use of existing workflows, Java programming language based, web server application	https://taverna.incubator.apache.org/
Galaxy	Framework for biocomputational tools, workflow editor, open source, web server application	https://usegalaxy.org/
Data repositories		
PRoteomicsIDEntification database (PRIDE)	Protein and peptide identifications, post-translational modifications, raw data	http://www.ebi.ac.uk/pride/
PeptideAtlas	Library of identified peptides	http://www.peptideatlas.org
Japan ProteOmeSTandard Repository/Database (jPOST)	Integrated proteome datasets	https://jpostdb.org/
MassIVE	Protein and peptide identifications	https://massive.ucsd.edu/ProteoSAFe/static/massive.jsp
iProX	Proteomic datasets	https://www.iprox.org/

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
Panorama Public	Skyline processed data	https://panoramaweb.org/project/Panorama%20Public/begin.view?
Open Proteomics Database (OPD)	Proteomic datasets	http://data.marcottelab.org/MSdata/OPD/
The Global Proteome Machine (GPM)	Metadata	http://gpmdb.thegpm.org/
Functional analysis resources		
Bioconductor/R	Statistical and graphical environment for analysis of high-throughput data	https://www.bioconductor.org/
MixOmics	Statistical package, visualization of analysis runs, requires the statistical software R	http://mixomics.org/
PANDA-view	Statistical analysis, data visualization, quantitative proteomics, requires the statistical software R but provides its own GUI, Microsoft Windows application	https://sourceforge.net/projects/panda-view/
Perseus	Post-analysis of MaxQuant data, data integration, statistical analysis, sample comparisons, Microsoft Windows application	http://coxdocs.org/doku.php?id=perseus:start
InterPro	Protein families, domains and functional sites,	http://www.ebi.ac.uk/interpro/
Gene Ontology (GO)	Hierarchically clustered annotations of functional terms that describe the biological process, molecular function or cellular component	http://www.geneontology.org
Database for Annotation, Visualisation, and Integrated Discovery (DAVID)	GO analysis, KEGG mapping, domain grouping, web application	https://david.ncifcrf.gov/

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
ClueGO	GO analysis, statistical analysis, hierarchical clustering, Cytoscape app	http://www.ici.upmc.fr/cluego/
ReviGO	GO term clustering, semantic analysis	http://revigo.irb.hr/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Pathway mapping, static maps	http://www.genome.jp/kegg/
ReactomeKnowledgeBase	Pathway mapping	http://www.reactome.org
Ingenuity Pathway Knowledge Base (IPA)	Pathway mapping, data clustering, enrichment analysis	https://www.qiagenbioinformatics.com/products/ingenuity-pathway-analysis/
Wikipathways	Database of pathway maps	http://wikipathways.org/index.php/WikiPathways
BioCyc Knowledge Library	Database, pathway mapping	http://biocyc.org/
RHEA	Biochemical reactions database	http://www.rhea-db.org
Protein ANalysisTHrough Evolutionary Relationships (PANTHER)	Signaling pathways	http://www.pantherdb.org
PathVisio	Pathway drawing and pathway analysis tool	https://www.pathvisio.org/
IMPaLA	Pathway analysis, web application	http://impala.molgen.mpg.de/
Molecular Interaction Database (IntAct)	Database, protein-protein interactions, protein-compound interactions	http://www.ebi.ac.uk/intact
Molecular Interaction Database (MINT)	Protein–protein interactions	http://mint.bio.uniroma2.it/

Table continued on following page

TABLE 2 **Data resources and typical software solutions used in MS protein assignments, proteomics workflows, downstream analysis, data repositories and functional analyses (Continued)**

Name	Scope	URL
Search Tool for the Retrieval of Interacting Genes/Proteins (STRING)	Protein–protein interactions	http://string-db.org
JasparDB	Transcription factors and regulatory sites	http://api.bioinfo.no/wsdl/JasparDB.wsdl
Online Mendelian Inheritance in Man (OMIM)	Online catalogue of human genes and genetic disorders	https://www.omim.org/
DisGeNET	Human gene-disease associations, animal models, database, web-query, Ctoscape implementation	http://www.disgenet.org/
Babelomics	Correlations, clustering, ontologies, pathways, heatmaps	http://www.babelomics.org/
Cytoscape	Network analysis environment, graph drawing	https://cytoscape.org/

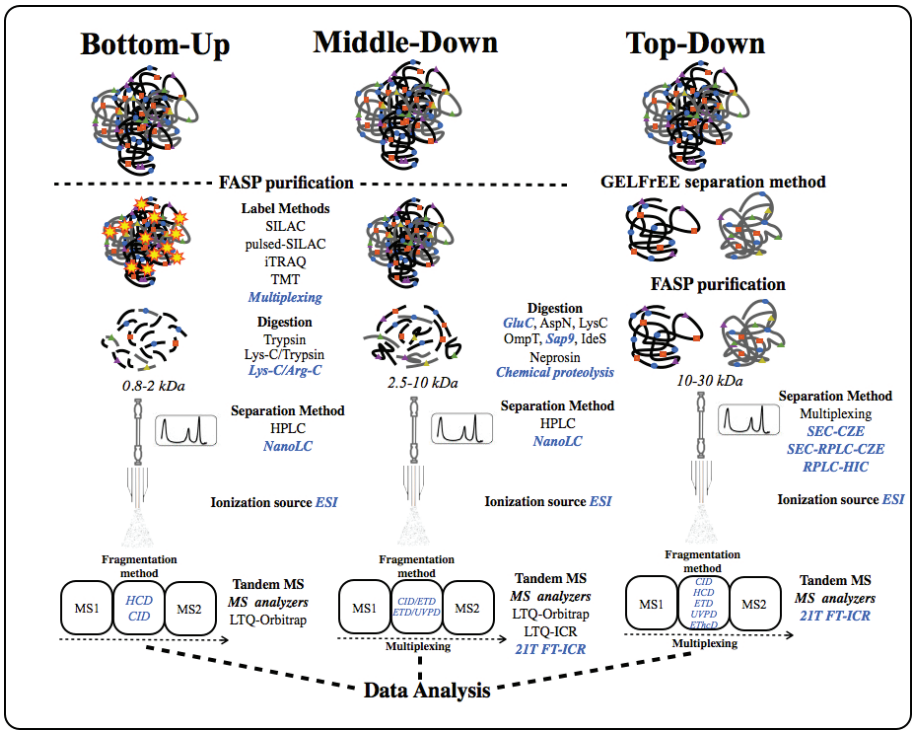


Figure 2 Bottom–up, middle–down and top–down proteomic high-throughput approaches. A general view of each of the approaches and the essential steps to follow are shown from top to bottom. Up-to-date tools, methodologies and techniques most commonly and successfully applied for high-throughput proteomic analyses are highlighted in blue for each of the approaches.

in the ProteomeXchange Consortium via the PRIDE partner repository for sharing, general availability and further study (44, 45).

Top–down data analysis

In TD data analysis, Proteome Discoverer is commonly used to process raw data files, and through its ProSight tools, as well as through MascotTD, identification and characterization of intact proteins can be achieved (46, 47). A database search using ProSight against specific databases (UniProt, SwissProt and RefSEQ) leads to top–down data interpretation and also identifies PTMs within a protein sample (48, 49). Furthermore, deconvolution is crucial for data interpretation, and it is commonly achieved through Xtract, MS-Deconv and YADA (within ProLuCID), among other tools (50). Additionally, in order to give meaning to the identified intact proteins/proteoforms and analyze them more deeply, an integrated network approach can be followed. As an example, "Proteoform" Suite has been recently used for dataset identification and proteoform integration. By assessing its function using gene ontology (GO) analysis, it also enables the visualization of

association and abundance within networks through Cytoscape (51, 52). Although top–down proteomics is still rapidly evolving, the complexity of the analysis and technological issues remain, preventing it to be a typical method to follow when studying high-throughput PTMs.

Middle–down data analysis

MD approaches are based mainly on ESI, where multiple peaks of charged fragment ions are generated. Therefore, it is essential to perform deconvolution prior to MS spectra interpretation. Several tools have been described for this purpose, such as Xtract and YADA (within ProLuCID) or Proteome Discoverer (45, 53). The subsequent dataset analysis and database searches are usually performed using Mascot or Sequest (44, 45, 53). Moreover, new software tools are under development to filter Mascot and Sequest results, such as isoScale, where Mascot results are imported and peptides with confidently assigned combinatorial PTMs are identified, which means that all the modifications are uniquely validated by ions that determine and confirm the localization of a PTM site (45, 53, 54). Since MD proteomics research has a considerable impact on PTM research, specific software tools have been created to analyze PTMs and relevant data. Among these tools are the previously mentioned isoScale software and the Skyline software (55). MD proteomics is still lacking established and standardized tools suitable for data interpretation, and although algorithms and software tools remain under constant development and improvement, this issue is mainly overcome by using TD proteomics tools instead. However, due to a different focus (no proteolytic peptides), such MD analyses are prone to error.

DATA HANDLING AND WORKFLOWS

The general process of data analysis, shown in Figure 3, involves procedures of raw data conversion, deconvolution, normalization, spectral identification, peak alignments, validation, statistical modeling, peptide identification, abundance measurements, protein inference, data storage (raw and processed), data visualization, eventual further data analysis steps such as dataset comparisons and ultimately deposition of data into public data repositories.

Data processing software

A vast amount of computational solutions have been developed to handle and analyze proteomic MS data, ranging in thousands of applications, add-ons and scripts, covering every single aspect of data conversion, deconvolution, normalization and alignment, as listed in website (https://omictools.com/proteomics-category). Currently, the main problem is to find the most appropriate and suitable analysis tool rather than to find a way to analyze the experimental MS spectra. A good overview of the software landscape of such tools can be found in Ref. (56), which also poignantly describes the incompatibility issues when faced with such

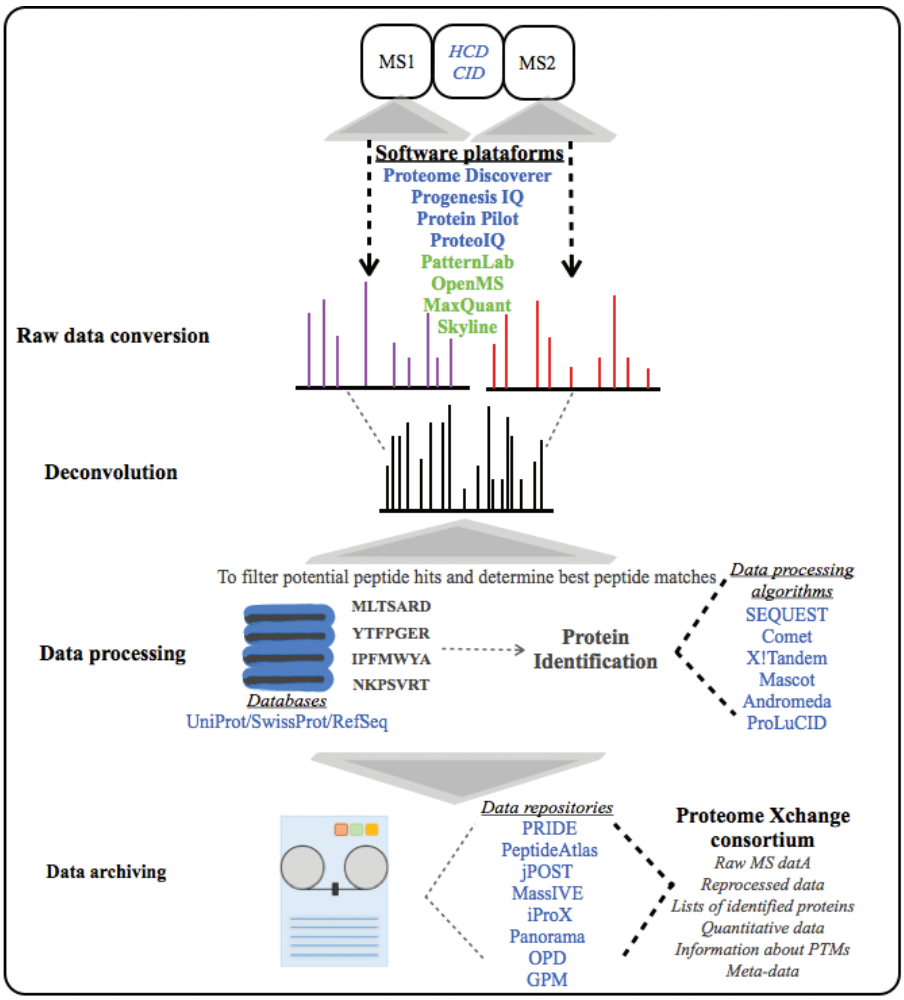


Figure 3 Data processing workflow from raw MS spectra to identified biomolecules. Raw data conversion, deconvolution, data processing and data archiving are the main steps illustrated. The most popular tools within each of them are highlighted in blue. Commercial (blue) and open-source integrated software platforms (green) for the analysis of proteomic data are included. They all encompass modules to manage raw spectral file data, peptide identification using search engines, clustering and sample comparison, identification of PTMs, quantification, statistical analysis and visualization tools. The most common data processing algorithms, data bases and data repositories to release data into the public domain are also shown.

a huge array of computational tools. Therefore, a focused view of the most general, yet commonly used software solutions is summarized in Table 2.

Integrated pipelines

The exuberance of programs and applied algorithms in data processing led to a fragmented landscape of often incompatible steps needed to perform MS data analysis, and the obvious solution was to integrate these various steps into one single workstream implemented in platform tools. A considerable amount of reviews of existing platform software programs are available (57–59). Common amongst many platform solutions is that they usually have one or more of the aforementioned protein/peptide search engines embedded in the workflow. All major MS system manufacturers also provide integrated software solutions for the analysis of proteomic data and specific applications, thereby eliminating the need of having separate software solutions for data acquisition and data processing; however, it needs to be noted that MS instrument control might still require vendor-specific applications. As a consequence, data formats of raw MS data are specific for the manufacturer of the MS equipment, and inter-operability of software solutions is severely hampered and sometimes impossible. This lock-in has understandable commercial reasons, but quite a number of open-source solutions have also been made available over the years. One of the main differences between commercial and open-source solutions is user friendliness, where open-source programs might require specialist computing skills in order to implement the various components of the software programs. However, in recent developments, more user-friendly platforms have been generated that integrate these open-source solutions or algorithms. Therefore, most of these open-source applications feature a modular design, where individual algorithms and procedures are combined to form the entire workflow.

DATA INTERPRETATION AND FUNCTIONAL ANALYSIS

One of the key aspects in proteomic research is the downstream analysis, whereby lists of molecules are interrogated using a variety of software tools in order to put biological meaning into such lists, extract statistically evaluated parameters or match them against other known assemblies (Figure 4). These steps generally involve the use of other databases that hold specific information for each molecule, such as functionality, disease association or pathway data. More than 300 software tools to accomplish various aspects are listed at this website (https://www.ms-utils.org/) alone, and thousands more have been developed and used in proteomics research over the last 20 years. Table 2 lists some of the most common tools used in proteomic downstream analysis.

Statistical approaches

The large-scale nature of proteomic data, which reflects not only the biological factors but also the technical and experimental factors, often requires algorithms to reduce the dimensionality. Statistical tools are an essential part in the analysis of such data, ranging from outlier detection methods and imputation of missing data to expression profiling and group comparisons, including networks and

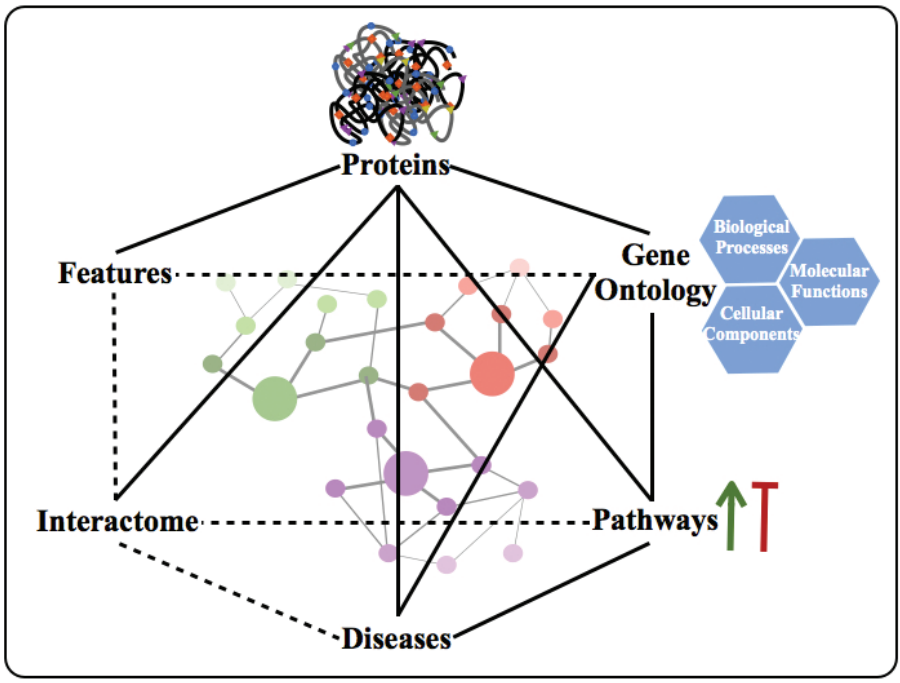


Figure 4 Downstream data analysis in proteomics research and relationships of analysis scopes. Full lines depict direct information flow between the analysis or data types, and dashed lines depict indirect associations.

protein cluster detection (60). A vast number of these procedures have been implemented as scripts in the statistical open-source tool R, or in one of its derivatives such as Bioconductor, where a number of packages were written specifically for use in proteomics applications and data analysis (61). A basic first step in the analysis of large MS-derived datasets can also involve a possible enrichment of specific protein families or domains. The InterPro database is an integrated documentation resource for protein domains, families and functional sites, incorporating other databases with similar scope, namely ProDom, PROSITE, PRINTS and Pfam (62). Analysis of the protein landscape using the InterPro resource is generally a practical and efficient way to interrogate proteomic datasets.

Gene ontology

One of the most prominent and heavily used data resource for downstream analysis is the GO database, whose aim is to generate a dynamic, yet controlled vocabulary that can be used in all eukaryotes as the knowledge of gene and protein roles in cells is growing and constantly changing (63). The database unifies similar prior approaches from other databases and describes molecules in terms of their involvement in biological processes, their molecular function and their sub-cellular location in a hierarchical way. Originally, this process of annotating functionality

tags to molecules was done manually, but nowadays it is performed mainly through computational tools. There are many software solutions that make use of the GO database, and surprisingly, depending on the algorithm used in GO-analysis, the results can vary drastically (64). Therefore, one needs to carefully evaluate which tools to use and which are trustworthy in their analysis outcomes.

Pathway analysis

Other high-quality, manually curated databases that extend the knowledge of molecular functionalities are comprised of pathway databases, and more than 600 databases within this scope are currently listed at this website (http://pathguide.org). Pathways, such as signaling and metabolic cascades, can be used to physically link proteins in a concatenated manner to a series of events with a measurable outcome, thereby reducing the complexity of the protein-centric view to a more meaningful one through identification of functional biological processes (65). They can also be used to bridge or integrate data from one omics stream such as proteomics and another like metabolomics. Additionally, many signaling cascades, in particular gene-activation pathways, terminate at the point where gene expression is induced or repressed, thereby breaking the information flow from one signaling event to another via an intermediary step of gene modulation. In order to fill this gap, it is necessary to identify potential transcription factors, their DNA binding sites and the targeted genes (66). Such information can be used for both down-stream pathway mapping and up-stream analysis, thereby enabling the exploration of causes leading to the observed proteomic profile changes, as well as the consequences of such changes.

Interactomes

An additional aspect to consider is that most proteins do not act alone and independently, but rather as an assembly of multiple proteins to perform specific actions by forming transient or stable complexes. Examples are scaffolders that bring proteins into close proximity in protein signaling cascades, protein regulatory networks and structural components. Based on the composition of such complexes, a specific protein might be involved in a function that is fundamentally different from the same molecule participating in an assemblage with other proteins. Therefore, in order to gain a better understanding of the biological data from MS-derived experimentation, the use of protein–protein interaction databases can be particularly helpful (67). Most protein–protein interaction databases contain literature-based interaction data that were manually curated and assessed, whereas some resources use literature mining tools to populate the database, and their data are therefore not necessarily based on experimental observations, but rather predicted interactions.

Disease mapping

Further contextualization of proteomic data can also be achieved by interrogating disease databases, where disease terms are linked to a collection of associated genes derived from the literature. Two such examples are the Online Mendelian

Inheritance in Man (OMIM) (68) and DisGeNET (69). Both are expert-curated databases that analyze text-mined data to establish a link between phenotypes and genes and both have their own web interface to query the databases. While OMIM and its derivative table of gene-disease associations termed MorbidMap are only covering human genes and disease conditions, DisGeNET additionally includes data from animal disease models. Although they are comparable in scope, they both do not use the exact same medical term dictionary, which can cause problems comparing and fusing results using both databases

Integrated frameworks

The diverse nature of biological questions to be answered by proteomics can make it difficult for non-experts in data analysis to make the right selection of analysis tools, and together with specific requirements such as R scripting or programming skills, it can become a daunting endeavor. Yet, new tools are emerging that bring together various data downstream processing procedures most commonly used in omics research such as Babelomics and Cytoscape that will help researchers to put meaning into large-scale datasets, and some of the tools described before have also been integrated into these software solutions as well. Babelomics, although in principle more useful in gene and array analysis, can also be used in the functional characterization of proteomic datasets and other downstream analysis steps (70). It includes a comprehensive suite of modules to perform differential expression profiling, enrichment analysis, GO and pathway analysis, text mining and protein interaction analysis. It is implemented as a web-based application and is freely available and accessible. Cytoscape is an open-source and freely available software framework for interaction network analysis and is offered as a desktop application or a web-plugin (71). In itself, it provides basic functionalities such as graph drawing and network layout and construction and enables linking to large databases. It is extendable by providing a run-time environment for other data analysis plug-ins. Currently, approximately 350 additional apps are available.

CONCLUSION

MS, and in particular the LC-MS/MS shotgun proteomics workflow, is widely used to identify and quantify sample peptides and proteins. The methodology, however, still poses several challenges for large-scale use, such as the MS-manufacturer dependent diverse raw data file formats, the relatively large false-positive peptide assignment rate and the disconnect between observed peptides and originating sample proteins. There are still quite a number of issues to be resolved concerning proteomics in general, such as missing data or data depth, where the sensitivity of the mass spectrometer is insufficient to reliably detect low-abundance molecules, or where the very nature of the molecules under investigation prohibits certain applications, which is commonly encountered with transmembrane spanning proteins. Problems that arise due to masking effects, particularly encountered with high-abundance molecules that raise the detection threshold, are more of a technical issue that can be overcome with improvement

of methodologies, whereas database drift, which is associated with underlying reference databases where accession numbers are lost over time due to various reasons, can pose real problems in the long term.

While many elegant software solutions of data acquisition to spectral data analysis exist, the field is rather fragmented and disjointed when it comes to downstream data analysis such as integrating or merging results derived from pathway mapping, terminology clustering and disease analysis. Yet, tremendous efforts have already begun to pay off in collating and merging individual applications and algorithms into a more cohesive framework. One such framework, the Pan-omics Analysis Database (PADB) initiative, has been in existence for more than 15 years and has been successfully used to address proteomic and genomic large-scale data analysis in various disease areas (72). Another obvious solution is the reuse of existing pipelines and workflows generated in other omics-streams, in particular from the genomics and transcriptomics fields. These tools can be helpful in many ways in proteomics data analysis, yet they might also confuse the picture of available tools and analysis workstreams.

Nevertheless, it is very apparent that since proteomics entered the mainstream and has become an accepted standard in large-scale biological investigations, many breakthroughs were achieved that were unthinkable before. A very new view of the small-scale world has opened and, although the most obvious impact at that moment was how little we understand in terms of molecular flux and interplay, enabled us to start interrogating biological processes on an unprecedented scale. In particular, disease analysis, understanding of abnormal phenotypes and how to pharmacologically interfere with the protein landscape at various stages of disease progression, has started to bear fruit and will continue to do so in the foreseeable future.

Conflict of interest: The authors declare that they have no conflicts of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: We confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s). All original sources have been appropriately acknowledged and/or referenced.

REFERENCES

1. Krishna RG, Wold F. Post-translational modifications of proteins. In: Imahori K, Sakiyama F, editors. Methods in protein sequence analysis. Boston, MA: Springer US; 1993. p. 167–72.
2. Pandey A, Mann M. Proteomics to study genes and genomes. Nature. 2000 Jun 15;405(6788):837–46. http://dx.doi.org/10.1038/35015709
3. Boersema PJ, Kahraman A, Picotti P. Proteomics beyond large-scale protein expression analysis. Curr Opin Biotechnol. 2015 Aug;34:162–70. http://dx.doi.org/10.1016/j.copbio.2015.01.005
4. Aslam B, Basit M, Nisar MA, Khurshid M, Rasool MH. Proteomics: Technologies and their applications. J Chromatogr Sci. 2017 Feb;55(2):182–96. http://dx.doi.org/10.1093/chromsci/bmw167
5. Nilsson T, Mann M, Aebersold R, Yates JR, Bairoch A, Bergeron JJM. Mass spectrometry in high-throughput proteomics: Ready for the big time. Nat Methods. 2010 Sep;7(9):681–5. http://dx.doi.org/10.1038/nmeth0910-681

6. Moore JB, Weeks ME. Proteomics and systems biology: Current and future applications in the nutritional sciences. Adv Nutr. 2011 Jul;2(4):355–64. http://dx.doi.org/10.3945/an.111.000554
7. Carnielli CM, Winck FV, PaesLeme AF. Functional annotation and biological interpretation of proteomics data. Biochim Biophys Acta. 2015 Jan;1854(1):46–54. http://dx.doi.org/10.1016/j.bbapap.2014.10.019
8. Goh WWB, Wong L. Design principles for clinical network-based proteomics. Drug Discov Today. 2016 Jul;21(7):1130–8. http://dx.doi.org/10.1016/j.drudis.2016.05.013
9. Egertson JD, Kuehn A, Merrihew GE, Bateman NW, MacLean BX, Ting YS, et al. Multiplexed MS/MS for improved data-independent acquisition. Nat Methods. 2013 Aug;10(8):744–6. http://dx.doi.org/10.1038/nmeth.2528
10. Xu J, Wang L, Li J. Biological network module-based model for the analysis of differential expression in shotgun proteomics. J Proteome Res. 2014 Dec 5;13(12):5743–50. http://dx.doi.org/10.1021/pr5007203
11. Allmer J. A call for benchmark data in mass spectrometry-based proteomics. J Integr OMICS. 2012;2(2):1–5. http://dx.doi.org/10.5584/jiomics.v2i2.113
12. Colangelo CM, Chung L, Bruce C, Cheung K-H. Review of software tools for design and analysis of large scale MRM proteomic datasets. Methods. 2013 Jun 15;61(3):287–98. http://dx.doi.org/10.1016/j.ymeth.2013.05.004
13. Husi H, Albalat A. Proteomics. In: Padmanabhan S, editor. Handbook of pharmacogenomics and stratified medicine. 1st ed. London: Academic Press; 2014. p. 147–79. http://dx.doi.org/10.1016/B978-0-12-386882-4.00009-8
14. Gutstein HB, Morris JS, Annangudi SP, Sweedler JV. Microproteomics: Analysis of protein diversity in small samples. Mass Spectrom Rev. 2008 Jul–Aug;27(4):316–30. http://dx.doi.org/10.1002/mas.20161
15. Thakur D, Rejtar T, Wang D, Bones J, Cha S, Clodfelder-Miller B, et al. Microproteomic analysis of 10,000 laser captured microdissected breast tumor cells using short-range sodium dodecyl sulfate-polyacrylamide gel electrophoresis and porous layer open tubular liquid chromatography tandem mass spectrometry. J Chromatogr A. 2011 Nov 11;1218(45):8168–74. http://dx.doi.org/10.1016/j.chroma.2011.09.022
16. Feist P, Hummon A. Proteomic challenges: Sample preparation techniques for microgram-quantity protein analysis from biological samples. Int J Mol Sci. 2015 Feb 5;16(2):3537–63. http://dx.doi.org/10.3390/ijms16023537
17. Drissi R, Dubois M-L, Boisvert F-M. Proteomics methods for subcellular proteome analysis. FEBS J. 2013 Nov;280(22):5626–34. http://dx.doi.org/10.1111/febs.12502
18. Shoemaker LD, Kornblum HI. Neural stem cells (NSCs) and proteomics. Mol Cell Proteomics. 2016 Feb;15(2):344–54. http://dx.doi.org/10.1074/mcp.O115.052704
19. Zhao W, Li J, Mills GB. Functional proteomic characterization of cancer cell lines. Oncoscience. 2017 Jun 10;4(5–6):41–2. http://dx.doi.org/10.18632/oncoscience.351
20. Phelan K, May KM. Basic techniques in mammalian cell tissue culture: Basic techniques in mammalian cell tissue culture. Curr Protoc Cell Biol. 2015 Mar 2;66:1.1.1–22. http://dx.doi.org/10.1002/0471143030.cb0101s66
21. Schwaid AG, Krasowka-Zoladek A, Chi A, Cornella-Taracido I. Comparison of the rat and human dorsal root ganglion proteome. Sci Rep. 2018 Sep 7;8(1):13469. http://dx.doi.org/10.1038/s41598-018-31189-9
22. Russo C, Lewis EEL, Flint L, Clench MR. Mass spectrometry imaging of 3D tissue models. Proteomics. 2018 Jul;18(14):e1700462. http://dx.doi.org/10.1002/pmic.201700462
23. Velic A, Macek B, Wagner CA. Toward quantitative proteomics of organ substructures: Implications for renal physiology. Semin Nephrol. 2010 Sep;30(5):487–99. http://dx.doi.org/10.1016/j.semnephrol.2010.07.006
24. Gonneaud A, Asselin C, Boudreau F, Boisvert FM. Phenotypic analysis of organoids by proteomics. Proteomics. 2017 Oct;17(20). http://dx.doi.org/10.1002/pmic.201700023
25. Bendixen E. Animal models for translational proteomics. Proteomics Clin Appl. 2014 Oct;8(9–10):637–9. http://dx.doi.org/10.1002/prca.201470054

26. Skardal A, Murphy SV, Devarasetty M, Mead I, Kang HW, Seol YJ, et al. Multi-tissue interactions in an integrated three-tissue organ-on-a-chip platform. Sci Rep. 2017 Aug 18;7(1):8837. http://dx.doi.org/10.1038/s41598-017-08879-x
27. Khalid N, Arif S, Kobayashi I, Nakajima M. Lab-on-a-chip techniques for high-throughput proteomics and drug discovery. In: Santos HA, Dongfei Liu D, Zhang H, editors. Micro and nano technologies, microfluidics for pharmaceutical applications. Norwich, NY: William Andrew Publishing; 2019. p. 371–422. https://doi.org/10.1016/B978-0-12-812659-2.00014-4
28. Raimondo F, Morosi L, Chinello C, Magni F, Pitto M. Advances in membranous vesicle and exosome proteomics improving biological understanding and biomarker discovery. Proteomics. 2011 Feb;11(4):709–20. http://dx.doi.org/10.1002/pmic.201000422
29. Guay C, Regazzi R. Exosomes as new players in metabolic organ cross-talk. Diabetes Obes Metab. 2017 Sep;19(Suppl 1):137–46. http://dx.doi.org/10.1111/dom.13027
30. Kwasnik A, Tonry C, Ardle AM, Butt AQ, Inzitari R, Pennington SR. Proteomes, their compositions and their sources. In: Mirzaei H, Carrasco M, editors. Modern proteomics – Sample preparation, analysis and practical applications. Cham: Springer International Publishing; 2016. p. 3–21.
31. Bladergroen MR, van der Burgt YEM. Solid-phase extraction strategies to surmount body fluid sample complexity in high-throughput mass spectrometry-based proteomics. J Anal Methods Chem. 2015;2015:250131. http://dx.doi.org/10.1155/2015/250131
32. Hsieh S-Y, Chen R-K, Pan Y-H, Lee H-L. Systematical evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling. Proteomics. 2006 May;6(10):3189–98. http://dx.doi.org/10.1002/pmic.200500535
33. Sköld K, Alm H, Scholz B. The impact of biosampling procedures on molecular data interpretation. Mol Cell Proteomics. 2013 Jun;12(6):1489–501. http://dx.doi.org/10.1074/mcp.R112.024869
34. Kelstrup CD, Bekker-Jensen DB, Arrey TN, Hogrebe A, Harder A, Olsen JV. Performance evaluation of the Q exactive HF-X for shotgun proteomics. J Proteome Res. 2018 Jan 5;17(1):727–38. http://dx.doi.org/10.1021/acs.jproteome.7b00602
35. Sadygov RG, Cociorva D, Yates JR. Large-scale database searching using tandem mass spectra: Looking up the answer in the back of the book. Nat Methods. 2004 Dec;1(3):195–202. http://dx.doi.org/10.1038/nmeth725
36. Pandeswari PB, Sabareesh V. Middle-down approach: A choice to sequence and characterize proteins/proteomes by mass spectrometry. RSC Adv. 2019 Jan 2;9:313–44. http://dx.doi.org/10.1039/C8RA07200K
37. Zhang Q, Ma C, Gearing M, Wang PG, Chin L-S, Li L. Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer's disease. Acta Neuropathol Commun. 2018 Mar 1;6(1):19. http://dx.doi.org/10.1186/s40478-018-0524-2
38. Yu Y, Bekele S, Pieper R. Quick 96FASP for high throughput quantitative proteome analysis. J Proteomics. 2017 Aug 23;166:1–7. http://dx.doi.org/10.1016/j.jprot.2017.06.019
39. Spanka D-T, Konzer A, Edelmann D, Berghoff BA. High-throughput proteomics identifies proteins with importance to postantibiotic recovery in depolarized persister cells. Front Microbiol. 2019 Mar 6;10:378. http://dx.doi.org/10.3389/fmicb.2019.00378
40. Dams M, Dores-Sousa JL, Lamers R-J, Treumann A, Eeltink S. High-resolution nano-liquid chromatography with tandem mass spectrometric detection for the bottom-up analysis of complex proteomic samples. Chromatographia. 2018 Nov 7;82(1):101–10. http://dx.doi.org/10.1007/s10337-018-3647-5
41. Sinitcyn P, Rudolph JD, Cox J. Computational methods for understanding mass spectrometry–based shotgun proteomics data. Annu Rev Biomed Data Sci. 2018 Jul;1:207–34. http://dx.doi.org/10.1146/annurev-biodatasci-080917-013516
42. Thomas S, Hao L, Ricke WA, Li L. Biomarker discovery in mass spectrometry-based urinary proteomics. Proteomics Clin Appl. 2016 Apr;10(4):358–70. http://dx.doi.org/10.1002/prca.201500102
43. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res. 2017 Jan 4;45(D1):D362–8. http://dx.doi.org/10.1093/nar/gkw937
44. Cristobal A, Marino F, Post H, van den Toorn HW, Mohammed S, Heck AJ. Toward an optimized workflow for middle-down proteomics. Anal Chem. 2017 Mar 21;89(6):3318–3325. http://dx.doi.org/10.1021/acs.analchem.6b03756

45. Greer SM, Sidoli S, Coradin M, Schack Jespersen M, Schwämmle V, Jensen ON, et al. Extensive characterization of heavily modified histone tails by 193 nm ultraviolet photodissociation mass spectrometry via a middle-down strategy. Anal Chem. 2018 Sep 4;90(17):10425–33. http://dx.doi.org/10.1021/acs.analchem.8b02320
46. Zhang H, Ge Y. Comprehensive analysis of protein modifications by top-down mass spectrometry. Circ Cardiovasc Genet. 2011 Dec;4(6):711. http://dx.doi.org/10.1161/CIRCGENETICS.110.957829
47. McCool EN, Chen D, Li W, Liu Y, Sun LL. Capillary zone electrophoresis-tandem mass spectrometry with ultraviolet photodissociation (213 nm) for large-scale top–down proteomics. Anal Methods. 2019 May 7;11:2855–61. http://dx.doi.org/10.1039/C9AY00585D
48. Cleland TP, DeHart CJ, Fellers RT, VanNispen AJ, Greer JB, LeDuc RD, et al. High-throughput analysis of intact human proteins using UVPD and HCD on an orbitrap mass spectrometer. J Proteome Res. 2017 May 5;16(5):2072–9. http://dx.doi.org/10.1021/acs.jproteome.7b00043
49. Skinner OS, Haverland NA, Fornelli L, Melani RD, Do Vale LHF, Seckler HS, et al. Top-down characterization of endogenous protein complexes with native proteomics. Nat Chem Biol. 2018 Jan;14(1):36–41. http://dx.doi.org/10.1038/nchembio.2515
50. Tholey A, Becker A. Top-down proteomics for the analysis of proteolytic events – Methods, applications and perspectives. Biochim Biophys Acta Mol Cell Res. 2017 Nov;1864(11 Pt B):2191–9. http://dx.doi.org/10.1016/j.bbamcr.2017.07.002
51. Cesnik AJ, Shortreed MR, Schaffer LV, Knoener RA, Frey BL, Scalf M, et al. Proteoform suite: Software for constructing, quantifying, and visualizing proteoform families. J Proteome Res. 2018 Jan 5;17(1):568–78. http://dx.doi.org/10.1021/acs.jproteome.7b00685
52. Schaffer LV, Rensvold JW, Shortreed MR, Cesnik AJ, Jochem A, Scalf M, et al. Identification and quantification of murine mitochondrial proteoforms using an integrated top-down and intact-mass strategy. J Proteome Res. 2018 Oct 5;17(10):3526–36. http://dx.doi.org/10.1021/acs.jproteome.8b00469
53. Sidoli S, Lu C, Coradin M, Wang X, Karch KR, Ruminowicz C, et al. Metabolic labeling in middle-down proteomics allows for investigation of the dynamics of the histone code. Epigenetics Chromatin. 2017 Jul 6;10(1):34. http://dx.doi.org/10.1186/s13072-017-0139-z
54. Sidoli S, Garcia BA. Middle-down proteomics: A still unexploited resource for chromatin biology. Expert Rev Proteomics. 2017 Jul;14(7):617–26. http://dx.doi.org/10.1080/14789450.2017.1345632
55. Moradian, A, Franco C, Sweredoski, MJ, Hess, S. Middle-down electron capture dissociation and electron transfer dissociation for histone analysis. J Anal Sci Technol. 2015 Dec;6:21. http://dx.doi.org/10.1186/s40543-015-0060-7
56. Krappmann M, Luthardt M, Lesske F, Letzel T. The software-landscape in (prote)omic research. J Proteomics Bioinform. 2015;8(7):164–75. http://dx.doi.org/10.4172/jpb.1000365
57. Navarro P, Kuharev J, Gillet LC, Bernhardt OM, MacLean B, Röst HL, et al. A multicenter study benchmarks software tools for label-free proteome quantification. Nat Biotechnol. 2016 Nov;34(11):1130–6. http://dx.doi.org/10.1038/nbt.3685
58. Välikangas T, Suomi T, Elo LL. A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief Bioinform. 2018 Nov 27;19(6):1344–55.
59. Chen Y, Wang F, Xu F, Yang T. Mass spectrometry-based protein quantification. Adv Exp Med Biol. 2016;919:255–79. http://dx.doi.org/10.1007/978-3-319-41448-5_15
60. Urfer W, Grzegorczyk M, Jung K. Statistics for proteomics: A review of tools for analyzing experimental data. Proteomics. 2006 Sep;6(Suppl 2):48–55. http://dx.doi.org/10.1002/pmic.200600554
61. Gatto L, Christoforou A. Using R and bioconductor for proteomics data analysis. Biochim Biophys Acta. 2014 Jan;1844(1 Pt A):42–51. http://dx.doi.org/10.1016/j.bbapap.2013.04.032
62. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. 2001 Jan 1;29(1):37–40. http://dx.doi.org/10.1093/nar/29.1.37
63. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25–9. http://dx.doi.org/10.1038/75556
64. Khatri P, Drăghici S. Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics. 2005 Sep 15;21(18):3587–95. http://dx.doi.org/10.1093/bioinformatics/bti565

65. Wu X, Hasan MA, Chen JY. Pathway and network analysis in proteomics. J Theor Biol. 2014 Dec 7;362:44–52. http://dx.doi.org/10.1016/j.jtbi.2014.05.031
66. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: An open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D91–4. http://dx.doi.org/10.1093/nar/gkh012
67. Koh GC, Porras P, Aranda B, Hermjakob H, Orchard SE. Analyzing protein-protein interaction networks. J Proteome Res. 2012 Apr 6;11(4):2014–31. http://dx.doi.org/10.1021/pr201211w
68. Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. Nucleic Acids Res. 2015 Jan;43(Database issue):D789–98. http://dx.doi.org/10.1093/nar/gku1205
69. Piñero J, Queralt-Rosinach N, Bravo À, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015 Apr 15;2015:bav028. http://dx.doi.org/10.1093/database/bav028
70. Medina I, Carbonell J, Pulido L, Madeira SC, Goetz S, Conesa A, et al. Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. Nucleic Acids Res. 2010 Jul;38(Web Server issue):W210–13. http://dx.doi.org/10.1093/nar/gkq388
71. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003 Nov;13(11):2498–504. http://dx.doi.org/10.1101/gr.1239303
72. Husi H. NMDA receptors, neural pathways, and protein interaction databases. Int Rev Neurobiol. 2004;61:49–77. http://dx.doi.org/10.1016/S0074-7742(04)61003-8

9

Cheminformatics and Computational Approaches in Metabolomics

Marco Fernandes[1,2] • Bela Sanches[3] • Holger Husi[2,4]

[1]Department of Psychiatry, Warneford Hospital, Translational Neuroscience and Dementia Research, Oxford University, Oxford, UK; [2]Institute of Cardiovascular and Medical Sciences, BHF Glasgow Cardiovascular Research Centre, University of Glasgow, Glasgow, UK; [3]Strathclyde Institute of Pharmacy & Biomedical Sciences (SIPBS), University of Strathclyde, Glasgow, UK; [4]Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, UK

Author for correspondence: Holger Husi, Division of Biomedical Sciences, Centre for Health Science, University of Highlands and Islands, Inverness, United Kingdom.
Email: Holger.Husi@uhi.ac.uk

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch9

Abstract: Metabolomics can be viewed as an evolved form of chemical analysis, which required an early instrumental revolution in which the technological core of spectroscopy and spectrometry was developed. This was followed by the advent of high-throughput and high-performance liquid chromatography, together with the establishment of compound libraries and database systems. The ease in the use of metabolomics platforms was coupled with an implementation of data mining methods and bioinformatics tools using machine learning approaches. Cheminformatics makes use of software packages and tools to convey workflows and to streamline data analysis. On the other hand, computational biology offers the contextual approach to the functional characterization of metabolite profiles from a dataset, providing ontologies and annotations. In this chapter, we discuss the main technical procedures used in metabolomics data acquisition, data processing and pipelines, followed by data mining and statistical approaches including machine learning, and ultimately how metabolomics data can aid in elucidating aberrant pathways and metabolic dysfunctions in disease.

Keywords: cheminformatics; computational biology; functional annotation; machine learning; metabolomics

In: *Computational Biology.* Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

The metabolome is the genome's final product, which is defined as the total quantitative group of small molecular weight compounds (metabolites) present in a cell or organism that is involved in metabolic reactions (1). Metabolites are small molecules that are chemically transformed during metabolism, providing functional information of the cellular state, which serves as direct signatures of biochemical activity. Therefore, they are easy to correlate with phenotypes when compared to genes and proteins, whose function is subject to epigenetic regulation and post-translational modifications, respectively (2). Metabolomics (Figure 1) is part of the omics strategies (genomics, proteomics and transcriptomics) that aim to describe the metabolome qualitatively and quantitatively by applying various analytical platforms and methods (3).

Metabolomics combines analytical chemistry strategies and is based on several technological platforms such as mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy with streamline data analysis (4). In metabolomics, the choice of platforms and techniques is less evident, whereas in genomics and proteomics this appears to be more intuitive to implement for a given study, such as the use of next-generation sequencing (NGS) or microarrays and in-gel or in-solution MS, respectively (1). Nevertheless, MS and NMR are usually the preferred choices for metabolome investigations (5). Data generated through these acquisition platforms need to be further processed using different open-source software or commercial software, such as MZmine (6), Mnova, MetAlign (7), MathDAMP (8), MS-DIAL (9), and XCMS (10) (Table 1). The software can be jointly used with other online or commercially available libraries and databases, depending on the purpose of the study, like the Dictionary of Natural Products (DNP), ChemSpider (11), MarinLit (12), or in-house/custom databases, to

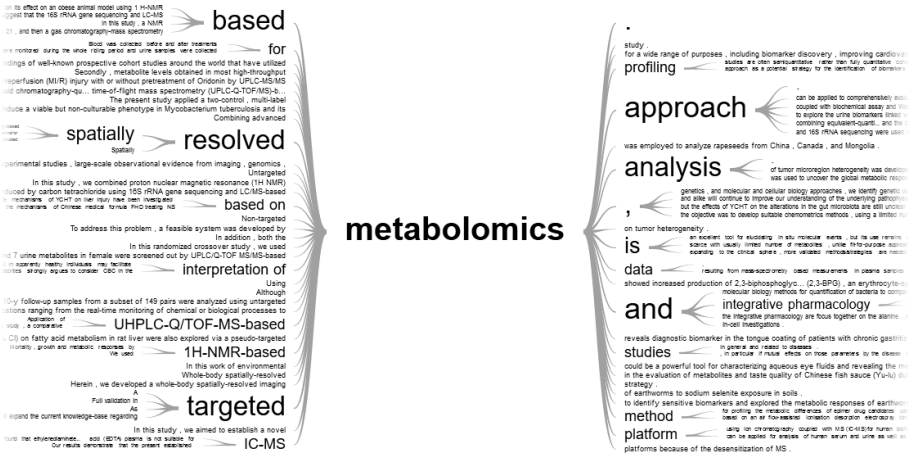


Figure 1 Tree mapping of the most frequent terms in the metabolomics field. Data mining from abstracts indexed in PubMed using as primary key word – "metabolomics" in "Pub-tree" available at https://esperr.github.io/pub-trees/.

TABLE 1 **Software solutions for acquisition and pre-processing data across metabolomics platforms**

Software package	Selected features	Platform	Distribution	Ref.
MS-DIAL	Built-in DIA analysis, annotation and visualization	GC/LC/MS	open-source	(9)
XCMS	User-friendly, retention time correction, statistical analysis	LC/MS		(10)
MZmine2	Batch mode, deconvolution, statistical analysis, visualization	LC/MS		(45)
Mnova	Single suite for processing and visualization	NMR, GC/LC/MS	commercial	—
speaq 2.0	Peak picking and grouping; multivariate statistical functions	1D NMR	open-source	(47)
MetaboAnalyst	Modules for integrative data analysis	NMR, GC/LC/MS		(48)
rDolphin	Enhances ROI by estimation of baseline and signal parameters to maximize fitting of the signals	1H-NMR		(49)
BATMAN	Concentration estimates for known compounds from raw spectra	NMR		(50)
rNMR	Visualisation of NMR signals from multiple spectra concurrently by assigning chemical shift ranges	NMR		(51)

ROI, regions of interest; NMR, nuclear magnetic resonance; IR, infrared Raman; XRF, X-ray fluorescence; DIA, data independent acquisition.

identify secondary metabolites based in the information of the chemical structure of known natural products. Accordingly, the processed data are further subjected to multivariate statistical analysis applying, for example, soft independent modeling by class analogy (SIMCA), which uses unsupervised clustering such as partial component analysis (PCA) or supervised clustering like orthogonal partial least squares discriminant analysis (OPLS-DA), to provide information on the putative bioactive metabolite at the first fractionation step or detect putative biomarkers in a cellular process (13).

Screening for new compounds of pharmacological interest for a specific disease or a disease class has a long history of success cases. For instance, the use of high-throughput screening (HTS) methods for early-stage drug discovery directly yielded cyclosporin A (14), a fungal-derived immunosuppressant medication, and mevastatin, a mold-derived agent, used to normalize cholesterol levels (15). Likewise, drug discovery using structure-based drug design (SBDD) led to the development of new drug candidates such asdorzolamide (16), which is a topical ophthalmic agent applied in the treatment of glaucoma. This method was also

used to develop imatinib, a cancer chemotherapy agent for specific treatment of many leukemia subtypes. Other examples include vemurafenib, a BRAF inhibitor used as chemotherapeutic agent in late stages of melanoma (17). Although it becomes apparent that an ideal workflow for earlier drug discovery should rely on a whole range of tools, from detection and analytical platforms, used either coupled or in parallel, through to computational and statistical steps (Figure 2). This will not only assist in the investigation of novel compounds, but accelerate the discovery stage or even to boost drug repurposing programs (18). This becomes even more apparent when costs are factored in, since the development of a new drug, from target identification to the availability of a final product including approval for prescription to the general public by a governmental or local state authority, involves a multi-step procedure, which can easily take around 12 to 15 years, and is associated with extremely high costs for companies (19). This process starts with basic research that includes lead identification, synthesis scale-up and in vitro pharmacology. This is followed by preclinical development which includes assessing of in vitro toxicity and measuring specific activities by conducting studies of absorption, distribution, excretion and metabolism, and activity of relevant enzymes (20). Alternatively, if the lead target such as a protein is known and its 3D-structure has been elucidated, in silico approaches to predict drug–enzyme interactions can be pursued, using docking algorithms and other well-established computational structure-based approaches (21).

METABOLOMICS DATA ACQUISITION AND PRE-PROCESSING

This section gives an overview of common detection methodologies in metabolomics, conversion of machine data to spectral files and mapping to both known and putative libraries, and ultimately construction of discovery matrices allowing peak-metabolite pairing and quantitative measures. Acquiring raw data from metabolomics analytical platforms and their conversion to extracted data, such as peak lists and spectral bins, requires specific software packages that in many cases need proprietary licenses that are often tied to the platform manufacturer (Table 1).

Mass spectrometry analytical platform

Mass spectrometry (MS) is the analytical technique of choice in metabolomics for identification and/or quantification of varied classes of metabolites, consisting in the production of gas-phase ions that are then detected and characterized by their mass and charge (22). Basically, a mass spectrometer consists of a sample inlet, an ion source, a mass analyzer and a detector and, in that order, functions by introducing the sample into the mass spectrometer, generates gas-phase ions via an ionization technique, separates the ions according to their mass-to-charge ratio (m/z) and generates an electric current from the incident ions that is proportional to their abundances (22). Moreover, the combination of separation techniques such as gas chromatography (GC), high performance liquid chromatography (HPLC), and capillary electrophoresis (CE) allows improved metabolite identification and quantification by MS, which is particularly beneficial when dealing with complex biological samples (5). The recent introduction of a reengineered

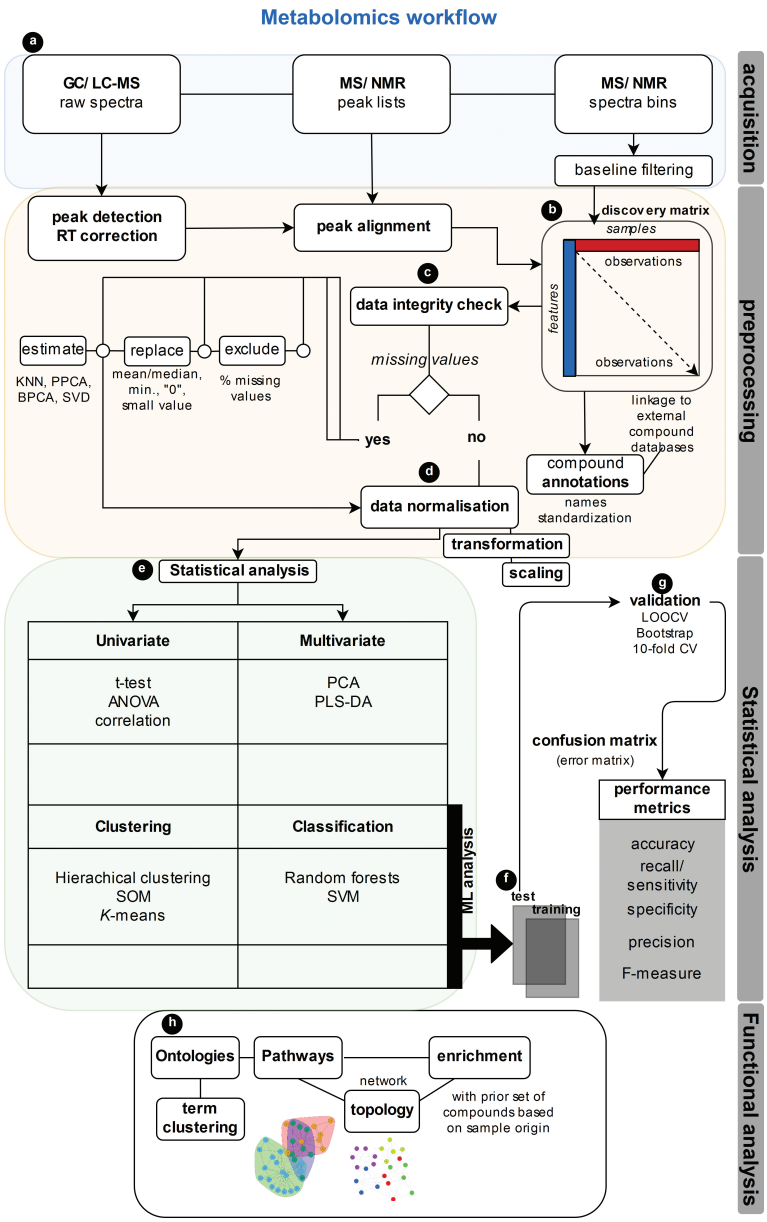


Figure 2 Metabolomics workflows. Data acquisition (a), pre-processing steps including discovery matrix generation (b), data integrity check (c) and data normalisation (d). Followed by statistical analysis (e), machine learning (ML) approaches (f) and validation based of randomisation (g), and functional analysis (h). **Abbreviations**: Gas Chromatography (GC), Liquid Chromatography (LC), Mass Spectrometry (MS), Nuclear Magnetic Resonance (NMR), Leave-One-Out Cross Validation (LOOCV), *n*-times/fold Cross Validation (CV), k-nearest neighbours (KNN), probabilistic principal component analysis (PPCA), Bayesian principal component analysis (BPCA), singular value decomposition (SVD), analysis of variance (ANOVA) , Principal Component Analysis (PCA), Partial Least Squares (PLS).

chromatographic technology such as ultra-high-pressure liquid chromatography (UHPLC) has led to enhanced resolution, higher throughput, lower running times and better cost-effectiveness than traditional HPLC. The use of MS in metabolomics has important advantages such as requiring small sample volumes and provides highly sensitive detection and metabolite identification via interpretation of the spectra and comparison of molecular formula determination via precise mass measurements (23). Additionally, MS is also destructive, and therefore an analyzed sample is not recoverable, and is a relatively slow detection methodology, unlike NMR spectroscopy (23).

Nuclear magnetic resonance analytical platform

NMR spectroscopy is a widely used technique for metabolomics studies with many benefits, such as being specific and at the same time non-selective and non-destructive, and requires no separation or derivatization, is fast and offers highly reproducible and quantitative analyses (1). ANMR spectrum is specific and unique to each compound and provides valuable structural information about the components of the analyzed sample. It combines the information of chemical shift (the nature of the chemical environment), signal multiplicities (neighboring signals), homonuclear and heteronuclear coupling constants, integrals of the signals (number of protons), spin–spin coupling (number and nature of neighbors and connectivity information), and relaxation or diffusion (size of molecule and large-scale environment of location) (24). Although one-dimensional (1D) proton (H) and carbon (C) NMR is one of the most used modes, currently alternative techniques are available, offering additional chemical and structural information, since, in some cases, 1H and 13C NMR are insufficient to provide enough information to entirely characterize metabolites (5) and resolve their identity. To complement the 1D experiments, it is possible to perform two-dimensional (2D) correlation spectroscopy such as 1H-1H COSY, 1H-13C HMBC, 1H-13C HMQC, 1H-13C HSQC,1H-1H ROESY, and 1H-1H NOESY, which enables the elucidation of complex structures. Additionally, samples can be reused, since this technique is non-destructive and does not require pre-selection of analysis conditions like ion source, which is a pre-requisite of MS, or chromatographic operating conditions such as stationary phase, mobile phase, and temperature (1).

Metabolite identification strategies, libraries and algorithms

The metabolomics field has been evolving according to the need for chemical characterization of the composition of biological matrices and extracts from a diverse range of organisms. A fundamental task and simultaneously one of the major bottlenecks in many research areas that use metabolomics workflows is to accurately identify unknown small molecules from the MS and NMR spectra data. Therefore, libraries containing reference spectra with peak assignment to metabolites from previous experiments are being collated and maintained in spectral and compound databases. NMR-based spectral databases are SDBS (13C-NMR, ESR and Raman spectra) (25) (13C-NMR, ESR and Raman spectra), BioMagResBank (26), NMRShiftDB2 (27) and The Birmingham Metabolite Library Nuclear Magnetic Resonance database (BML-NMR) (28). On the other hand, MS-based spectral databases consist of METLIN (29), NIST (30), GMD (31) and MassBank (32).

The Madison Metabolomics Consortium Database (MMCD) (33), The Human Metabolome Database (HMDB) (34) and MetaboLights (35) cover both MS and NMR spectra. Splitting by analytical platform and type of content, either selecting only by spectral data or selecting only by compound annotations, is rather conceptual, since many "modern" metabolomics databases aim to implement both contents in an integrative way. Despite the steady increase in the number of metabolite identities across databases, many cannot be detected through this strategy of database matching due to the absence of their spectral information. Conventional approaches for the identification of these unknowns require reduction of sample complexity by successive steps of fractionation, in order to isolate the target metabolite or compound from the complex mixture, which poses several technical challenges and is highly time-consuming. However, it often does not guarantee identification of low-abundance metabolites via NMR or other spectroscopic techniques (36). Instead, either using the raw or crude sample mixture or even partial sample fractionations can achieve elucidation of the metabolite structure. Then software implementations such as MetFrag2 (37) and CSI:FingerID (38) are available, where MS2 (MS/MS) LC-MS/MS spectra of an unknown experimental metabolite is compared with the in silico generated MS2 fragmentation spectra of putative metabolite structures to find a best match. Other approaches include the use of NMR chemical shifts, in a straight analogy with the previously mentioned strategy, comparing in this case the deconvoluted experimental chemical shifts of unknown metabolites with predictions to yield a best match, where deconvolution is a process to remove instrument-specific signal distortions (39). Recently, the possibility to perform joint analysis with complementary platforms such as NMR and MS was suggested to solve the current paradigm of identification of unknown metabolites (40). Hybrid strategies, such as the SUMMIT MS/NMR (41), primarily resolves all the chemical formulas of the sample detected in the MS1 spectra and then generates all the possible structure permutations. This follows a prediction of NMR chemical shifts for each structural rearrangement and comparison with experimental records acquired to consistently identify molecular structures from both platforms. Other groups used oversimplistic approaches by correlating signal intensities from peak lists from NMR and LC-MS data as proof of principle for the identification of individual metabolites in a sample (42).

Data pre-processing

This step aims to generate a matrix that typically comprehends features (rows) and samples (columns) with each pair coding for an observation from primary raw data. Here, the analysis cascade usually is performed in a stepwise manner and also involves other pre-processing workflows for quality control (QC) dependent on the nature of the acquisition platform, for instance, deconvolution of overlapping peaks, peak picking, integration and alignment (43).

One of the initial steps in the analysis of mass spectrometry data is to convert the vendor-specific binary files to an open or universal format. Thus, LC-MS raw data can be split by ionization mode (positive and negative) using, for instance, the ProteoWizardmsConvert tool (44), and then imported and processed using the open-source MZmine2 (45) toolbox or other software solutions displayed in

Table 1. MZmine2 can carry out peak detection, alignment, deconvolution (decomposing overlapping peaks), peak picking and deisotoping, filtering (e.g., removing low-intensity peaks) and gap-filling when, for instance, peaks were detected in some runs or scans but not in others. Additionally, this allows the prediction of putative molecular formulas for each feature by minimizing mis-assignment of features by stepwise removing adducts and complexes (45). This is followed by verifying how novel the "new" compound is by applying dereplication methods, which are particularly relevant for the discovery of new compounds derived from natural product metabolomic data, since it filters from the analysis all the known ones (46). Similar approaches can be found in subtractive and differential genome analysis. DEREPLICATOR+is such an improved algorithm for the dereplication task of core importance in natural products discovery (46).This algorithm assembles theoretical spectra of peptides from non-ribosomal peptide synthetases and ribosomally synthesized post-translationally modified peptide synthetases by first generating a decoy database of peptidic natural products. It then builds predicted spectra for all peptidic products within the database, thereby generating and attributing a score for each peptide and associated spectrum matches, calculating P-values and correction for multiple testing using false discovery rate for the former pairs matching and infer the initial seed of peptidic products by spectral network approaches. Customized libraries with relevant peptidic products can be created by applying dereplication algorithms and further explored or reused by coupling with state-of-the-art software toolboxes such as MZmine2.

On the other hand, the acquired NMR data can be processed with the commercial MestReNova (Mnova) software to confirm and elucidate chemical structures. The 1D and 1H spectrums are processed using the following steps: The baseline is corrected by manual phasing and by using the Whittaker Smoother, and Gaussian is set to 1 Hz for apodization. The chemical shifts are given in ppm and the coupling constants are given in Hz. Chemical shifts in ppm are used to generate the unique primary ID while there are no other secondary IDs considered. It is possible to add the integral number that gives information about the number of hydrogens present in the structure and the multiplicity indicating the neighboring number of hydrogens, thereby allowing a positive assignment of measured data and structure information.

DATA MINING APPROACHES, STATISTICAL ANALYSIS AND ML METHODS

This section will give a brief description of some ML algorithms and performance metrics with examples from the literature of their implementation in the analysis workflow of metabolomics datasets. This includes the initial use of dimensionality reduction methods for visual inspection or data summarization tasks, additional feature selection through filtering metabolites that show higher variability across samples and further computational downstream analysis (Table 2). The popularity and choice of ML algorithms is highly dependent on the domain of science, availability, computational cost, model complexity and interpretability. The eternal model trade-off between "too simple," yet highly biased, and "too complex," yet

TABLE 2 Machine learning methods and algorithms

Class	Description	Implementation/toolbox	Weka	KNIME	TensorFlow	Caret
Association rule learning algorithms	Rules extraction to explain variables association	Apriori and Eclat algorithms	+++	++	+	+++
Artificial neural network algorithms including deep learning	Neural networks construction	Perceptron, back-propagation, Hopfield network, [a]RBFN, CNN, stacked auto-encoders	++	++	+++	++
Bayesian algorithms	Bayes' theorem for classification and regression problems	Naive Bayes, Gaussian Naive Bayes, Bayesian Network, McMC	+++	++	+++	+++
Dimensionality reduction	Unsupervised and supervised approaches to resolve multidimensional data structures	[b]PCA, CCA, PLS, OPLS, MDS, LDA, MDA, QDA, FDA	+++	+++	+++	+++
Ensemble algorithms	Composite of multiple models trained independently in which their individual predictions are fused to yield enhanced overall predictions	Boosting, bootstrapped aggregation (bagging), AdaBoost, stacked generalization (blending), [c]GBM, GBRT, random forests (RF)	+++	++	+++	+++
Decision tree	Trained on the data for classification and regression problems providing a flowchart-like structure model where nodes denote tests on an attribute with each branch represents outcome of a test and each leaf node holds a class label	Classification and regression tree, C4.5 and C5.0, decision stump, regression tree	+++	++	+	++
Regularization	Penalization measures to convey simple models	[d]LASSO, ridge, elastic net	+++	++	++	
Instance based	Comparison of test samples with train samples	[e]kNN, SOM, SVM	+++	++	+++	
Regression	Model relationship between features and sample, error as measure	[f]OLSR, LOESS, linear regression	+++	+++	+++	

Standalone software or analysis framework solutions are available (Weka (W), KNIME (K), TensorFlow (T) library and Caret R package) and can perform most of algorithmic tasks described. Natively supports (+++), supports with add-ons/plugins or extensions (++), or not available or poorly described (+).

[a]Radial Basis Function Network (RBFN), Deep Boltzmann Machine (DBM), Deep Belief Networks (DBN), Convolutional Neural Network (CNN).

[b]Principal Component Analysis (PCA), Canonical Correspondence Analysis (CCA),Partial Least Squares (PLS),Orthogonal PLS (OPLS), Multidimensional Scaling (MDS), Linear Discriminant Analysis (LDA),Mixture Discriminant Analysis (MDA), Quadratic Discriminant Analysis (QDA),Flexible Discriminant Analysis (FDA).

[c]Gradient Boosting Machines (GBM), Regression Trees (GBRT), random forests (RF).

[d]Least Absolute Shrinkage and Selection Operator; (LASSO).

[e]k-Nearest Neighbors (kNN), Self-Organizing Map (SOM), Support Vector Machine (SVM).

[f]Ordinary Least Squares Regression (OLSR), Locally Estimated Scatterplot Smoothing (LOESS).

highly variable, is a core concept in statistics and ML. Standard ML performance metrics such as area under the curve (AUC) are derived from receiver operating characteristic curves (ROC), R2/Q2 ratios, and *k*-fold cross-validation. This also includes concepts like sensitivity, the ratio of the proportion of true positives and the sum of the proportion of false negatives and true positives, which in medical sciences could be interpreted as the proportion of individuals with disease whose test is positive. This is in contrast to specificity, the ratio of the proportion of true negatives and the sum of the proportion of true negatives and false positives is the proportion of individuals without disease whose test returned negative.

Dimensionality reduction and multivariate analysis

Today, an extensive variety of statistical methods is available, ranging from unsupervised methods, such as principal components analysis (PCA), or hierarchical clustering (HCA) to supervised methodologies like partial least squares (PLS), partial least squares discriminant analysis (PLS-DA) and orthogonal partial least squares discriminant analysis (OPLS-DA) (52). Processed MS and NMR data usually are in the form of a matrix of signal intensities signal origins, and, since both are in the same format, it is possible to apply standard analysis techniques to both (53). The first step in metabolomics data analysis is using PCA as an initial exploratory and visualization method that gives an overview of the variability of the dataset as the samples are grouped based on similarity or differences within the group of samples. This enables the detection of trends, groups and outliers, and it is possible to visualize the data as a score plot and a loading plot. In the score plot, each point represents an individual sample, while the loading plot gives information about which variables have the greatest contribution to the positioning of the samples on the scores plot and are responsible for the clustering of samples (24). PCA analysis is followed by supervised pattern recognition techniques, which applies class information of the samples to maximize the separation between different groups of samples and detect the metabolic signatures that contribute to the classifications (24). OPLS-DA is the most used supervised methodology, which has the same predictive power as PLS but gives better interpretation of the relevant variables. This methodology provides information about the causes for class separation (54). In metabolomics, most of the analysis workflows are bespoke procedures, thus requiring implementation of individual software solutions for a given task. For instance, MS-derived MZmine Ids can be combined with ionization mode (positive and negative) to generate a unique primary ID, while other variables like retention time (RT), m/z and molecular weight (MW) should be considered as secondary IDs. Then, using OPLS-DA to compare among groups, it is possible to discriminate and rank metabolites according to their variable importance in projection (VIP) value, ranging from 0 to 1. This is achieved by applying Pareto scaling, which is similar to autoscaling (55), and models can be validated based on multiple correlation (R^2) and cross-validation (Q^2) coefficients as well as by permutation tests for the supervised method.

Kernel methods

Support vector machine (SVM) is the best well-known classification algorithm within machine learning kernel methods, which is the gathering of kernel functions able to map any two points in the initial space representation based on the distances between them into the new space representation, avoiding the computational burden to compute all data point coordinates into the new space. SVM is broadly applied to many classification problems, and a boost in its use was observed with the rise of omics high-throughput data since in most setups it performs well with multidimensional and noisy data. Conceptually, it aims at solving classification problems by finding optimal decision margins between two sets of points belonging to two distinct categories. A decision margin can be described as a line on a surface separating training data into two spaces corresponding to two categories. The classification of new data points is to verify which side of the decision margin they fall on. The data are mapped to a new high-dimensional representation where the decision margin can be expressed as a hyperplane, which is a straight line in any case of dealing with only two dimensions. An optimal decision margin is computed by trying to maximize the distance between the hyperplane and the nearby data points from each class, a procedure named margin maximization, which allows generalization to new samples outside of the training dataset (56). Thereby, data points nearby the maximum margin hyperplane that sit on the margin are so-called support vectors. SVM is a good generalization classifier and has shown good performance using metabolomics data. For instance, Mahadevan et al. (57) did show that SVM gives better predictive models for diagnosis of pneumonia among individuals based on NMR spectral data measured in urine, yielding a classification accuracy greater than 99% using only 30 features selected via recursive feature elimination (RFE). On the other hand, traditional PLS-DA achieved >98% accuracy using 50 features ranked by VIP score. Others built classifiers using SVM with LOO cross-validation for the diagnostic purpose of ovarian cancer with an accuracy superior to 90% using LC/TOF-MS metabolic data detected in serum samples (58). Similarly, using ultra performance (UP) liquid chromatography (LC) with tandem MS for the detection of serum metabolites in early-stage ovarian cancer, the authors claim that using only 16 features selected by SVM-RFE, they are able to discriminate early ovarian cancer (N=46) from healthy controls (N=49) with perfect performance metrics in accuracy, sensitivity and specificity (59).

Ensemble algorithms, decision trees and random forests

Popular ensemble algorithms are bagging and boosting. The first trains each unconstrained model in parallel and the latter trains constrained models in series, learning from the previous ones, and thus evolving overtime. In ML, random forests (RF) (60) is a widely used ensemble algorithm that combines the output of multiple randomly generated decision trees into a composite averaged tree model. RF is applied in many domains of science in classification and regression tasks since it is easy to train and does not require complex tuning adjustments. Additionally, RF yields accurate and robust predictions and is recognized to be less prone to over fitting, a term used to describe the generation of a statistical

model that fits too well to the test or investigation data and fails subsequently in fitting subsequent data, since the rise in the number of each independent randomized tree in an ensemble model would be less likely to increase the generalization error (60). In metabolomics, RF has proven its value in many classification tasks, for instance, by building classifiers to distinguish colorectal cancer (CRC) patients and healthy individuals, as well as pre-surgical against post-surgical CRC patients based on the GC-MS measured urinary metabolome (61). After evaluation of the classification performance, RF, compared with LDA, SVM and PLS via AUC, R2/Q2 and 10-fold cross-validation, outperformed in all of those metrics. Ranking each metabolite through the RF Gini score, and further selection of those with a score >50, yielded, among others, homovanillate and lysine, which are able to discriminate healthy and CRC cases in an early-stage discovery study. Other examples of applicability of this ML algorithm are the development of classifiers able to discriminate among a large set of individuals infected with Zika virus with a specificity and sensitivity over 95% through the use of previously built RF classifiers containing 42 spectral signatures measured in blood using high-resolution mass spectrometry (62).

Functional annotation and biological interpretation of metabolomic data

At this stage, it is expected that a set of compounds or metabolites are identified in at least one chemical database. This simplifies further analysis since most of the available functional and enrichment analysis tools require different database identifiers. Thereby, once identified in any database, it becomes relatively trivial to cross-map compounds to other databases. Additionally, if information of sample concentration or expression is known and allows comparison across sample groups, for example, case versus control, this should be incorporated in the analysis. After having generated a list or matrix with annotated metabolites or compounds and their expression, concentration or ratio metric quantitative values, one can perform enrichment analysis, over-representation analysis, topology-based pathway analysis and activity profiling within pathways. This can be accomplished by using KEGG mapper web server functionality (https://www.genome.jp/kegg/mapper.html). However, this requires that the input is KEGG accession IDs that can be converted from chemical names using web solutions such as the CTS (https://cts.fiehnlab.ucdavis.edu) or MetaboAnalyst (48) ID converter functionality. The final output of the analysis however is only a list of pathways with the number of "hits" found. For a more formal statistical determination of pathway importance modules, Metabo Analyst can be used for enrichment or topological pathway analysis (48). Network-based analysis can be performed using Cytoscape (63), a standalone Java application, which provides multidimensional representations of large-scale networks. This platform supports directed, undirected and weighted graphs, filtering functionalities, merging and extensions for searching active sub-networks and pathway modules, and also incorporates a built-in statistical analysis of the network parameters. Several plug-ins are available for specific tasks, such as metabolomics integration with genomics and proteomics which is implemented in the

MetScape app (64). Additionally, Cytoscape allows interfacing with R and Python, which is useful for scaling and automation of tasks. For pathway editing and mapping metabolites or joint integrative analysis with genomics and proteomics, PathVisio (65) enables visualization and pathway statistical inference using firstly BridgeDb (66) to cross map molecular identifiers and then relies on curated collections of pathways from Wiki Pathways (67) and Reactome (68). In this tool, estimation of over-representation of pathways is based on a Z-score statistical procedure under the hypergeometric distribution and a P-value ranking based on a permutation procedure, which compares actual and permuted Z-scores.

CONCLUSION

The metabolomics field is rapidly evolving and appears to be catching up with genomics and proteomics approaches, which are more established in the research for disease biomarkers. Nevertheless, to establish foundations, protocols and standard operating procedures (SOPs), a more detailed evaluation of how to handle missing data is required through the assessment of the effects of imputation of missing values by means of statistical analysis across analytical metabolomics platforms and by the type of biological matrix. Inclusion and integration of other contextual biological counterparts such as genomics and proteomics will support a global overview of the system in study. Currently, matching experimental spectral data requires the query of many individual database resources to enable the best coverage and maximize compound identification. Ideally, those resources should cover both spectral and compound chemical characteristics, along with biological activities aggregated from many sources, and records would preferentially be manual-annotated and corrected to ensure the highest quality. Structural elucidation of new compounds is a complex, challenging and time-consuming task, but computational-assisted tools and algorithms will reduce such burden and potentiate in-line joint analysis of higher dimensional NMR experiments with high-resolution MS to achieve accurate identifications (24). In the years to come, we will undoubtedly see advances in the development of comprehensive metabolite spectral libraries, algorithms and bioinformatics tools for functional characterization and biological interpretation of metabolite profiles, thereby not only improving our understanding of biology and etiology of disease but also having an impact on drug discovery and personalized medical therapies.

Conflict of interest: The authors declare that they have no conflicts of interest with respect to research, authorship and/or publication of this chapter.

Copyright and permission statement: We confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s). All original sources have been appropriately acknowledged and/or referenced.

REFERENCES

1. Dunn WB, Bailey NJ, Johnson HE. Measuring the metabolome: Current analytical technologies. Analyst. 2005;130(5):606–25. http://dx.doi.org/10.1039/b418288j
2. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: The apogee of the omics trilogy. Nat Rev Mol Cell Biol. 2012;13(4):263–9. http://dx.doi.org/10.1038/nrm3314
3. Lopes AS, Cruz EC, Sussulini A, Klassen A. Metabolomic strategies involving mass spectrometry combined with liquid and gas chromatography. Adv Exp Med Biol. 2017;965:77–98. http://dx.doi.org/10.1007/978-3-319-47656-8_4
4. Griffiths WJ, Koal T, Wang Y, Kohl M, Enot DP, Deigner HP. Targeted metabolomics for biomarker discovery. Angew Chem Int Ed Engl. 2010;49(32):5426–45. http://dx.doi.org/10.1002/anie.200905579
5. Villas-Boas SG, Mas S, Akesson M, Smedsgaard J, Nielsen J. Mass spectrometry in metabolome analysis. Mass Spectrom Rev. 2005;24(5):613–46. http://dx.doi.org/10.1002/mas.20032
6. Katajamaa M, Oresic M. Processing methods for differential analysis of LC/MS profile data. BMC Bioinformatics. 2005;6:179. http://dx.doi.org/10.1186/1471-2105-6-179
7. Lommen A. MetAlign: Interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing. Anal Chem. 2009;81(8):3079–86. http://dx.doi.org/10.1021/ac900036d
8. Baran R, Kochi H, Saito N, Suematsu M, Soga T, Nishioka T, et al. MathDAMP: A package for differential analysis of metabolite profiles. BMC Bioinformatics. 2006;7:530. http://dx.doi.org/10.1186/1471-2105-7-530
9. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nat Methods. 2015;12(6):523–6. http://dx.doi.org/10.1038/nmeth.3393
10. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: Processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78(3):779–87. http://dx.doi.org/10.1021/ac051437y
11. Williams AJ, Tkachenko V, Golotvin S, Kidd R, McCann G. ChemSpider – Building a foundation for the semantic web by hosting a crowd sourced databasing platform for chemistry. J Cheminform. 2010;2(Suppl 1):O16. http://dx.doi.org/10.1186/1758-2946-2-S1-O16
12. Blunt J, Munro M. MarinLit database. Canterbury: University of Canterbury; 2012.
13. Robotti E, Marengo E. Chemometric multivariate tools for candidate biomarker identification: LDA, PLS-DA, SIMCA, ranking-PCA. Methods Mol Biol. 2016;1384:237–67. http://dx.doi.org/10.1007/978-1-4939-3255-9_14
14. Kreis W, Soricelli A. Cyclosporins: Immunosuppressive agents with antitumor activity. Experientia. 1979;35(11):1506–8. http://dx.doi.org/10.1007/BF01962813
15. Brown AG, Smale TC, King TJ, Hasenkamp R, Thompson RH. Crystal and molecular structure of compactin, a new antifungal metabolite from Penicillium brevicompactum. J Chem Soc Perkin 1. 1976;(11):1165–70. http://dx.doi.org/10.1039/p19760001165
16. Sugrue MF, Mallorga P, Schwam H, Baldwin JJ, Ponticello GS. Preclinical studies on L-671,152, a topically effective ocular hypotensive carbonic anhydrase inhibitor. Br J Pharmacol. 1989;98(Suppl):820P. http://dx.doi.org/10.3109/02713689008999600
17. Halaban R, Zhang W, Bacchiocchi A, Cheng E, Parisi F, Ariyan S, et al. PLX4032, a selective BRAF(V600E) kinase inhibitor, activates the ERK pathway and enhances cell migration and proliferation of BRAF melanoma cells. Pigment Cell Melanoma Res. 2010;23(2):190–200. http://dx.doi.org/10.1111/j.1755-148X.2010.00685.x
18. Roy A. Early probe and drug discovery in academia: A minireview. High Throughput. 2018;7(1):pii: E4. http://dx.doi.org/10.3390/ht7010004
19. Mohs RC, Greig NH. Drug discovery and development: Role of basic biological research. Alzheimers Dement (NY). 2017;3(4):651–7. http://dx.doi.org/10.1016/j.trci.2017.10.005
20. Pereira F, Aires-de-Sousa J. Computational methodologies in the exploration of marine natural product leads. Mar Drugs. 2018;16(7):236. http://dx.doi.org/10.3390/md16070236

21. Issa NT, Wathieu H, Ojo A, Byers SW, Dakshanamurthy S. Drug metabolism in preclinical drug development: A survey of the discovery process, toxicology, and computational tools. Curr Drug Metab. 2017;18(6):556–65. http://dx.doi.org/10.2174/1389200218666170316093301
22. Becker S, Kortz L, Helmschrodt C, Thiery J, Ceglarek U. LC-MS-based metabolomics in the clinical laboratory. J Chromatogr B Analyt Technol Biomed Life Sci. 2012;883–884:68–75. http://dx.doi.org/10.1016/j.jchromb.2011.10.018
23. Lindon JC, Nicholson JK, Wilson ID. Directly coupled HPLC-NMR and HPLC-NMR-MS in pharmaceutical research and development. J Chromatogr B Biomed Sci Appl. 2000;748(1):233–58. http://dx.doi.org/10.1016/S0378-4347(00)00320-0
24. Dona AC, Kyriakides M, Scott F, Shephard EA, Varshavi D, Veselkov K, et al. A guide to the identification of metabolites in NMR-based metabonomics/metabolomics experiments. Comput Struct Biotechnol J. 2016;14:135–53. http://dx.doi.org/10.1016/j.csbj.2016.02.005
25. Yamamoto O, Someno K, Wasada N, Hiraishi J, Hayamizu K, Tanabe K, et al. An integrated spectral data base system including IR, MS, ^{1}H-NMR, ^{13}C-NMR, ESR and Raman Spectra. Anal Sci. 1988;4(3):233–9. http://dx.doi.org/10.2116/analsci.4.233
26. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, et al. BioMagResBank. Nucleic Acids Res. 2008;36(Database issue):D402–8. http://dx.doi.org/10.1093/nar/gkm957
27. Kuhn S, Schlorer NE. Facilitating quality control for spectra assignments of small organic molecules: Nmrshiftdb2 – A free in-house NMR database with integrated LIMS for academic service laboratories. Magn Reson Chem. 2015;53(8):582–9. http://dx.doi.org/10.1002/mrc.4263
28. Ludwig C, Easton JM, Lodi A, Tiziani S, Manzoor SE, Southam AD, et al. Birmingham metabolite library: A publicly accessible database of 1-D 1 H and 2-D 1 H J-resolved NMR spectra of authentic metabolite standards (BML-NMR). Metabolomics. 2012;8(1):8–18. http://dx.doi.org/10.1007/s11306-011-0347-7
29. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: A metabolite mass spectral database. Ther Drug Monit. 2005;27(6):747–51. http://dx.doi.org/10.1097/01.ftd.0000179845.53213.39
30. Linstrom PJ, Mallard WG. The NIST Chemistry WebBook: A chemical data resource on the internet. J Chem Eng Data. 2001;46(5):1059–63. http://dx.doi.org/10.1021/je000236i
31. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, et al. GMD@CSB.DB: The golm metabolome database. Bioinformatics. 2005;21(8):1635–8. http://dx.doi.org/10.1093/bioinformatics/bti236
32. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: A public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703–14. http://dx.doi.org/10.1002/jms.1777
33. Cui Q, Lewis IA, Hegeman AD, Anderson ME, Li J, Schulte CF, et al. Metabolite identification via the Madison Metabolomics Consortium Database. Nat Biotechnol. 2008;26(2):162–4. http://dx.doi.org/10.1038/nbt0208-162
34. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R, et al. HMDB 4.0: The human metabolome database for 2018. Nucleic Acids Res. 2018;46(D1):D608–d17. http://dx.doi.org/10.1093/nar/gkx1089
35. Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, et al. MetaboLights: An open-access database repository for metabolomics data. Curr Protoc Bioinformatics. 2016;53:14.3.1–8. http://dx.doi.org/10.1002/0471250953.bi1413s53
36. Koehn FE, Carter GT. The evolving role of natural products in drug discovery. Nat Rev Drug Discov. 2005;4(3):206–20. http://dx.doi.org/10.1038/nrd1657
37. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. J Cheminform. 2016;8:3. http://dx.doi.org/10.1186/s13321-016-0115-9
38. Duhrkop K, Shen H, Meusel M, Rousu J, Bocker S. Searching molecular structure databases with tandem mass spectra using CSI: FingerID. Proc Natl Acad Sci U S A. 2015;112(41):12580–5. http://dx.doi.org/10.1073/pnas.1509788112
39. Komatsu T, Ohishi R, Shino A, Kikuchi J. Structure and metabolic-flow analysis of molecular complexity in a 13C-labeled tree by 2D and 3D NMR. Angew Chem Int Ed. 2016;55(20):6000–3. http://dx.doi.org/10.1002/anie.201600334

40. Marshall DD, Powers R. Beyond the paradigm: Combining mass spectrometry and nuclear magnetic resonance for metabolomics. Prog Nucl Magn Reson Spectrosc. 2017;100:1–16. http://dx.doi.org/10.1016/j.pnmrs.2017.01.001
41. Bingol K, Bruschweiler-Li L, Yu C, Somogyi A, Zhang F, Bruschweiler R. Metabolomics beyond spectroscopic databases: A combined MS/NMR strategy for the rapid identification of new metabolites in complex mixtures. Anal Chem. 2015;87(7):3864–70. http://dx.doi.org/10.1021/ac504633z
42. Li X, Luo H, Huang T, Xu L, Shi X, Hu K. Statistically correlating NMR spectra and LC-MS data to facilitate the identification of individual metabolites in metabolomics mixtures. Anal Bioanal Chem. 2019;411(7):1301–9. http://dx.doi.org/10.1007/s00216-019-01600-z
43. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. Anal Chem. 2006;78(2):567–74. http://dx.doi.org/10.1021/ac051495j
44. Kessner D, Chambers M, Burke R, Agus D, Mallick P. ProteoWizard: Open source software for rapid proteomics tools development. Bioinformatics. 2008;24(21):2534–6. http://dx.doi.org/10.1093/bioinformatics/btn323
45. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395. http://dx.doi.org/10.1186/1471-2105-11-395
46. Mohimani H, Gurevich A, Shlemov A, Mikheenko A, Korobeynikov A, Cao L, et al. Dereplication of microbial metabolites through database search of mass spectra. Nat Commun. 2018;9(1):4035. http://dx.doi.org/10.1038/s41467-018-06082-8
47. Beirnaert C, Meysman P, Vu TN, Hermans N, Apers S, Pieters L, et al. speaq 2.0: A complete workflow for high-throughput 1D NMR spectra processing and quantification. PLoS Comput Biol. 2018;14(3):e1006018. http://dx.doi.org/10.1371/journal.pcbi.1006018
48. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: Towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018;46(W1):W486–w94. http://dx.doi.org/10.1093/nar/gky310
49. Canueto D, Gomez J, Salek RM, Correig X, Canellas N. rDolphin: A GUI R package for proficient automatic profiling of 1D (1)H-NMR spectra of study datasets. Metabolomics. 2018;14(3):24. http://dx.doi.org/10.1007/s11306-018-1319-y
50. Hao J, Astle W, De Iorio M, Ebbels TM. BATMAN – An R package for the automated quantification of metabolites from nuclear magnetic resonance spectra using a Bayesian model. Bioinformatics. 2012;28(15):2088–90. http://dx.doi.org/10.1093/bioinformatics/bts308
51. Lewis IA, Schommer SC, Markley JL. rNMR: Open source software for identifying and quantifying metabolites in NMR spectra. Magn Reson Chem. 2009;47(Suppl 1):S123–6. http://dx.doi.org/10.1002/mrc.2526
52. Shi L, Westerhuis JA, Rosen J, Landberg R, Brunius C. Variable selection and validation in multivariate modelling. Bioinformatics. 2019;35(6):972–80. http://dx.doi.org/10.1093/bioinformatics/bty710
53. Spicer R, Salek RM, Moreno P, Canueto D, Steinbeck C. Navigating freely-available software tools for metabolomics analysis. Metabolomics. 2017;13(9):106. http://dx.doi.org/10.1007/s11306-017-1242-7
54. Westerhuis JA, van Velzen EJ, Hoefsloot HC, Smilde AK. Multivariate paired data analysis: Multilevel PLSDA versus OPLSDA. Metabolomics. 2010;6(1):119–28. http://dx.doi.org/10.1007/s11306-009-0185-z
55. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142. http://dx.doi.org/10.1186/1471-2164-7-142
56. Duan K, Keerthi SS, Poo AN. Evaluation of simple performance measures for tuning SVM hyperparameters. Neurocomputing. 2003;51:41–59. http://dx.doi.org/10.1016/S0925-2312(02)00601-X
57. Mahadevan S, Shah SL, Marrie TJ, Slupsky CM. Analysis of metabolomic data using support vector machines. Anal Chem. 2008;80(19):7562–70. http://dx.doi.org/10.1021/ac800954c
58. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinformatics. 2009;10:259. http://dx.doi.org/10.1186/1471-2105-10-259

59. Gaul DA, Mezencev R, Long TQ, Jones CM, Benigno BB, Gray A, et al. Highly-accurate metabolomic detection of early-stage ovarian cancer. Sci Rep. 2015;5:16351. http://dx.doi.org/10.1038/srep16351
60. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. http://dx.doi.org/10.1023/A:1010933404324
61. Chen T, Cao Y, Zhang Y, Liu J, Bao Y, Wang C, et al. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. Evid Based Complement Alternat Med. 2013;2013:298183. http://dx.doi.org/10.1155/2013/298183
62. Melo C, Navarro LC, de Oliveira DN, Guerreiro TM, Lima EO, Delafiori J, et al. A machine learning application based in random forest for integrating mass spectrometry-based metabolomic data: A simple screening method for patients with zika virus. Front Bioeng Biotechnol. 2018;6:31. http://dx.doi.org/10.3389/fbioe.2018.00031
63. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13(11):2498–504. http://dx.doi.org/10.1101/gr.1239303
64. Gao J, Tarcea VG, Karnovsky A, Mirel BR, Weymouth TE, Beecher CW, et al. Metscape: A cytoscape plug-in for visualizing and interpreting metabolomic data in the context of human metabolic networks. Bioinformatics. 2010;26(7):971–3. http://dx.doi.org/10.1093/bioinformatics/btq048
65. Fried JY, van Iersel MP, Aladjem MI, Kohn KW, Luna A. PathVisio-faceted search: An exploration tool for multi-dimensional navigation of large pathways. Bioinformatics. 2013;29(11):1465–6. http://dx.doi.org/10.1093/bioinformatics/btt146
66. van Iersel MP, Pico AR, Kelder T, Gao J, Ho I, Hanspers K, et al. The BridgeDb framework: Standardized access to gene, protein and metabolite identifier mapping services. BMC Bioinformatics. 2010;11:5. http://dx.doi.org/10.1186/1471-2105-11-5
67. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, et al. WikiPathways: A multifaceted pathway database bridging metabolomics to other omics research. Nucleic Acids Res. 2018;46(D1):D661–d7. http://dx.doi.org/10.1093/nar/gkx1064
68. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–d55. http://dx.doi.org/10.1093/nar/gkx1132

10

Feature Selection in Microarray Data Using Entropy Information

Ali Reza Soltanian[1] • Niloofar Rabiei[2] • Fatemeh Bahreini[3]

[1]Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran; [2]Department of Biostatistics, School of Public Health, Hamadan University of Medical Sciences, Hamadan, Iran; [3]Department of Molecular Medicine and Genetics, Faculty of Medicine, Hamadan University of Medical Sciences, Hamadan, Iran

Author for correspondence: Ali Reza Soltanian, Modeling of Noncommunicable Diseases Research Center, Hamadan University of Medical Sciences, Hamadan, Iran.
Email: soltanian@umsha.ac.ir

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch10

Abstract: Researchers in biological sciences and genetics are faced with high-dimensional data, such as the microarray data, and the analysis and proper interpretation of these data are very important in bioinformatics and systems biological sciences. In such types of data, the number of variables, for example, the genes, is many times greater than the number of samples. Therefore, the dimension of the data must be reduced at the primary point. Then, the analysis, for example, clustering, is performed on the compacted data. This process is called data summarization. There are various ways to summarize high-dimensional data, which depends on the nature of the data. The aim of data summarization is to remove unnecessary features so that the data are classified more accurately. Shannon's entropy information is a common method for clustering genes in microarray data and selecting a set of disease-related genes. This chapter introduces and illustrates statistical inference concepts of entropy in microarray data clustering to select a set of the most important genes associated with a disease.

Keywords: data mining; entropy; genetics; microarray; system biology.

In: Computational Biology. Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

To analyze high-dimensional data, many mathematical and statistical models have been developed. Most of these models focus on eliminating the unnecessary and unimportant features of a dataset, so that the clustered data are accurate. A popular source of clustering and modeling of high-dimensional data is the microarray data. The concept of systems biology has become more prominent in biological sciences (1). Systems biology is the science of summarizing data and detecting patterns among datasets. In other words, systems biology is the computational modeling of biosystems to interpret high and complex genetic data and other complex biological systems (2–4). Systems biology incorporates computational science, mathematics and statistics in the modeling of genetic and biological data (Figure 1). Entropy is one of the mathematical concepts that can be used in the modeling of systems biology data. In entropy, there are two concepts: entropy and information. Researchers usually do not distinguish between the two concepts. Entropy represents the irregularity (i.e., uncertainty) of a system, while information represents the difference between the maximum and the actual value of entropy of a system. In other words, information shows the correlation between two systems (e.g., two genes), which is derived from the entropy of the two systems and their subscription (5). Entropy application is a kind of mathematical challenge in analyzing biological data that can be important in determining relationships and clustering of results. Researchers have used entropy techniques to model cellular systems and study changes in gene expression patterns. In this chapter, the role of entropy to model the expression of genes in microarray data is discussed with emphasis on clustering, refinement and Shannon's entropy theory.

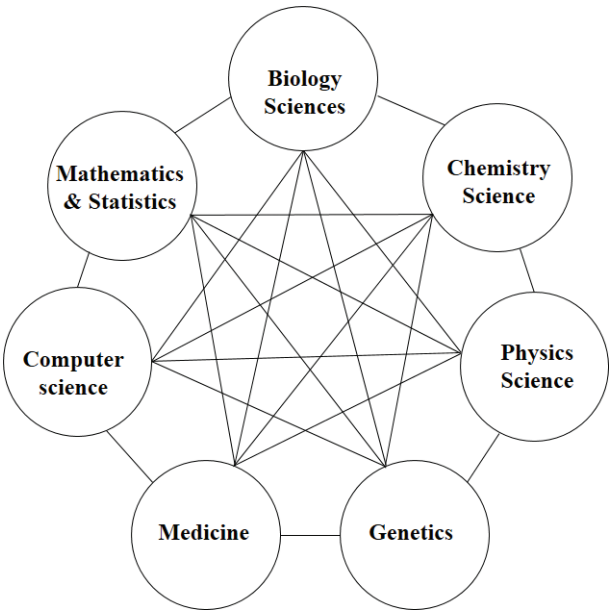


Figure 1 Schematic diagram for the concept of communication in systems biology.

DATA NORMALISATION METHOD

Data refinement is very important in the analysis of complex systems such as the microarray data. The calculation of gene expression in the microarray technique is based on the coloration of the genes, and problems associated with coloration are not uncommon. The occurrence of such problems leads to an unreasonable or artificial increase or decrease in the expression level of genes. A simple method to avoid outliers is to use *mean* ± *3SD* and, occasionally, *mean* ± *2SD* intervals. This approach eliminates the values outside of these ranges. Another approach to avoid staining errors in the microarray data is the fold-change criterion. Usually, this criterion is obtained based on the expression of a gene in healthy and diseased samples as follows:

$$Fold-change = 2^{\left|\log_2\left(\frac{Ave(C)}{Ave(N)}\right)\right|},$$

where the mean of gene expression levels in healthy and patient samples is indicated by *Ave*(*C*) and *Ave*(*N*), respectively. A cut-off is considered for the obtained fold-changes, so that fold-changes less than the cut-off are usually left out of the analysis process.

SHANNON'S ENTROPY

In the analysis of high-dimensional data, there are two approaches to estimating parameters and effects: the classical approach (i.e., frequentist) and the Bayesian approach. Entropy is a classical approach, and it indicates the degree of irregularity or uncertainty in a system. Uncertainty exists in many of the learning stages of high-dimensional data (6). Although the concept of entropy is used and defined in physics and mathematical sciences, we have attempted to determine a set of coordinates with the least irregularity in a signaling complex using the concept of entropy. Entropy is based on the concept of uncertainty, which means one is unconfident about the occurrence of a process. Therefore, increasing the uncertainty of a system means reducing the entropy of that system. In fact, evaluation, measurement and modeling of uncertainty that affects the whole process of data analysis have a significant impact on the learning performance of high-dimensional data. Without considering this uncertainty, the performance of learning strategies is sharply reduced. Claude E. Shannon, an American mathematician, introduced Shannon's entropy and information theory in 1948 under the title "A mathematical theory of communication" (7). In the concept of entropy, Shannon refers to the degree of uncertainty in the received information and expresses it with probability theory. Shannon's entropy in information theory is the criteria for measuring the uncertainty expressed by a probability distribution.

Note, information theory is the expectation value of information (i.e., mean) contained in each variable which can also be a gene. In other words, the entropy of each variable is the amount of its uncertainty. To calculate the uncertainty of a system, we must be able to formulate the probability of events in that system. Let us consider a

random experiment X (e.g., microarray data) with m events x_1, x_2,…, x_m, with the occurrence probabilities $p(x_1)$, $p(x_2)$,…, $p(x_m)$, respectively. In this case, we can consider x_i as the i^{th} gene, in a microarray dataset. The uncertainty of X (i.e., entropy) is represented by $H(X)$, and the function must depend only on the $p(x_i)$, $i = 1, 2,\ldots, m$. The formulation of $H(X)$ function should be the following properties.

Property 1: The desired function should not be dependent on the sequence of events (e.g., genes), hence:

$$H(p_1, p_2,\ldots, p_i, p_{i+1},\ldots, p_m) = H(p_1, p_2,\ldots, p_{i+1}, p_i,\ldots, p_m).$$

Note, the $H(X)$ function on any $p_i = p(x_i)$, $i = 1,2,\ldots,m$ must be continuous.

Property 2: Since the entropy function is continuous, so with a slight change in the probabilities $p(x_1)$, $p(x_2)$,…, $p(x_m)$, the amount of uncertainty (i.e., entropy value) will also change.

Property 3: If an event divides into two events, the original $H(X)$ function must be the sum of the weighed $H(X)$ functions.

Property 4: The entropy function $H(p_1, p_2,\ldots,p_i, p_{i+1},\ldots, p_m)$ should be established in the following equation:

$$H(p_1, p_2,..., p_m) = H(p_1 + p_2, p_3,..., p_m) + (p_1 + p_2)H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

Property 5: Let, two experiments X and Y with m and n events ($n < m$), respectively. If occurrence probability of the events in the two experiments is $\frac{1}{m}$ and $\frac{1}{n}$, then:

$$H\left(\frac{1}{m},...,\frac{1}{m}\right) \geq H\left(\frac{1}{n},...,\frac{1}{n}\right).$$

It can easily be shown that the entropy function has the above properties. Shannon's entropy can be defined for a random variable with a discrete or continuous distribution (7). In this section, we try to mention both together and illustrate the concept of entropy by several examples. Let a discrete random variable such as $X = \{x_1, x_2,\ldots, x_m\}$ with a probability mass function $p(x)$. The entropy of X is:

$$H(X) = \Sigma_{i=1}^{n}\ p(x_i)log_2\left(\frac{1}{p(x_i)}\right) = -\Sigma_{i=1}^{n}\ p(x_i)log_2 p(x_i) = -E[log(p_x)],$$

where $p(x_i) = \Pr(X = x_i)$ and is the probability of the i^{th} value of the random variable X.

Now, let a continuous random variable X. Usually, entropy for the continuous random variables is called the differential entropy. The entropy value for the continuous random variable X with the probability density function $f(X)$ is:

$$H(X) = H(f(x)) = E[-log(f(x))] = -\int f(x)log(f(x))dx,$$

where $0log0 = 0$.

The entropy may have a logarithmic base 2, 10 or Euler's number *e*. If the logarithmic base is 2 or *e*, then the entropy unit is "bit" or "nat," respectively. Here we should note that some physicists and mathematicians such as Lazare Carnot, Ludwig Eduard Boltzmann, and Rudolf Clausius have tried to introduce the concept of entropy theory, and others such as Claude Elwood Shannon were leading the introduction of entropy information theory (4, 7, 8).

Let, two random variables *X* and *Y* with probability density functions *f*(*X*) and *f*(*Y*) from the support regions *S* and *T*, respectively. The three entropies (i.e., *H*(*X*), *H*(*Y*), and *H*(*X, Y*)), the mutual information *I*(*X; Y*), and the entropy of *X* conditioned on *Y* and vice versa (i.e., *H*(*X*|*Y*) and *H*(*Y*|*X*)), are shown graphically in Figure 2.

The above features will be described below. The entropy function is fundamentally different from the maximum likelihood function. To further understand the entropy concept compared to maximum likelihood, let a Bernoulli random variable X with parameter p. The entropy value of the Bernoulli random variable is:

$$H(X) = -\Sigma_{X=0}^{1} p^X (1-p)^{1-X} log_b \left(p^X (1-p)^{1-X} \right) = -(1-p) log_b (1-p) - p log_2 (p),$$

where $b = 2,10$ *or* e; then,

$$H(X) = log_b \left(\frac{(1-p)^{p-1}}{p^p} \right).$$

The Bernoulli entropy function has the highest value when $p = \frac{1}{2}$. Here are some simple examples for understanding entropy and calculating it.

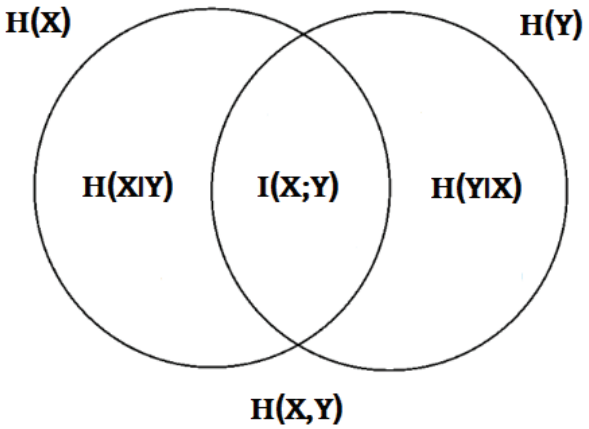


Figure 2 Features of Shannon's entropy function.

Example 1

Let us assume X is the presence or absence of a *G* allele in the CAPN10 gene, which associate with type 2 diabetes mellitus. The *G* and *A* alleles are detected with the probability *p* and (1–*p*), respectively, that is,

$$X = \begin{cases} 1 & \text{with prob } p; \text{ for "G" allele} \\ 0 & \text{with prob } (1-p); \text{ for "A" allele} \end{cases}$$

For various values of *p*, that is, *G* allele frequency, the entropy value for the random variable X is shown in Table 1. In fact, when *p* is closer to 0.5, the uncertainty level over the *G* allele is increased, and thus, the amount of information about the test will be increased. In this example, we obtain an entropy *G* allele with $p = 0.25$.

$$H(X) = -\sum_{x=0}^{1} p^x (1-p)^{1-x} ln\left(p^x (1-p)^{1-x}\right)$$

$$= (0.25^1 \ (1-0.25)^{1-1} \ ln[0.25^1(1–0.25)^{1-1}])+$$

$$(0.25^0)(1-0.25)^{1-0} \ ln[0.25^0(1\text{-}0.25)^{1-0}]) = 0.56 \ nat \ .$$

Example 2

In this example, we will show how to calculate the Shannon's entropy information, which is a kind of dependency between variables using discrete expression profile. Now, suppose the discrete expression profile for two genes *A* and *B* is [1, 1, 0,-1, 0] and [1, -1, 0, 1, 1], respectively. The occurrence probability of each mode for the genes is presented in Table 2.

Therefore, the amount of entropy for the gene *A* and gene *B* is:

$$H(A) = -\Sigma_{x=0}^{3} P_x ln(p_x) = -\left[\left(\frac{2}{5}\times ln\left(\frac{2}{5}\right)\right)+\left(\frac{2}{5}\times ln\left(\frac{2}{5}\right)\right)+\left(\frac{1}{5}\times ln\left(\frac{1}{5}\right)\right)\right] = 1.05 \ nat,$$

$$H(B) = -\Sigma_{x=0}^{3} P_x ln(p_x) = -\left[\left(\frac{3}{5}\times ln\left(\frac{3}{5}\right)\right)+\left(\frac{1}{5}\times ln\left(\frac{1}{5}\right)\right)+\left(\frac{1}{5}\times ln\left(\frac{1}{5}\right)\right)\right] = 0.95 \ nat.$$

TABLE 1 **The five various values and entropies of G allele of CAPN10 gene**

p:	0	0.25	0.50	0.75	1
H(*X*):	0	0.56	0.69	0.56	0

TABLE 2 Frequency distribution of the expression profile of A and B genes

Gene:	$P(1)$	$P(0)$	$P(-1)$
"A"	2/5	2/5	1/5
"B"	3/5	1/5	1/5

To calculate $H(A,B)$, the nine possible combinations with respect to the joint probabilities $P(A,B)_s$ should be considered as follows:

$$P(1,1)=\frac{1}{5};\quad P(1,0)=0;\quad P(1,-1)=\frac{1}{5},$$

$$P(0,1)=\frac{1}{5};\quad P(0,0)=\frac{1}{5};\quad P(0,-1)=0,$$

$$P(-1,1)=\frac{1}{5};\quad P(-1,0)=0;\quad P(-1,-1)=0,$$

then $H(A, B) = 1.61$. Finally, the mutual information between the two expression profiles A and B is:

$$I(A, B) = H(A) + H(B) - H(A, B) = 1.05 + 0.95 - 1.61 = 0.39\ nat.$$

Note, high levels of mutual information suggest similarity between two expression profiles.

In addition to the mentioned concepts, one of the important entropy rules for random variables (*iid*) is the *asymptotic equipartition property* (AEP) theorem, which points out that the joint probability of a sequence of random variables, that is, $p(X_1, X_2,\ldots, X_n)$, is very close to $2^{-nH(X)}$.

Let $X_1, X_2,\ldots, X_n$ be a sequence of *iid* random variables with a probability of density function $f(X)$, then:

$$-\frac{1}{n}\log\left(f\left(X_1,X_2,\ldots,X_n\right)\right)\xrightarrow{p}E\left(-\log\left(f(X)\right)\right)=H(X).$$

The above definition leads to the definition of a typical set $A_{\epsilon}^{(n)}$; so that $\epsilon > 0$ and $\forall$n, the usual set $A_{\epsilon}^{(n)}$ is defined as follows:

$$A_{\epsilon}^{(n)}=\left\{\left(x_1,x_2,\ldots x_n\right)\in S^n:\left|-\frac{1}{n}\log\left(f\left(x_1,x_2,\ldots,x_n\right)\right)-H(X)\right|\le\epsilon\right\},$$

where

$$f\left(x_1,x_2,\ldots,x_n\right)=\Pi_{i=1}^{n}f\left(x_i\right).$$

In the context of entropy, we encounter another concept called joint and conditional differential entropy. In other words, the differential entropy for a set of

n random variables $X_1, X_2, \ldots, X_n$ with the density function $f(x_1, x_2, \ldots, x_n)$ is defined as follows:

$$H(X_1, X_2, \ldots, X_n) = -\int f(x_1, x_2, \ldots, x_n)\ log(f(x_1, x_2, \ldots, x_n))\ dx_1\, dx_2 \ldots dx_n.$$

For example, suppose that n random variables $X_1, X_2, \ldots, X_n$ have a multivariate normal distribution with mean vector $\mu_{n\times 1}$ and a variance–covariance matrix Σ, then the entropy of a multivariate normal distribution is:

$$H\left(X_1, X_2, \ldots X_n\right) = \frac{1}{2}\log\left(2\pi e\right)^n |\Sigma|,$$

where $|\Sigma|$ is the determinant of variance–covariance matrix Σ.

On the other hand, if X and Y are two random variables (e.g., two genes) with a joint density function $f(X, Y)$, then their conditional differential entropy indicated by $H(X, Y)$ is defined as follows:

$$H(X|Y) = -\int f(x,y)\log(f(x|y))\ dxdy.$$

Since

$$f\left(x|y\right) = \frac{f\left(x,y\right)}{f\left(y\right)}$$

can be written as

$$H(X|Y) = H(X, Y) - H(Y).$$

In choosing a set of random variables (e.g., a set of related genes), we must use two concepts of relative entropy and mutual information, which are referred to next. The relative entropy for continuous random variables X and Y with probability density functions $f(X)$ and $g(Y)$ is equal to:

$$D\left(f||g\right) = E_f\left[\log\left(\frac{f(x)}{g(x)}\right)\right] = \int f(x)\log\left(\frac{f(x)}{g(x)}\right)dx.$$

Note that relative entropy is always non-negative, that is,

$$D(f||g) \geq 0.$$

The mutual information for the two continuous random variables X and Y with the joint probability density function $f(X, Y)$ is:

$$I\left(X;Y\right) = E\left(log\left(\frac{f(X,Y)}{f(X)f(Y)}\right)\right) = \int f\left(x,y\right) log\left(\frac{f\left(x,y\right)}{f(x)f\left(y\right)}\right)dxdy.$$

For simplicity, the mutual information can be written as follows:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X),$$

hence:

$$I(X; Y) = D(f(x, y)||f(x)f(y)).$$

Note,

$$I(X; Y) \geq 0,$$

$$H(X|Y) \leq H(X).$$

In $H(X|Y) \leq H(X)$, equality will be achieved if and only if X and Y are independent.

Example 3

Now, let two gene expressions corresponding to A and B genes as two random variables, which they have bivariate normal distribution as follows:

$$\begin{pmatrix} A \\ B \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma\right), \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}.$$

Suppose that the gene expressions of two genes A and B for three tumor tissues are:

Then, the measures of entropies $H(A)$, $H(B)$ and $H(A,B)$ are:

$$H(A) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(A-\mu_1)2}{2\sigma_1^2}} ln\left(\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(A-\mu_1)^2}{2\sigma_1^2}}\right) dA$$

$$= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{(A-\mu_1)^2}{2\sigma_1^2}} \left(-ln\left(\sqrt{2\pi\sigma_1^2}\right) - \frac{(A-\mu_1)^2}{2\sigma_1^2}\right) dx$$

$$= \ln\left(\sqrt{2\pi\sigma_1^2}\right) + \frac{1}{2\sigma_1^2} \int_{-\infty}^{\infty} \frac{(A-\mu_1)^2}{2\sigma_1^2} e^{-\frac{(A-\mu_1)^2}{2\sigma_1^2}} dx$$

$$= \frac{1}{2} ln\left(2\pi\sigma_1^2\right) + \frac{\sigma_1^2}{2\sigma_1^2} = \frac{1}{2}\left(2\pi e\sigma_1^2\right)$$

$$= \frac{1}{2} + \frac{1}{2} ln(2\pi) + ln\left(\sigma_1^2\right)$$

Note, the above equation shows that the high variance increases the measure of entropy or uncertainty.

Consider the gene expression levels in Table 3 for the two genes and three tissues. Therefore, $\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}=\begin{pmatrix}2.48\\ 2.40\end{pmatrix}$, and

$$\Sigma=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}=\begin{pmatrix}1.66 & (0.98)(1.29)(1.49)\\ (0.98)(1.29)(1.49) & 2.22\end{pmatrix}=\begin{pmatrix}1.66 & 1.88\\ 1.88 & 2.22\end{pmatrix}$$

In this case, $H(A) = 1.93$, $H(B) = 2.22$ *nat*. In addition, $H(A, B)$ calculate as follows:

$$H(A,B)=\frac{1}{2}ln\left[(2\pi e)^2|\Sigma|\right]=1+ln(2\pi)+ln(\sigma_1\sigma_2)+\frac{1}{2}ln(1-\rho^2)=1.88\ nat.$$

Gel'fand and Yaglorn (9) showed that an exact relationship between entropy information, $I(A,B)$, and the correlation coefficient for A and B gene, r is:

$$I(A,B)=H(A)+H(B)-H(A,B)=-\frac{1}{2}ln(1-\rho^2)=1.61.\ nat.$$

The important limitation of entropy information is that its upper limit is unknown, that is, $I(X, Y) \in (0, +\infty)$ Therefore, an index to measure the correlation of two random variables based on entropy information should be introduced, which does not have this limitation. The normalized mutual information, $U(X,Y)$, has such property. The normalized mutual information concept, $U(X,Y)$, is used to choose a set of correlated variables using the uncertainty function, which is shown for two random variables (e.g., two genes) X and Y as follows:

$$U(X,Y)=\frac{2I(X,Y)}{H(X)+H(Y)},$$

where $0 \le U(X,Y) \le 1$. The value $U(X,Y)$ close to zero means that the two random variables X and Y have a high mutual relevance, that is, relation, while the value $U(X,Y)$ close to 1 means that the two random variables have a low mutual

TABLE 3 The gene expression levels for three tumor tissues

	Tissue 1	Tissue 2	Tissue 3
Gene "*A*"	1.12	2.65	3.68
Gene "*B*"	0.98	2.28	3.95

relevance, that is, independence (4, 10). Therefore, $U(A,B)$ with respect to data in example 3 is 0.78 *nat*, which it is a low mutual relevance. For further study on entropy and its properties, we suggest two books: *Handbook of Statistical System of Biology* and Elements of Information Theory (4, 8).

APPLICATION

In this section, we use the results of Bahreini et al. (11) which extracted the information (i.e., gene expression) from the study of Notterman et al. (12). In their study, 18 adenocarcinoma colon and 18 normal tissue samples from the Cooperative Human Tissue Network were evaluated. In that research, the mean (±SD) age of the patients was 67.56 (±14.09) years. Of the total 7465 available cDNAs, only 3,228 genes had fold change more than one and they were selected for analysis. Shannon's entropy method was used to select an appropriate set of genes associated with colon cancer, and 29 genes with the highest amount of information were finally selected.

Before using entropy to select a gene set associated with colon cancer based on gene expressions, the hierarchical method was used for clustering of the genes (Figure 3). The figure shows that 3128 selected genes are shared in three clusters. However, the hierarchical cluster analysis dendrogram shows that the frequency

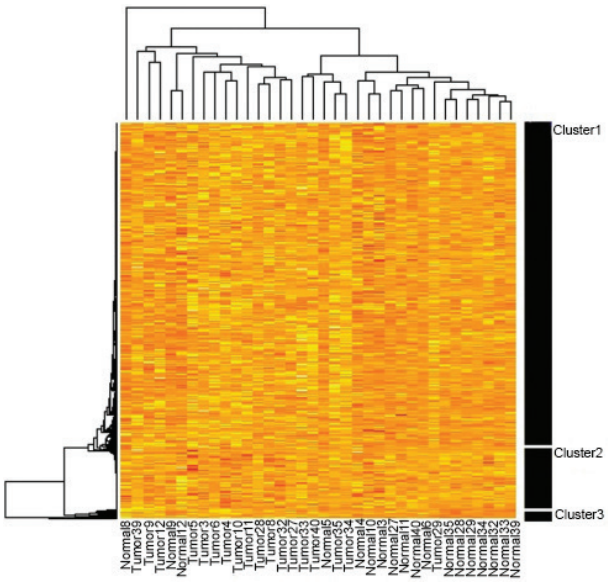


Figure 3 Cluster map derived from a two-way cluster analysis by the hierarchical method. Approximately 3000 common genes in tumor tissues and paired normal tissues were combined in a matrix. Clustering was performed on this matrix. Each color patch on the cluster map indicates the expression intensity level of the associated gene in that tumour and normal tissue samples. The color patches on the cluster map have continuity on expression levels from yellow (highest) to red (lowest) (11).

of the yellow points (i.e., high gene expression) in tumor samples is higher than normal samples, but it is simply not possible to identify a set of most relevant genes with colon cancer recognition. In other words, in Figure 3 there is no specific visible pattern in the color spectrum. Usually, clustering is appropriate when a specific spectrum of colors can be found in normal and tumor samples. Therefore, although the genes are shared into three clusters in Figure 3, the obtained result is not accurate. One of the problems may be the lack of refinement of the levels of gene expression. In studies on gene expression analysis, data refinement process and the removal of outlier values are very important.

Figure 4 demonstrates the importance of refining the data. The data refinement methods are numerous and varying. For example, in analyzing microarray data, the gene expression levels obtained may be very large. In this case, fold change can be used to refine the data. Due to the choice of a suitable cut-off point in the fold-change index, we can omit the outside domain data from the analysis to yield more accurate results. It should be emphasized that we were not able to find a proper and accurate statistical method for choosing the fold-change critical point. In Bahreini et al.'s study (11), 29 genes were selected from 3128 genes after performing Shannon's entropy information to determine a collection of the most relevant genes associated with colon cancer. Usually, for graphical representation of the gene expression levels, a dendrogram plot was used. Figure 4 shows that the 29 selected

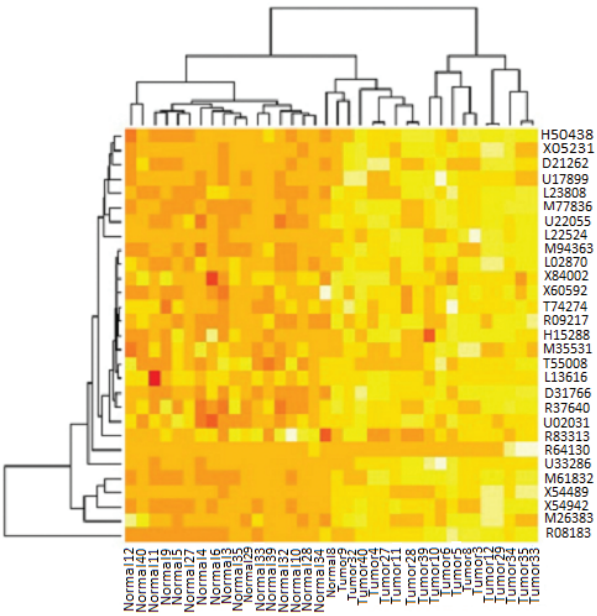


Figure 4 Cluster map derived from two-way cluster analysis with the hierarchical method. We combined 29 common genes in tumor and normal tissues in a matrix. Clustering was performed on this matrix. Each color patch on the cluster map indicates the expression intensity level of the associated gene in that tumor and normal tissue samples. The color patches on the cluster map have continuity on expression levels from yellow, that is, highest, to red, that is, lowest (11).

genes are shared into two clusters by Shannon's entropy method. By comparing two dendrograms (Figures 3 and 4), it can be seen easily that in the second dendrogram (Figure 4), the gene expression in tumor samples is far more than in normal samples, while such a difference was not obvious in the first dendrogram (Figure 3).

CONCLUSION

To reduce dimension in the microarray data and to prevent common errors in statistical modeling, many methods have been introduced, and entropy is one of the most widely used concept in medical and genetic sciences. Entropy was introduced by Nicholas Georgescu-Roegen in 1971 and later developed by scientists based on the principles established by Shannon. Shannon had a major role in introducing entropy information, which has been widely used in high-dimensional studies. One of the advantages of entropy is that calculation of values is based on theoretical forms, not the empirical and personal concepts. These values give small or large weights, proportional to the small or large actual values. Where researchers seek to estimate the risk from an agent, the level of uncertainty is the basis of the computational form of the risk value (Risk = Uncertainty + Damage, where "Damage" in the equation shows the measure of the loss). The function of conventional feature selection algorithms is based more on the choice of the ones that have the most connection with the target class and the least redundancy among the selected features. The major disadvantage of these algorithms is that they ignore the dependencies between the candidate and the unselected feature. However, based on Shannon's entropy information, we can introduce a theoretical algorithm that does not display such disadvantages. Although entropy is often used as a feature of the information concept, it is crucially dependent on the probability model.

Conflict of Interest: The authors declare no potential conflict of interest with respect to research, authorship, and/or publication of this chapter.

Copyright and permission statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s).

REFERENCES

1. Stadtländer CTK-H. Systems biology: Mathematical modeling and model analysis. J Biol Dyn. 2018;12(1):11–15. http://dx.doi.org/10.1080/17513758.2017.1400121
2. Westerhoff HV, Winder C, Messiha H, Simeonidis E, Adamczyk M, Verma M, et al. Systems biology: The elements and principles of life. FEBS Lett. 2009;583:3882–90. http://dx.doi.org/10.1016/j.febslet.2009.11.018
3. Breitling R. What is systems biology? Front Physiol. 2010;1:9. http://dx.doi.org/10.3389/fphys.2010.00009
4. Stumpf MPH, Balding DJ, Girolami M. Handbook of statistical systems biology. 1st ed. Chichester: Johan Wiley & Sons, Ltd., The Atrium; 2011.

5. Adami C. Information theory in molecular biology. Phys Life Rev. 2004;1:3–22. http://dx.doi.org/10.1016/j.plrev.2004.01.002
6. Wang X-Z, Xing H-J, Li Y, Hua Q, Dong C-R, Pedrycz W. A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning. IEEE Trans Fuzzy Syst. 2015;23(5):1638–54. http://dx.doi.org/10.1109/TFUZZ.2014.2371479
7. Shannon CE. A Mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423. http://dx.doi.org/10.1002/j.1538-7305.1948.tb01338.x
8. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons; 2012.
9. Gel'fand IM, Yaglom AM. Calculation of amount of information about a random function contained in another such function. Am Math Soc Transl. 1957;2(12):199–246. English translation of original in Uspekhi Matematicheskikh Nauk 12(1):3–52.
10. Cover TM, Thomas JA. Elements of information theory. New York: John Wiley & Sons, Inc.; 1991. http://dx.doi.org/10.1002/0471200611
11. Bahreini F, Soltanian AR. Identification of a gene set associated with colorectal cancer in microarray data using the entropy method. Cell J. 2019;20(4):569–75.
12. Notterman DA, Alon U, Sierk AJ, Levine AJ. Transcriptional gene expression profiles of colorectal adenoma, adenocarcinoma, and normal tissue examined by oligonucleotide arrays. Cancer Res. 2001;61(7):3124–30.

11

Template-Based and Template-Free Approaches in Cellular Cryo-Electron Tomography Structural Pattern Mining

Xindi Wu[1] • Xiangrui Zeng[1] • Zhenxi Zhu[2] • Xin Gao[3] • Min Xu[1]

[1]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, USA; [2]Beijing University of Posts and Telecommunications, Beijing, China; [3]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal, Saudi Arabia

Author for correspondence: Min Xu, Computational Biology Department, Carnegie Mellon University, Pittsburgh, 15213, PA, USA. Email: mxu1@cs.cmu.edu

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch11

Abstract: Cryo-electron tomography (Cryo-ET) has made possible the observation of cellular organelles and macromolecular complexes at nanometer resolution in native conformations. Without disrupting the cell, Cryo-ET directly visualizes both known and unknown structures in situ and reveals their spatial and organizational relationships. Consequently, structural pattern mining (a.k.a. visual proteomics) needs to be performed to detect, identify and recover different sub-cellular components and their spatial organization in a systematic fashion for further biomedical analysis and interpretation. This chapter presents three major Cryo-ET structural pattern mining approaches to give an overview of traditional methods and recent advances in Cryo-ET data analysis. Template-based, supervised deep learning-based and template-free approaches are introduced in detail. Examples of recent biological and medical applications and future perspectives are provided.

Keywords: cryo-electron tomography; deep learning; macromolecular complexes; structural pattern mining; visual proteomics

In: *Computational Biology*. Holger Husi (Editor), Codon Publications, Brisbane, Australia. ISBN: 978-0-9944381-9-5; Doi: http://dx.doi.org/10.15586/computationalbiology.2019

INTRODUCTION

Cryo-electron tomography (Cryo-ET) is a powerful diagnostic and research tool that combines specimen cryo-fixation and multi-angle electron microscopy imaging (1), which enables structural biologists to produce three-dimensional (3D) volume reconstructions of near-native state cells and determine the structure of sub-cellular structures with molecular-scale resolution (2). Those images contain information about the 3D cell structure projected into a single plane. In order to recover the actual 3D arrangement of components in the specimen, the information in 2D projection images should be integrated computationally. Cryo-EM has experienced a dramatic increase in the attainable resolution of 3D reconstructions. Complexes with high intrinsic contrast, such as ribosomes, have been successfully analyzed. The discrete conformation of membrane receptors can be recognized, which provides a theoretical basis for exploring the structural basis of signals in the whole cell.

In recent years, the amount of information about molecular roles involved in cellular processes has increased dramatically, and it has become possible to detect or obtain cellular tomograms with information about macromolecular complexes, their structures and spatial positions in the cell. Proteomics, based on genomics and mass spectrometry, has carried out a comprehensive analysis of the cell proteome. Nevertheless, it is still very challenging to discover the structure of unknown complexes in tomograms. Due to various shapes, sizes, cellular abundance of unlabeled complexes, high crowding levels, limitations of template libraries, low signal-to-noise ratio (SNR) and the limited range of tilt angles (3), the structural discovery can be detected by structure pattern mining methods. The methods for molecular separation and purification for structural and functional studies have been a great success. This chapter focuses on three major Cryo-ET structural pattern mining approaches, giving an overview of traditional methods and recent advances in Cryo-ET data analysis. Template-based, supervised deep learning-based and template-free approaches are included. Examples of recent applications in biology and medical field along with future perspectives are discussed.

TEMPLATE-BASED STRUCTURAL PATTERN MINING

Template search/match has been the most popular template-based method for detecting spatial location and orientation of a known structure of interest. The visual proteomics method is capable of identifying individual protein complexes in intact cells (4). A guide on how to implement the visual proteomics method was proposed by Förster and colleagues (5). There are three main processing steps included in this method. First, a library containing the reference structure of the target protein complexes resampled to the relevant electron optical conditions is assembled. Second, for all possible positions and directions, the local cross-correlation coefficient between each reference structure and tomogram is calculated and stored in the cross-correlation volume. Finally, the distribution of

cross-correlation values in these volumes is translated into a list of locations by peak extraction and statistical methods. For data acquisition, an experimental setup is required. It is desirable to obtain the highest quality frozen electron tomography in terms of SNR, and the dosage spent on the specimen during the acquisition should be well controlled. Recommendations on choosing acquisition parameters after analyzing the different factors on signal content in Cryo-ET can be found in this work (5).

The template-matching process includes four parts: handling MolMatch, creating motif lists, scoring and visualization of molecular atlases. The scoring function (SF) for visual proteomics (6) relies on three different knowledge-based, empirical readouts. Besides, they also discuss other technical improvements. To detect low-abundance protein complexes with confidence, data acquisition and post-processing should be paid enough attention. The signal in visual proteomics is due to the contrast given by the surrounding solvent, so a better dosage control during the data acquisition is unavoidable. Also, phase plates may help to obtain a better performance. The electron dose that can be applied to the specimen limits the resolution that can be achieved by Cryo-ET, which leads to the limitation of the Cryo-ET application in visual proteomics.

It is important to assess whether the subtomogram, which is a small 3D cubic subvolume containing one macromolecular complex, or the recovered structure can be matched to a particular known structure. A template-match calculates the structural correlation between a subtomogram or a recovered structure with a known structural template. However, a simple correlation score cannot fully conclude the template matching. Rigorous statistical tests need to be carried out. Wang et al. (7) proposed a Monte Carlo sampling hypothesis testing framework based on generative adversarial network modeling for assessing template matching results. First, a generative adversarial network is constructed by using known structures to generate the structural distribution of macromolecular complexes. The structural generator is trained to the extent that the discriminator cannot distinguish between a known structure and a pseudo one. Second, a large number of pseudo macromolecular complexes are generated from the learned structural distribution in a Monte Carlo sampling fashion. Third, the subtomogram or recovered structure of interest is compared to the known structure and pseudo structure to assess the statistical confidence of template matching. This method computes not only a correlation score of template matching but also the P-value of whether the structure is significantly close to the template. Such a statistical assessment provides rigorous evidence of template matching and reduces its false-positive rate.

SUPERVISED SUBTOMOGRAM CLASSIFICATION AND SEGMENTATION

Since 2017, supervised deep learning approaches, including classification and semantic segmentation, have been applied to Cryo-ET.

Semantic segmentation using convolutional neural networks

The first deep learning-based semantic segmentation framework proposed for Cryo-ET data (8) classifies tomogram 2D slices in a voxel-wise binary fashion. With training data voxel-labeled manually, it predicts the segmentation mask of ribosomes, mitochondrial membrane, microtubules, and vesicles. To facilitate the prediction of membrane structures in different orientations, data augmentation was integrated into the training process with a moderately increased computational cost.

3D ConvNet (9) is a 3D semantic segmentation model for Cryo-ET data based on the U-Net architecture. 3D ConvNet predicts the segmentation mask of ribosomes, membrane and membrane-bound ribosomes in a multi-class fashion. As a result, the computational time does not increase linearly with the increase of class numbers (8).

Two 3D semantic segmentation convolutional neural networks (SSN3D) and their variants segmenting the main structural region from subtomograms have been proposed (10). This is a very useful step in subtomogram analysis because masking out neighboring structures can significantly reduce the structural bias for further analysis such as averaging and classification. Inspired by encoder-decoder type segmentation networks and fully connected networks, the networks are designed to be an encoder connected to a decoder inputting both high-level features and low-level features. The encoder is designed with alternation of convolution layers and max-pooling layers. The decoder is designed with alternation and convolution layers and upsampling layers. By combining different types of layers and combining both high-level and low-level feature information, the two segmentation networks can achieve high accuracy in 3D subtomogram semantic segmentation tasks.

Subtomogram classification

Similarly, a deep learning-based particle subdivision approach (11) proposes two convolutional neural networks, namely Inception3D network and DSRF3D network, for subdividing heterogeneous subtomograms into some homogeneous subsets. After extracting features from subtomograms using the Inception3D or DSRF3D network, unsupervised clustering can be applied. Furthermore, it was demonstrated that the generalization ability of models of novel structures that do not exist in the training data can still be discovered and clustered. Based on this work, Che et al. (12) proposed three convolutional neural networks with promising classification performance for datasets of extremely low SNR (0.01): (i) DSRF3Dv2, an extended version of DSRF3D (Deep Small Receptive Field 3D); (ii) RB3D, a 3D residual block-based neural network; and (iii) CB3D, a convolutional 3D (C3D)-based model, with improved classification accuracy. Among these, the CB3D achieves the best performances and yields accuracy close to 0.9 for normal datasets.

Model compression

Guo et al. (13) proposed a model compression approach for Cryo-ET data. Based on the deep neural network employed for the classification of

subtomograms, knowledge distillation to compress such networks was used. In previous works related to model compression on 2D images (14, 15), a model compression approach through knowledge distillation was proposed in order to speed up the separation of macromolecules at the prediction stage. The DSRF3D-v2 (Deep Small Receptive Field) model was chosen for compression considering the processing time and performance among the pre-existing models (12). Three student networks have been proposed to reduce the number of layers and parameters. The student network is a simplification of the teacher network. The compression includes compressing convolutional layers, pooling layers and eliminating one of the two fully connected layers. Reduction of the number of filters leads to simpler convolutional layers. DSRF3D-v2-s1 achieved the best performance by increasing the pooling size and stride from 2×2×2 to 3×3×3 and dismissing the dropout layers and one fully connected layer. Usually, a higher compression rate will lead to a greater loss of accuracy. Distilled models proposed in this study reduce the number of parameters, time and cost, and improve accuracy.

Domain adaptation

For Cryo-ET, it is usually time-consuming and computationally intensive to create valid training data due to a massive demand for labeled training data. Obtaining training data from a separate data source where the annotations are readily available or can be executed in a high-throughput manner would be beneficial. The challenge is that the cross-data source prediction is often biased due to the different image-intensity distributions (a.k.a. domain shift). Domain adaptation has been shown to be beneficial for addressing this challenge. Lin et al. (16) adopted an adversarial domain adaptation framework called 3D-ADA for the structural classification of macromolecules captured by Cryo-ET. In order to obtain a robust model for a cross-data source macromolecular structural classification, this framework utilizes 3D convolutional neural networks and adversarial learning, mapping subtomograms into a latent space shareable between separate domains. Also, the training-feature extractions on multiple source domains can extend 3D-ADA to utilize multiple training data sources of Cryo-ET. Covariate shift is a typical case of domain shift (17). Compared to the original adversarial domain adaptation method (18), they have several modifications: (i) 2D CNNs to 3D with new 3D network architectures for Cryo-ET; (ii) two independent feature extractors to extract features from both source and target domains, making target domain features more robust; (iii) independent feature extractor for target data to enable the target domain feature to be more flexible and robust; and (iv) gradient forwarding of adversarial loss function. To avoid a local minimum for the model, the adversarial loss uses the proper domain supervision information for both the adversarial discriminator and the feature extractor training that avoids gradient vanishing in back-propagation.

Simultaneous classification and segmentation by multi-task learning

Built on the above semantic segmentation and classification model, a multi-task learning neural network model was proposed (19) to perform semantic

segmentation, classification and coarse structural recovery (regression) simultaneously. The feature extraction layers are shared, which later split into three networks to perform each individual task independently; the loss of each task is linearly combined. This network design allows the training of the three tasks to mutually reinforce each other for better feature extraction and therefore higher accuracy. The accuracy of this model for classification and semantic segmentation outperformed single-task models (10, 11).

TEMPLATE-FREE STRUCTURAL PATTERN MINING

Template-based methods have their own shortcomings due to the possibility that the template structure can misfit its targets. If the template structures come from different organisms, there will be different conformations in the template structures. Also, the conformational changes or additional bound components to the structure in vivo can be challenging to template-based methods. Under such circumstances, several template-free structural pattern mining methods have been proposed recently.

De novo structural pattern mining via multi-pattern pursuit

A framework called multi-pattern pursuit (MPP) was designed for discovering frequently occurred structural patterns in Cryo-ET (3). It formulates the template-free visual proteomics analysis as a de novo pattern mining problem. The aim of MPP is to cluster, detect and estimate the abundance of large-scale complexes inside single cells automatically.

In this framework, first, after subtomograms are detected using template-free particle picking methods (20, 21), feature patterns are initialized. Second, initialized feature patterns are assessed for adding to the pattern library. Third, patterns are selected and aligned into common frames. Fourth, subtomograms are aligned with the candidate patterns and redundant patterns are discarded. The steps are iterated until high-quality patterns are distinguished and further refined individually. Therefore, representative and abundant patterns in a tomogram can be discovered without templates of known structures. Moreover, after patterns are successfully discovered, they can be embedded in the tomogram to visually present or statistically analyze their spatial distributions and interactions.

After the above subtomogram data processing steps, one of the most crucial steps is to average and cluster the subtomograms. To recover the structure of macromolecular complexes inside subtomograms which are heavily distorted by the noise and missing wedge effects, the use of a large number of subtomograms (usually more than a thousand) containing the same structure and averaging is recommended. The averaging process includes rotating and translating each subtomogram in a reference-free fashion because guidance from the known structures may bias the structural recovery. Only if the macromolecular complexes in the subtomograms are rotated and translated to a homogeneous orientation, and centered, the underlying true structure can be fully recovered

from averaging. The task becomes more challenging when there are multiple classes of subtomograms, meaning that different subtomograms may contain different macromolecular complexes. Averaging as well as clustering need to be refined during each iteration (22–24).

Autoencoder for mining abundant and representative features

A convolutional autoencoder-based unsupervised approach has been proposed (25) for coarse selection of subtomograms of interest from a large number of subtomograms (scale ranging from thousands to millions). After subtomograms are extracted from the tomogram using particle picking methods, an optional pose normalization approach is provided to adjust the particle orientation and center for better clustering of the same structures of different orientations, which simplifies the process of structural mining. It also assists the image-features characterization in a less orientation-dependent way. A convolutional neural network is designed for encoding each subtomogram into a feature vector and decoding the feature vector to reconstruct the tomogram. K-means clustering algorithms and autoencoder networks are combined together to cluster Cryo-ET small-subvolumes into sets with homogeneous image features. Subtomogram-cluster centers are decoded and plotted for visual guidance for selection. Therefore, selecting among a large number of subtomograms becomes selecting among a few (usually less than a hundred) clusters. Interesting clusters, such as clusters of membrane features or globule features can be selected for further analysis. In addition, we designed a weakly supervised semantic segmentation convolutional neural network to which results from the convolutional autoencoder can be applied.

As illustrated in Figure 1, the autoencoder and MPP can be integrated into an unsupervised pipeline for template-free recovery of representative structures. Although the current template-free structural pattern mining approach is still more of proof-of-concept, when more and more unsupervised methods are developed in the future, we expect a powerful system of template-free methods to accurately and efficiently detect and recover both the known and unknown representative structures in a systematic fashion (26).

CRYO-ET BIOLOGICAL AND MEDICAL APPLICATIONS

Due to the ability of Cryo-ET that can reveal the native structure and arrangement of macromolecular complexes inside intact cells, a lot of investigations have used this method to better understand the structural information of cells. Recently, there has been growing interest in establishing Cryo-ET as a diagnostic approach to complement conventional methods. Some recent examples of structural discovery and medical application using Cryo-ET are discussed in the following section. The increasing number of Cryo-ET medical applications demonstrates the potential of establishing Cryo-ET as a powerful diagnostic tool.

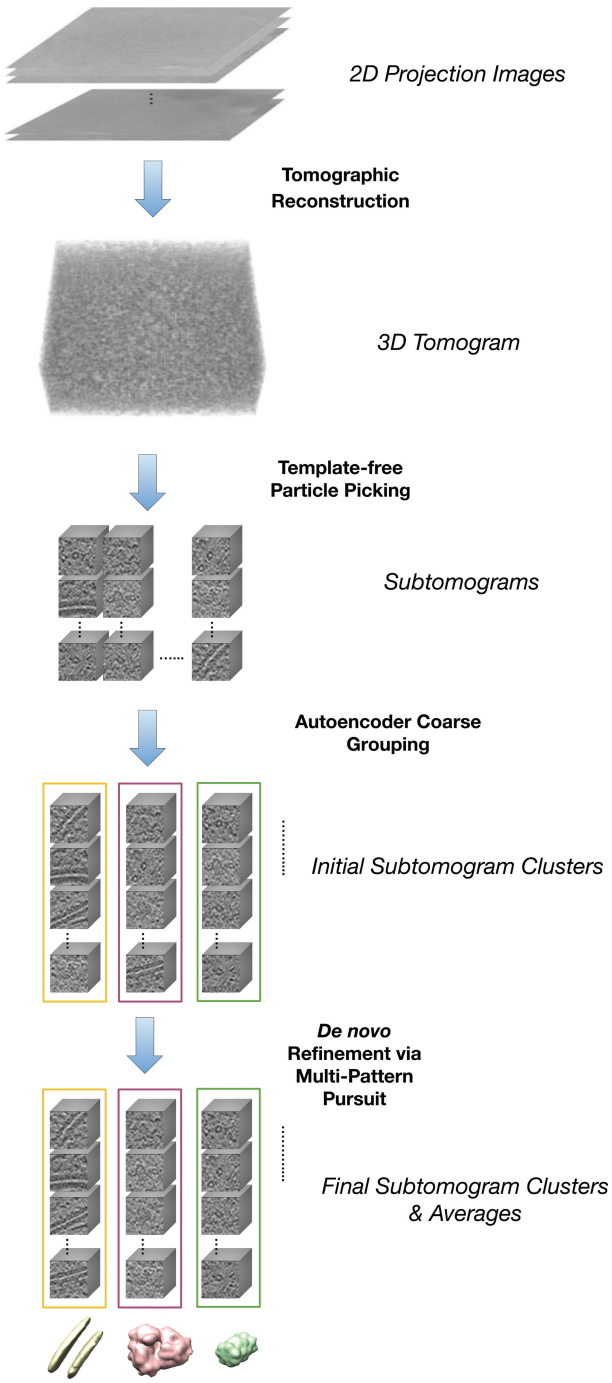


Figure 1 Unsupervised structural pattern mining pipeline integrating the autoencoder and multi-pattern pursuit approach.

Bacterial cell biology

Cryo-ET is capable of discovering detailed structure of bacterial cells in their native environment. Bacteria are viewed as structurally complex assemblies of macromolecular machines rather than undifferentiated bags of enzymes. This organization includes highly ordered arrays of chemosensory components (27, 28). Cryo-ET further enables the characterization of microcompartments for optimizing metabolism and storing nutrients (29–33). A visual inspection of more than 15,000 tomograms of intact frozen-hydrated cells belonging to 88 different bacterial species and several uncharacterized features in these tomograms has been reported (34). This has greatly improved our understanding of the complexity of bacterial cells. The advent of cryogenic focused ion beam (FIB) milling has extended the domain of Cryo-ET to include regions even deep within thick eukaryotic cells.

Medical diagnosis

Since 2014, there has been a growing interest in the research community to establish Cryo-ET as a medical diagnostic tool to help resolve molecular differences between healthy and diseased states. Cryo-ET was applied to human clinical samples to elucidate human ciliary structural defects in patients with primary ciliary dyskinesia, where the conventional diagnosing tool EM failed 30% of the time (35). Later, Wang et al. (36) demonstrated the effectiveness of using Cryo-ET as a non-invasive tool to identify ovarian cancer patients by imaging their platelets. They built a simple model using the number of mitochondria and length of microtubules in Cryo-ET images and correctly predicted 20 out of 23 cases. Other studies have identified cellular structural changes in disease states such as Leigh syndrome (37), Huntington's disease (38), and virus infection (39).

CONCLUSION

Cryo-ET led to a revolution in in-situ structural biology. However, due to the low SNR and structural complexity, it also poses challenges to the subsequent computational analysis. Template-based methods have enabled the systematic detection of known structures. To reduce the computational cost, a supervised deep learning-based approach was proposed to classify and segment cellular components. However, the success of such a supervised approach depends heavily on the availability of a large amount of properly labeled training data. The template-free approach has made it possible to automatically recover representative structures and the discovery of even unknown structures. Although the template-free approach has opened up promising new possibilities, the current accuracy and efficiency still has a large room for improvement. With an increasing amount of data being collected and an increasing amount of robust computational methods being developed, Cryo-ET has a large potential to advance structural biology and medical diagnosis progressively.

Acknowledgement: This work was supported in part by U.S. National Institutes of Health (NIH) grant P41 GM103712. XZ was supported by a fellowship from Carnegie Mellon University's Center for Machine Learning and Health. XG acknowledges the support by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research (OSR) under Award No. BAS/1/1624, FCC/1/1976-18, FCC/1/1976-23, FCC/1/1976-25, and FCC/1/1976-26.

Conflict of Interest: The authors declare no potential conflicts of interest with respect to research, authorship, and/or publication of this chapter.

Copyright and Permission Statement: To the best of our knowledge, the materials included in this chapter do not violate copyright laws. All original sources have been appropriately acknowledged and/or referenced. Where relevant, appropriate permissions have been obtained from the original copyright holder(s)

REFERENCES

1. Koning RI, Koster AJ, Sharp TH. Advances in cryo-electron tomography for biology and medicine. Ann Anat. 2018;217:82–96. http://dx.doi.org/10.1016/j.aanat.2018.02.004
2. Chang YW, Chen S, Tocheva EI, Treuner-Lange A, Löbach S, Søgaard-Andersen L, et al. Correlated cryogenic photoactivated localization microscopy and cryo-electron tomography. Nat Methods. 2014;11(7):737. http://dx.doi.org/10.1038/nmeth.2961
3. Xu M, Singla J, Tocheva EI, Chang YW, Stevens RC, Jensen GJ, et al. De Novo structural pattern mining in cellular electron cryotomograms. Structure. 2019;27(4):679–91.e14. Available from: http://www.sciencedirect.com/science/article/pii/S096921261930005X
4. Nickell S, Förster F, Linaroudis A, Del Net W, Beck F, Hegerl R, et al. TOM software toolbox: Acquisition and analysis for electron tomography. J Struct Biol. 2005;149(3):227–34. http://dx.doi.org/10.1016/j.jsb.2004.10.006
5. Förster F, Han BG, Beck M. Visual proteomics. Methods Enzymol. 2010;483:215–43. http://dx.doi.org/10.1016/S0076-6879(10)83011-3
6. Beck M, Malmström JA, Lange V, Schmidt A, Deutsch EW, Aebersold R. Visual proteomics of the human pathogen Leptospira interrogans. Nat Methods. 2009;6(11):817. http://dx.doi.org/10.1038/nmeth.1390
7. Wang K, Zeng X, Liang X, Huo Z, Xing E, Xu M. Image-derived generative modeling of pseudo-macromolecular structures – Towards the statistical assessment of Electron CryoTomography template matching. In: British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3–6, 2018; 2018. p. 130. Available from: http://bmvc2018.org/contents/papers/0532.pdf
8. Chen M, Dai W, Sun SY, Jonasch D, He CY, Schmid MF, et al. Convolutional neural networks for automated annotation of cellular cryo-electron tomograms. Nat Methods. 2017;14(10):983. http://dx.doi.org/10.1038/nmeth.4405
9. Moebel E, Martinez A, Larivière D, Ortiz J, Baumeister W, Kervrann C. 3D ConvNet improves macromolecule localization in 3D cellular cryo-electron tomograms. 2018. https://hal.inria.fr/hal-01966819/document
10. Liu C, Zeng X, Lin R, Liang X, Freyberg Z, Xing E, et al. Deep learning based supervised semantic segmentation of electron cryo-subtomograms. In: 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE; 2018. p. 1578–82.
11. Xu M, Chai X, Muthakana H, Liang X, Yang G, Zeev-Ben-Mordehai T, et al. Deep learning-based subdivision approach for large scale macromolecules structure recovery from electron cryo tomograms. Bioinformatics. 2017;33(14):i13–i22. http://dx.doi.org/10.1093/bioinformatics/btx230

12. Che C, Lin R, Zeng X, Elmaaroufi K, Galeotti J, Xu M. Improved deep learning-based macromolecules structure classification from electron cryo-tomograms. Mach Vision Appl. 2018;29(8):1227–36. http://dx.doi.org/10.1007/s00138-018-0949-4
13. Guo J, Zhou B, Zeng X, Freyberg Z, Xu M. Model compression for faster structural separation of macromolecules captured by cellular electron cryo-tomography. In: International Conference Image Analysis and Recognition. Springer; 2018. p. 144–52.
14. Han S, Mao H, Dally WJ. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:151000149. 2015.
15. Meng W, Gu Z, Zhang M, Wu Z. Two-bit networks for deep learning on resource-constrained embedded devices. arXiv preprint arXiv:170100485. 2017.
16. Lin R, Zeng X, Kitani K, Xu M. Adversarial domain adaptation for cross data source macromolecule *in situ* structural classification in cellular electron cryo-tomograms. Bioinformatics. 2019;35(14):i260–8. http://dx.doi.org/10.1093/bioinformatics/btz364
17. Patel VM, Gopalan R, Li R, Chellappa R. Visual domain adaptation: A survey of recent advances. IEEE Sig Process Mag. 2015;32(3):53–69. http://dx.doi.org/10.1109/MSP.2014.2347059
18. Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, et al. Domain-adversarial training of neural networks. J Mach Learn Res. 2016;17(1):2096–30.
19. Liu C, Zeng X, Wang K, Guo Q, Xu M. Multi-task learning for macromolecule classification, segmentation and coarse structural recovery in cryo-tomography. In: British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3–6, 2018; 2018. p. 271. Available from: http://bmvc2018.org/contents/papers/1007.pdf
20. Xu M, Beck M, Alber F. Template-free detection of macromolecular complexes in cryo electron tomograms. Bioinformatics. 2011;27(13):i69–76. http://dx.doi.org/10.1093/bioinformatics/btr207
21. Zhou B, Guo Q, Wang K, Zeng X, Gao X, Xu M. Feature decomposition based saliency detection in electron cryo-tomograms. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 2467–73.
22. Zhao Y, Zeng X, Guo Q, Xu M. An integration of fast alignment and maximum-likelihood methods for electron subtomogram averaging and classification. Bioinformatics. 2018;34(13):i227–36. http://dx.doi.org/10.1093/bioinformatics/bty267
23. Bharat TA, Scheres SH. Resolving macromolecular structures from electron cryo-tomography data using subtomogram averaging in RELION. Nat Protocols. 2016;11(11):2054. http://dx.doi.org/10.1038/nprot.2016.124
24. Scheres SH, Melero R, Valle M, Carazo JM. Averaging of electron subtomograms and random conical tilt reconstructions through likelihood optimization. Structure. 2009;17(12):1563–72. http://dx.doi.org/10.1016/j.str.2009.10.009
25. Zeng X, Leung MR, Zeev-Ben-Mordehai T, Xu M. A convolutional autoencoder approach for mining features in cellular electron cryo-tomograms and weakly supervised coarse segmentation. J Struct Biol. 2018;202(2):150–60. http://dx.doi.org/10.1016/j.jsb.2017.12.015
26. Doerr A. Template-free visual proteomics. Nature Methods. 2019;16: 285
27. Briegel A, Li X, Bilwes AM, Hughes KT, Jensen GJ, Crane BR. Bacterial chemoreceptor arrays are hexagonally packed trimers of receptor dimers networked by rings of kinase and coupling proteins. Proc Natl Acad Sci. 2012;109(10):3766–3771. http://dx.doi.org/10.1073/pnas.1115719109
28. Liu J, Hu B, Morado DR, Jani S, Manson MD, Margolin W. Molecular architecture of chemoreceptor arrays revealed by cryoelectron tomography of Escherichia coli minicells. Proc Natl Acad Sci. 2012;109(23):E1481–8. http://dx.doi.org/10.1073/pnas.1200781109
29. Beeby M, Cho M, Stubbe J, Jensen GJ. Growth and localization of polyhydroxybutyrate granules in Ralstoniaeutropha. J Bacteriol. 2012;194(5):1092–9. http://dx.doi.org/10.1128/JB.06125-11
30. Comolli LR, Kundmann M, Downing KH. Characterization of intact subcellular bodies in whole bacteria by cryo-electron tomography and spectroscopic imaging. J Microscopy. 2006;223(1):40–52. http://dx.doi.org/10.1111/j.1365-2818.2006.01597.x
31. Iancu CV, Ding HJ, Morris DM, Dias DP, Gonzales AD, Martino A, et al. The structure of isolated Synechococcus strain WH8102 carboxysomes as revealed by electron cryotomography. J Mol Biol. 2007;372(3):764–73. http://dx.doi.org/10.1016/j.jmb.2007.06.059

32. Pšenčík J, Collins AM, Liljeroos L, Torkkeli M, Laurinmäki P, Ansink HM, et al. Structure of chlorosomes from the green filamentous bacterium Chloroflexusaurantiacus. J Bacteriol. 2009;191(21):6701–8. http://dx.doi.org/10.1128/JB.00690-09
33. Schmid MF, Paredes AM, Khant HA, Soyer F, Aldrich HC, Chiu W, et al. Structure of Halothiobacillusn eapolitanuscarboxysomes by cryo-electron tomography. J Mol Biol. 2006;364(3):526–35. http://dx.doi.org/10.1016/j.jmb.2006.09.024
34. Dobro MJ, Oikonomou CM, Piper A, Cohen J, Guo K, Jensen T, et al. Uncharacterized bacterial structures revealed by electron cryotomography. J Bacteriol. 2017;199(17):e00100–17. http://dx.doi.org/10.1128/JB.00100-17
35. Lin J, Yin W, Smith MC, Song K, Leigh MW, Zariwala MA, et al. Cryo-electron tomography reveals ciliary defects underlying human RSPH1 primary ciliary dyskinesia. Nat Commun. 2014;5:5727. http://dx.doi.org/10.1038/ncomms6727
36. Wang R, Stone RL, Kaelber JT, Rochat RH, Nick AM, Vijayan KV, et al. Electron cryotomography reveals ultrastructure alterations in platelets from patients with ovarian cancer. Proc Natl Acad Sci. 2015;112(46):14266–71. http://dx.doi.org/10.1073/pnas.1518628112
37. Siegmund SE, Grassucci R, Carter SD, Barca E, Farino ZJ, Juanola-Falgarona M, et al. Three dimensional analysis of mitochondrial crista ultrastructure in a patient with leigh syndrome by in situ cryo-electron tomography. iScience. 2018;6:83–91. http://dx.doi.org/10.1016/j.isci.2018.07.014
38. Bäuerlein F, Mishra A, Dudanova I, Hipp M, Klein R, Hartl F, et al. Structural characterization of mutant huntingtin inclusion bodies by cryo-electron tomography. Microsc Microanal. 2016;22(S3):80–1. http://dx.doi.org/10.1017/S1431927616001252
39. Cao S, Maldonado JO, Grigsby IF, Mansky LM, Zhang W. Analysis of human T-cell leukemia virus type 1 particles by using cryo-electron tomography. J Virol. 2015;89(4):2430–5. http://dx.doi.org/10.1128/JVI.02358-14

Index

Numbers and Symbols

A

B

C

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

Y

Z

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ind