PERSONALIZED EPIGENETICS

Trygve O. Tollefsbol

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
125 London Wall, London EC2Y 5AS, UK
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloging-in-Publication Data
A catalog record for this book is available from the British Library

ISBN: 978-0-12-420135-4

Acquisition Editor: Catherine Van Der Laan
Editorial Project Manager: Lisa Eppich
Production Project Manager: Karen East and Kirsty Halterman
Designer: Mark Rogers

Typeset by TNQ Books and Journals
www.tnq.co.in

Printed and bound in the United States of America

Contents

I
OVERVIEW

II
EPIGENETIC VARIATIONS AMONG INDIVIDUALS

III

BIOINFORMATICS OF PERSONALIZED EPIGENETICS

IV

DIAGNOSTIC AND PROGNOSTIC EPIGENETIC APPROACHES TO PERSONALIZED MEDICINE

V

ENVIRONMENTAL PERSONALIZED EPIGENETICS

VI

PHARMACOLOGY AND DRUG DEVELOPMENT OF PERSONALIZED EPIGENETICS

VII

PERSONALIZED EPIGENETICS OF DISORDERS AND DISEASE MANAGEMENT

VIII

CHALLENGES AND FUTURE DIRECTIONS

Contributors

Vanessa Aguiar-Pulido School of Computing & Information Sciences, Florida International University, Miami, FL, USA; Department of Information & Communication Technologies, University of A Coruña, A Coruña, Spain

Alexandre F. Aissa Department of Clinical Analyses, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, São Paulo, Brazil

Seena K. Ajit Department of Pharmacology & Physiology, Drexel University College of Medicine, Philadelphia, PA, USA

Lusânia M.G. Antunes Department of Clinical Analyses, Toxicology and Food Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo (USP), Ribeirão Preto, São Paulo, Brazil

Elisabeth B. Binder Max Planck Institute of Psychiatry, Translational Research, Munich, Germany

Christopher Chang Division of Rheumatology, Allergy and Clinical Immunology, University of California at Davis, CA, USA

Samit Chattopadhyay National Centre for Cell Science, Pune University Campus, Ganeshkhind, Pune, India

Arpankumar Choksi National Centre for Cell Science, Pune University Campus, Ganeshkhind, Pune, India

Fabio Coppedè Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy

Jose M. Eirin-Lopez Department of Biological Sciences, Florida International University, North Miami, FL, USA

Ammad A. Farooqi Laboratory for Translational Oncology and Personalized Medicine, Rashid Latif Medical College, Lahore, Pakistan

Domniki Fragou Laboratory of Forensic Medicine and Toxicology, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Rebecca C. Fry Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC, USA; Curriculum in Toxicology, School of Medicine, University of North Carolina, Chapel Hill, NC, USA

Gunnar H. Heine Department of Internal Medicine IV, Nephrology and Hypertension, Saarland University Medical Center, Homburg, Germany

Anke Hoffmann Max Planck Institute of Psychiatry, Translational Research, Munich, Germany

Muhammad Ismail IBGE, Islamabad, Pakistan

Leda Kovatsi Laboratory of Forensic Medicine and Toxicology, School of Medicine, Aristotle University of Thessaloniki, Thessaloniki, Greece

Anders M. Lindroth Graduate School of Cancer Science and Policy, National Cancer Center, Goyang-si, Republic of Korea

Ying Liu Department of Dermatology, Hunan Key Laboratory of Medical Epigenomics, Second Xiangya Hospital, Central South University, Hunan, China

Angela Lopomo Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy; Doctoral School in Genetics, Oncology, and Clinical Medicine, University of Siena, Siena, Italy

Kenneth Lundstrom PanTherapeutics, Lutry, Switzerland

Qianjin Lu Department of Dermatology, Hunan Key Laboratory of Medical Epigenomics, Second Xiangya Hospital, Central South University, Hunan, China

Emily Machiela Laboratory of Aging and Neurodegenerative Disease, Van Andel Research Institute, Grand Rapids, MI, USA

Lucia Migliore Department of Translational Research and New Technologies in Medicine and Surgery, University of Pisa, Pisa, Italy

Giri Narasimhan School of Computing & Information Sciences, Florida International University, Miami, FL, USA

Monica D. Nye Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC, USA

Joo H. Park Department of Nutritional Science and Food Management, Ewha Womans University, Seoul, Republic of Korea

Yoon J. Park Department of Nutritional Science and Food Management, Ewha Womans University, Seoul, Republic of Korea

Sonal Patel National Centre for Cell Science, Pune University Campus, Ganeshkhind, Pune, India

Jacob Peedicayil Department of Pharmacology and Clinical Pharmacology, Christian Medical College, Vellore, India

Javier Pereira Department of Information & Communication Technologies, University of A Coruña, A Coruña, Spain

Anthony Popkie Laboratory of Cancer Epigenomics, Van Andel Research Institute, Grand Rapids, MI, USA

Muhammad Z. Qureshi Department of Chemistry, GCU, Lahore, Pakistan

Nicole C. Riddle Department of Biology, The University of Alabama at Birmingham, Birmingham, AL, USA

Lorenzo F. Sempere Laboratory of MicroRNA Diagnostics and Therapeutics, Van Andel Research Institute, Grand Rapids, MI, USA

Dietmar Spengler Max Planck Institute of Psychiatry, Translational Research, Munich, Germany

Victoria Suarez-Ulloa Department of Biological Sciences, Florida International University, North Miami, FL, USA

D. Syndercombe Court Faculty of Biological Sciences and Medicine, King's College London, London, UK

Trygve O. Tollefsbol Department of Biology, University of Alabama at Birmingham, Birmingham, AL, USA; Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, AL, USA; Comprehensive Center for Healthy Aging, University of Alabama at Birmingham, Birmingham, AL, USA; Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, AL, USA; Comprehensive Diabetes Center, University of Alabama at Birmingham, Birmingham, AL, USA

Athina Vidaki Faculty of Biological Sciences and Medicine, King's College London, London, UK

Louis P. Watanabe Department of Biology, The University of Alabama at Birmingham, Birmingham, AL, USA

Hehuang Xie Department of Biological Sciences, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

Yeongran Yoo Department of Nutritional Science and Food Management, Ewha Womans University, Seoul, Republic of Korea

Andrew E. Yosim Department of Environmental Sciences and Engineering, University of North Carolina, Chapel Hill, NC, USA

Adam M. Zawada Department of Internal Medicine IV, Nephrology and Hypertension, Saarland University Medical Center, Homburg, Germany

Peng Zhang Department of Dermatology, Hunan Key Laboratory of Medical Epigenomics, Second Xiangya Hospital, Central South University, Hunan, China

Christoph A. Zimmermann Max Planck Institute of Psychiatry, Translational Research, Munich, Germany

CHAPTER

6

Computational Methods in Epigenetics

Vanessa Aguiar-Pulido[1,3,*], *Victoria Suarez-Ulloa*[2,*], *Jose M. Eirin-Lopez*[2], *Javier Pereira*[3], *Giri Narasimhan*[1]

[1]School of Computing & Information Sciences, Florida International University, Miami, FL, USA; [2]Department of Biological Sciences, Florida International University, North Miami, FL, USA; [3]Department of Information & Communication Technologies, University of A Coruña, A Coruña, Spain

OUTLINE

[*] The first two authors contributed equally to this chapter.

Personalized Epigenetics
http://dx.doi.org/10.1016/B978-0-12-420135-4.00006-1

1. INTRODUCTION

Epigenetics is defined as the heritable changes in gene expression resulting from modifications in chromatin structure, without involving changes in the genetic information stored in DNA [1]. The understanding of epigenetics requires integrative analysis of disparate heterogeneous data to generate global interpretations and biological knowledge [2]. The epigenome arises from interactions among different epigenetic mechanisms, including discrete biomolecules (e.g., nucleic acid–protein interactions) as well as chemical modifications (i.e., DNA methylation and protein post-translational modifications, or PTMs), thus possessing an elaborate combinatorial complexity [1]. Peculiarities of each of the different epigenetic marks must be taken into account to understand the diverse nature of the data resulting from epigenomic studies. The analysis of each one of these marks involves specific techniques and work flows, resulting in different types of data (Figure 1).

The heterogeneity and scale of data of epigenomics studies pose serious challenges for computational analyses and information management, similar to those created by other complex systems (e.g., markets, social dynamics) [3]. Developing more efficient analyses to keep up with the pace of data production and expand the frontiers in epigenetics is the current task of bioinformatics. Fortunately, a number of efforts are aimed at standardizing and integrating heterogeneous data sources in such a way that they become suitable for meta-analysis and are openly available to the scientific community [4].

In this chapter, we describe the main characteristics of the various types of data generated during epigenetic studies, providing a description of the most common computational approaches used for their integrative analysis. Additionally, we cover substantial advances in biomedical research that are illustrated by the production of online resources and by the

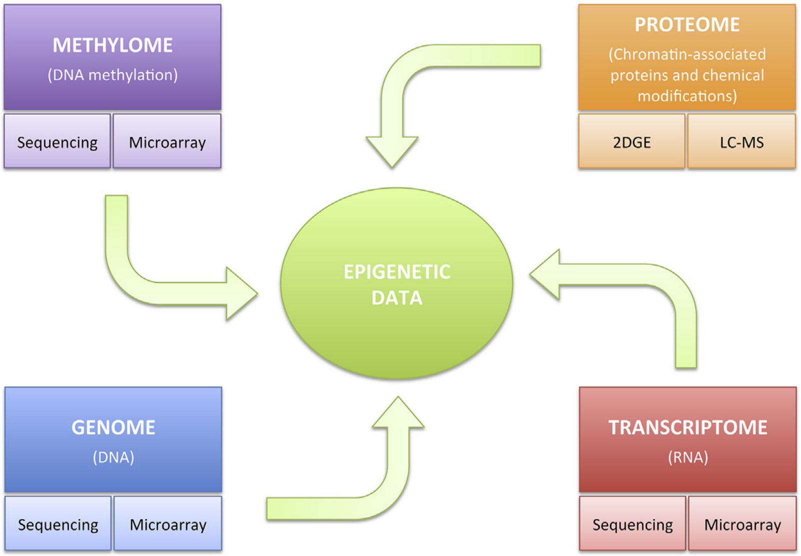

FIGURE 1 **Heterogeneous types of data generated from various "omics" and specific techniques are included as epigenetic data.** High-throughput techniques such as DNA sequencing and microarray-based analyses are currently dominant in functional genomics as well as in the study of the methylome, producing specific data formats that must be processed and standardized before being compared. Proteomics comprises all chromatin-related proteins of interest, most notably the histone family, including their possible chemical modifications and interactions. In proteomics, gel-based methods and mass spectrometry represent the main sources of high-throughput data. The combination of all these disparate data types constitutes the bulk of epigenetic information.

establishment of worldwide consortia for the standardization of molecular data. These contributions are critical to effectively linking genetic and epigenetic data with clinical information and pave the way toward a realistic future of personalized medicine.

2. EPIGENETIC PROFILING: HETEROGENEOUS DATA AND PARTICULAR CHALLENGES OF DIFFERENT EPIGENETIC FACTORS

Epigenetic profiling involves the coordinated study of diverse biological marks responsible for the transmission of epigenetic information, including but not limited to DNA methylation, histone variants and their PTMs, and chromatin remodeling complexes. Subsequently, some of the most relevant biological aspects of epigenetics research as well as the main characteristics of the data that they involve are addressed.

2.1 Methylation Patterns

Methylation is the best studied epigenetic mark, especially pertaining to CpG islands in the case of mammals [5]. A variable percentage of the CpG dinucleotides of the genome, ranging from 60% to 90%, is actually methylated. The remaining portion of CpGs are free of methylation, largely constituting the so-called CpG islands, which are usually associated with gene promoter regions [6]. The presence of methylation marks on DNA has been widely associated with repressive states of the chromatin. Moreover, it is hypothesized that its evolutionary origin was the neutralization of invading DNA by blocking its ability to be expressed [7]. The actual effect of the methylation transformation of DNA varies depending on a number of factors: the proximity of methylated CpG islands to the gene promoter, the density of those methylation marks, and the strength of the promoter itself [5].

Furthermore, the specific location of the methylation marks in relation to the promoter may display contradictory effects. This has been labeled as the “methylation paradox” [8]: although methylation of CpG islands in the promoter region is strongly associated with inhibition of transcription, it has been found that methylation of CpG islands downstream from the site of transcription initiation shows no inhibition effect. Furthermore, it has been reported that this may actually enhance the levels of transcription [9]. Although the specific molecular effects of methylation events may be difficult to predict, it is widely accepted that the patterns of methylation marks are nonrandomly associated with diseases (such as cancer), displaying a good level of specificity between tumor types in many cases [10].

The previous findings highlight the importance of an accurate mapping of methylated sites on the genome at a single-base resolution. Unbiased epigenetic mapping and annotation of genomes requires high-throughput sequencing techniques; however, more directed methods such as microarray platforms could also be used. The information of which nucleotides (cytosines, C’s) are methylated on the genome is lost during PCR amplification and additional experimental steps are required, most notably the selective bisulfite transformation of unmethylated cytosine into thymine (C→T transformation). Each different procedure requires optimized bioinformatic data-processing techniques to produce standardized data summaries that allow comparisons across experiments; that is, the differentially methylated regions (DMRs) table. To produce these tables, raw data must be processed and controlled for quality and then mapped onto a reference genome, and last, an appropriate statistical analysis must be carried out to obtain a list of significantly differentially methylated genomic regions along with the absolute values of methylation levels.

The set of nucleic acid methylation modifications in an organism's genome or in a particular cell is referred to as the methylome. Methylome data processing must address common general problems that are posed by high-throughput sequencing or microarray-based techniques. Most notably, the bioinformatic analysis of bisulfite-treated DNA is challenging owing to the decrease in complexity of the sequences after C→T transformation. Genome-wide sequencing or enriched DNA libraries produce a collection of reads that must be aligned to a reference genome. To allow the alignment of bisulfite-transformed reads that account for C–T mismatches, different approaches can be used. One possible approach involves wildcard aligners, which change C's in the sequence to the degenerate IUPAC symbol "Y" (equivalent to both C and T). Another existing approach modifies the scoring matrix used by the alignment algorithm to prevent penalization of C–T mismatches. As a third possible approach, all C's can be converted into T's on both reads as well as on the reference genome. This way, the alignment is worked on a three-letter alphabet for both the template and the complementary strand sequences of the genome. All these methods present some bias owing to lower complexity in the sequence and consequently a lack of specificity in the alignment, causing useful good quality reads to be discarded.

To improve mapping efficiency, additional steps such as local realignment, analysis of sequence quality scores, and the application of statistical models of allele distribution can be carried out [11]. The visualization of this information can be performed with any available genome browser. Methylated positions are often represented using color codes and quantitative methylation data with bar charts are superposed. Once mapped, quantification of methylation at a single-base resolution is carried out and then associated with each genomic position. There are a number of well-established protocols for producing bisulfite reads that use high-throughput sequencing technologies such as methylC-seq [12] and reduced representation bisulfite sequencing [13]. Specific bioinformatics tools that focus in processing this type of data were developed [14–16]. Alternatively, protocols such as methylated DNA immunoprecipitation (Me-DIP) use a different approach for methylome analysis based on the use of specific antibodies for 5-methylcytosine. This protocol has been adapted for the use of both sequencing techniques (Me-DIP-seq [17]) and microarray technologies (Me-DIP-chip [18]). When dealing with microarray data, the critical steps to optimize the accuracy of the method include image processing and data normalization [19–21]. The final parameters calculated are the β-value and the M-value. The β-value represents a ratio calculated between the intensity of the methylated probe and the sum of the intensities of the methylated and unmethylated probes, while the M-value is defined as the $\log_2$ transform of the ratio between the intensities of the methylated and the unmethylated probes [22]. Many pipelines and specific software packages have been developed over time to carry out this initial processing of data. The R-Bioconductor repository is particularly useful since it offers up-to-date analysis packages written using the R statistical language [23].

Ultimately, the most common goal of methylome analysis is to find significant differences in methylation patterns when comparing several groups (i.e., diseased samples vs healthy ones, different tumor types, different developmental stages, etc.). Multiple hypothesis testing is required; methods such as *t* test or Wilcoxon rank sum test are commonly utilized as a basis. Afterward, an adjustment of the *p*-values obtained with these methods is carried out, always at the cost of reducing the statistical power of the analysis. The false discovery rate is currently the most used. Finally, a ranked list of DMRs is obtained for further analysis and interpretation.

2.2 Histone Proteins and Their Chemical Modifications

The study of histones has come a long way since they were simply considered as structural proteins. Although initially believed to be a simple physical support for the DNA within the cell nucleus, it is now clear that histones play critical functional roles [24]. Histone proteins are highly conserved throughout the different branches of the tree of life, being ubiquitous in eukaryotes and represented in some Archaea groups. In addition to canonical histone types (H1, H2A, H2B, H3, H4), several specialized histone variants have arisen during evolution. The recruitment of these variants into nucleosomes modulates the physicochemical properties of the chromatin structure, regulating the access of the transcription machinery to target genes [24].

In addition to the characterization of histone variants, posttranslational modifications in histones are considered fundamental epigenetic marks. Two main approaches can be used for the production of histone mark data: unbiased identification and quantitation of histone modifications using mass spectrometry methods, or genome-wide mapping of specific modified histones using chromatin immunoprecipitation (ChIP) techniques. Taken together, histones and their modifications offer an overwhelming range of possible combinatorial effects that remains to be fully understood. Several efforts have been made to tackle this complex regulatory mechanism, including the "histone code" hypothesis [25]. To unravel this hypothetical code, innovative analytical techniques in proteomics seek to find the patterns that would work as biomarkers of specific molecular processes.

In addition to the identification of histone genes and the genome-wide mapping of histones and PTMs, there are two main aspects of interest when considering histone proteomics: first, the quantification and comparison of expression levels and modification levels; second, the structural characterization and the simulation of the protein dynamics under specific conditions. In this section, these two issues are addressed, and an overview of the different types of data that can be produced in this context is provided.

2.2.1 Quantitative Protein Analysis

Current advances in proteomic techniques, using both gel-based and gel-free methods, allow high-throughput quantitative analysis of proteins. To date, the most widely used technique in proteomic analysis has been two-dimensional gel electrophoresis (2DGE). From a computational perspective, the analysis of 2DGE requires the development of image processing algorithms that allow the accurate reading and comparison of the gels, requiring the processing of thousands of spots that correspond to the various proteins separated by the technique [26]. The intensity of the spot correlates with the amount of protein present in the sample. Therefore, differential expression analysis can be carried out by aligning the spots obtained from a problem sample with those spots from a control sample with the aim of finding the correspondence and subsequently calculating the difference in their intensity levels [27].

Gel-based methods, however, are rapidly being displaced by "shotgun" methods, mainly involving liquid chromatography coupled to mass spectrometry (LC–MS) [28]. The nature of the data obtained from LC–MS analyses is substantially different from the data observed from the 2D gels. Mass spectrometry ionizes molecules and separates them according to their mass-to-charge ratio under an electrical field. Results are then recorded in the form of a mass spectrum. Mass spectra provide a graphical representation of the different masses of ions detected in the analysis. Different mass spectrum peaks correspond to different mass/charge ratios (usually the charge equals 1, thus the peak simply represents the mass of the ion), and the area under the peak corresponds to the quantity of ions detected. Automated analyses of mass spectra are possible by comparing the observed patterns of peaks with those stored in databases. Once the elements of interest are identified, absolute quantification analyses are possible using calibration curves, as well as relative quantification analyses.

LC–MS techniques do not allow just the identification of the proteins, but also the identification of their chemical modifications. Alternatively, ChIP-seq data on histone marks can also be considered quantitative, thus allowing differential analyses. Computational methods like that one described by Xu et al. compare the differences in read count between two sequencing libraries, using hidden Markov models (HMMs), with the objective of finding differential histone modification sites [29].

2.2.2 Structural Modeling and Dynamic Simulation

The traditional method for modeling the three-dimensional structure of proteins that have a well-defined crystal structure from X-ray diffraction or nuclear magnetic resonance analysis is comparative modeling (also known as homology modeling). In this method, an initial known structure of a homologous protein is used as a template to build the predicted structure of the target protein. This is possible because the protein structure is evolutionarily more conserved than the corresponding genetic coding sequence [30]. The structures of template proteins can be obtained in protein data bank (pdb) format files from the RCSB Protein Data Bank [31]. Three-dimensional modeling can be useful in the prediction of the dynamic effects caused by the structural and physicochemical variations. An application of these techniques in epigenetics would involve the prediction of changes in nucleosome and chromatin structure resulting from the replacement of canonical histones by histone variants. However, the study of histone variants poses specific challenges since the majority of the structural differences tend to accumulate in the most external and dynamic part of the protein chain [32], that is, the tails, which are especially difficult to study by X-ray diffraction methods and therefore difficult to model.

For dynamic studies, molecular dynamics simulations are frequently used [33,34]. In these methods, the known structure of the protein is translated into an array of coordinates for every atomic nucleus present. These are allowed to vibrate under a simulated force field while the instantaneous kinetic and potential energies for each atomic bond are calculated *ab initio*. These methods produce a very visual output by which the movement of the protein and the development of chemical processes can be observed, conveying a very important application for novel drug discovery and design [35,36]. These methods are applicable for the computational prediction of condensation levels of the DNA in the nucleosomes and, subsequently, to gauge the transcriptional accessibility of the genetic material, providing insights of great value for further experimental validation.

2.3 Nucleosome Positioning

Nucleosomes are the fundamental subunits of the chromatin, constituted by segments of DNA wrapped around a core structure of histone proteins. Nucleosome structure and positioning have direct implications for gene transcription since the majority of transcription factors cannot bind DNA packed by nucleosomes. Additionally, the positioning of nucleosomes can be actively modified by ATP-dependent remodeling factors that dynamically reorganize the structure of the chromatin [37]. Therefore, the accurate genome-wide mapping of nucleosomes and the analysis of their dynamics convey critical information about the epigenetic state of the cell.

The genome-wide analysis of specific chromatin components has been traditionally carried out using ChIP-based technologies (ChIP-seq or

ChIP-chip) [38]. ChIP-based techniques target specific proteins (usually histone variants including PTMs) using appropriate antibodies; that protein is then precipitated together with its associated DNA. More recently, the combination of a digestion step with DNases and high-throughput sequencing methods has been successfully used for this purpose, allowing a single-nucleotide resolution for the mapping of nucleosome positions [39]. With either one of these techniques, the bioinformatic analysis starts with the processing of high-throughput sequencing data followed by the mapping of those sequences on a reference genome to specify the protein-binding loci. The computational challenges of these methods are the common issues of short-read alignments [40,41].

When experimental data is lacking, however, it is still possible to perform computational predictions of nucleosome positioning based on motifs found on the genome sequence and thermodynamic properties of the chromatin [42,43].

2.4 Noncoding RNA

Part of what was considered years ago as "junk DNA" is nowadays an important focus of research for the scientific community. Although the inclusion of the ncRNA as an epigenetic mark at the same level of the methylome or the chromatin remains controversial, there is a general trend toward its acceptance [44–49], specifically regarding the role of long noncoding RNA in the epigenetic regulation of gene expression through interactions with chromatin-modifying proteins [50]. One of the most particular computational methods used in this field is the one that serves as the basis for comparative genomics. This approach is known as the "guilt by association" method and it implies functional inference through the observation of consistent coexpression events. In general, the construction of interaction networks in epigenetics is a problem that involves all different factors mentioned here as well as others that have been overlooked, such as chromatin-associated proteins (which may have a direct influence in the chromatin structure and dynamics) [51]. This problem involves difficulties well beyond the data processing related to the extraction of knowledge from heterogeneous data.

3. EPIGENETIC DATA INTEGRATION AND ANALYSIS

Data integration refers to the process by which a system combines information from different sources to make meaningful interpretations and produce relevant outcomes. Epigenetics is a field of research in which data integration is especially relevant, given the complex nature and interactions among the mechanisms responsible for various epigenetic marks. Data from various epigenomic studies, which may be obtained through different techniques, must be analyzed from a holistic perspective. Additional marks encompassing potential epigenetic relevance that can be considered include other chromatin-associated biomolecules such as histone-modifying enzymes and remodeling factors.

The number of existing bioinformatics resources created for this general purpose has steadily increased. These resources include tools, services, databases, standards, or even terminologies for each specific domain or area of expertise. Many publications highlight the importance and usefulness of properly integrating different types of biomedical data [52]. The large amount of information and diverse technology platforms raise multiple challenges, regarding not only data access, but also data processing [53]. More specifically, in the context of epigenetics, it is very likely that data-integrating approaches, with the aim of identifying functional genetic variability, represent a possible solution to the challenge of interpreting meaningfully the results of genome-wide profiling [54].

3.1 Quality and Format Standards for Data Integration

The most significant barrier for adopting a holistic perspective on data obtained from different sources is probably the standardization of methods and formats. Standardization in file formats allows the machine to find the specific pieces of information required for each step of the processing algorithm. This issue has been recognized and addressed through the establishment of consortia that specify necessary standards for the scientific community. The ultimate goal is to make all the produced data suitable for integrative analyses. The standards often refer to the metadata of the data sets (i.e., “data about the data”) that are generated, information about the experiments that produced those data sets, and other general characteristics. For biomedical and molecular biology research, the MIBBI project is an important global endeavor with the objective of establishing data standards to aid collaboration and meta-analyses [4]. Consisting of many subprojects, MIBBI suggests protocols to standardize reporting of data. From the well-established MIAME project for gene expression data obtained with microarray technologies [55], to the more recent MIAPE for proteomics data reports [56], it also addresses data standards relevant for epigenetics. In each of these protocols for standards compliance, all the various possible technologies and preprocessing methods for the generation of data sets are considered. Specific data exchange formats are designed to allow automated interpretation of the various data sets, mainly using tag-based structures such as the widely known XML.

Data standardization is crucial in bioinformatics and there exist multiple emerging standards that go beyond the scope of this chapter. Chervitz et al. provide a detailed review of the most relevant standards that may be applied to epigenetics [57].

3.2 Data Resources

A database may be defined as a searchable collection of interrelated data. The amount of high-throughput biological data generated has increased exponentially [58]. Furthermore, these types of data (including sequences, microarray data sets, gel imaging files, etc.) are not published in a conventional manner anymore; they are stored in databases. In fact, many journals require researchers to upload their data to specific repositories prior to trying to publish their research work.

The need for storing and linking large-scale data sets has also consistently increased. Archiving, curating, analyzing, and interpreting all of these data sets represent a major challenge. Therefore, the development of methods that allow the proper storage, searching, and retrieval of information becomes critical. Databases represent the most efficient way of managing this glut of data. The construction of databases and tools that allow accessing the data will enable the scientific community to manage and share vast amounts of high-throughput biological information. Hence, support for large-scale analysis is essential. Access to data must be facilitated and data must be periodically updated. Specifically in epigenetics, browsers have been made available to the scientific community facilitating the access and integration of the data generated by some of the initiatives described in the next section. Among these, we can find the UCSC Genome Browser [59], the Roadmap Epigenomics Visualization Hub (VizHub), and the Human Epigenome Browser at Washington University [60] (Table 1). Finally, knowledge extracted from various fields involving, among others, different disciplines within epigenetics and general biology, as well as clinical medicine, must be linked.

Given all of the above, it is clear that databases have become vital for carrying out successful bioinformatics research. They make data available to researchers in a format that is understandable by a machine. Hence, analyses can be carried out automatically with computers, managing great amounts of data and providing user-friendly interfaces. Data will be stored in predefined formats making possible the automatic retrieval of information. Ultimately, valuable extra information could be extracted if data are properly linked to external resources. However, an important

TABLE 1 Epigenetic Browsers

Browser name	Web site
UCSC Genome Browser	http://www.epigenomebrowser.org
Human Epigenome Browser at Washington University	http://epigenomegateway.wustl.edu/
VizHub	http://vizhub.wustl.edu

challenge that must be taken into account is the proper anonymization of data, to protect the privacy of the subjects who participate in the studies from which the data are collected.

3.2.1 Large-Scale Projects and Consortia

There exist numerous initiatives that manage huge amounts of diverse biological data, such as the Encyclopedia of DNA Elements (ENCODE) [61] or the Human Epigenome Project (HEP) [62]. The first project involves researchers from all over the world and can be considered a continuation of the Human Genome Project. Its objective is to identify all functional elements in the human genome and is funded by the National Human Genome Research Institute. Under the ENCODE project, many computational approaches were developed to handle epigenomic data [63–66]. The second project, on the other hand, seeks the identification and classification of genome-wide DNA methylation patterns for all human genes, studied in different tissues, linking this information to diseases and environmental conditions. This project is an international endeavor of global interest, and it is funded by public funds as well as private investment via a consortium of genetic research organizations.

Other relevant initiatives include the NIH Roadmap Epigenomics Mapping Consortium [67], which was launched in 2008. The goal of this project was to develop publicly available resources (more specifically, reference epigenome maps from a variety of cell types) of human epigenomic data to foster basic biology and disease-oriented research. On this basis, two data repositories were made available: the National Center for Biotechnology Information (NCBI) Epigenome Gateway and the Epigenome Atlas (see Table 2 in the next section).

The NIH Roadmap Epigenomics Program is a member of the International Human Epigenome Consortium [68], a growing international effort to coordinate worldwide epigenome mapping and to disseminate experimental standards for epigenome characterization, officially presented in 2010.

Other U.S. initiatives include the Epigenetic Mechanisms in Cancer Think Tank [69], sponsored by the National Cancer Institute in 2004, and the American Association for Cancer Research Human Epigenome Task Force [70], which emerged from a series of workshops and included scientists from all over the world.

Epigenetic research has been funded by several entities outside the United States as well. The European Union has dedicated significant amount of resources (more than €50M) over the years. Numerous initiatives have been funded, such as the above-mentioned HEP, High-Throughput Epigenetic Regulatory Organization in Chromatin, and Epigenetic Treatment of Neoplastic Disease, to focus on general questions such as DNA methylation, chromatin profiling, and treatment of neoplastic disease,

TABLE 2 Epigenetic Resources

Resource name	Web site
Cancer Methylome System	http://cbbiweb.uthscsa.edu/KMethylomes/
DBCAT	http://dbcat.cgm.ntu.edu.tw/
CREMOFAC	http://www.jncasr.ac.in/cremofac/
EpimiR	http://bioinfo.hrbmu.edu.cn/EpimiR/
HEMD	http://mdl.shsmu.edu.cn/HEMD/
Histome	http://www.actrec.gov.in/histome/
HistoneHits	http://histonehits.org
Histone Database	http://research.nhgri.nih.gov/histones/
Human Epigenome Atlas	http://www.genboree.org/epigenomeatlas/
Human lincRNA Catalog	http://www.broadinstitute.org/genome_bio/human_lincrnas/
MeInfoText	http://bws.iis.sinica.edu.tw:8081/MeInfoText2/
MethBase	http://smithlabresearch.org/software/methbase/
MethDB	http://www.methdb.net/
MethyCancer	http://methycancer.psych.ac.cn/
MethyLogiX	http://www.methylogix.com/genetics/database.shtml.htm
NCBI Epigenomics Gateway	http://www.ncbi.nlm.nih.gov/epigenomics/
NGSMethDB	http://bioinfo2.ugr.es/NGSmethDB/
NONCODE	http://www.noncode.org/
PEpiD	http://wukong.tongji.edu.cn/pepid
PubMeth	http://www.pubmeth.org/

respectively. Furthermore, the European Commission created in 2004 the Epigenome Network of Excellence [71] with the objective of studying major epigenetic questions in the postgenomic era.

Asia has focused mainly on disease epigenomes through the organization of various international meetings, as well as the creation of the Japanese Society for Epigenetics. Australia has also contributed to the Human Epigenome Project by creating in 2008 the Australian Alliance for Epigenetics and holding several workshops. Finally, Canada tried to position itself at the forefront of international efforts by creating the Canadian Epigenetics, Environment and Health Research Consortium, which is funded by the Canadian Institutes of Health Research and multiple Canadian and international partners.

As a worldwide initiative, and with the aim of joining all possible efforts, the Alliance for the Human Epigenome and Disease [72] was created. The aim of this project was to provide high-resolution reference epigenome maps, which will be useful in basic and applied research, will have an impact on how many diseases are understood, and, ultimately, will lead to the discovery of new ways of controlling these diseases.

3.2.2 *Data Models*

3.2.2.1 Traditional Database Models

Traditionally, most databases have followed what is known as the "entity-relationship model." This model tries to describe the data using entities, which correspond to concepts or objects, and relationships that may exist between these. In general, this model leads to a relational database implementation.

Most databases are usually offered as part of a tool or a service, which in most cases is presented through a Web interface. Most of them provide free access via the Internet and/or allow researchers the visualization or downloading of data. There exist numerous resources of this type in the field of epigenetics [73] and, in many cases, they have been constructed as a result of text mining analyses (see Section 3.3.3). Some of these widely used resources are listed in Table 2.

The NCBI, the European Bioinformatics Institute (EMBL-EBI), and the DNA Data Bank of Japan represent the three most important and largest available resources regarding biomedical databases. Major databases included as part of the first resource are GenBank (for DNA sequences), Gene Expression Omnibus, and PubMed (bibliographic database of biomedical literature). The EMBL-EBI provides major bioinformatics resources such as Ensembl, UniProt, ArrayExpress, and Reactome, among others, as well as tools and services to browse and analyze these databases.

The existing epigenetic databases can be broadly classified into several categories according to the type of data they store. The first category includes DNA methylation databases. These databases are useful for studying the covalent modification of a cell's genetic material. Among these, we can find DBCAT, MethBase, MethDB, MethyLogiX, and NGSMethDB.

The second category contains all of those databases related to histone data. Histone databases are important for research in the compaction and accessibility of eukaryotic and probably Archaeal genomic DNA. Some examples of this type of database are Histome, HistoneHits, and Histone Database.

The third category comprises databases related to chromatin-associated factors. Although the molecules involved in these processes are not directly part of the chromatin, they do interact with it. Within this category, databases including chromatin remodeling factors or noncoding RNA data can be found: CREMOFAC, Human lincRNA Catalog, and NONCODE.

The fourth category includes epigenetic databases related to cancer. Most of them are cancer methylation databases, which are helpful for analyzing irregular methylation patterns correlated with cancer. Some of these databases are Cancer Methylome System, MeInfoText, MethyCancer, PEpiD, and PubMeth.

Finally, other more general databases including all types of epigenetic information can be found. Some examples are EpimiR, HEMD, and NCBI Epigenomics Gateway. All the resources mentioned are listed in Table 2.

3.2.2.2 Nontraditional Models

Recently, new models involving semistructured or nonstructured data have gained more attention in scientific fields. Big companies such as Google or Amazon have put a lot of effort into NoSQL resources. NoSQL, also known as Not Only SQL, provides a mechanism for storing and retrieving data that is modeled different from the tabular relations used in relational databases. In this regard, the biomedical field has taken advantage of the Semantic Web and, thus, has focused on developing resource description framework (RDF)-based solutions.

The Semantic Web can be seen as an extension of the World Wide Web. It allows people to share content by providing a standardized way of representing the relationships between Web pages. Thus, machines will be capable of understanding the meaning of hyperlinked information and the information will be given well-defined meaning.

In this context, the RDF model should be highlighted. The RDF, although it was originally designed as a metadata model, is being utilized as a general manner of describing concepts or modeling information (being widely implemented in Web resources) and represents the immediate future [74].

Based on the use of RDF, many approaches in computational bioinformatics have been developed. As of this writing, mashups are the most frequent ones. These Web pages or applications, taking advantage of the Semantic Web, use content from different data sources with the aim of creating one unique service that will be displayed by means of a single graphical interface. Therefore, the main objective is to make searches easier and data more useful.

To integrate and standardize different databases, there have been approaches, such as Bio2RDF [75], that try to help solve the problem of knowledge integration in bioinformatics by developing a mashup application. Other authors developed an ontology-driven mashup that integrates two resources of genomic information and three resources containing information of biological pathways [76], proving that Semantic Web technologies provide an effective framework for information integration in life sciences. Cheung et al. demonstrate the power of the Semantic Web by applying different tools to two different scenarios, concluding that these could be used by people without programming experience to accomplish useful data mashup over the Web [77].

3.3 Knowledge Discovery in Databases

The aim of knowledge discovery in databases (KDD) is to make sense out of data. Traditionally, this task was carried out manually. Nowadays, however, given the amount of data involved, various computational methods and techniques have become indispensable, taking advantage of the processing power offered by computers and turning a tedious process into a largely automatized procedure.

Trying to extract knowledge from data is a nontrivial process. KDD involves several stages, each with its own complexity: data acquisition and storage, data preprocessing and transformation, data mining, and data postprocessing (see Figure 2). Hence, this includes data preparation and selection, data cleaning, incorporating prior knowledge in data sets, and interpreting accurate solutions from the observed results. The use of data warehouses, understood as central repositories of information obtained from the integration of data from diverse data sources, can make some of these tasks easier to deal with. This type of data resource was designed to facilitate massive data analyses and, therefore, the reporting of results. However, the still increasing volume of data available introduces a new challenge, that is, algorithms must be scaled to support massive data analysis and management.

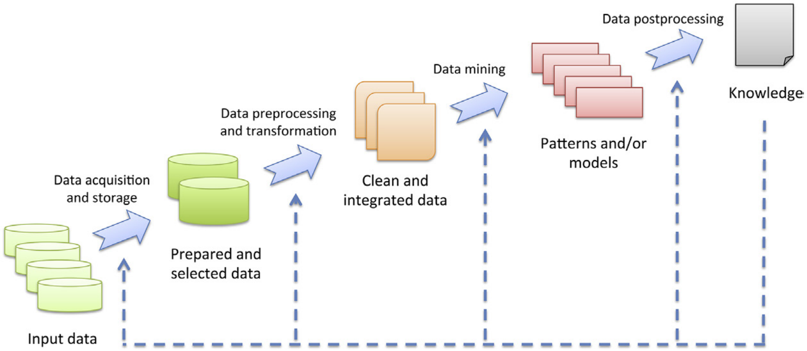
Data postprocessing
Knowledge
Data mining
Patterns and/or models
Data preprocessing and transformation
Clean and integrated data
Data acquisition and storage
Prepared and selected data
Input data

FIGURE 2 **The knowledge discovery in databases process. The various stages of this process and the products obtained as a result of each stage are shown.** Data are acquired and stored from different heterogeneous data sources. As part of this stage, it may be necessary to prepare and obtain a selection of the available input data. After that, the data goes through a preprocessing and transformation phase, in which they will be cleaned and integrated, obtaining a consistent data resource. Over this resource, data mining techniques can be applied, obtaining as a result different patterns and/or models. Finally, these results are postprocessed to extract the final product of the whole process, that is, knowledge.

3.3.1 Data Preprocessing

The data preprocessing stage involves not only selecting the data sources, but also handling missing data issues and altering the data if necessary. Data integration is not trivial, a convenient model must be chosen because this will have a direct impact on the performance of the bioinformatics pipeline.

3.3.1.1 Data Enrichment

Countless efforts have been made to integrate various resources and data existing over the Internet, especially in the biomedical field, and, for this purpose, standards and/or terminologies, such as ontologies, have been developed.

Ontologies can be defined as a set of concepts (terms) and the relationships among them as representing the consensual knowledge of a specific domain. Ontologies can be represented as graphs (with the nodes representing terms and the edges representing relationships) or as trees (with the nodes as terms and the branches representing hierarchical relationships).

Ontologies enable a clear and unified machine-readable vision of a domain that enables sharing, reusing (partially or totally), and extending knowledge. They are currently the most utilized form for representing biomedical knowledge. Furthermore, it is a field in which a lot of effort is being dedicated, with an increasing rate of usage. Within the Semantic Web (described in the previous section), they have become a key element since they make knowledge representation easier. They are also playing a major role regarding linked data, that is, a way of publishing structured data so that it can be interconnected and, hence, more useful. There currently exist more than 300 biomedical ontologies, and BioPortal [78] is the most important resource, providing access and tools for working with them.

Most studies involving epigenetics use Gene Ontology (GO) [79] to enrich data and draw conclusions from it. The GO project is a collaborative effort to address the need for consistent descriptions of gene products across databases. The GO project offers three structured, controlled vocabularies (ontologies) that describe gene products in terms of their associated (1) biological processes, (2) cellular components, and (3) molecular functions. Using GO terms across databases enables one to obtain more uniform queries. Apart from maintaining and developing these ontologies, other aims of the project are annotating genes and gene products, as well as assimilating and disseminating these data, or providing tools that facilitate accessing and utilizing the data and that allow functional interpretation of experimental data using GO, for example, via enrichment analysis.

GO-based analyses have been used to obtain clusters and to do functional analyses by looking for over- and underrepresentation of GO terms, possibly after combining genomic and transcriptomic data [80–83]. KEGG pathways have also been used in this context [84–86].

Finally, other existing approaches include Chromatin Regulation Ontology siRNA Screening, a new method that has been developed to identify writers and erasers of epigenetic marks [87]. In this work, the authors use this method to identify chromatin factors involved in histone H3 methylation and conclude that it facilitates the identification of drugs targeting epigenetic modifications.

3.3.1.2 Massive Data Set Analysis

Analysis of large data sets usually implies having to preprocess the data or represent it in certain ways. For instance, dimensionality reduction may be carried out to reduce the number of initial variables and obtain a more manageable data set [88]. Many techniques model data using graphs, networks, or matrices. These approaches are very powerful and allow hidden relationships to be found, as well as giving a new perspective of the data. Here, some examples of this are included.

Goh and Wong [89] present four scenarios in which building networks from proteomics data improves the results. They find that networks are convenient for identifying primary causes of cancer, given that they can reflect a structured hierarchy of molecular regulations. The typical network-based analysis framework for proteomics would include several stages. The first would involve data preprocessing, in which the data would be transformed into a network. The second would involve the usage of supervised or unsupervised methods such as those described above. Finally, the third stage would involve interpreting and evaluating the results obtained. Another example of this type of data representation is that proposed by Zheng et al. [90], in which they encode histone modification data as a Bayesian network for gene-regulatory network reconstruction.

Principal component analysis (PCA) has also been used to reduce data dimensionality. For example, Dyson et al. [91] and Figueroa et al. [92] use PCA on gene expression array data, while Volkmar et al. [93] use it on DNA methylation data. Cieślik and Bekiranov [63] take advantage of nonnegative matrix factorization to reduce the dimensionality of epigenetic data. Clustering techniques may also be used for reducing dimensionality [94].

3.3.2 Data Mining

The data mining stage, as part of the KDD process, involves choosing the most appropriate method or technique to be used for searching underlying patterns, as well as the creation of explicative and/or predictive models [95]. This stage comprises deciding which models and parameters might be appropriate for the overall KDD process. Also, searching for patterns of interest in a specific representational form (such as trees or rules) and applying tools such as regression or clustering are part of the data mining stage. Finally, the most important part is the interpretation of the relevant knowledge that can be drawn from the obtained results. This knowledge may then be used and/or incorporated into the bioinformatic analysis pipeline.

Two distinct approaches can be considered in data mining techniques: supervised and unsupervised methods. Supervised methods try to obtain relationships between a set of independent variables and a dependent variable. Therefore, the objective of these methods will be to infer a function from labeled training data. In this case, the computer's task will be the extraction of patterns from the input data to get the dependent or target variable. For this approach, two different types of problem can be identified. On one hand, the first type corresponds to those cases in which the dependent variable is categorical or nominal, usually referred to as a "class." On the other hand, the second type corresponds to those problems in which the dependent variable can take infinite numeric values, that is, a continuous variable. Thus, we call the first type classification problems and the second type regression problems.

Schäfer et al. present a Bayesian model for carrying out integrative analyses using "omics" data [96]. A Bayesian mixture model is utilized to compare and classify measurements of histone acetylation in order to identify DNA fragments obtained from ChIP analyses. Mo et al. [97] propose a framework for joint modeling of discrete and continuous variables obtained from integrated genomic, epigenomic, and transcriptomic profiling. Within this framework, the authors developed iCluster+. This method is capable of performing pattern discovery that integrates binary, categorical, and continuous data. It is based on different types of regression (linear and lasso regression), and it is used to extract novel biological information from integrated cancer genomic data for tumor classification and cancer gene identification. As a last example, Gonzalo et al. [98] apply logistic regression adjusted for different factors to colorectal cancer data, with the aim of evaluating the difference in DNA methylation data between two independent groups.

Unlike supervised techniques, when using unsupervised methods, no target variable is specified. In this case, instead of asking the computer to predict the value of a dependent variable out of a given data set (which corresponds to the independent variables), the question will be "which are the best four groups that can be made out of the data?" or "which variables are most likely to occur together?" On this basis, two types of problems can be identified. The first is clustering and it consists in grouping similar items together. The second is association analysis and it entails finding which features are most frequently found together.

Hierarchical clustering is by far the most popular method when trying to analyze epigenetic data. This technique groups data by creating a hierarchy of clusters, known as a cluster tree or dendrogram. Additionally, *k*-means is also frequently chosen to carry out this type of analyses. This technique partitions the observations into *k* clusters, in such a way that each observation will belong to the cluster that has the closest mean, known as a centroid. After that, centroids are updated as the mean value of the observations that belong to the cluster. This process is repeated iteratively until the centroids do not change.

DeltaGseg [99] is an R package that applies hierarchical clustering for preprocessing signals to perform estimations from multiple replicated series. Hence, molecular biologists/chemists will be able to gain physical insight into the molecular details that are not easily accessible by experimental techniques. Unsupervised hierarchical clustering was also utilized by Busche et al. [100] to cluster methylation levels and by Towle et al. [86] to analyze methylation patterns.

Zeller et al. [83] applied both hierarchical clustering and *k*-means to DNA methylation profiles to validate the results. Another study used both techniques to cluster transcription factors [101]. Clifford et al. [102] compared hierarchical clustering to other clustering techniques (*k*-means, *k*-medoids, and fuzzy clustering) to determine the most appropriate one for analyzing Illumina methylation data. Since no significant difference was found between the methods, a combination was proposed; the final output will be given by the method that achieves the best results in each case. McGaughey et al. applied *k*-means to methylation data and observed that genome-wide methylation signals can reliably distinguish tissues [103].

In contrast to these two widely used methods, Jung et al. developed a density-based PIWI-interacting RNA (piRNA) clustering algorithm named piClust [104]. This algorithm is provided as a Web service through a graphical interface. piClust works as follows: first of all, it determines the clustering parameters carrying out a *k*-dist analysis; then it clusters preprocessed and previously aligned reads; last, it scores and validates candidate piRNA clusters. Ucar et al. [105] present an unsupervised subspace-clustering algorithm, named "coherent and shifted bicluster identification," which was designed to identify combinatorial patterns of chromatin modification across a specific epigenome. It was believed that applying this tool to the epigenome would help in the understanding of the role of chromatin structure in gene expression regulation. Yu et al. [106] proposed an algorithm, named GATE, for clustering genomic sequences based on spatiotemporal epigenomic information. This algorithm is based on a probabilistic model that was developed to annotate the genome using temporal epigenomic data. Each cluster obtained, which was modeled as an HMM, represented a time series of related epigenomic states. Steiner et al. [107] used an artificial neural network, more specifically a self-organizing map, to perform clustering, multidimensional scaling, and visualization of epigenetic patterns.

Last, Bayesian methods are on the rise, becoming an interesting alternative for unsupervised classification. Zhang et al. [108] developed an adaptive clustering algorithm aimed at analyzing ovarian cancer genome-wide gene expression, DNA methylation, microRNA expression, and copy number alteration profiles following an integrative approach. The method proposed combines an adaptive algorithm based on the Bayesian information criterion with another deep clustering algorithm, which was published previously ("super *k*-means"). Finally, Wockner et al. [109] used a recursively partitioned mixture model to cluster DNA methylation data, with the aim of obtaining profiles that could be used as a future prognostic indicator of schizophrenia. This model combines a fuzzy clustering algorithm with a level-weighted version of the Bayesian information criterion.

3.3.3 *Text Mining*

Within the scope of data mining techniques, text mining or text data mining can be defined as the process of deriving high-quality information from text. This way, the information is obtained by observing patterns and/or trends through the application of various techniques, such as statistically based ones. Text mining usually entails structuring the input text (by parsing it, adding and/or removing linguistic features, and inserting the result into a database), deriving patterns within the preprocessed data, and, last, evaluating and interpreting the output. Therefore, it can be considered as a special case of data mining. Text mining is very useful in the process of building databases and allows automatizing literature search, which is usually done manually and is time-consuming [110–114].

Kolářik et al. [115] proposed an approach designed for the identification of histone modifications in biomedical literature with conditional random fields and for the resolution of known histone modification term variants by term standardization. As part of their work, these authors also developed a histone modification term hierarchy to be used in a semantic text retrieval system. They concluded that this approach significantly improves the retrieval of articles that describe histone modifications. Bin Raies et al. [116] presented an innovative text mining methodology based on the concept of position weight matrices for text representation and feature generation. This concept was applied in combination with the document-term matrix, with the purpose of accurately extracting associations between methylated genes and diseases from free text. This methodology is offered also as a Web tool called DEMGD.

Ongenaert and Dehaspe [117] proposed a tool for automatic literature retrieval and annotation of DNA methylation data named GoldMine. This tool, taking into account data introduced by the user (a list of genes, keywords, and highlighting terms), carries out a search over PubMed and then processes the results. Li and Liu [118] also perform text mining over PubMed, but in this case, to obtain a list of candidate biomarkers.

4. CONCLUSIONS AND FUTURE TRENDS

The need for processing and interpreting the huge volume of biological data being produced in the postgenomic era (especially that pertaining to the mechanisms underlying the epigenetic transmission and regulation of heritable information) is currently being addressed by the development of big international projects and standardization endeavors. With the fast growth of the field of bioinformatics, a new landscape of possibilities for massive generation of biological knowledge is in sight. The computational modeling of systems considering simultaneously all relevant factors in a time-resolved manner is indeed the ultimate frontier for epigenetic knowledge. The development of databases and specialized algorithms and software for dynamic simulations should enhance modeling and prediction in epigenetics. Given the amount of data involved and its predictable exponential growth, it is essential that researchers divide the work and that decentralized data storage is used. Multidisciplinary collaboration seems to be the most adequate way to cover all the possible research perspectives. Researchers usually have at their service high-performance computing systems, such as clusters, to carry out computationally expensive tasks. However, it will be essential to continue working on different types of data representations and making algorithms more efficient so that they are scalable for big data analysis. Not only will a decentralized approach be required to achieve this objective, but also the parallelization of the algorithms must be strongly considered. In this sense, technologies such as grid computing appear to be a very promising approach. This type of computing involves many networked, heterogeneous, and geographically dispersed computers, which will probably be loosely coupled, acting together to perform large tasks.

Although a considerable number of epigenetic resources are currently available, these repositories of information are not usually linked. This could be considered an example of what is known as “the functional silo syndrome.” To avoid this and with the aim of integrating and linking as much information as possible to take the best out of it, the RDF represents an interesting solution. In combination with ontologies, and what is known as ontology-based data mining, this will very probably be involved in the future of computational epigenetics. Multiple initiatives have been created to move forward in this direction.

LIST OF ACRONYMS AND ABBREVIATIONS

2DGE	Two-dimensional gel electrophoresis
ChIP	Chromatin immunoprecipitation
DMR	Differentially methylated region
EMBL-EBI	European Bioinformatics Institute
GO	Gene Ontology
HEP	Human Epigenome Project
HMM	Hidden Markov model
KDD	Knowledge discovery in databases
LC–MS	Liquid chromatography coupled to mass spectrometry
NCBI	National Center for Biotechnology Information
ncRNA	Noncoding RNA
NIH	National Institutes of Health
PCA	Principal component analysis
piRNA	PIWI-interacting RNA
RDF	Resource description framework

References

[1] Allis CD, Jenuwein T, Reinberg D. Epigenetics. New York: Cold Spring Harbor Laboratory Press; 2007.

[2] Gomez-Cabrero D, Abugessaisa I, Maier D, Teschendorff A, Merkenschlager M, Gisel A, et al. Data integration in the era of omics: current and future challenges. BMC Syst Biol 2014;8(Suppl. 2):I1.

[3] Marx V. Biology: the big challenges of big data. Nature 2013;498(7453):255–60.

[4] Taylor CF, Field D, Sansone SA, Aerts J, Apweiler R, Ashburner M, et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat Biotechnol 2008;26(8):889–96.

[5] Bird A. The essentials of DNA methylation. Cell 1992;70(1):5–8.

[6] Ng HH, Bird A. DNA methylation and chromatin modification. Curr Opin Genet Dev 1999;9(2):158–63.

[7] Bestor TH. DNA methylation – evolution of a bacterial immune function into a regulator of gene-expression and genome structure in higher eukaryotes. Philo Trans R Soc Lond Ser B Biol Sci 1990;326(1235):179–87.

[8] Jones PA. The DNA methylation paradox. Trends Genet 1999;15(1):34–7.

[9] Liang G, Salem CE, Yu MC, Nguyen HD, Gonzales FA, Nguyen TT, et al. DNA methylation differences associated with tumor tissues identified by genome scanning analysis. Genomics 1998;53(3):260–8.

[10] Costello JF, Fruhwald MC, Smiraglia DJ, Rush LJ, Robertson GP, Gao X, et al. Aberrant CpG-island methylation has non-random and tumour-type-specific patterns. Nat Genet 2000;24(2):132–8.

[11] Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet 2012;13(10):705–19.

[12] Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, et al. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 2008;133(3):523–36.

[13] Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 2011;6(4):468–81.

[14] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for bisulfite-seq applications. Bioinformatics 2011;27(11):1571–2.

[15] Lutsik P, Feuerbach L, Arand J, Lengauer T, Walter J, Bock C. BiQ analyzer HT: locus-specific analysis of DNA methylation by high-throughput bisulfite sequencing. Nucleic Acids Res 2011;39(Suppl. 2):W551–6.

[16] Ryan DP, Ehninger D. Bison: bisulfite alignment on nodes of a cluster. BMC Bioinforma 2014;15:337.

[17] Zhao MT, Whyte JJ, Hopkins GM, Kirk MD, Prather RS. Methylated DNA immunoprecipitation and high-throughput sequencing (MeDIP-seq) using low amounts of genomic DNA. Cell Reprogr 2014;16(3):175–84.

[18] Hsu YW, Huang RL, Lai HC. MeDIP-on-Chip for methylation profiling. Methods Mol Biol 2015;1249:281–90.

[19] Do JH, Choi DK. Normalization of microarray data: single-labeled and dual-labeled arrays. Mol Cells 2006;22(3):254–61.

[20] Wang X, Ghosh S, Guo SW. Quantitative quality control in microarray image processing and data acquisition. Nucleic Acids Res 2001;29(15):E75–5.

[21] Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 2002;30(4):e15.

[22] Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, et al. Comparison of beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinforma 2010;11:587.

[23] Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 2004;5(10):R80.

[24] Talbert PB, Henikoff S. Histone variants–ancient wrap artists of the epigenome. Nat Rev Mol Cell Biol 2010;11(4):264–75.

[25] Jenuwein T, Allis CD. Translating the histone code. Science 2001;293(5532):1074–80.

[26] Hoffmann F, Kriegel K, Wenk C. An applied point pattern matching problem: comparing 2D patterns of protein spots. Discrete Appl Math 1999;93(1):75–88.

[27] Dowsey AW, English JA, Lisacek F, Morris JS, Yang GZ, Dunn MJ. Image analysis tools and emerging algorithms for expression proteomics. Proteomics 2010;10(23):4226–57.

[28] Lambert JP, Ethier M, Smith JC, Figeys D. Proteomics: from gel based to gel free. Anal Chem 2005;77(12):3771–87.

[29] Xu H, Wei CL, Lin F, Sung WK. An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. Bioinformatics 2008;24(20):2344–9.

[30] Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A. Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.

[31] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic Acids Res 2000;28(1):235–42.

[32] Eirín-López J, González-Romero R, Dryhurst D, Méndez J, Ausió J. Long-term evolution of histone families: old notions and new insights into their mechanisms of diversification across eukaryotes. In: Pontarotti P, editor. Evolutionary biology. Springer Berlin Heidelberg; 2009. p. 139–62.

[33] Biswas M, Voltz K, Smith JC, Langowski J. Role of histone tails in structural stability of the nucleosome. PloS Comput Biol 2011;7(12).

[34] Ettig R, Kepper N, Stehr R, Wedemann G, Rippe K. Dissecting DNA-histone interactions in the nucleosome by molecular dynamics simulations of DNA unwrapping. Biophys J 2011;101(8):1999–2008.

[35] Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol 2011;9.

[36] Borhani DW, Shaw DE. The future of molecular dynamics simulations in drug discovery. J Comput Aided Mol Des 2012;26(1):15–26.

[37] Narlikar GJ, Sundaramoorthy R, Owen-Hughes T. Mechanisms and functions of ATP-dependent chromatin-remodeling enzymes. Cell 2013;154(3):490–503.

[38] Jiang C, Pugh BF. Nucleosome positioning and gene regulation: advances through genomics. Nat Rev Genet 2009;10(3):161–72.

[39] Henikoff JG, Belsky JA, Krassovsky K, MacAlpine DM, Henikoff S. Epigenome characterization at single base-pair resolution. Proc Natl Acad Sci USA 2011;108(45):18318–23.

[40] 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.

[41] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 2009;25(14):1754–60.

[42] Balasubramanian S, Xu F, Olson WK. DNA sequence-directed organization of chromatin: structure-based computational analysis of nucleosome-binding sequences. Biophys J 2009;96(6):2245–60.

[43] Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, Altschuler SJ, et al. Genome-scale identification of nucleosome positions in *S. cerevisiae*. Science 2005;309:626–30.

[44] Brown JD, Mitchell SE, O'Neill RJ. Making a long story short: noncoding RNAs and chromosome change. Heredity 2012;108(1):42–9.

[45] Lee JT. Epigenetic regulation by long noncoding RNAs. Science 2012;338(6113):1435–9.

[46] Magistri M, Faghihi MA, St Laurent III G, Wahlestedt C. Regulation of chromatin structure by long noncoding RNAs: focus on natural antisense transcripts. Trends Genet 2012;28(8):389–96.

[47] Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol 2013;20(3):300–7.

[48] Morlando M, Ballarino M, Fatica A, Bozzoni I. The role of long noncoding RNAs in the epigenetic control of gene expression. ChemMedChem 2014;9(3):505–10.

[49] Whitehead J, Pandey GK, Kanduri C. Regulation of the mammalian epigenome by long noncoding RNAs. Biochimica Biophysica Acta General Subj 2009;1790(9):936–47.

[50] Backofen R, Vogel T. Biological and bioinformatical approaches to study crosstalk of long-non-coding RNAs and chromatin-modifying proteins. Cell Tissue Res 2014;356(3):507–26.

[51] Suarez-Ulloa V, Fernandez-Tajes J, Aguiar-Pulido V, Rivera-Casas C, Gonzalez-Romero R, Ausio J, et al. The CHROMEVALOA database: a resource for the evaluation of okadaic acid contamination in the Marine environment based on the chromatin-associated transcriptome of the mussel *Mytilus galloprovincialis*. Mar Drugs 2013;11(3):830–41.

[52] Goh WW, Wong L. Computational proteomics: designing a comprehensive analytical strategy. Drug Discov Today 2014;19(3):266–74.

[53] Hawkins RD, Hon GC, Ren B. Next-generation genomics: an integrative approach. Nat Rev Genet 2010;11(7):476–86.

[54] Heyn H. A symbiotic liaison between the genetic and epigenetic code. Front Genet 2014;5:113.

[55] Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29(4):365–71.

[56] Martinez-Bartolome S, Binz PA, Albar JP. The minimal information about a proteomics Experiment (MIAPE) from the proteomics standards initiative. Methods Mol Biol 2014;1072:765–80.

[57] Chervitz SA, Deutsch EW, Field D, Parkinson H, Quackenbush J, Rocca-Serra P, et al. Data standards for omics data: the basis of data sharing and reuse. Methods Mol Biol 2011;719:31–69.

[58] Shakya K, O'Connell MJ, Ruskin HJ. The landscape for epigenetic/epigenomic biomedical resources. Epigenetics Official J DNA Methylation Soc 2012;7(9):982–6.

[59] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome Res 2002;12(6):996–1006.

[60] Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, et al. The human epigenome browser at Washington university. Nat Methods 2011;8(12):989–90.

[61] Consortium EP. The encode (ENCyclopedia of DNA Elements) project. Science 2004;306(5696):636–40.

[62] Rakyan VK, Hildmann T, Novik KL, Lewin J, Tost J, Cox AV, et al. DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. PLoS Biol 2004;2(12):e405.

[63] Cieslik M, Bekiranov S. Combinatorial epigenetic patterns as quantitative predictors of chromatin biology. BMC Genomics 2014;15:76.

[64] Benveniste D, Sonntag HJ, Sanguinetti G, Sproul D. Transcription factor binding predicts histone modifications in human cell lines. Proc Natl Acad Sci USA 2014;111(37):13367–72.

[65] Rosenbloom KR, Dreszer TR, Long JC, Malladi VS, Sloan CA, Raney BJ, et al. ENCODE whole-genome data in the UCSC Genome Browser: update 2012. Nucleic Acids Res 2012;40(Database issue):D912–7.

[66] Podlaha O, De S, Gonen M, Michor F. Histone modifications are associated with transcript isoform diversity in normal and cancer cells. PLoS Comput Biol 2014;10(6):e1003611.

[67] Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH roadmap epigenomics mapping consortium. Nat Biotechnol 2010;28(10):1045–8.

[68] Bae JB. Perspectives of international human epigenome consortium. Genomics Inform 2013;11(1):7–14.

[69] Sogn JA, Anton-Culver H, Singer DS. Meeting report: NCI think tanks in cancer biology. Cancer Res 2005;65(20):9117–20.

[70] Jones PA, Martienssen R. A blueprint for a human epigenome project: the AACR human epigenome workshop. Cancer Res 2005;65(24):11241–6.

[71] Akhtar A, Cavalli G. The epigenome network of excellence. PLoS Biol 2005;3(5):e177.

[72] American Association for Cancer Research Human Epigenome Task F, European Union NoESAB. Moving AHEAD with an international human epigenome project. Nature 2008;454(7205):711–5.

[73] Lim SJ, Tan TW, Tong JC. Computational epigenetics: the new scientific paradigm. Bioinformation 2010;4(7):331–7.

[74] Wang X, Gorlitsky R, Almeida JS. From XML to RDF: how semantic web technologies will change the design of 'omic' standards. Nat Biotechnol 2005;23(9):1099–103.

[75] Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Informatics 2008;41(5):706–16.

[76] Sahoo SS, Bodenreider O, Rutter JL, Skinner KJ, Sheth AP. An ontology-driven semantic mashup of gene and biological pathway information: application to the domain of nicotine dependence. J Biomed Informatics 2008;41(5):752–65.

[77] Cheung KH, Yip KY, Townsend JP, Scotch M. HCLS 2.0/3.0: health care and life sciences data mashup using Web 2.0/3.0. J Biomed Inform 2008;41(5):694–705.

[78] Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 2011;39(Web server issue):W541–5.

[79] 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;25(1):25–9.

[80] Wippermann A, Klausing S, Rupp O, Albaum SP, Buntemeyer H, Noll T, et al. Establishment of a CpG island microarray for analyses of genome-wide DNA methylation in Chinese hamster ovary cells. Appl Microbiol Biotechnol 2014;98(2):579–89.

[81] Triff K, Konganti K, Gaddis S, Zhou B, Ivanov I, Chapkin RS. Genome-wide analysis of the rat colon reveals proximal-distal differences in histone modifications and proto-oncogene expression. Physiol Genomics 2013;45(24):1229–43.

[82] Kalari S, Jung M, Kernstine KH, Takahashi T, Pfeifer GP. The DNA methylation landscape of small cell lung cancer suggests a differentiation defect of neuroendocrine cells. Oncogene 2013;32(30):3559–68.

[83] Zeller C, Dai W, Curry E, Siddiq A, Walley A, Masrour N, et al. The DNA methylomes of serous borderline tumors reveal subgroups with malignant- or benign-like profiles. Am J Pathol 2013;182(3):668–77.

[84] Bajpai M, Kessel R, Bhagat T, Nischal S, Yu Y, Verma A, et al. High resolution integrative analysis reveals widespread genetic and epigenetic changes after chronic in-vitro acid and bile exposure in Barrett's epithelium cells. Genes Chromosomes Cancer 2013;52(12):1123–32.

[85] Akulenko R, Helms V. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. Hum Mol Genet 2013; 22(15):3016–22.

[86] Towle R, Truong D, Hogg K, Robinson WP, Poh CF, Garnis C. Global analysis of DNA methylation changes during progression of oral cancer. Oral Oncol 2013;49(11):1033–42.

[87] Baas R, Lelieveld D, van Teeffelen H, Lijnzaad P, Castelijns B, van Schaik FM, et al. A novel microscopy-based high-throughput screening method to identify proteins that regulate global histone modification levels. J Biomol Screen 2014;19(2):287–96.

[88] Reutlinger M, Schneider G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. J Mol Graph Model 2012;34:108–17.

[89] Goh WW, Wong L. Networks in proteomics analysis of cancer. Curr Opin Biotechnol 2013;24(6):1122–8.

[90] Zheng J, Chaturvedi I, Rajapakse J. Integration of epigenetic data in Bayesian network modeling of gene regulatory network. In: Loog M, Wessels L, Reinders MT, de Ridder D, editors. Pattern recognition in bioinformatics. Berlin Heidelberg: Springer; 2011. p. 87–96.

[91] Dyson MT, Roqueiro D, Monsivais D, Ercan CM, Pavone ME, Brooks DC, et al. Genome-wide DNA methylation analysis predicts an epigenetic switch for GATA factor expression in endometriosis. PLoS Genet 2014;10(3):e1004158.

[92] Figueroa ME, Wouters BJ, Skrabanek L, Glass J, Li Y, Erpelinck-Verschueren CA, et al. Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. Blood 2009;113(12):2795–804.

[93] Volkmar M, Dedeurwaerder S, Cunha DA, Ndlovu MN, Defrance M, Deplus R, et al. DNA methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. EMBO J 2012;31(6):1405–26.

[94] Loss LA, Sadanandam A, Durinck S, Nautiyal S, Flaucher D, Carlton VE, et al. Prediction of epigenetically regulated genes in breast cancer cell lines. BMC Bioinforma 2010;11:305.

[95] Aguiar-Pulido V, Seoane JA, Gestal M, Dorado J. Exploring patterns of epigenetic information with data mining techniques. Curr Pharm Des 2013;19(4):779–89.

[96] Schafer M, Lkhagvasuren O, Klein HU, Elling C, Wustefeld T, Muller-Tidow C, et al. Integrative analyses for omics data: a Bayesian mixture model to assess the concordance of ChIP-chip and ChIP-seq measurements. J Toxicol Environ Health Part A 2012;75(8–10):461–70.

[97] Mo Q, Wang S, Seshan VE, Olshen AB, Schultz N, Sander C, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. Proc Natl Acad Sci USA 2013;110(11):4245–50.

[98] Gonzalo V, Lozano JJ, Alonso-Espinaco V, Olshen AB, Schultz N, Sander C, et al. Multiple sporadic colorectal cancers display a unique methylation phenotype. PloS One 2014;9(3):e91033.

[99] Low DH, Motakis E. deltaGseg: macrostate estimation via molecular dynamics simulations and multiscale time series analysis. Bioinformatics 2013;29(19):2501–2.

[100] Busche S, Ge B, Vidal R, Spinella JF, Saillour V, Richer C, et al. Integration of high-resolution methylome and transcriptome analyses to dissect epigenomic changes in childhood acute lymphoblastic leukemia. Cancer Res 2013;73(14):4323–36.

[101] Tian R, Feng J, Cai X, Zhang Y. Local chromatin dynamics of transcription factors imply cell-lineage specific functions during cellular differentiation. Epigenetics Official J DNA Methylation Soc 2012;7(1):55–62.

[102] Clifford H, Wessely F, Pendurthi S, Emes RD. Comparison of clustering methods for investigation of genome-wide methylation array data. Front Genet 2011;2:88.

[103] McGaughey DM, Abaan HO, Miller RM, Kropp PA, Brody LC. Genomics of CpG methylation in developing and developed zebrafish. G3 2014;4(5):861–9.

[104] Jung I, Park JC, Kim S. piClust: a density based piRNA clustering algorithm. Comput Biol Chem 2014;50:60–7.

[105] Ucar D, Hu Q, Tan K. Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering. Nucleic acids research 2011;39(10):4063–75.

[106] Yu P, Xiao S, Xin X, Song CX, Huang W, McDee D, et al. Spatiotemporal clustering of the epigenome reveals rules of dynamic gene regulation. Genome Res 2013;23(2):352–64.

[107] Steiner L, Hopp L, Wirth H, Galle J, Binder H, Prohaska SJ, et al. A global genome segmentation method for exploration of epigenetic patterns. PloS One 2012;7(10):e46811.

[108] Zhang W, Liu Y, Sun N, Wang D, Boyd-Kirkup J, Dou X, et al. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. Cell Reports 2013;4(3):542–53.

[109] Wockner LF, Noble EP, Lawford BR, Young RM, Morris CP, Whitehall VL, et al. Genome-wide DNA methylation analysis of human brain tissue from schizophrenia patients. Transl Psychiatry 2014;4:e339.

[110] Ongenaert M, Van Neste L, De Meyer T, Menschaert G, Bekaert S, Van Criekinge W. PubMeth: a cancer methylation database combining text-mining and expert annotation. Nucleic Acids Res 2008;36(Database issue):D842–6.

[111] Fang YC, Huang HC, Juan HF. MeInfoText: associated gene methylation and cancer information from text mining. BMC Bioinforma 2008;9:22.

[112] Fang YC, Lai PT, Dai HJ, Hsu WL. MeInfoText 2.0: gene methylation and cancer relation extraction from biomedical literature. BMC Bioinforma 2011;12:471.

[113] Harmston N, Filsell W, Stumpf MP. What the papers say: text mining for genomics and systems biology. Hum Genomics 2010;5(1):17–29.

[114] Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol 2010;593:341–82.

[115] Kolarik C, Klinger R, Hofmann-Apitius M. Identification of histone modifications in biomedical text for supporting epigenomic research. BMC Bioinforma 2009;10 (Suppl. 1):S28.

[116] Bin Raies A, Mansour H, Incitti R, Bajic VB. Combining position weight matrices and document-term matrix for efficient extraction of associations of methylated genes and diseases from free text. PloS One 2013;8(10):e77848.

[117] Ongenaert M, Dehaspe L. Integrating automated literature searches and text mining in biomarker discovery. BMC Bioinforma 2010;11(Suppl. 5):O5.

[118] Li H, Liu C. Biomarker identification using text mining. Comput Math Methods Med 2012;2012:4.