Translational Bioinformatics 19
Series Editor: Xiangdong Wang, MD, PhD, Prof

Kang Ning *Editor*

# of Multi-Omics Data Integration and Data Mining

Techniques and Applications

Springer

Translational Bioinformatics

Volume 19

Series Editor

Xiangdong Wang, Shanghai Institute of Clinical Bioinformatics, Zhongshan Hospital Institute of Clinical Science, Fudan University Shanghai Medical College, Shanghai, China

Translational bioinformatics is defined as the development of storage-related, analytic, and interpretive methods to optimize the transformation of increasingly voluminous biomedical data, and genomic data in particular, into proactive, predictive, preventive, and participatory health. Translational bioinformatics includes research on the development of novel techniques for the integration of biological and clinical data and the evolution of clinical informatics methodology to encompass biological observations. The end product of translational bioinformatics is the newly found knowledge from these integrative efforts that can be disseminated to a variety of stakeholders including biomedical scientists, clinicians, and patients. Issues related to database management, administration, or policy will be coordinated through the clinical research informatics domain. Analytic, storage-related, and interpretive methods should be used to improve predictions, early diagnostics, severity monitoring, therapeutic effects, and the prognosis of human diseases.

Kang Ning
Editor

Methodologies of Multi-Omics Data Integration and Data Mining

Techniques and Applications

Springer

Editor
Kang Ning
Department of Bioinformatics and Systems Biology, Center of AI biology, College of Life Science and Technology
Huazhong University of Science and Technology
Wuhan, China

ISSN 2213-2775 ISSN 2213-2783 (electronic)
Translational Bioinformatics
ISBN 978-981-19-8209-5 ISBN 978-981-19-8210-1 (eBook)
https://doi.org/10.1007/978-981-19-8210-1

Preface

Many biomedical and clinical questions can now be answered using the wealth of multi-omics data that has become available in the age of omics. In the process, however, it has also created hurdles in the integration, mining, and comprehensive understanding of omics data.

Several biomedical applications are of special interest among various applications. The first is about cancer omics, which is always at the forefront of omics data analysis. Previous cancer omics research has focused on genomics and transcriptomics, whereas current multi-omics analysis would undoubtedly be the focus of in-depth mining of cancer progression principles. The second is about inflammation disease omics, which has piqued the interest of the research community, in part due to the growing proportion of patients suffering from inflammatory diseases such as arthritis. Multi-omics research, particularly on the dynamics of multi-omics, would shed light on a better understanding of the development of inflammation disease. The third is about the microbiome, which is a current hot topic: microbial communities are now thought to be linked to a variety of diseases, including T2D, IBD, and others. And, as with so many other questions, the principle governing the regulation of the microbiome on these various diseases remains a mystery. As a result, metagenomic data mining and explanations would be extremely valuable in the omics field. The fourth topic is omics data integration, which is related not only to databases and online data analysis pipelines, as well as visualization tools, but also to the development of various methods for multi-omics data correlation analysis or even causal or dynamic pattern discovery in the data integration procedure. Only through such high-level data integration could a solid foundation for data mining be built. Finally, method development is critical for a better understanding of hidden principles that can only be recovered by novel creative artificial intelligence tools.

This book has covered not only multi-omics big-data integration and data-mining techniques, but also cutting-edge researches in the principles and applications of several omics, including cancer omics, inflammation disease omics, and microbiome research. (1) Multi-omics big-data integration and data-mining techniques: Data

integration and data-mining techniques will be introduced, along with illustrative examples and figures, to provide a better understanding of the essence of the definitions of both multi-omics and data mining, as well as how they can be combined to gain the most insights from these omics data. (2) Advancement in concrete research on multi-omics big-data: readers will learn the fundamental procedures for conducting representative and concrete multi-omics studies: given a set of omics data, how data-mining techniques can be applied to meet the needs of specific biological questions of interest. (3) Cutting-edge research in applications such as cancer omics, inflammation disease, and microbiome research: three topics would be highlighted out of many applications, one on cancer omics data analysis and explanations, another on inflammation disease, and another on specifically featured microbiome applications such as those related to T2D and IBD. (4) Contemporary data resources, tools, and analytical platforms will also be featured for readers to gain hands-on experience.

Intended as a book on the biomedical big-data expedition in the omics age, this book focuses on data integration and data-mining methods for multi-omics researches, explaining the "What," "Why," and "How" of the topic in detail and with supporting examples. It is an attempt to bridge the gap between biomedical multi-omics big data and data-mining techniques to obtain optimal practices in contemporary bioinformatics and in-depth insights into biomedical and clinical questions.

Wuhan, China Kang Ning

About the Book

This book features multi-omics big-data integration and data-mining techniques. In the omics age, the paramount of multi-omics data from various sources is the new challenge we are facing, but it also provides clues for several biomedical or clinical applications. For multi-omics research, this book discusses in detail and with examples how to integrate data and performed data mining. This book focuses on data integration and data-mining methods for multi-omics research, which explains in detail and with supportive examples the "What," "Why," and "How" of the topic. The contents are organized into eight chapters, out of which one is for the introduction, followed by four chapters devoted to omics integration techniques and data-mining methods, and three chapters devoted to the applications of multi-omics analyses, where data-mining methods are used to demonstrate how multi-omics analyses can be used in practice. This book is an attempt to bridge the gap between biomedical multi-omics big data and the data-mining techniques, for the best practice of contemporary bioinformatics and the in-depth insights for the biomedical questions. It would be of interest to the researchers and practitioners who want to conduct multi-omics studies in cancer, inflammation disease, and microbiome researches.

Contents

About the Editor

Kang Ning Professor, PI of Microbial Bioinformatics Group, Director of Department of Bioinformatics and Systems Biology, School of Life Science and Technology, Huazhong University of Science and Technology, China.

Kang obtained his BS in Computer Science from USTC, and PhD in Bioinformatics from NUS. He has had his Post-Doc training in Bioinformatics from the University of Michigan.

Kang has more than 20 years of experiences in bioinformatics for omics big-data integration, microbiome analyses, and single-cell analyses. His current research interests include AI method for multi-omics especially metagenomics data mining, as well as their applications. He is also interested in synthetic biology and high-performance computation.

Kang is the leading or corresponding author of over 100 papers and reviews on leading journals including PNAS, Gut, Genome Biology, Nucleic Acids Research, Briefings in Bioinformatics and Bioinformatics, which have more than 3,000 citations. He has been the committee member of several national bioinformatics and biology big-data committees in China. He serves as an editorial board member of several journals including Genomics Proteomics and Bioinformatics, Microbiology Spectrum and Scientific Reports, and served as reviewers for several international funding agencies including UK-BBSRC and UK-NERC. He has collaborations with biologists, doctors, and statisticians in many countries and has given talks on international conferences for more than hundred times. For details, please refer to his official website as: http://www.microbioinformatics.org/.

Chapter 1
Introduction to Multi-Omics

Kang Ning and Yuxue Li

The rapid development of technologies and informatics tools for producing and interpreting massive biological data sets (omics data) has resulted in a paradigm shift in how we approach biomedical challenges (Manzoni et al. 2018). Large data sets are typically generated during genomics, transcriptomics, proteomics, microbiomics, and metabolomics research (Osier et al. 2017). With the advancement of these omics investigations, multi-omics research has emerged as one of the most promising venues for a deeper understanding of biological problems (Sun and Hu 2016). As the name suggests, multi-omics encompasses all digital genetic resources relevant to the research objectives, and its related research will automatically generate more comprehensive information to achieve the purpose of the research.

Multi-omics studies typically include omics data from multiple sources, including genomics, transcriptomics, proteomics, epigenomics, and microbiomics (Chung and Kang 2019). Genomics refers to omics data derived from DNA materials (Manzoni et al. 2018). Transcriptomics is the study of omics data derived from RNA materials (Manzoni et al. 2018). Proteomics is the collection of omics data from protein materials (Manzoni et al. 2018). Epigenomics refers to omics data derived from the whole range of epigenetic alterations on genetic material (Casadesús and Noyer-Weidner 2013). Microbiomics refers to omics data derived from a microbial community's entire set of genetic materials (Kumar 2000). Each of these omics represents a different part of the research goal, and when combined, they could disclose the regulatory patterns and principles that govern how genetic materials regulate genotypes (Fig. 1.1). On a more generalized scope, the omics can also

K. Ning (✉) · Y. Li
Key Laboratory of Molecular Biophysics of the Ministry of Education, Hubei Key Laboratory of Bioinformatics and Molecular-imaging, Center of Artificial Intelligence Biology, Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan, China
e-mail: ningkang@hust.edu.cn

K. Ning (ed.), *Methodologies of Multi-Omics Data Integration and Data Mining*,
Translational Bioinformatics 19, https://doi.org/10.1007/978-981-19-8210-1_1

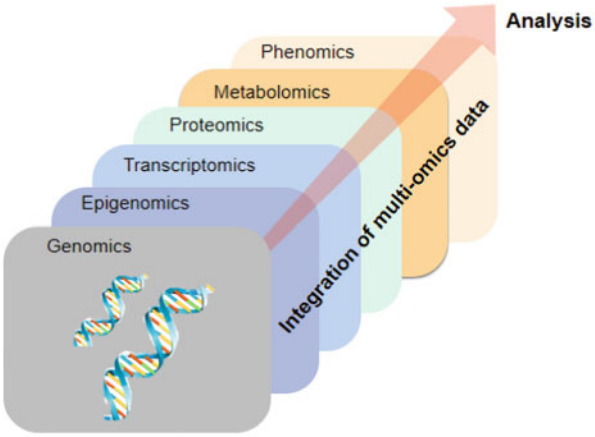

Fig. 1.1 The generalized definition of multi-omics. Large-scale acquisition of omics data from different molecular levels such as genome, transcriptome, proteome, epigenome, metabolome, microbiome, etc., and integrated analysis to achieve a deeper understanding of biological processes and molecular mechanisms

include those data from bioimaging, biosensors, and even social networks (Antonelli et al. 2019; Sriram and Subrahmanian 2020; Loizou 2016).

1.1 The History of Omics

The omics studies have quite a long history. Back in 1958, the first sequencing technique emerged, as Frederick Sanger has invented the protein sequencing methods, especially the amino acid sequence of insulin (Heather and Chain 2016). However, sequencing technology did not develop significantly during the next twenty to thirty years. DNA was originally extracted in 1869, it was not until more than a century later that the first genomes were sequenced, making genomics a relatively new field that truly began in 1970s.

1.1.1 1971–1910: Discovery of DNA

In 1871, Friedrich Miescher published a paper identifying the presence of nuclein and associated proteins in the nucleus. This is what we now call DNA, which is the foundation of the field of genomics.

Walter Sutton and Theodor Boveri discovered in 1904 that chromosomes appeared in pairs, with one inherited from each parent., which is known as the theory of chromosome inheritance. In 1910, Albert Kossel discovered the five nucleotide bases: adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U).

1.1.2 1950–1968: Development of Knowledge about DNA

Erwin Chargaff discovered the base pairing of adenosine, cytosine, guanine, and thymine nucleotides in 1950. He discovered that the concentrations of thymine and adenine or cytosine and guanine in DNA samples are always equal. As a result, he concluded that adenosine and thymine form a chromosome pair, while cytosine and guanine form a chromosome pair.

In 1952, Alfred Hershey and Martha Chase proved through a series of experiments that it was DNA, not protein, that carried inherited genetic features. The following year, the double helix structure of DNA was discovered by James Watson and Francis Crick (Portin 2014). A research team led by Marshall Nirenberg and Har Gobind Khorana discovered what is now known as DNA "codons" in 1961.

1.1.3 1977–Present: Sequencing of DNA Related Stories

Frederick Sanger developed a DNA sequencing technology in 1977 to sequence the first complete genome, known as the phiX174 virus, which opened the door to new possibilities in genomics. In 1983, Dr. Kary Mullis invented the polymerase chain reaction (PCR) technique for amplifying DNA (García-Quesada et al. 2021). The first bacterial genome sequence, Haemophilus influenza, was completed in 1995 (Fraser and Rappuoli 2004). The yeast genome was completed one year later (Zhang 1999). Dolly the sheep, the first cloned animal, was also born at this time (Elster 1999).

In 1990, the Human Genome Project was launched to sequence 3 billion letters of the human genome. As part of this project, chromosome 22 was sequenced as the first chromosome in 1999. The project was finished in 2003, and it confirmed that humans have between 20,000 and 25,000 genes.

In 2007, there was a breakthrough in DNA sequencing technology that increased the output of DNA sequencing by 70 times in 1 year. This prompted the launch of the 1000 Gene Project in 2008, intending to sequence the genomes of a large population of 2500 people.

In general, advances in DNA technology have aided the development of omics. Figure 1.2 depicts a brief timeline of recent developments in multi-omics research.

1.2 Omics: DNA, RNA, Protein, and Microbiome

There are now many hot omics studies such as transcriptomics for RNA research, proteomics for protein research, and microbiomics for microbiome research (Yu et al. 2018). Transcriptomics is the study of gene transcription in cells as well as transcriptional regulation in general (Dong and Chen 2013). Proteomics is the study of the composition of cells, tissues, or biological proteins and their changing

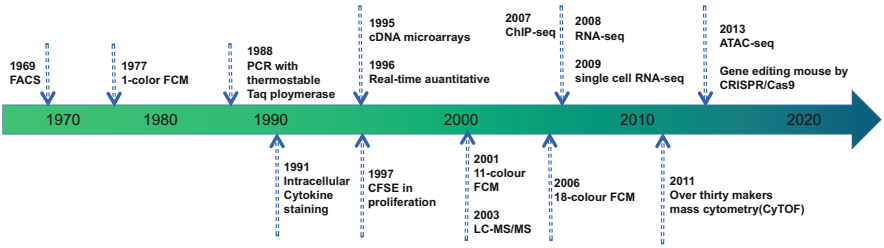

Fig. 1.2 The timeline for the development of multi-omics researches. The development history of multi-omics is actually the process of development and innovation of different omics data acquisition technologies. With the enrichment and cross-use of omics data, multi-omics analysis has gradually been applied

laws using protein as the research object. The term "microbiome" refers to the genomes of microorganisms (bacteria, archaea, lower or higher eukaryotes, and viruses) as well as their entire environment (Marchesi and Ravel 2015).

Multi-omics has emerged with the accumulation of various omics datas, and has become a research focus in recent years due to its importance in basic research and clinical application (Chakraborty et al. 2018). A series of disease-related differences are typically generated for omics data. These data can be used as disease process markers as well as insights into biological pathways or process differences between the disease and the control group. However, only one type of data analysis has limited relevance, primarily reflecting the reaction process rather than causality. The integration of various omics data types is typically used to clarify potential pathogenic changes or treatment targets that cause the disease, which can then be tested further. Multi-omics research, when compared to a single type of omics research, can better understand the basic information flow of diseases.

In recent years, sequencing technologies have generated a large amount of multi-omics data worldwide, but also brings many problems. First, because the growth rate of multivariate data was unimaginable ten years ago, large public databases have used cloud facilities to store these data. Secondly, the cost of generating multi-omics data has decreased rapidly, leading to a further increase in the amount of multi-omics data as well (NHGRI 2021).

1.3 Databases and Tools for Omics Studies

When confronted with multi-omics data, there is an increasing demand for computational methods that can rationally integrate and accurately analyze heterogeneous multi-omics data. To date, numerous databases and analytical tools have been developed to aid in the analysis of these multi-omics datasets (Tables 1.1 and 1.2).

Table 1.1 Representative databases for multi-omics research

Database	Functionality	Web link	Reference
ChEBI	Metabolomics database and ontology	http://bigd.big.ac.cn/databasecommons/database/id/364	Degtyarenko et al. (2007)
E.coli metabolome database (ECMDB)	Annotated metabolomics and metabolite pathway database	https://ecmdb.ca/	Guo et al. (2012)
FlyBase	Genes and RNA-seq data of different drosophila	https://flybase.org/	Thurmond et al. (2018)
GenBank (database)	Proteomics database open access annotated collection of all publically available nucleotide sequences and their protein transitions.	https://www.uniprot.org/database/DB-0028	Benson et al. (2017)
Human Metabolome Database (HMDB)	Human metabolite and pathway database	https://hmdb.ca/	Wishart et al. (2017)
KEGG	Collection of databases dealing with genomes biological pathways, disease, drugs and chemical substances	https://www.kegg.jp/	Kanehisa et al. (2016)

Table 1.2 Representative analytical tools for multi-omics research

Tool/Method	Tool/Method approach	Tool/Method link	Reference
PARADIGM	Probabilistic graphical models using directed factor graphs	http://paradigm.five3genomics.com/	Gluth et al. (2013)
iCluster	Joint latent variable model-based clustering method	https://cran.r-project.org/web/packages/iCluster/index.html	Shen et al. (2009)
iClusterPlus	Generalized linear regression for the formulation of the joint model	http://www.bioconductor.org/packages/release/bioc/html/iClusterPlus.html	Pierre-Jean et al. (2019)

1.4 Multi-Omics Applications

Multi-omics research has been successfully applied to many biological problems such as cancer omics, inflammatory disease omics and microbiome research.

The applications of multi-omics in disease (Hasin et al. 2017), including the integration of genome, epigenome, transcriptomics, proteomics, metabolomics and microbiome, as well as their interrelationships. Figure 1.3 depicts the various omics data types and disease research methods from an article. Each layer represents an omics data type. The omics data is gathered across the entire molecular pool, which is represented by circles. Except for the genome, all data layers reflect genetic regulation and the environment, and the environment's impact on each molecule may differ. In a recent work, cancer genomic profiling of 78 clinical tumor samples (Rusch et al. 2018) using three-platform sequencing of the whole genome, whole

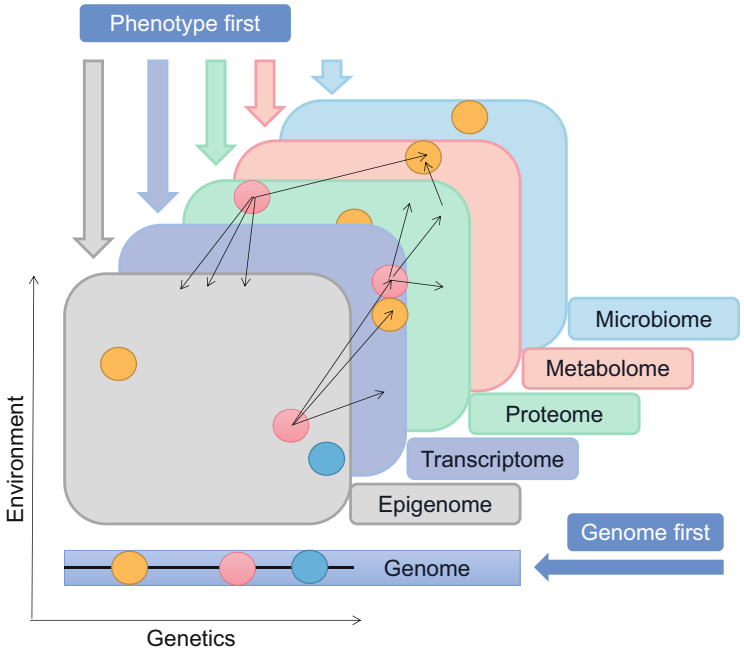

Fig. 1.3 Different omics data and corresponding analytical methods for disease research. Except for the genome, all data layers reflect both gene regulation and environment, which may affect each molecule to varying degrees. Potential interactions or correlations detected between molecules in different layers are represented by thin arrows

exome, and transcriptome to identify tumor-related structure variation (SV), somatic cell mutation, and pathogenic mutation, among other things. Sequencing, variant detection, variant classification, group review, and report generation are all covered in this the clinical three-platform sequencing design. In another research, multi-omics approaches to study secondary metabolites biosynthesis in microbes (Palazzotto and Weber 2018).

In general, multi-omics data resources are rapidly growing, and their analysis tools and platforms are maturing. Multi-omics research has made remarkable achievements in cancer and biological problems. Some applications of multi-omics research are listed below.

1. **Multi-omics approaches to cancer** (Aure et al. 2013) tracked genetic associations caused by breast cancer using complete genome-wide copy number and expression data. The author proposed a method for analyzing in-cis correlated genes in biological processes that is not biased towards particular types or functional processes. The goal of this method is to find cis-regulated genes whose expression correlation with other genes supports the role of network interference in cancer. This method was used to examine the genome-wide