CENTRO INTERNACIONAL DE ESTUDOS
DE DOUTORAMENTO E AVANZADOS
DA USC (CIEDUS)

TESE DE DOUTORAMENTO

Topological Data Analysis of High-dimensional Correlation Structures with Applications in Epigenetics

Sara Prada Alonso

ESCOLA DE DOUTORAMENTO INTERNACIONAL
PROGRAMA DE DOUTORAMENTO EN MATEMÁTICAS

SANTIAGO DE COMPOSTELA
2020

DECLARACIÓN DEL AUTOR/A DE LA TESIS

D./Dña. **Sara Prada Alonso**

Título da tese: **Topological Data Analysis of High-dimensional Correlation Structures with Applications in Epigenetics**

Presento mi tesis, siguiendo el procedimiento adecuado al Reglamento y declaro que:

1) La tesis abarca los resultados de la elaboración de mi trabajo.
2) De ser el caso, en la tesis se hace referencia a las colaboraciones que tuvo este trabajo.
3) Confirmo que la tesis no incurre en ningún tipo de plagio de otros autores ni de trabajos presentados por mí para la obtención de otros títulos.

Y me comprometo a presentar el Compromiso Documental de Supervisión en el caso que el original no esté depositado en la Escuela.

En **Vigo**, **29 de octubre de 2020**.

Firma electrónica

AUTORIZACIÓN DEL DIRECTOR/TUTOR DE LA TESIS

D./Dña. **Antonio Gómez Tato**

En condición de: **Director/a**

Título de la tesis: **Topological Data Analysis of High-dimensional Correlation Structures with Applications in Epigenetics**

INFORMA:

Que la presente tesis, se corresponde con el trabajo realizado por D/Dña **Sara Prada Alonso**, bajo mi dirección/tutorización, y autorizo su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director/tutor de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

En **Santiago de Compostela, 29 de octubre de 2020**

Firma electrónica

AUTORIZACIÓN DEL DIRECTOR/TUTOR DE LA TESIS

D./Dña. **Maria de los Ángeles Casares de Cal**

En condición de: **Director/a**

Título de la tesis: **Topological Data Analysis of High-dimensional Correlation Structures with Applications in Epigenetics**

INFORMA:

Que la presente tesis, se corresponde con el trabajo realizado por D/Dña **Sara Prada Alonso**, bajo mi dirección/tutorización, y autorizo su presentación, considerando que reúne los requisitos exigidos en el Reglamento de Estudios de Doctorado de la USC, y que como director/tutor de esta no incurre en las causas de abstención establecidas en la Ley 40/2015.

En **Santiago de Compostala, 29 de octubre de 2020**

Firma electrónica

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Abstract

This thesis comprises a comprehensive study of the correlation of high-dimensional datasets from a topological perspective. Derived from a lack of efficient algorithms of big data analysis and motivated by the importance of finding a structure of correlations in genomics, we have developed two analytical tools inspired by the topological data analysis approach that describe and predict the behavior of the correlated design. Those models allowed us to study epigenetic interactions from a local and global perspective, taking into account the different levels of complexity. We applied graph-theoretic and algebraic topology principles to quantify structural patterns on local correlation networks and, based on them, we proposed a network model that was able to predict the locally high correlations of DNA methylation data. This model provided with an efficient tool to measure the evolution of the correlation with the aging process. Furthermore, we developed a powerful computational algorithm to analyze the correlation structure globally that was able to detect differentiated methylation patterns over sample groups. This methodology aimed to serve as a diagnostic tool, as it provides with selected epigenetic biomarkers associated with a specific phenotype of interest. Overall, this work establishes a novel perspective of analysis and modulation of hidden correlation structures, specifically those of great dimension and complexity, contributing to the understanding of the epigenetic processes, and that is designed to be useful for non-biological fields too.

Keywords: topological data analysis, correlation, high-dimensional data, epigenetics.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Acknowledgments

A partial time thesis requires a great ability to balance academic with professional development, being both strict and demanding. This section is dedicated to all the people that helped me to overcome both situations during this hard but beautiful path.

Firstly, I would like to express my gratitude to my supervisors, Antonio Gómez Tato and María de los Ángeles Casares de Cal, for their continuous support during these exciting research years. Their guide and experience were fundamental for the correct thesis development, understanding also the constraints generated by the distance and the partial time dedication. I am truly glad of working with them.

I would like to thank my company Clinipace, especially to my role models Verónica de Lázaro and Charlene Dark, for trusting on me. Leaving the company during these months was not an easy choice, but I have found in them all the possible understanding and energies.

My family, especially my parents Antonio and Carmen and my sister Laura, were always supportive of any decisions I took, so I need to thank them for what I am. I remember that I started to be curious about genomics thanks to our lively philosophical talks at lunchtime. "Total thanks" to my life partner, Mariano, for his unconditional love that gave me the strength to continue. I also need to mention my baby nephew Alex, who was able to make me forget the thesis by just being with him.

I am lucky to have friends like Leire and Marta who remind me what are the important things in life. I was absent during many moments, but they know that I am coming back. Thanks also to Natalia, who was an enormous psychological support, re-organizing my mind when it was becoming chaotic.

Last but not least, I would like to express how thankful I feel to all the public institutions and public research that allowed me to learn more than ever on my own during this research career. There is no science without science.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Contents

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Introduction

There is a crack in everything,
that's how the light gets in.

Anthem, Leonard Cohen.

The capacity of the current humankind to generate and store big quantities of data in multiple areas provokes an increasing data analysis need, leading to the excitement of an army of mathematicians. We are, more than our previous generations, dependent on the data we analyze and therefore on the predictions we generate. The new era of artificial intelligence is indeed supported by the availability of novel computational algorithms based, ideally, on founded mathematical theories.

Among this increasing data offer, there is one specific data source that has a special interest for us due to its importance on human evolution: *genomics* (or generally, omics studies). In particular, the study of the genomic modifications that do not alter the DNA sequence and are produced by the interaction between subjects and their environment, called *epigenetics*, is increasingly being studied as part of the *epigenomics* field. Epigenetics gives a different perspective to human life, as it supports the idea that our lifestyle decisions play a major role in ourselves and the future generations [1]. It also opened the door to a different way of doing medicine, as epigenetic disorders may be reversed. Epigenetic alterations were already linked to different diseases like cancer, or to different conditions as the aging process. Generally, the study of the epigenetic modifications and their overall functioning within the genome is not an easy task. It has many different levels of complexity, and maybe some of them are still unknown.

Additionally, the relationship between epigenetics and the spatial design of the DNA within the nucleus of the cell (what is normally referred as "the three-dimensional structure" of the genome) is under current study, adding one level more of complexity to the whole system pointing out novel discoveries. This spatial design is capable to "join" distant genomic regions that may end in a common functioning or a genomic alteration.

The success of epigenetic studies is increasingly dependent on the development of efficient mathematical models and computational algorithms that are able to accurately analyze that big complex data collection for its posterior interpretation. Indeed, the research field called "mathematical biology" focuses on the use of mathematical tools to study biological systems and investigate their structure, development, and behavior. From the first simpler applications of the Fibonacci series or the golden ratio, the use of complex mathematical realization applied or motivated by biological problems is increasingly growing above all from the genome sequencing (with "systems biology"). We are at the dawn of a new revolutionary way of understanding our biology and future medicine. Constructing intellectual bridges from abstract mathematical formulation to real problems is needed to analyze all the complexity levels and do what mathematicians know: to travel to the depths of the reasoning cave.

In light of this situation and the mathematical challenges derived, this thesis focuses on providing with novel epigenetic data analysis tools through a topological perspective. But, what is really "data analysis"? and "topology"?

A brain is a perfect machine of data analysis. From the first life steps, it receives information trough the sensory system and analyzes it in a less to more sophisticated way. The first phase of this analytical thinking process that occurs naturally is usually data segmentation. Human babies, for instance, differentiate the flavor of a lemon from the flavor of milk, but they do not know the reason of the difference. Next time they taste a citrus aliment, it will "go" directly to the "lemon side" (sensory memory). This kind of automatic learning, that goes prior to the use of the language and communication, could be considered an "unsupervised learning".

However, once we start to be more conscious about ourselves and our environment, we tend to label the information learned and so the groups or clusters that we have previously detected. This may be considered then a "supervised learning". Indeed, there is a very typical childhood phase, where we continuously ask for names and reasons (what is that? why is that?). Every time we knew something novel, the information goes directly through the processing chain, that quickly labels it and stores it. For example, a three years old baby is able to distinguish a giraffe from a dolphin and even is able to distinguish terrestrial animals from aquatic animals. Sometimes this label is incorrect, but the own learning machinery will probably fix it. Having all of this into account, one can easily imagine how the process chain and the labeling gets affected by an alteration of the sensory system that is related to by many neurological disorders.

Interestingly, in this second phase of cognitive development, the sensory system is also able to adopt a geometric perspective, arranging the space according to the

elementary relations of similarity, proximity, separation, enclosure or continuity. This first way of detecting highly abstract mathematical structures is present in low aged kids, that may be indirectly drawing geometric structures from a topological perspective through their main topological invariants. For instance, they may not be drawing very differently a circle from a triangle, but they would be drawing quite accurately the fact that two figures are non-intersecting.

Indeed, *topology* is the field of mathematics that studies the properties of geometrical objects unaffected by continuous transformations of shape or size of figures. The capacity to deform a donut into a cup, or to define how different looks letter "D" from the letter "C" (holes/non-holes). Topological deformations seems to be aquatic and natural, and they describe the world from a very particular perspective. Being studied for centuries, topology is of interest for many distinct fields as biology, physics or even art.

Figure 1: Mobius Strip II (Red Ants) (1963), by M.C. Escher. All M.C. Escher works © 20120 The M.C. Escher Company - the Netherlands. All rights reserved. Used by permission. www.mcescher.com

Coming back to our brain processing machinery, technological and digital development add huge quantities of information to the processing chain, but their labeling into a cluster becomes complicated and the chain may fail. Sometimes we can train our brain to be able to do an approximate labeling (with analytical training, for example), but most of the times the amount or level of complexity of data leads us to ask for help. The data analysis comes then from this natural failure and tries to imitate our brain's processing chain. Thus, we initially go through a first exploratory unsupervised phase to segment the data and detect clusters that will be afterwards used to classify new information in a supervised phase. That is how *mathematical data analysis* is born.

There is currently a lack of standard and efficient analytical tools to deal with

the great quantities and varieties of high-dimensional data (also referred as "High Dimension, Low Sample Size" (HDLSS) data, which means, much more variables than independent observations are taken into account) as the genetic one. Particularly, the analysis of big-dimensional correlation structures is a pending topic in the epigenetics field. In our opinion, knowing the epigenetic correlation structure is crucial to understand completely the biological interactions that may be deregulated and manifested by a specific phenotype (as a disease), and could be associated with the described three-dimensional architecture of the genome. As we sill show in this thesis, the topological analysis of the large correlation structures contributes greatly to their understanding and interpretation.

Generally, the application of algebraic topology (field of topology that combines abstract algebra with the study of topological spaces) in data analysis through *topological data analysis* (TDA) provides with an efficient perspective, as the study of the "shape" of the data is key to extract underlying data characteristics doing minimal prior assumptions about their distribution. The application of the main aspects of Morse theory and homology groups theory allows to obtain the principal data characteristics in different ways. Moreover, topology may extract the topological invariants of the data space itself, which helps with the accurate design of mathematical models to describe it and predict it. Furthermore, the dimensionality reduction is key to treat high-dimensional datasets with thousands of variables. For instance, TDA techniques reduce the data dimension through the creation of simplicial complexes that reproduce those main data characteristics and provide with a visualization tool. As a powerful methodology of computational topology, TDA has a machine learning design. TDA is indeed the theoretical inspiration of the present work.

Motivated by the biological need of understanding better the complex epigenetic interactions, we developed a research work where the study of the correlation of such huge datasets is the principal target. We aimed to extract the main topological elements from the correlation structure of the epigenetic modifications (particularly of DNA methylation) to understand and model the correlation design and establish a link between epigenetic structure and functionality. We developed novel mathematical approaches and analytical tools to help on a higher understanding of the methylation correlation patterns and serve as an open door for other non-biological applications.

Using the topological data analysis idea, our main proposal is to study the correlation through the topological properties of the associated correlation networks, which represents a novel method to describe and model those structures. This analysis was done locally and globally to cover distinct complexity levels, and designing different methodologies for each aim. We generated a model to describe the local correlation structure and developed a computational algorithm to study

the global correlation and the alterations of the DNA methylation with different sample conditions (as aging or disease status). The main idea of the designed algorithm is to do big data analysis in an easy and quick way, automatizing the process of dealing with thousands of variables. Both methods are novel ways of studying the epigenetic landscape and they pointed out new promising biological discoveries in this field.

This thesis has therefore a high transversality, where different mathematical techniques were used to overcome biological problems. Main aspects of algebraic topology, graph theory, statistics, and computer science were used to develop efficiently the analyses done. Moreover, it could be considered a multidisciplinary work as we analyze deeply biological structures that are currently being investigated. A comprehensive understanding of complex genomic structures was needed to propose a rational mathematical hypothesis and model. Nevertheless, our ideas and technical proposals could be applied to any other field where high-dimensional data can be gathered.

Derived from this transversality, this thesis opens the door to different research ways, from a more theoretical path centered on the description of the correlation networks and their topological properties; to a more applicable direction focused on the evolution and contributions of the algorithm created, together with the biological interpretation of results.

The present work is hence divided into four main parts. The first part is an introduction to the problem statement and methodology. The second and third parts present our developed work and, finally, the last part includes an extensive description of the conclusions and open research. Specifically, this thesis contains:

1. The first part is an introduction to the biological and mathematical challenges to solve. It is divided in three chapters that contain:

   (a) The first chapter contains a description of the principal biological concepts that we use, and that are needed to understand the approach of the problem and relevance of the work. We describe extensively the epigenetic regulation, focusing on DNA methylation and its alterations with several features as the aging process or diseases as cancer. We explain here the characteristics of the methylation datasets we analyze. Please note that public data from humans was always used as a basis of the present work.

   (b) The second chapter includes the mathematical hypothesis derived from the biological problem and its argumentation. Also, we specify the first exploratory analyses done based on this hypothesis. We introduce the mathematical challenges that current techniques are unable to solve (as

the analysis of big correlation structures) and present the need of novel analytical tools in the direction of the topological data analysis approach.

(c) The chapter 3 of this first part is a deep description of TDA methodology, that inspired us, and the mathematical theory behind. We also include an extensive description of two of the main TDA techniques, Mapper and persistent homology, presenting examples in the epigenetics field and introducing our analytical proposal.

2. The second part of this work contains two chapters and consists on the deep analysis of the correlation structure of DNA methylation from a local perspective. We study the spatial distribution of the correlation within CpG islands and its main topological properties. To do it, we represent the correlation matrices as correlation graphs that have a special modulated design based on this spatial distribution. The chapter 4 comprises an extensive study of the graph's characteristics based on graph theory principles and persistent homology. The chapter 5 describes a model of the correlation structure of the CpG islands based on the properties studied. This model estimates successfully the interactions between CpG sites with high correlation. Besides, the model is able to distinguish distinct correlation behaviors among different age sample groups, so we are able to measure the potential evolution of the correlation design with the aging process.

3. The third part includes four chapters presenting the computational algorithm developed, called *MultiNet*, and the derived results. The main description of the algorithm and its implementation is specified in chapter 6. The chapter 7 contains a guide for the needed parameters selection, and we present the contributions of the algorithm to the epigenetics study in chapter 8. We propose here a model to detect the correlation structure globally that improves the current algorithms used to analyze high-dimensional datasets. This could be viewed as an extension of part II, using the same topological perspective but with a wider design. The algorithm detects, moreover, methylation trends with different sample groups of interest (as case/control studies) including prediction models to detect epigenetic biomarkers that may be useful in diagnosis. We use the algorithm to study epigenetic modifications with the aging process and with cancer, obtaining results in line with public information plus novel promising findings. Computationally, it was designed to be fast and easy to use by diverse professionals. The algorithm was developed in R, as stated in chapter 9.

4. The last part is composed by the conclusions of the work. We summarize the novel contributions and the interpretation of the results as a whole. We also highlight the different research pathways that could be opened after this investigation.

Objectives

The general goal of this thesis is to develop novel analytical tools to study high-dimensional epigenetic correlation structures based on founded algebraic topology theories represented by the topological data analysis approach. In particular, we aim:

- To develop mathematical models and computational algorithms to describe and predict the non-random design of the DNA methylation correlation structure, detecting epigenetic patterns and markers associated with a sample condition or a phenotype, like the age or a disease.
- To understand better the local and global correlation structure of the epigenetic processes potentially linked to the genome architecture.
- To measure the evolution of the correlation with the age.
- To present mathematical strategies that could be useful beyond the biological context.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Part I

Biological Context and Mathematical Methodology

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 1

Main Biological Concepts

Biology is a huge field of science with many areas of study as evolution, ecology, and genetics, which contain themselves a lot of research specializations. Among them, what is the importance of epigenetics and the role of mathematical data analysis on it? As we have a biological motivation to be solved mathematically, within this chapter we will introduce the needed biological concepts [2], centered on epigenetics, to understand the context and relevance of the present work and the generation of the related mathematical hypothesis. We will also introduce the design and content of the databases that we analyze in this work.

1.1 Basic Genetics

Cells are the basic unit of life of known organisms. There is a huge variety of cell types but they all present a few common characteristics: the genetic material, the cell membrane and the cytoplasm. The nuclear membrane and genetic material make a nucleus, that is rounded by the cytoplasm. Around the cytoplasm we find the cell membrane, also known as the plasma membrane, that acts as the boundary of the cell and separates it from the external environment. Cells are basically divided into two types as per their structure: those with their genetic material in a nuclear membrane (called *eukaryotic cells*), and those without a nuclear membrane (*prokaryotic cells*).

The genetic material in each cell is found in the form of molecules known as *DNA*, or deoxyribonucleic acid. The total set of DNA in an organism is the *genome*, also called the "book of life", and it contains all the information about the organism biology. DNA molecules are the cornerstone of genetics and their structure and behavior were subject of continuous study, even nowadays. The study of genetics is key to completely understand ourselves and our environment, and correctly treat or cure diseases. As a summary, DNA contains all the information that keeps a cell working and keeps an organism alive with certain characteristics.

DNA was observed firstly by Frederich Miescher in 1869, a German biochemist. In 1953 James Watson, Francis Crick, Maurice Wilkins and Rosalind Franklin discovered the structure of the DNA and realized that it could contain important biological information. Some of them were, in fact, awarded with the Nobel Prize in Medicine in 1962.

RNA (ribonucleic acid) is another essential nucleic acid molecule very important for its role in translating the information from DNA molecules, a process called *transcription*. DNA molecules are transcribed into RNA molecules which then translate the information to cells through the messenger RNA (mRNA) to build the proteins. Said with other words, the DNA chain would be the alphabet and RNA our capacity to read or talk. Nucleic acids are made by joining many smaller molecules called *nucleotides* (that would be the letters of the "DNA alphabet"). Each nucleotide molecule contains a sugar, a base and a phosphate group. There are two key differences between DNA and RNA molecules: DNA has a sugar called *deoxyribose* while RNA molecules have a sugar called *ribose*. As for the nucleotide bases, both DNA and RNA have four types of bases. Three of them are found in both: adenine (A), guanine (G), and cytosine (C). The difference is in the fourth base, where DNA has a base called thymine (T), RNA has a base called uracil (U). Those "letters" encode the genetic information present in the DNA or RNA. Our DNA chain is composed of billions of nucleotides making an "alphabet" of around two meters of length contained in each one of our cells. Despite we share most of this code, a single modification of one base may produce really significant and unexpected changes in our biology.

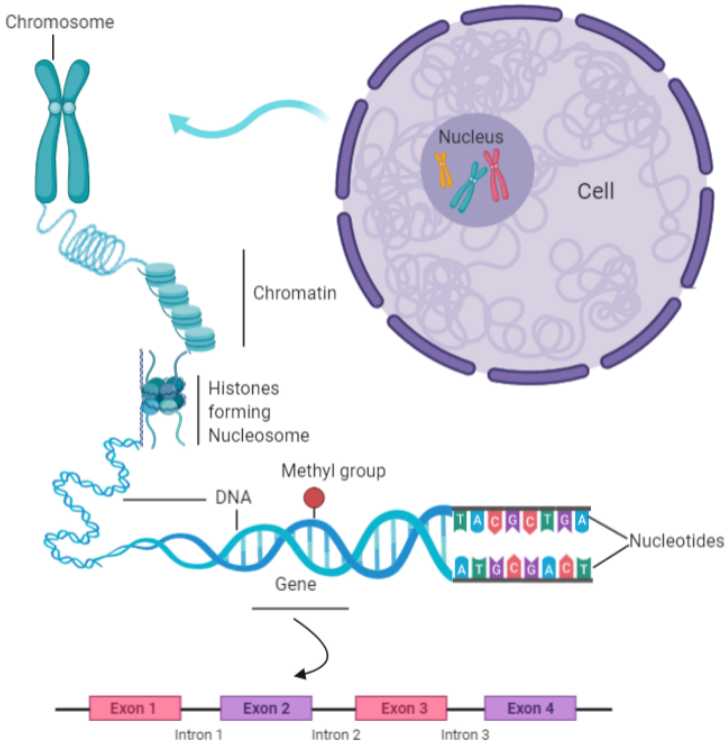
DNA molecules come in pairs. Each strand of DNA is bonded to a second strand and forms a *double helix*. They are bounded via the four bases and form defined base pairs: adenine always bonds with thymine and cytosine always bonds to guanine. Not all DNA is functional DNA, which means that not all DNA is potentially converted into proteins or has a specific known function. There is also a great part of non-coding DNA, that refers to genome segments which have no selected-effect function (that we know). During the reproduction of a cell, also known as cell division, the DNA of the cell is replicated so that both new cells have the same DNA. DNA replications occur by splitting apart the two strands of a DNA double helix and using each strand as a template to create new DNA molecules.

Chromosomes are DNA molecules with part of the genetic material. They are formed in the nucleus of eukaryotic cells during cell division. They are composed of the chromatin fiber, that is made of nucleosomes (histone octamers with DNA enveloped around it). Chromatin fibers are packaged by proteins into a compacted structure called *chromatin*. This high-level condensation of the chromatin allows the very long DNA molecule to fit into the cell nucleus. A complete set of chromosomes

in a cell nucleus contains all the DNA of a cell. Please see figure 1.1 for a visualization of the DNA structure in the cell nucleus.

Chromosomes are then relatively large microscopic structures. The number of chromosomes that make a complete set varies between species. Humans have 46 chromosomes (each body cell contains identical copies of 23 pairs of chromosomes), some species have more than 100, and others only two. The two chromosomes in a pair are almost equal, with the exception of the sex chromosome that has two types, X and Y: XX for woman and XY for men, being both who determines the sex.

Figure 1.1: Representation of the DNA structure within the cell.



A distinct and unique section of DNA that provides the information for one specific action or functionality is called a *gene*. It is also the basic unit of genetic inheritance. A lot of genes are constituted by codifying regions (exons) interrupted by non-codifying regions (introns) that are removed in the RNA process to form proteins. After the gene promoter regions, genes reading goes from left (5'UTR part) to right (3'UTR part). The quantity of genes per organism is really variable, for instance homo sapiens have around 25,000 genes. Gene mutations occur when the number or order of bases in a gene is disrupted. In addition, gene expression can be altered through different mechanisms without altering the DNA sequence. Those mutations may cause diseases or even death, despite most of them are benign and the base of evolution.

Each different form of the same gene is known as an *allele*. In sexual reproduction, offspring receive one allele of each gene from each parent. Gregor Mendel was the first person understanding the process of inheritance. Around the 1800s, several experiments were developed to test how traits are passed from parents to their offspring. Traits can be influenced by multiple genes and the environment altering their expression.

Early efforts at sequencing genes were intensive, from the genome sequencing of Haemophilus influenzae in 1995 to a first published human genome sequencing in 2001. The sequencing of the human genome was one of the most important milestones of the XXI century, promising a revolution in the way we understand the behavior and evolution of all species in the world. Since 2001, millions of scientific research were developed trying to decode and understand the huge amount of data that DNA sequencing pointed out. The scientific community is still on the early steps of the long way we have until understanding completely the whole genome information and its potential uses. The data is there, we just need to analyze it.

Different techniques were used since then, being more and more sophisticated and faster with time. In addition, the interest and investment from the research community increased during the last years, from the public and private sides. In fact, nowadays, the personalized DNA sequencing for humans is one of the increasing business, where a lot of companies offer reasonable prices for sequencing your DNA and give you information about your food preferences or diseases likelihood. The design of genetic-based drugs is also on an evolution process that promises important results.

The main approaches followed to obtain and analyze the genome data are the called genome-wide association studies (GWAS), that are able to find biological biomarkers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular pattern or disease. Such studies are especially useful in finding genetic variations that contribute to complex diseases, such as cancer, diabetes, heart disease and neurological disorders. GWAS studies compare the differences in the sampling frequency of single nucleotide polymorphisms, frequently called SNPs, that is the most common type of genetic variation among people. However, despite the success of those methods to find disease-related genes, there are several disadvantages that make them inefficient from a biological, computational, and financial perspective [3].

The "microarray" technology developed in the last years allows to measure the levels of RNA for hundreds of thousands of genes at the same time. This provides with an efficient way of determining the gene pattern expression in multiple tissue types pointing out new challenges on classifying the information and doing statistical

analysis. The databases generated with this technology are the ones we will analyze.

More recently, a new era is being developed and applied since 2013, based on genome editing or genome engineering. Following the National Institute of Health (NIH), *genome editing* is a group of technologies that give scientists the ability to change an organism's DNA. These technologies allow to alter the genetic material of the DNA sequence on purpose at particular locations in the genome. Several approaches as *CRISPR* (Clustered Regularly Interspaced Short Palindromic Repeats) have been developed in the last few years. Despite this is recent technology with an ongoing ethical debate, it is definitively an important step forward in medical science [4].

The National Human Genome Research Institute (NHGRI) supports the public research consortium named *ENCODE*, the Encyclopedia of DNA Elements, to identify all functional elements in the human and mouse genomes. ENCODE has produced vast amounts of data since it started in 2003 that can be accessed through the ENCODE Portal [5]. The National Center for Biotechnology Information (NCBI) [6] contains a great amount of genomic information and data too.

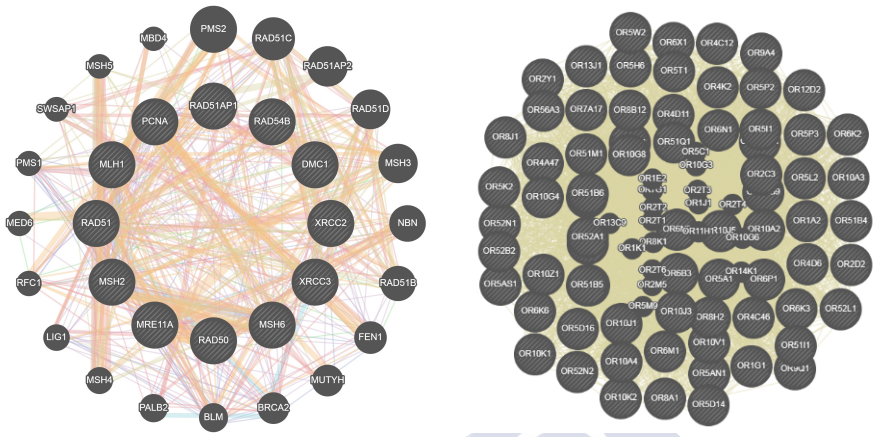
1.1.1 Gene Networks and Biological Pathways

A network that has been inferred from gene expression data is called a *gene regulatory network*, briefly denoted as GRN [7]. GRNs provide information about regulatory interactions between regulators and their potential targets; gene-gene interactions, protein-protein interactions. The increasing detection of GRNs, specifically associated to different disease conditions, makes possible their mathematical analysis [8, 9] to enrich our understanding about them with important applications in drug development.

Related to gene networks are the *biological pathways*. As specified by the NIH, biological pathways could be defined as a series of actions among molecules in a cell that leads to a certain outcome or a change in the cell. There are many types of biological pathways. Among the most well-known are pathways involved in metabolism, in the regulation of genes and in the transmission of signals. It is not easy to define when a biological pathway starts and ends, and its complexity is higher than initially thought. A *biological network* is then formed from the interaction of many biological pathways. Still, several biological pathways remain to be discovered or their functionality is not totally understood.

The KEGG pathways database is a collection of pathway maps representing our knowledge on the molecular interaction, reaction and relation networks for human diseases, genetic information processing, drug development, cellular processes,

Figure 1.2: Example of gene interaction networks. Each node is a gene and they are connected based on their expression relationship. The second network is heavily dense and contains some of the genes from the olfactory receptor (OR) gene family.



environmental information processing, metabolism and organisms systems. It is the most known official database of biological pathways and can be found in a web page [10] or be accessible trough programming environments as R (as for example, with the package *clusterProfiler*, and the function *enrichKEGG*).

The study of gene networks and biological pathways associated to different diseases has been increased in the recent years as a result of a wider analysis, where the complete map of biological interactions could be understood. Sometimes, the separated study of genes or proteins does not give substantial results that indeed can be found when digging into the related gene networks or biological pathways. In addition, a biological pathway could be related to more than one alteration or disease, leading to potential common medical solutions or a deeper knowledge of the condition.

1.2 Epigenetics

Up to early 90s, it was thought that the genetic inheritance present in the DNA sequence was the unique piece of information about ourselves and, therefore, our destiny. During the 90s decade, a new trend emerged based on rethinking the genetic procedures and understanding that both the environment and individual lifestyle can also directly interact with the genome. *Epigenetics* was born as a result of that wider view, describing the changes produced in our genome due to the interaction with the environment. It was also known that those changes occur during all life and can be transferred to the offspring through epigenetic inheritance [11, 12, 13, 14].

The term epigenetics, which was coined by Waddington in 1942 [15], was derived from the Greek word "epigenesis" which originally described the influence of genetic processes on development (the Greek prefix epi- ["over"] means "on top of" or "in addition to" the traditional genetic notion). Waddington already anticipated in his study:

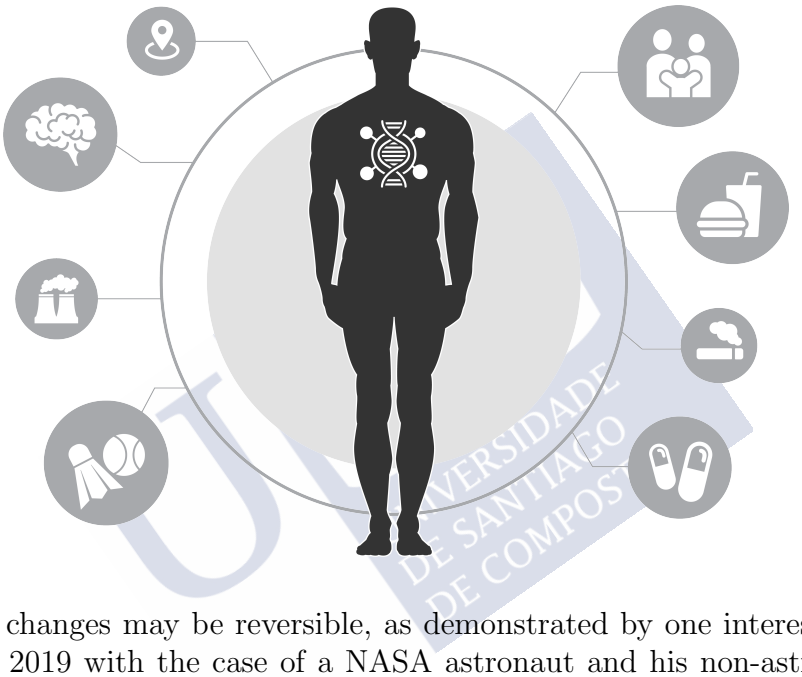
> [..] We might use the name "epigenetics" for such studies, thus emphasizing their relation to the concepts, so strongly favorable to the classical theory of epigenesis, which have been reached by the experimental embryologists. We certainly need to remember that between genotype and phenotype, and connecting them to each other, there lies a whole complex of developmental processes. It is convenient to have a name for this complex: "epigenotype" seems suitable [...].

The field of epigenetics describes therefore the genomic changes caused by our environment that do not involve modifications in the DNA sequence itself. In fact, the called *epigenome* refers to the group composed by the genome and its corresponding epigenetic profile. Although our epigenetic marks are more stable during adulthood than during early development, they are still thought to be dynamic and modifiable by lifestyle choices and environmental influence. Pollution, stress, wrong diet, smoking, etc. [16] are examples of life decisions or situations that may alter our genome through epigenetic changes and even have an influence on embryonic development. The increasing interest in epigenetics has revealed discoveries about the relationship between epigenetic changes and diseases like cancer, or several neurological disorders.

Epigenetic experiments started with twins samples. Monozygotic (MZ) twins have exactly the same DNA sequence and, however, they develop in a variety of different ways and may have different diseases. Therefore, they are particularly interesting because they allow us to determine the level of divergence among inherited genetics and epigenetics (or, generally speaking, genotype and phenotype). So, how do

epigenetic differences arise in MZ twins that lived in the same environment during a big part of their life? Several studies concluded that genetically identical individuals became epigenetically non-identical as they age: epigenetic differentiation among both starts in the utero, but those differences become more pronounced with age and exposure to different environments [17]. Those analyses re-enforce the idea that the epigenome reflects environmental differences.

Figure 1.3: Visual definition of epigenetics.



Epigenetic changes may be reversible, as demonstrated by one interesting article published in 2019 with the case of a NASA astronaut and his non-astronaut twin [18]. By comparing blood samples from the twins before, during and after the 1-year NASA mission, researchers found that astronaut's gene expression was exceedingly altered after spending a year in space, and that his DNA suffered damage probably due to radiation exposure. Of course, the astronaut genetic code did not change, but several epigenetic modifications lead to the gene expression alteration recognized. Even more surprising and interesting was the fact that within six months of returning to Earth, about 90% of astronaut's affected genes returned to their normal expression levels. We could say that this is an example of an accelerated epigenetic change, resulting from a totally abnormal exposure (and adaptation) to different environmental conditions.

X chromosome inactivation is an interesting epigenetic regulation process that was first introduced by Mary Lyon around 1960. Females have two copies of the X chromosome, determining their sex. However, only the information contained in one of these two X chromosomes is finally used, leading to the inactivation of the

other one. This is what we call *X inactivation*. X inactivation could be considered the epigenetic phenomenon par excellence, as the X chromosome is not inactivated by mutation of the DNA sequence. It is a natural process to not duplicate the genetic information contained on this chromosome that has about 1,300 genes. The interesting thing about this inactivation process is the complexity and importance. We are taking about deactivating more than one thousand of genes irreversibly. This occurs early in development and randomly, meaning that we cannot predict which one of the two chromosomes (maternal or paternal) will be inactivated. Moreover, this process is heritable and demonstrates the capacity of the cell to count the number of X chromosomes (for that reason this process does not happen in men). Its importance is coming from the consequences of the non-inactivation, that is basically the death of the coming live. This amazing process that seems to be taken from a science fiction book is really happening in every cells of a future woman. There are many publications about this process and ongoing research is still being developed in order to completely understand and describe the mechanism [19].

Generally speaking, there was a change in the way we understand genetics and therefore in the way we investigate animal and human genomes. The "epigenetics revolution" was a breaking point on the whole genetics field, moving forward to a wider view of the complex biological mechanisms behind. Of course, the overall epigenetic regulation remains diffuse, but the studies published in the last years were a big progress on the area. They provided with a detailed idea on how the environment modifies our and future generations' genome, and how it may be connected to several complex diseases. This was, at the same time, a big step forward on the development of precision medicine [20] with new drugs that can regulate the epigenetic alteration or revert it. The complete understanding or current and future disease's functioning is the only way to, not only cure their symptoms, but genetically eradicate them.

1.2.1 Epigenetic Modifications

There are different types of epigenetic modifications that indirectly modify our genome functioning without altering the DNA sequence itself. All those mechanisms can be studied separately despite they should of course be also viewed as a complete work chain. Currently, DNA methylation is one of the most broadly studied and well-characterized epigenetic modifications but there are other alterations to be studied including gene expression, chromatin remodeling, histone modifications, and non-coding RNA mechanisms. As we mentioned at the beginning of the present document, the specific mechanisms of the overall epigenetic regulation and its consequences on evolution and diseases remain unclear.

Similarly to GWAS, epigenome-wide association studies (EWAS) [21] quantify epigenetic marks, such as DNA methylation, in different individuals to derive associations between epigenetic variation and a particular identifiable phenotype. While the genetic risk of a disease is currently unmovable, there is hope that epigenetic risk may be reversible and/or modifiable. During the last decades, research is more and more focused on the creation and manipulation of high-dimensional arrays of epigenetic marks that speed up epigenetic analysis. However, there are still a few challenges to overcome as tissue variability or low sample sizes. Several statistical and computational techniques raised for their analysis, however they are not optimal or standard from distinct points of views. For instance, most studies are limited to the study of local epigenetic patterns, whereas methods for analyzing large-scale organizations are still lacking.

DNA Methylation

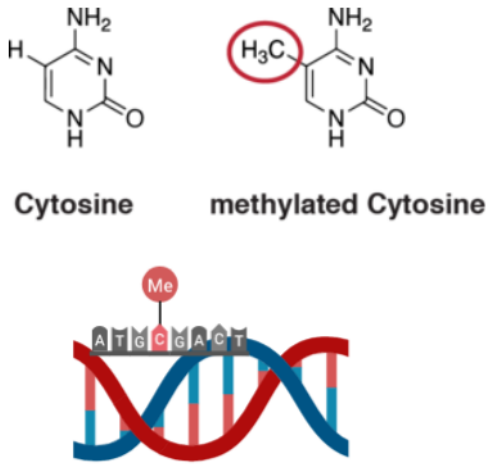
In particular, the present work studies the mechanisms and patterns behind DNA methylation as an epigenetic feature. DNA methylation is an epigenetic mark by which methyl groups attach to the DNA strand. The base C gets methylated to form 5-methylcytosine. This methylation reaction is carried out in our cells by one of the three enzymes called DNA methyltransferases: DNMT1, DNMT3A or DNMT3B. The base C, that is sensitive to be methylated, may be followed by a G in the linear sequence of bases forming a *CG* site or *CpG site* (from cytosine-phosphate-guanine).

In mammals, around 80% of CpG cytosines are methylated. This chemical group added does not modify the underlying DNA sequence, but methyl groups specify new signals to guide the reading of the genetic information, potentially impacting gene expression. DNA methylation is essential for normal development, and it plays a very important role in several key processes including genomic imprinting or X chromosome inactivation. DNA methylation at gene promoters is important for transcriptional regulation, with dense promoter hypermethylation around the transcription start site being associated with gene repression or silencing. Generally, its dysregulation, that may occur more frequently that a genetic mutation, contributes to risk of diseases like cancer.

CpG dinucleotides frequently occur in *CpG islands*. CpG islands are large DNA sequences that deviate significantly from the average genomic pattern by being enriched in "GC" (CpG-rich) and predominantly non-methylated. The formal definition specify them as a stretch of DNA of at least 200 base pairs (BP or bp) long with at least 50% of CpG content. They could be associated with the start of the gene (promoter regions) and are typically unmethylated when the nearby gene is active or expressed. Indeed, the presence of a CpG island is used to predict genes' location.

CpG islands have been widely investigated due to the mentioned regulatory importance. However, their overall functioning and interaction or the reason of their alterations are points not completely determined. For us, they will be an important source of study, as they represent interesting regions with a special correlation design.

Figure 1.4: DNA methylation process.



DNA methylation has a complex structure mainly due to the variability of its patterns with distinct tissues, sample characteristics, or genomic regions. DNA methylation works together with other epigenetic modifications as histone alterations. The heavily repressed areas of the genome present high levels of DNA methylation and are very condensed. The DNA becomes very tightly wound up, inaccessible to enzymes that transcribe genes. Repressive histone modifications attract DNA methyltransferases, which deposit DNA methylation near to those histones. This methylation in turn attracts more repressive histone modifying enzymes, creating a loop that leads to an increasingly hostile region for gene expression [13].

Repressive histone modifications can therefore attract DNA methylation to the promoter of a tumor suppressor gene. As a consequence, an alternative therapy could be based on finding the right epigenetic enzymes to re-express the tumor suppressor genes without the need of change the DNA methylation.

There are several methods available to determine the methylation status of DNA samples or differential methylated regions, and their use depends on the matter of study [26]. Those comprise bisulfite sequencing, methyl-sensitive cut counting, or array technology. The last one is the most used and the one we use. It takes methylated DNA fractions of the genome, identifying differentially-methylated re-

gions non-uniformly distributed in the genome.

DNA methylation patters should be carefully interpreted taking into account the underlying functional aspects. For example, it is known that the methylation levels may change with the tissue or cell type where they were measured [22, 23], with the existence of identified tissue-dependent or cell-type specific CpG sites. On the other side, demographic characteristics as race/ethnicity, age, gender, or lifestyle decisions (smoking, diet habits, drugs abuse, stress, etc.) were associated with different methylation levels [24, 25].

We can extract methylation samples from multiple tissues: brain (different cell types), blood (different cell types), cord blood, sperm, saliva, breast, cartilage, colon, head, neck, heart, kidney, liver, skin, placenta, lung, stomach, etc. The comparison of the results found with different tissue-based arrays could be correlated but has to be done carefully.

Due to the dimension, complexity, and variability of the mechanisms beyond the DNA methylation, and the functional factors just described, it is very difficult to find standard analytical tools of patterns' identification that obtain accurate results. For that reason, one of the aims of this thesis is to provide with alternative mathematical methods of methylation analysis.

Analysis of DNA Methylation

The mathematical methods developed in this thesis are conducted to analyze DNA methylation patterns and their correlation structure, associated to different factors as the aging process or certain diseases.

The type of data that is normally used in genomic studies is High Dimension, Low Sample Size (HDLSS) data, which means, much more variables than independent observations are taken into account. Infinium HumanMethylation27 or the newer HumanMethylation450 Bead Chip array by Illumina are the main arrays used for publications and both are HDLSS (we call them simply "high-dimensional datasets"). The difference among both is the number of CpG marks analyzed per subject: about 27,000 or 485,000 individual CpG sites in 99% of known genes, and different regions. Currently, there is a novel array available with more than 850,000 CpGs but it is still on the first steps of application.

To date, two parameters have been proposed to measure the methylation level of the CpG marks. The first one is called *Beta-value* (or β value) and has been widely used to measure the percentage of methylation. It is defined as the ratio of the methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities). This is the one recommended by Illumina. The

Beta-value follows a beta distribution and for an i^{th} interrogated CpG site is defined as:

$$\beta_i = \frac{max(y_{i,meth}, 0)}{max(y_{i,unmeth}, 0) + max(y_{i,meth}, 0) + \alpha} \tag{1.1}$$

where $y_{i,meth}$ and $y_{i,unmeth}$ are the intensities measured by the i^{th} methylated and unmethylated probes, respectively. The α parameter is an Illumina recommendation (by default, $\alpha = 100$) to regularize Beta-value when both methylated and unmethylated probe intensities are low. The Beta-value results in a number between 0 (completely unmethylated) and 1 (completely methylated).

The second one, the M-value, is calculated as the log2 ratio of the intensities of the methylated probe versus unmethylated probe as follows:

$$M_i = log_2 \left(\frac{max(y_{i,meth}, 0) + \alpha}{max(y_{i,unmeth}, 0) + \alpha} \right)$$

The Beta-value has to be used with the awareness that it violates the Gaussian distribution assumption used by many statistical methods. However, the M-value can be appropriately analyzed with these methods [27]. The Beta-value can be converted into a M-value (ignoring α), and in the other way around, as follows:

$$\beta_i = \frac{2^{M_i}}{2^{M_i} + 1}$$

$$M_i = log_2 \left(\frac{\beta_i}{1 - \beta_i} \right)$$

In our case, β levels of the HumanMethylation450 Bead Chip array are going to be analyzed. The array used can be viewed as a matrix with more than 485,000 rows and as many columns as determined by the sample size. Let be N the number of individuals in our sample $\{X_1, ..., X_N\}$. Supposing that our array contains n CpG sites or variables $\{CG_1, ..., CG_n\}$, then our starting point is the following methylation matrix of dimension $n \times N$:

$$B = \begin{pmatrix} \beta_{CG_{1,X_1}} & \cdots & \beta_{CG_{1,X_N}} \\ \vdots & \ddots & \vdots \\ \beta_{CG_{n,X_1}} & \cdots & \beta_{CG_{n,X_N}} \end{pmatrix} \tag{1.2}$$

Together with the methylation arrays just described, there is also available the metadata for each one the CpG sites analyzed and the sample selected. The metadata related to the sample is quite variable dependent on the dataset, but normally

contains the main demographic characteristics of the sample (gender, age, ethnicity, disease group, etc.) together with the place, date, tissue and type of sample collected. The metadata related to each one of the CpGs contained in the 450K array is more standard and is given by the called Infinium Methylation450K manifest. Among all the information given, the variables presented in the table 1.1 are the most used, and the ones we use:

Table 1.1: Infinium methylation 450K manifest content

Column header	Description
IlmnID	Illumina CG dataset ID
CHR	Chromosome containing the CpG
MAPINFO	Chromosomal coordinates of the CpG
UCSC_RefGene_Name	Target gene name(s)
UCSC_RefGene_Group[1]	Gene region category describing the CpG position
UCSC_CpG_Islands_Name	Chromosomal coordinates of the island
Relation_to_UCSC_CpG_Island[2]	The location of the CpG relative to the CpG island
DMR[3]	Differentially methylated regions
Regulatory_Feature_Group	Regulatory feature: gene-associated, non-gene associated, promoter, unclassified

The line [1] above could be divided in: TSS200 = 0–200 bases upstream of the transcriptional start site (TSS); TSS1500 = 200–1500 bases upstream of the TSS; 5'UTR = Within the 5' untranslated region, between the TSS and the start site; Body = Between the start site and stop codon; 3'UTR = after the stop codon.

The line [2] is categorized in: Island, N-Shelf, N-shore, OpenSea, S-Shelf, S-Shore. They mean: Shore = 0–2 kb (1 kb = 1000 bp) from the island, Shelf = 2–4 kb from the island; N = upstream (5') of the CpG island, and S = downstream (3') of the CpG island. OpenSea region corresponds then to areas not classified as islands or surroundings.

The variable [3] has the different categories determined experimentally: DMR = Differentially Methylated Region. CDMR = Cancer-specific Differentially Methylated Region. RDMR = Reprogramming-specific Differentially Methylated Region.

Current Methylation Analysis Techniques

During the last years, the use of the 450K array increased significantly being analyzed by different statistical approaches and computational algorithms. Several techniques [28, 29, 30] were raised to study deeply the potential methylation patterns and their association with case/control samples. They follow similar analysis lines to detect methylation patterns in form of differentially methylated positions (DMPs) related to a phenotype covariate, differentially methylated regions (DMRs) for a specific genome region, or differentially methylated CpG sites (DMCs or DMSs) without regional restriction. However, most of them analyze methylation data on a site-by-site basis, ignoring correlation effects and limiting the statistical power and biological interpretation of results.

Some of those methods analyze differentiated patterns between regions, but they have difficulties to detect global non-regional patterns. They normally use the term "co-methylation", that describes proximal CpG probes with similar DNA methylation values that are detected with different techniques [31, 32, 33]. Co-methylation analyses were relatively successful to detect DMRs but not to differentiate sample proves or larger inter-chromosomal interactions. Some examples of implemented regional statistical analyses are adjacent site clustering [34], Bumphunter [35], blockFinder, Probe Lasso [36, 37], or DMRcate among others [38, 39, 40, 41].

To detect DMPs or DMCs, different techniques based on a regression analysis per CpG site are commonly used to model methylation levels from different sample groups, using afterwards a t-test to reject or accept the null hypothesis of null model coefficients. The obtained p-values are adjusted for multiplicity with false discovery rates (FDRs) normally. However, those methods force to test thousands of different hypotheses with the consequent power reduction of the p-value adjustment method.

Generally, the mentioned methods assume normality, which is not really true for microarray data. Besides, linear models do not work well with very variable data, increasing the false positive rate. In addition, they are unable to detect potential data confounders as the difference in cell-type proportions that may lead to a different methylation status. Several methods providing with alternative solutions are not so efficient from a computational perspective, as Hidden Markov models [42, 43, 44] used to identify CpG islands, chromatin states [45, 46], DMRs or DMCs [47, 48].

There are several R packages [29] defined to detect regions or sites differentially methylated, being one of the most used the called *Minfi* [49]; but also *IMA*, *FastDMA*, *ICDMR*, *QDMR*, *COHCAP*, *BumpHunting*, *WFMM*, and *MethylKit*. There is not specific prof that one method performs better than the others, it will

depend on the sample and data characteristics and the aim of the study.

Improvements in the statistical methodology to provide with enough power, robustness, efficiency, and non-biased results are needed to be developed [28]. Ideally, the standardization of the methodologies used depending on the study aim could facilitate the analytical process. Moreover, the analysis of the correlation structure has to be studied deeply in order to increase the understanding of the methylation regulation networks and their impact. The main problem is that we would need to calculate an enormous correlation matrix of dimension $485,000 \times 485,000$. That is really a non-manageable computational time, in case our computer is able to do it. Moreover, results would be highly uninterpretable. Therefore, usual techniques of exploring the data correlation cannot be used and we need to find other options.

Those challenges motivated us to present analytical solutions described in this work.

1.2.2 Epigenetics of Aging

Aging is a biological process known in all species and was always a matter to study. The understanding of the mechanisms behind the aging process and derived diseases is crucial for several aspects. Specifically, the epigenetic functioning of this process is interesting to understand the changes produced in our genome as years go by, in order to revert them, and dig into the longevity processes too.

As mentioned in previous sections, epigenetics may serve to explain why the pattern of aging is different between two genetically identical individuals, such as identical twins, or, in the animal kingdom, between animals with identical genetic makeup, such as queen bees and worker bees [54]. Several studies published during the last decade, in humans and other species, indicate that epigenetic changes have a huge influence on the aging process altering the local access to the genetic material and leading to genomic instability. There is, therefore, a potential difference among the biological or epigenetic age and the chronological age of an individual [50, 51]. In fact, it is currently possible to calculate the epigenetic age of a subject based on specific DNA methylation marks thanks mainly to the work done by Horvath [52]. This work and related ones are extremely important in forensics medicine [53].

Some of the methylation changes that occur with age are directional and involve specific regions of the genome, with a generalized CpG hypomethylation trend more present in CpG islands. However, hypermethylation also occurs at specific CpG sites of the genome, presumably to repress the expression of specific genes. Global hypomethylation at the repetitive regions and site-specific hypermethylation at certain promoters have also been reported during cancer, suggesting a potential

connection between age-dependent DNA methylation changes and increased cancer risk observed in the elderly population. Hence, a better understanding of the reasons behind the changes in the DNA methylome during aging may help us to also understand the causes of cancer. This altered pattern was theoretically related to the onset of other common human diseases where age is a known risk factor (type 2 diabetes, cardiovascular diseases, renal diseases, mental disorders), but was not always demonstrated. The complete understanding of the epigenetic changes associated with longevity is still under research, with the main aim of delaying aging and cure age-related diseases.

One of the main biological objectives of this work is to clarify some of the aging mechanisms studying the evolution of the correlation structure of DNA methylation over the years.

1.2.3 Epigenetic Markers in Cancer

The influence of epigenetic modifications on several diseases has been studied intensively during the last years. Even, sometimes, several studies pointed out that an epigenetic alteration could be part of the disease pathogenesis. Among all of them, there is one important disease that can be considered the most studied one: cancer. Following official figures from the World Health Organization (WHO), cancer is the second leading cause of death globally (after cardiovascular diseases), and is responsible for an estimated 9.6 million deaths in 2018. There are several risk factors associated with cancer, from age or diet characteristics to tobacco or alcohol abuse. The early diagnosis is crucial for its cure, but normally the late-diagnosis and invasive or inaccessible treatment are increasing the death rates. Cancer incidence is similar between men and women, but there is a difference among the types of cancer incidence by gender: being lung, prostate and colorectal cancer most common in men (in that order); and breast, colorectal and lung more common in women (in that order).

Cancer is a disease of the genome, produced by the creation and spread of abnormal cells. If a proto-oncogene (genes that promote cell proliferation) becomes over-active, it may push a cell towards a cancerous state. Conversely, if a tumor suppressor gene (genes that prevent proliferation) gets inactivated, it will no longer act as a brake on cell division. The outcome in both cases is a too rapid proliferation. In addition, if cells divide too quickly, they form structures called benign tumors. Those are in principle benign if they do not press an organ or increase significantly, but their presence increases anyway the likelihood of developing a future cancer. This disease is, therefore, a multi-step process that can take years up to be manifested, and that cannot be easily generalized and studied. Cells accumulate defects as they move

increasingly close to becoming cancerous, and those defects can be also inherited. Here it lies the complexity of finding a cure or efficient treatment.

There are many different types of alterations, also dependent on the tissue, and the potential cure seems to be more "personalized" than "generalized". The application of precision medicine with personalize treatments based on genomic profiles will probably increase the overall survival of the targeted population. The studies based on understanding the genetic and epigenetic mechanisms behind are then crucial to find this personalized approach. Great progress were done on the understanding of the hidden cancer mechanisms and detecting related biomarkers, but, unfortunately, a lot of effort is still to be done in order to improve prevention, diagnosis and treatment.

The majority of tumor-suppressor gene silencing in cancer occurs through epigenetic mechanisms, specifically via focal hypermethylation in the promoter of the related CpGs [55]. There are several examples known of epigenetic modifications causing cancer [56, 57]. For instance, more than 10% of breast cancer patients without family history was found to have a hypermethylation profile in the CpG island related to the BRCA1 gene. Similarly, colon or prostate [58] cancer patients were found to have high levels of promoter DNA methylation in many different genes simultaneously. Besides cancer, other disorders as autism have been linked to epigenetic alterations [59, 60].

Drugs that inhibit the DNMT1 enzyme have been licensed by the FDA (US Food and Drug Administration) for clinical use in cancer patients. However, a long way is still to be walked with more safe and efficient studies that understand the disease and what are the treatment needs of each patient based on the individual's genetic profile. The discovery of new drugs based on the role of epigenetics in cancer is without a doubt a potential and reasonable way to follow.

We will study cancer methylation alterations trough the detection of differentially methylated CpG sites, genes, and regions in cancer samples.

1.3 The Genome Hierarchy

Compacting around two meters of DNA sequence in every nucleus cell requires an enormous compaction in different levels: DNA is packed inside the nucleus of the cell forming histones, that form nucleosomes, that form the chromatin (and so the chromosomes). Those levels define a hierarchy in the genome that goes away from a linear process and acquires a three-dimensional (3D) structure [61, 62].

There is a lot of ongoing research to investigate why and how this spatial structure is formed. Some studies suggest that the chromatin architecture might be just a stochastic process as a result of the intensive wrapping of the DNA sequence within the cell nucleus. However, most of the current investigations state that this structure is non-random but answers to internal cell forces related to biological functions, contributing to many cellular processes and gene expression regulation [63]. Moreover, the chromatin is not static, but there are traces of dynamic chromatin movements that also determine changes in the transcription process. This non-random hypothesis would be more in line with the fact that the chromosome territories are not random at all, but gene-rich chromosomes are located more in the center of the cell nucleus.

Thus, a main question to answer is how the 3D genome architecture has an impact on our biology and on the interactions among epigenetic marks or, summarizing, what is the relationship between structure and functionality. In mathematical words, we could ask: what is the topology of the chromatin and how is it varying with epigenetic changes?

One of the peculiarities of this architecture is that it joins genomic regions that initially would be very far apart in the linear sequence. That makes possible to establish a communication between long-range enhancer-promoter regions. Despite this is still a matter of debate, recent studies propose that this long-range communication could be also be produced with inter-chromosomal regions [64, 65]. The fact that two genes in different chromosomes develop similar functionalities or are located in a connected gene regulatory network, may be associated to a correlated design of the enhancer or promoter related regions. If so, epigenetic mechanisms would also reveal inter-chromosomal correlations with an influence on the genes co-regulation.

This chromatin design is especially interesting to investigate in the case of disease-related studies, where the creation/deletion of long-range interactions may be the basis of the altered functions related to a specific health condition. In fact, aberrant DNA methylation patterns have been associated with an alteration of the

DNA architecture related to cancer [66].

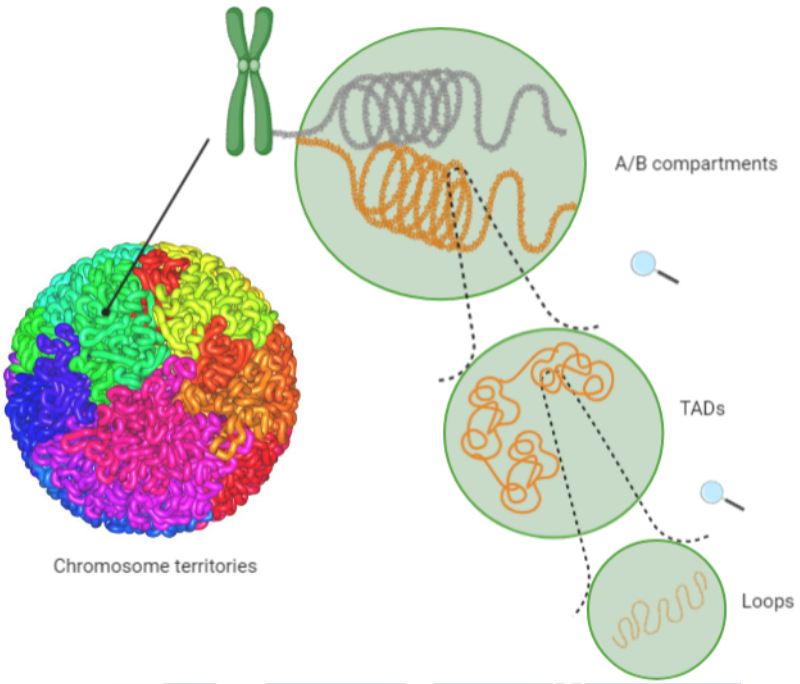
Chromatin topology gives rise to several biological features with different topological characteristics [67]:

- A/B compartments (typically hundreds to thousands of kb large): are differentiated in *euchromatic* (open chromatin, where most of the transcription occurs) and *heterochromatic* (condensed chromatin, where transcription is normally inhibited) regions, respectively. Chromosome territories are formed then on top of these domains and compartments. The proximity of different types of chromatin can influence gene expression [68]. In [69] they show how these compartments can be estimated using long-range correlations of methylation data. More recent articles go in the same direction [70, 71].
- Topologically associated domains or TADs (typically tens of kb large): are genomic regions containing DNA sequences with a higher level of intra-interaction. The functions of TADs are not fully understood, but most of the studies indicate that TADs are related to gene expression, and the deletion or inactivation of a TAD boundary can lead to inappropriate oncogene expression [72, 73, 74].
- Loops: chromatin loops are specific regions within the TADs that have a loop form interconnecting chromatin regions through their folding. Loops allow the interaction of distant segments of DNA, which could modify the expression of genes that are, in a first place, very faraway [75]. Loops may grow and more distant regions of the DNA can be brought into contact, generating a large-scale organization extended over tens of megabases or between different chromosomes [76].

The technologies called "Chromosome conformation capture" present a collection of methods to study the chromatin structure that identify close genomic locations (i.e., in physical proximity in space). High-throughput sequencing (Hi-C) protocols use techniques to provide genome-wide maps of DNA interaction. The information is summarized in a contact data matrix that measures the genomic locations that are in close three-dimensional proximity [77]. For instance, loops can be detected in the contact matrix as non-diagonal interactions.

As mentioned before, based on the spatial design of the chromatin both intra-chromosomal and inter-chromosomal long-range associations have been found and described, having a different functioning and frequency [78, 79, 80, 81]. Despite intra-chromosome interactions are more studied, also for CpG sites [82, 83], other studies start now to establish how long chromatin loops are enriched with highly significant

Figure 1.5: Chromatin topological elements.



correlated CpGs. They point out how that large DNA methylation nadirs (called *grand canyons*) can form long loops connecting anchor loci that may be dozens of megabases apart, as well as inter-chromosomal [84]. Determining how chromosomes are positioned and folded within the nucleus is critical to understand the role of chromatin topology in gene regulation. In conclusion, the new advances in the understanding of the three-dimensional structure of the chromatin and its impact on cellular processes may unravel novel findings and even implicate the revision of prior discoveries.

Indeed, as we will see in the chapter 2 and then in chapter 5 of the present work, inter-chromosomal correlations associated with DNA methylation were detected. For instance, CpG sites related to the genes OXR and OXTR were found to have a high level of intra-chromosomal interaction but also a long-range inter-chromosomal correlation (please see subsection 5.2.2).

Another important point is that our entire genome is evolving with age, or it is determining the aging process. Recently, several progresses were done on understanding the changes of the chromatin structure over the years [85], but there is still a lot to be explored. Generally, aging is associated with a loss of heterochromatin, leading to a chromatin that is less-condensed, changing the hierarchical structure, inducing genome instability, and disrupting the long-range

contacts or loops. This loss of condensation may unhide silenced genes, through lower DNA methylation, that become altered. Similarly, it may alter unexpected genomic regions increasing the levels of DNA methylation and silencing genes.

The increasing interest and knowledge about this 3D organization give us a rationale to think on a big structured network of interactions among DNA methylation sites, including short-range and long-range correlations, that could have an influence on gene expression and different diseases. In addition, last studies about the alteration of the chromatin model with age, give us the needed biological basis to test the mathematical hypothesis of a structural change of the methylation correlation with age to a more local (short-ranged) or less correlated design.

As far as we know, the deep analysis of the short-range/long-range correlations among DNA methylation marks, developing new analytical strategies and tools, was not totally covered in literature.

Chapter 2

The Mathematical Hypothesis

What is the mathematical challenge to solve? After the necessary biological concepts were introduced, we present in this chapter the main mathematical hypothesis generated from the described biological observation. In addition, we will present the initial analytical challenges of our work and the improved proposal to extract the maximum information about DNA methylation patterns and correlation.

2.1 The Correlation Hypothesis

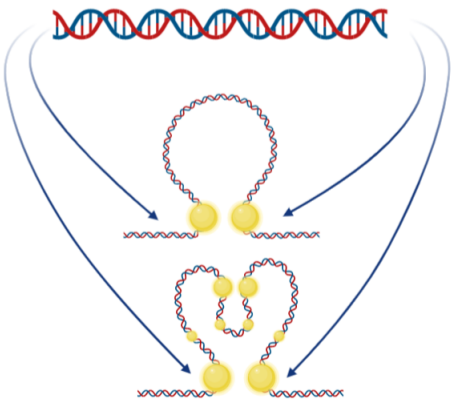
The joint analysis of the local (short-range) and global (long-range) correlation between DNA methylation marks was not extensively studied, limiting it almost exclusively to the local behavior of co-methylated trends. Generally, the analysis of correlation patterns of high-dimensional datasets requires sophisticated analytical tools that were not very developed. It is indeed what we would like to cover with the present work.

Having into account the genome structure just described in section 1.3 and the biological hypothesis of a chromatin spatial design that determines genomic functionalities, we develop a mathematical hypothesis to be tested. We hypothesizes that this structure also determines epigenetic interactions and has an influence in the correlation structure of DNA methylation.

The main hypothesis is then that the correlation between CpG sites is not random, but it reaches high significant levels in a structured and standard way for short-range and also for long-range genomic regions potentially derived from the loop design (please see figure 2.1 below as an example).

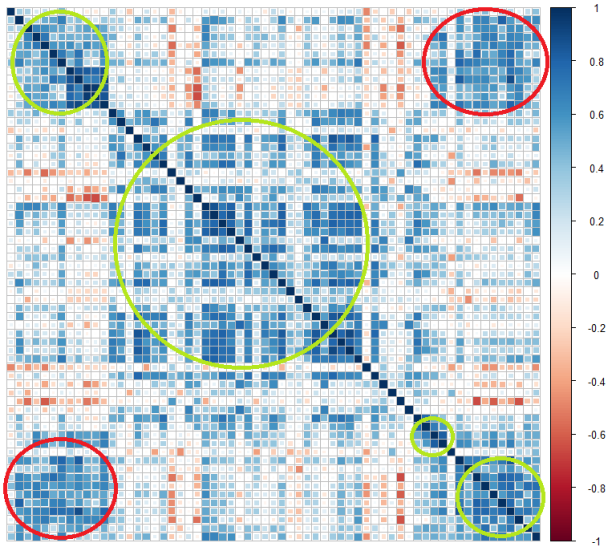
In analytical terms, the loop may be seen as a differentiated correlation cluster above or below the diagonal of the methylation correlation matrix, where CpGs are

Figure 2.1: Illustration of the loop design approaching distant genomic regions.



ordered by their genomic positions within the chromosome (as presented in figure 2.2). For instance, imagine that CpG sites CG_1 and CG_2 are located in the same chromosome with a genomic distance $d > \epsilon$ ($\epsilon > 0$) among them. The loop would reduce the spatial distance of those two CpG sites to $d < \epsilon$ increasing the likelihood of common biological functioning, methylation patterns and a higher correlation. Therefore, the Pearson correlation coefficient $\rho(CG_1, CG_2)$ may be higher than expected [82] having into count their genomic distance d.

Figure 2.2: Example of a methylation correlation matrix including short-range (green) and long-range (red) correlation clusters.



In the correlation matrix of a CpG island showed in figure 2.2 we find two types of correlated clusters, some of them in near genomic positions (over the

diagonal, marked with green circles) and others corresponding to distant genomic positions (over the lateral, marked with red circles). What we would like to test is that those clusters are not formed randomly but they form a structure that may be representing specific biological interactions. Of course, as we are considering a CpG island, the CpG sites are all nearer compared with other genomic regions.

It has been studied that part of the DNA methylation variability may be caused by the tissue used and the cell type decomposition within that tissue [86, 87]. Despite [88] points out that the majority of the high correlation is due to different technical and biological confounders, as cell-type sub-populations, we came to a different conclusion. We assume that there may be several CpG sites affected by the tissue and the cell-type leading to an increase of the methylation variability, but this does not necessarily mean a disappearance of a correlation structure and, moreover, it does not mean that the whole high correlation among epigenetic markers is exclusively due to this confounder. In fact, we have found significant high levels of correlation among randomly selected CpG sites (not necessarily nearly located or in the same chromosome) in different tissues and cell types, as we will describe in the next subsections.

Our first hypothesis is therefore the existence of significant correlation over any DNA methylation dataset, in an structured way. We want to test:

$$H_0 : \rho = 0$$

where ρ represents the population Pearson correlation coefficient of two X, Y random variables (or CpGs):

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

being cov the covariance, σ_X the standard deviation of X, and σ_Y the standard deviation of Y. We could then define the sample Pearson correlation coefficient as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where n is the sample size, x_i, y_i are the individual sample points indexed with i, and $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ is the sample mean (similarly for $\bar{y}$).

We will use the Fisher transformation of r to estimate the confidence intervals of the correlation coefficient. If we calculate the Fisher's transformation of the sample correlation r through the formula:

$$F(r) \equiv \frac{1}{2} ln \left(\frac{1+r}{1-r} \right) = \text{arctanh}(r),$$

then, under certain conditions, $F(r)$ follows approximately a normal distribution with mean $F(\rho) = \text{arctanh}(\rho)$ and standard error $SE = \frac{1}{\sqrt{n-3}}$ [89, 90]. With the approximation, z-score is, under the null hypothesis:

$$z = \frac{x - mean}{SE} = [F(r) - F(0)]\sqrt{n-3}$$

The confidence interval of ρ is then based on the confidence interval of $F(\rho)$. The normal cumulative distribution function Φ given the significance level $\alpha = 0.05$ is defined as:

$$\Phi(z) = P(Z \leq z) = 1 - \frac{\alpha}{2} = 0.975,$$
$$z = \Phi^{-1}(\Phi(z)) = 1.96$$

and we get:

$$0.95 = 1 - \alpha = P(-z \leq Z \leq z) =$$
$$= P(-1.96 \leq \frac{\text{arctanh}(r) - mean}{SE} \leq 1.96) =$$
$$= P(\text{arctanh}(r) - 1.96SE \leq mean \leq \text{arctanh}(r) + 1.96SE)$$

So the 95% confidence interval (CI) of $F(\rho)$ is then generally:

$$100(1-\alpha)\%CI : \text{arctanh}(\rho) \in [\text{arctanh}(r) \pm z_{\alpha/2}SE] \tag{2.1}$$

The inverse Fisher transformation brings the interval back to the correlation scale:

$$100(1-\alpha)\%CI : \rho \in [\tanh(\text{arctanh}(r) - z_{\alpha/2}SE), \tanh(\text{arctanh}(r) + z_{\alpha/2}SE)] \tag{2.2}$$

The defined CI will allow us to accept or reject the null hypothesis based on the results we are about to explain.

2.1.1 First Correlation Analysis

In order to have a first idea of the quantity of correlation that we may find in a methylation dataset, we calculated the correlation matrix of a 450K array by brute force. We use Beta levels as methylation descriptors, following the equation 1.1, using the public dataset coded as GSE40279 from human whole blood of 656 healthy samples aged between 19 and 101 years old.

The computational time is of course unmanageable (it took days to run in a normal computer without parallelization) but it is valid as a first method of hypothesis validation. We have found high positive and negative correlation among CpG sites belonging to the same and different chromosomes, and the most founded reason is that those global correlations can be caused by the three-dimensional

DNA structure.

We have determined in a matrix G the number of CpG sites with an absolute correlation greater than 0.5, 0.6, 0.7, and 0.8 with at least other CpG in the same chromosome or in a different chromosome. We called this matrix the "correlation grade" matrix, as the CpG grading is determined by the number of interactions (correlations) with other CpGs. For instance, the correlation grade matrix at 0.8 is a matrix of dimension 22×22, where each row and column correspond to a different chromosome. We would have similar matrices for the other weights 0.5, 0.6, and 0.7:

$$G_{0.8} = \begin{pmatrix} g_{1,1} & \cdots & g_{1,22} \\ \vdots & \ddots & \vdots \\ g_{22,1} & \cdots & g_{22,22} \end{pmatrix}$$

where $g_{v,w} = \{\#CG \,|\, |\rho(CG_i, CG_j)| > 0.8, \quad \forall\, CG_i \in v,\, CG_j \in w\}$, being v and w two of the 22 chromosomes (sex-related ones were not taken into account).

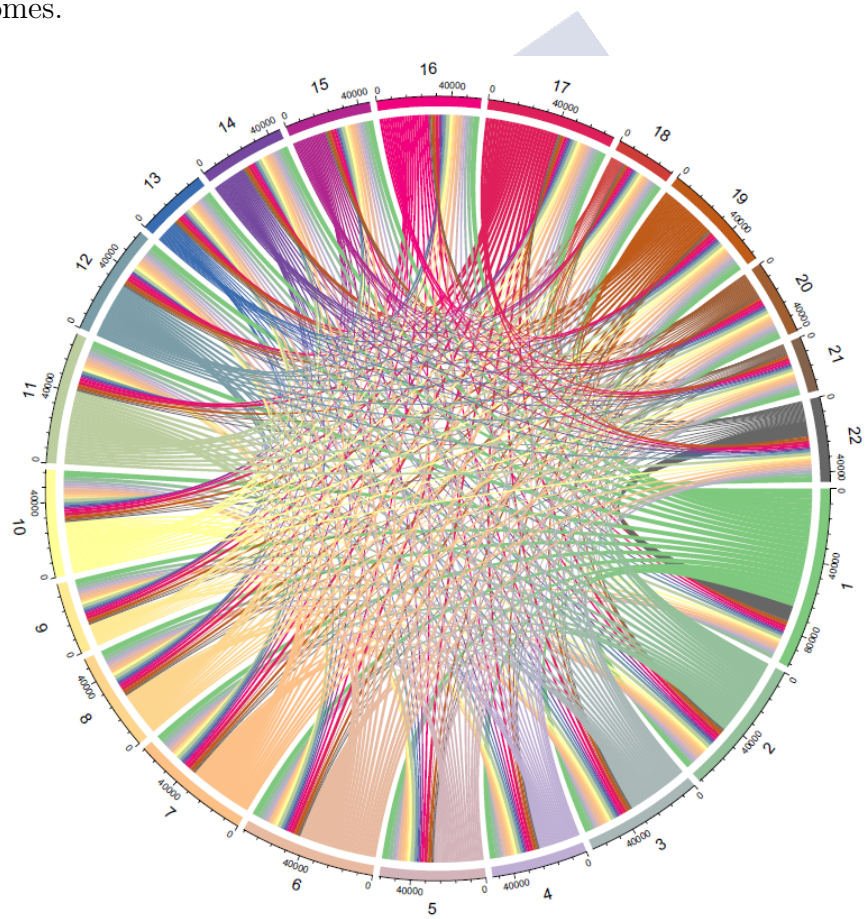
Among the 473,034 CpGs analyzed, around 180,000 had not any correlation greater than 0.5. Among the others, most of them had correlations mostly with other CpGs in the same chromosome, but also with other CpGs belonging to a different chromosome. Almost 293,000 had at least one correlation greater than 0.5 with another site (around 40-50% of the sites per chromosome), and almost 64,000 CpGs had at least one correlation greater than 0.8 with another site from the same or different chromosomes. From those, most of them belong to CpG islands (37%), followed by "OpenSea" regions (32%).

The CpG sites with at least one correlation greater than 0.8 are represented in the circle plot of figure 2.3, where each chromosome is represented by a color and where each line among two chromosomes represents a correlation greater than 0.8 between two CpG sites belonging to those chromosomes. This plot is programmed using the R package *circlize*. As we can observe, there are many inter-chromosome interactions, especially among chromosomes 1 and 22. Several studies pointed out that the majority of inter-chromosome interactions take place within the short chromosomes or from the short chromosomes to the larger ones, so this may be one of the reasons of this finding [65]. The correlation among the selected CpG sites and the sample age is generally low, with only a few values obtaining an absolute correlation coefficient near to 0.6.

Among the CpGs with correlations higher than 0.8, a group of around 2,000 present higher grades, i.e., they have high correlation levels with many (≥ 200) other CpG sites in the same and different chromosomes. Most of them (around 50%) belong

to chromosome 1 and present a high level of correlation (mostly positive). They are mainly present in body regions (45%) (that are also the majority in the array) not belonging to CpG islands, and only 200 of those sites are SNPs. Those methylation sites are not significantly correlated with age, having absolute correlation coefficients lower than 0.3. The genes related to those sites are CSF3R, FCER1G, GPR21, S100AB, S100A9, MIR626, or C5orf4 among others. They are related to metabolic functions marked by the KEGG biological pathways "Sphingolipid signaling pathway", "Platelet activation", or "Thyroid hormone signaling pathway". We might be then observing a set of methylation interactions motivated by common biological functions.

Figure 2.3: Circle plot of the correlations greater than 0.8 within and between chromosomes.



2.1.2 Correlation on Different Tissues and Cell Types

In order to validate the prior results in other datasets, tissues and cell types, we continued with a sampling process. We studied the high correlation (greater than a threshold δ) among a random sample of 100,000 CpG sites calculating the upper

diagonal of the correlation matrix among 2,000 CpGs randomly selected over 50 simulations. Please see the algorithm 1 below with the detailed process.

The results we are about to present improve (we obtain higher correlations) if those 2,000 sites are chosen as consecutive by genomic position. The number of simulations could be increased but results does not change significantly. The threshold δ varies between 0.5 and 0.8, as presented in the table 2.1 below. The resulting value represents then the percentage of correlation coefficients greater or equal than δ among the upper diagonal of the sampled correlation matrix.

We used 450K datasets of DNA methylation on different tissues: 5 datasets of blood tissue (GSE40279 whole blood, GSE41169 whole blood, GSE36064 peripheral blood leukocytes, GSE30870 cord blood and whole blood CD4+ cells, GSE34639 CD4+ cells), 4 datasets of brain tissue (GSE66351 glia and neuron cells frontal cortex, GSE50853 neuronal and non-neuronal cells orbitofrontal cortex, GSE41826 frontal cortex various cells, GSE15745 different brain parts), 2 of colon tissue (GSE32146, GSE48684), 1 of prostate tissue (GSE76938), 1 of buccal (GSE42700), and 1 of placenta tissue (GSE44667). For the cases with multiple cell types, we did the analysis for each cell-type sample and took the average of the resulting percentages. We always selected only the healthy or control samples from the case/control studies.

We restricted the sample sizes N of each case to samples containing about 10-30 healthy individuals in order to be able to compare results (as a higher sample size may lead to a decrease on the overall correlation). However, even for the cases with more than 100 samples, or even more than 600 samples, we have found a higher than expected quantity of correlation.

Algorithm 1: Sampling algorithm

Parameters: B matrix $485,000 \times N$ (as described in equation (1.2)), and threshold δ;
initialization;
while *simulation* $s = 1$ **do**
  Select a random sample s_1 of $n = 2,000$ CpG sites from the rows of B;
  if $\forall\, CG_i, CG_j \in s_1,\ i, j = 1, ..., n,\ j > i,$ *if* $\delta \leq |\rho(CG_i, CG_j)| < 1$ **then**
    $p_{\delta_{1_{ij}}} = 1$;
  else
    $p_{\delta_{1_{ij}}} = 0$;
  end
  $p_{\delta_1} = \sum_{ij} p_{\delta_{1_{ij}}}$;
end
Repeat simulation $s = 2, ..., 50$;
Result: $P_\delta = \frac{(\sum_s (p_{\delta_s})/50) \times 100}{n(n-1)/2}$

Table 2.1: Average P_δ results by tissue type.

Tissue	$\delta = 0.5$	$\delta = 0.6$	$\delta = 0.7$	$\delta = 0.8$
Blood	12%	6%	2%	0.5%
Brain	20%	12%	6%	3%
Colon	19%	11%	5%	2%
Prostate	8%	3%	1%	0.4%
Buccal	8%	3%	0.7%	0.1%
Placenta	7%	2%	0.7%	0.1%

For a population of 656 individuals, the Pearson correlation coefficient $r = 0.1$ would be already statistically significant rejecting the null hypothesis of a null correlation. If $r = 0.1$ then, as per equation (2.2), $\operatorname{arctanh}(r) = 0.1003$ (rounded), so the 95% confidence interval on the correlation scale is $\tanh(\operatorname{arctanh}(r) \pm 1.96/\sqrt{653})$, or approximately (0.024, 0.175). As the confidence interval does not contain the zero, we reject the null hypothesis. For a sample size of $n = 20$ and $r = 0.5$, the null hypothesis is also rejected.

We have discarded that this correlation is caused by the age of the samples as the mean correlation coefficient with age is normally lower than 0.1. The removal of sites related to the sex chromosomes and the SNPs did not alter the results.

Having now into account the percentages presented in the table 2.1, we could calculate the number of CpGs that have at least one correlation coefficient greater or equal than 0.5, i.e., how $P_{0.5}$ is distributed in the correlation matrix. We could use, for instance, the dataset containing 656 blood samples for comparing with the results of subsection 2.1.1.

We obtain that a 36% of the CpG sites (or correlation matrix rows) have at least one correlation greater or equal than 0.5 with other CpG. This percentage means that the correlation is not uniformly distributed for all CpG sites, but only some of them accumulate the high correlation coefficients. This percentage is a little bit lower but similar to the percentages of the grade matrix obtained in subsection 2.1.1 for the first threshold of 0.5, where about 40-50% (depending on the chromosome) of the sites per chromosome had a correlation greater than 0.5 with at least other CpG site. If we just use CpG islands as input to the algorithm 1, this calculated percentage increases up to a 44%, as the percentages of the table 2.1 increase too.

We conclude then that the fact of finding significant correlation for datasets of

different tissues and cell types demonstrates that this correlation is not random and is not completely determined by the tissue or the cell-type proportions. Moreover, a positional effect was observed when selecting continuous CpG sites and with the presence of sites located in CpG islands. These statements constitute a novel approach for the study of the methylation correlations globally, suggesting that there is indeed a network of correlations that were overlooked until now.

Overall, this study rejects our initial null hypothesis and pushes us to deeply study the structural correlation over intra and inter-chromosomal CpGs, which is the main biological objective of the present work. To do it, we will design specific analytical tools that aim to improve current methods and are in line with novel data analysis strategies, as we explain in the next sections.

2.2 The Correlation Analysis Challenge

Once we know that there is a potential correlation structure that deserves to be analyzed, the following question is: How do we do it? One of the main challenges when taking about high-dimensional correlation is the computational requirement to calculate the correlation matrix. How do you calculate a correlation matrix of dimension around $400,000 \times 400,000$? How do you visualize it and analyze it?

Overall, with the increasing availability of huge complex datasets, there is an increasing need of developing novel methodologies to analyze high-dimensional data. In this sense, dimensionality reduction techniques as principal component analysis (PCA) or multidimensional scaling (MDS) are a crucial step where the data exploratory visualization is not easy. However, those techniques are not always computational efficient for big data and they anyway need a posterior data modulation. Besides, sometimes they may be unstable to data perturbations and are unable to detect structured topological elements that are key to describe successfully the underlying data characteristics. Common statistical models may also be unsuccessful, as they require a prior data distribution knowledge limiting the data analysis flexibility.

During the last decade, correlation networks were also popularized in biology [91, 92]. Despite they present advantages to deal with high-throughput datasets and detect hidden structures, they may also be inefficient with an incorrect parameter selection, local noise or many genes. In addition, the development of network models and analytical tools to measure network's properties or differences among sample networks is still ongoing.

So although recent improvements in high-throughput technology enable the collection of large omics datasets for many biological fields, analysis methods to handle these data are still on development [93, 94]. Thereupon, the analysis of high-dimensional data finding non-biased correlation patterns is a mathematical challenge that we would like to overcome with the present work. We propose different analytical tools to reduce the dimension of the datasets creating related networks, to study their topology and build network models. First with a local design (each node represents a CpG), and afterwards with a wider global design (each node represent a cluster of CpGs), we aim to analyze all the information strata with distinct mathematical techniques. The inspiration of our proposal comes from the powerful technique described nextly.

2.2.1 New Methods: Topological Data Analysis

All the evolving and fascinating field of data analysis and computer science comes together with an interesting application of topology for data analysis, what can be called *computational topology*. During the last ten years, a general and powerful approach to analyze data has emerged from the foundations of algebraic topology: *Topological Data Analysis* (TDA) [95, 96, 97]. TDA is born from the need of developing novel analytical tools for big data, as described in the prior section, with greater flexibility to detect the "shape" of the data and its topological invariants.

It is increasingly important to obtain the hidden patterns in an enormous and complex data cloud without prior information about its distribution. TDA tends to provide with an efficient machine learning method to analyze and visualize the complex topological and geometric structure of the data, via the application of a topological model that reduces data dimension through the creation of simplicial complexes (we will define this extensively in the next chapter 3). Those simplicial complexes, that we will also call graphs or networks, are manageable mathematical elements easier to analyze and shape form represents the data distribution.

TDA can be used with other common statistical methodologies of data analysis because the main idea is not to compete with them but to extract more information about the data (being more flexible) and therefore to provide with a more complete outcome.

Coming back to our study, the analysis of DNA methylation and its correlation in a standard way and a reasonable computational time is crucial to understand this epigenetic regulation, epigenetic inheritance, and its power as a disease diagnosis mark. Novel techniques should be able to compare multiple groups or covariates, robustly identifying different methylation and correlation patterns or potential data confounders. Ideally, algorithms designed should also provide with parameter flexibility for its extended use (even on different research fields), and a visual tool for recognizing patterns. Moreover, we think that novel models have to be practical, so clinical staff from different non-technical areas could easily understand it and use it.

Our work aims to cover those requirements being inspired in the idea of topological data analysis: describe the "shape" of the data. We aimed to develop improved mathematical tools able to answer current analytical needs and that could be helpful to explain complex biological structures as the correlation of DNA methylation or its alteration with different conditions (as aging or several diseases). In the next chapter, we will describe deeper the TDA methodology and the underlying mathematical theory, which opens the door to the presentation of our work.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 3

Topological Data Analysis as Main Methodology

How could we do high-dimensional data analysis from a topological perspective? How could this be used to contribute to the epigenetics study? Once the main biological context and mathematical hypothesis were presented, we are prepared now to introduce the methodology that we use as inspiration to solve the mathematical challenges described previously. In this chapter, we will explain the main aspects of topological data analysis and the algebraic topology theory behind. Additionally, we will introduce the two principal techniques of topological data analysis and our novel application proposal of to DNA methylation data.

3.1 The Power of Topological Data Analysis

As we described above, topological data analysis is a novel technique developed to study the "shape" of the data that reduces the dataset dimension through different methods of analysis and visualization (as the creation of simplicial complexes from data). That would be equivalent to find descriptors (or topological invariants) of the high-dimensional dataset that are robust and resistant to noise, and that represent successfully the underlying data patterns.

Another of the main advantages of the TDA design is that it works as a machine learning algorithm making minimal assumptions about the data itself. The slogan is indeed "to let the data speak". Contrary to traditional statistical methodologies, the user just would need to define some general parameters that do not restrict the interpretation of the data cloud. Those parameters are easily changeable and can be automatized in sophisticated versions of the TDA algorithms. This is really helpful when dealing with huge datasets where you need to search for patterns and

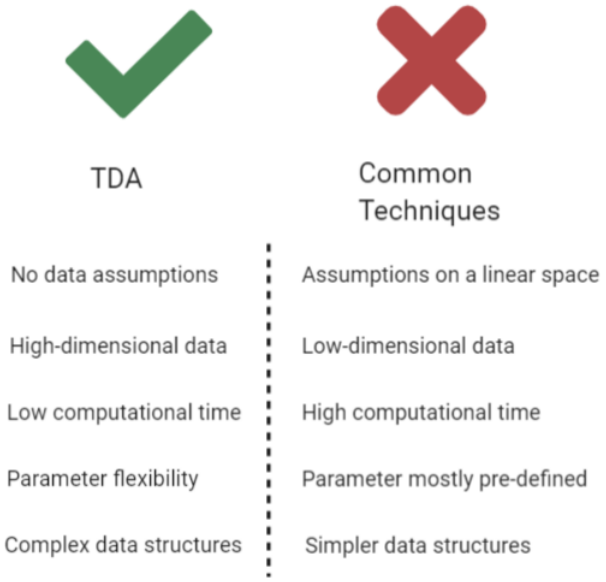
the exploratory analysis is difficult to visualize. Mapper and persistent homology are the two most popular methods in the field of TDA. First algorithms were already implemented in C++, Phyton [98] and R [99, 100].

TDA is in the middle of its golden age, as there was a significant development during the last recent years. As was pointed out by [101] in 2019:

> "[...]the translation of data into the language of algebraic topology opens many doors for analysis and subsequent insight. Those scientists with an understanding of this language can add a myriad of powerful tools to their analysis repository. Additionally, continued discussion between the mathematicians developing the tools and the scientists applying the tools will continue to spur methodical advances with biological questions as the driving force.[...]".

Several authors published that TDA is more sensitive than PCA, MDS, and cluster analysis in detecting both large and small-scale patterns, resulting in more sophisticated data interpretations [102].

Figure 3.1: TDA vs. common techniques of dimensionality reduction.



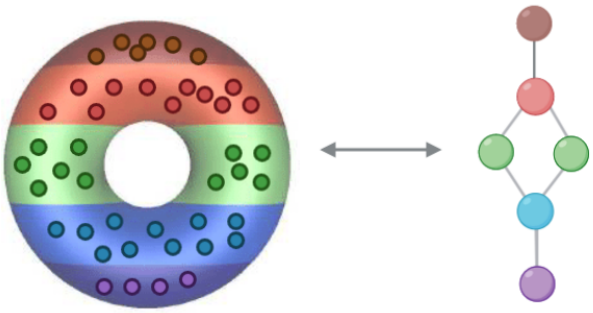
PCA coordinates were widely used as filter functions in the TDA technique called *Mapper*, which we will explain in the section 3.4. This approach is already a kind of machine learning algorithm, as we are "learning" from the data structure with simpler methods to afterwards sophisticate the design with TDA methodology to capture

better the data topology. Computational algorithms based on TDA are, moreover, faster than older techniques almost obsolete when dealing with high-dimensional data. We could say that TDA is a more exhaustive and standard method of analysis.

The increasing interest in TDA techniques is being demonstrated by the number of articles published in the last years about this topic and by the specialization of several private companies on this field, that put value on the future applications and financial advantages of the novel AI techniques. One of the most famous is the American company Ayasdi [103], founded in 2008 and focused on the development of machine learning algorithms. It offers private software to analyze big data and high-dimensional datasets with modern techniques as topological data analysis [104]. One of their founders is the mathematician Gunnar Carlsson, that did many instructive videos on TDA's idea, as [105]. Additionally, reference studies [106] were published in the last years applying TDA methodology to different fields from gene expression of breast cancer [107] to the analysis of brain imaging [108].

TDA is a novel and evolving methodology that may fulfill a lot of the current analytic needs of genomics. For this reason, we use it as a basis for our work, developing novel analytical tools over epigenetic data that could be useful in other research fields.

Figure 3.2: Example of a data cloud transcribed to a network that represents its shape.



In the next section, we will present a comprehensive description of the main algebraic topology theory that allows to define TDA algorithms.

3.2 Underlying Algebraic Topology Theory

The use of algebraic topology to do data analysis with TDA, requires a first understanding of several algebraic topology concepts as *simplicial complexes*, *homology groups* or the description of the main aspects of *Morse theory*. We will define them within this section.

Please refer to [109] for a more detailed description of the theory presented here.

3.2.1 Simplicial Complexes

Simplicial complexes are generalizations of graphs widely applied in biology. They allow to establish a relationship between structure and functionality in many real-world applications as neuronal networks, protein-interaction networks, or gene regulatory networks. The technical definition is based on a first description of the simplices:

Definition 3.2.1 *The k-simplex spanned by the distinct $k+1$ points $\{x_0, ..., x_k\} \in \mathbb{R}^n$ is the set of all points*

$$z = \sum_{i=0}^{k} a_i x_i, \quad \sum_{i=0}^{k} a_i = 1,$$

with $a_i > 0 \ \forall i$. For a given z, we refer to a_i as the ith barycentric coordinate.

For example, a 0-simplex is a point and a 2-simplex is a triangle with vertices the points $\{x_0, x_1, x_2\}$. The simplices are the basic pieces of a simplicial complex.

Definition 3.2.2 *For a simplex S spanned by the points $P = \{x_0, ..., x_k\}$, a face of S refers to any simplex spanned by a subset of P.*

Then, the union of all of the faces of S is the boundary of S, also denoted as $bd(S)$.

Definition 3.2.3 *A simplicial complex X in $\mathbb{R}^n$ is a set of simplices in $\mathbb{R}^n$ such that:*

1. *Every face of a simplex in X is also a simplex in X.*
2. *The intersection of two simplices in X is a face of each of them.*

We could also call *vertices* to the zero simplicies of a simplicial complex. Generally, the set of simplices up to dimension k is called the *k-skeleton* of the complex and is denoted as $X^{(k)}$. We will mainly use *finite simplicial complexes*, composed by a finite number of simplices.

Definition 3.2.4 *The geometric realization $|X|$ of a finite simplicial complex X is the topological space given by the union of simplices, with the relative topology as a subspace of $\mathbb{R}^n$.*

A simplicial complex that has only 0-simplices and 1-simplices represents a graph embedded in the Euclidean space, where the 0-simplices are the vertices and the 1-simplices are the edges.

We can also define continuous maps between simplicial complexes, as follows:

Definition 3.2.5 *Let X and Y be simplicial complexes. A simplicial map $f : X \to Y$ is specified by a map $X^{(0)} \to Y^{(0)}$ such that whenever $\{z_0, ..., z_k\} \subset X^{(0)}$ span a simplex of X, $\{f(z_0), ..., f(z_k)\}$ span a simplex of Y.*

The composition of two simplicial maps is a simplicial map.

Definition 3.2.6 *Let X and Y be simplicial complexes. An isomorphism of simplicial complexes is a simplicial map $f : X \to Y$ that is a bijection on 0-simplices and such that for any $k > 1$, a collection of vertices $\{x_1, ..., x_k\}$ specifies a simplex of X if and only if $\{f(x_1), ..., f(x_k)\}$ is a simplex of Y.*

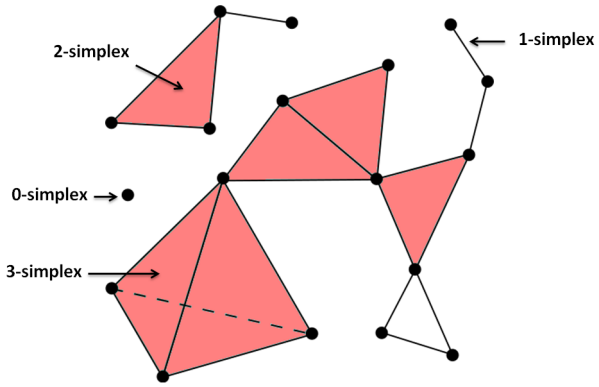
Thus, it is possible to form a category which objects are the simplicial complexes and which morphisms are the simplicial maps.

Lemma 3.2.1 *Geometric realization specifies a functor from the category of finite simplicial complexes and simplicial maps to the category of compact topological spaces and continuous maps.*

Abstract Simplicial Complexes

Simplicial complexes are also possible out of the Euclidean space $\mathbb{R}^n$, and this abstraction is called *abstract simplicial complex*. To the ones just presented above, we will then refer as *geometric simplicial complex*. We could define an abstract simplicial complex as the combinatorial counterpart of a geometric simplicial complex, as:

Figure 3.3: Example of a simplicial complex.



Definition 3.2.7 *An abstract simplicial complex is a set X of finite non-empty sets such that if σ is an element of X then so is every non-empty subset of σ.*

1. *Each element of X represents a simplex; we refer to elements of X as (abstract) simplices.*
2. *The dimension of an abstract simplex σ is $|\sigma| - 1$, where $|-|$ denotes the number of elements of a set.*
3. *Any non-empty subset of a simplex σ is a face of σ.*
4. *The vertices of X are the one-point sets in X.*
5. *More generally, we will denote the subset of X consisting of sets of cardinality $\leq k+1$ as $X^{(k)}$, the k-skeleton.*

As the ones presented for geometric simplicial complexes, similar definitions of maps and isomorphisms are available for abstract simplicial complexes.

There is a relationship among the geometric and the abstract simplicial complexes, that follows from the lemma and theorem below [109]:

Lemma 3.2.2 *Let X be a geometric simplicial complex spanned by the points $x_0, ..., x_k \subseteq \mathbb{R}^n$. Then there is an associated abstract simplicial complex specified by the collection of subsets of vertices of X which span a simplex in X.*

Theorem 3.2.3 *For every abstract simplicial complex S, there exists a geometric simplicial complex $\bar{S}$ such that S is associated to $\bar{S}$.*

This theorem allow us to define a geometric realization of an abstract simplicial complex though the geometric realization of the corresponding geometric simplicial complex. This abstraction is needed to extract abstract simplicial complexes from a high-dimensional data cloud, exploring the variables interaction more than their geometric realization.

The theorem above derives then in the following lemma [109]:

Lemma 3.2.4 *The geometric realization of the associated simplicial complex specifies a functor $|-|$ from the category of abstract simplicial complexes and simplicial maps to the category of topological spaces and continuous maps.*

Homology Groups

The homology groups are the central invariant used in TDA. Generally speaking, they are a collection of functors from the category of abstract simplicial complexes to the category of vector spaces over a field $\mathbb{F}$, $H_n : Simp \to Vect_{\mathbb{F}}$ with $n \geq 0$, that encode information about the simplicial complex structure. In other words, homology assigns a family of vector spaces (called homology groups in more general settings) to a simplicial complex. For a given dimension, the vector spaces capture the topological features in that dimension.

Before defining technically the homology group of a simplicial complex, we would need to be familiar with the *orientation* of a simplex, the *chain groups* and the *boundary homomorphism*, that provide algebraic encodings of the combinatorial information of a simplicial complex. We describe the case of coefficients in a field $\mathbb{F}$ (that will normally be $\mathbb{R}$), that is most relevant for topological data analysis.

Definition 3.2.8 *An orientation of the vertices of a simplex is an equivalence class of orderings of the vertices under the equivalence relation that two orderings are the same if they differ by an even permutation.*

We could denote $[v_0, ..., v_k]$ the oriented k-simplex in X specified by the vertices $\{v_0, ..., v_k\}$, where the orientation is specified by the ordering of the vertices.

Definition 3.2.9 *Let X be an abstract simplicial complex and σ a simplex. The k-chains $C_k(X; \mathbb{F})$ is the vector space with basis the set of oriented k-simplices. That is, elements of $C_k(X; \mathbb{F})$ are linear combinations of generators $\{g_\sigma\}$, where σ varies over the oriented k-simplices of X.*

For example, given an abstract simplicial complex $X = \{[v_0], [v_1], [v_0, v_1]\}$, the space of 0-chains $C_0(X; \mathbb{F})$ is a vector space which is isomorphic to $\mathbb{F} \oplus \mathbb{F}$ having elements of the form $a_0 v_1 + a_1 v_1,\ a_0, a_1 \in \mathbb{F}$. Similarly, the space of 1-chains $C_1(X; \mathbb{F})$ for X is a vector space which is isomorphic to $\mathbb{F}$ having elements of the form $a_0 g_{01}$, $a0 \in \mathbb{F}$, where g_{01} is a generator corresponding to the 1-simplex of X.

We therefore could define a linear transformation $\delta_k : C_k(X; \mathbb{F}) \to C_{k-1}(X; \mathbb{F})$ as the *boundary map*. This is an algebraic way to encode the boundary of a simplex.

Definition 3.2.10 *The linear transformation $\delta_k : C_k(X; \mathbb{F}) \to C_{k-1}(X; \mathbb{F})$ is specified on the generators as*

$$\delta_k([v_0, ..., v_k]) \mapsto \sum_{i=0}^{k} (-1)^i [v_0, ..., \hat{v_i}, ..., v_k]$$

where $\hat{v_i}$ means that vertex is deleted. The homomorphism is then specified by extending linearly to all $C_k(X; \mathbb{F})$.

The geometric interpretation of δ_k applied to a simplex is the alternating sum over the faces that make up the boundary of the simplex. For example, the boundary of the 1-simplex $[v_0, v_1]$ is $v_1 - v_0$. The boundary of the 2-simplex $[v_0, v_1, v_2]$ is then $[v_1, v_2] - [v_0, v_2] + [v_0, v_1]$.

The boundary map has the special property that applying it twice is 0, i.e., the composite $\delta_k \circ \delta_{k+1} = 0$ (the boundary of a boundary is 0). Therefore, we have that $im(\delta_{k+1}) \subseteq ker(\delta_k)$, where $ker(\delta_k)$ is the kernel of δ_k (i.e. the set of elements or "cycles" t with $\delta_k(t) = 0$), and $im(\delta_{k+1})$ is the image of δ_{k+1}.

With this definition, the homology group's idea is to take the subgroup of $C_k(X)$ of cycles, i.e $ker(\delta_k)$, and impose the equivalence relation that two chains c_1 and c_2 are *homologous* if their difference $c_1 - c_2$ is a boundary, i.e., if $c_1 - c_2$ is an element of $im(\delta_{k+1})$.

Definition 3.2.11 *Let X be a simplicial complex and k a natural number. The kth homology group with $\mathbb{F}$-coefficients $H_k(X; \mathbb{F})$ is defined to be the quotient group $ker(\delta_k)/im(\delta_{k+1})$, where δ_k is a linear transformation (or boundary map) among k-chains.*

We could think on the homology groups as a set of cycles in $C_k(X; \mathbb{F})$ that are not the boundaries of elements of $C_{k+1}(X; \mathbb{F})$. We could, therefore, say that:

1. H_0 is a measure of the path components of X.

2. H_1 is a measure of the one dimensional holes in X (i.e. loops).

3. Generally, H_k is a measure of k-dimensional geometric features of X, specifically, a count of the number of k-dimensional holes in X.

The dimensions of the vector spaces or homology groups are called *Betti numbers*, where β_k denotes the Betti number for dimension k.

For example, the abstract simplicial complex $X = \{[v_0], [v_1], [v_0, v_1]\}$ that we have presented previously has a boundary map $\delta_1 : C_1(X; \mathbb{F}) \to C_0(X; \mathbb{F})$. Therefore, taking into account the isomorphisms presented, $H_0(X) = \mathbb{F}$, since $ker(\delta_0)$ is all of $C_0(X, \mathbb{F})$ and the image of δ_1 is $\mathbb{F}$. The $H_1(X; \mathbb{F}) = 0$, as the $ker(\delta_1) = 0$. Consequently, all $H_i(X, \mathbb{F}) = 0$, for any $i > 1$. Geometrically, we could interpret that X represents a topological space that has only one path component without holes.

Of particular interest for topological data analysis is the fact that simplicial homology is algorithmically tractable as δ_k and can be expressed as a matrix where each column specifies the image in $C_{k-1}(S; \mathbb{F})$ of a generator of $C_k(S; \mathbb{F})$, where S is an abstract simplicial complex.

Simplicial Complexes Associated to Data

How can we construct a simplicial complex from a data cloud X? In order to exploit the tools of algebraic topology to study finite metric spaces, we develop an idea for assigning a topological space to (X, d_X).

Imagine that the sampled data X is generated by drawing from some probability distribution on a set embedded in $\mathbb{R}^n$. We need to fill in the gaps between the samples with, for example, the union of the balls around the points. The following definitions present some of those ideas:

Definition 3.2.12 (Union of balls) *Let $X \subset \mathbb{R}^n$ a finite subspace and fix $\epsilon \geq 0$. The union of balls is the union $\bigcup_{x \in X} B_\epsilon(x) \subset \mathbb{R}^n$.*

Definition 3.2.13 (Čech complex) *Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. The Čech complex $C_\epsilon(X, d_X)$ is the abstract simplicial complex with*

1. *vertices the points of X, and*

2. *a k-simplex $[v_0, ..., v_k]$ when a set of points $\{v_0, ..., v_k\} \subset X$ satisfies*

$$\bigcap_i B_\epsilon(v_i) \neq \varnothing$$

Definition 3.2.14 *The nerve $N(U_i)$ of a cover $\{U_i\}$ of X is the simplicial complex with:*

1. *vertices corresponding to the sets $\{U_i\}$, and*
2. *a k-simplex $[j_0, ..., j_k]$ when the intersection $U_{j_0} \cap ... \cap U_{j_k} \neq \varnothing$*

Proposition 1 *Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. There exists a homeomorphism*

$$\cup_{x \in X} B_\epsilon(x) \cong |C_\epsilon(X, d_X)|$$

between the union of balls and the geometric realization of the Čech complex.

Theorem 3.2.5 *Let X be a topological space. Let $\{U_i\}$ be an open cover of X such that all non-empty finite intersections $U_{j_1} \cap .... \cap U_{j_k}$ are contractible (homotopy equivalent to a point). Then the geometric realization $|N(U_i)|$ is homotopy equivalent to X.*

Now, knowing that a graph defined over a finite metric space is a 1-dimensional simplicial complex, we use a mid elaboration of this construction to define a simplicial complex associated to an arbitrary finite metric space (X, d_X).

Definition 3.2.15 (Vietoris-Rips complex) *Let (X, d_X) be a finite metric space and fix $\epsilon > 0$. The Vietoris-Rips complex $VR_\epsilon(X, d_X)$ is the abstract simplicial complex with*

1. *vertices the points of X, and*
2. *a k-simplex $[v_0, ..., v_k]$ when $d(v_i, v_j) \leq 2\epsilon$ for all $0 \leq i, j \leq k$.*

With other words, the Vietoris-Rips complex is the maximal simplicial complex determined by the vertices and 1-simplices specified by a graph G.

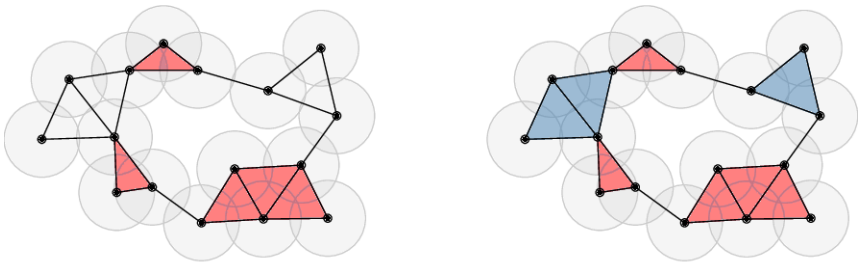
For a point cloud in $\mathbb{R}^n$, the Vietoris-Rips complex and the Čech complex can be different as the Vietoris-Rips complex is completely defined by its 1-skeleton, completing all the potential k-simplex present as observed in the prior figure. However, there is a strong relationship among both:

Lemma 3.2.6 *Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. There are natural simplicial inclusions*

$$C_\epsilon(X, d_X) \subseteq VR_\epsilon(X, d_X) \subseteq C_{2\epsilon}(X, d_X)$$

The Vietoris-Rips complex will be, among others, a way of recovering information about the underlying topological features of a simplicial complex homology study, that will be studied with a TDA method called *persistent homology*.

Figure 3.4: Example of a Čech complex and a Vietoris-Rips complex for the sample data cloud.



3.2.2 Morse Theory

Morse theory's ideas are the basis of another of the TDA methods studied, called *Mapper*. As we don't need the full Morse theory to understand the Mapper design, we will introduce here the basic principles of the theory that lead to constructions inspired on its approach, as the Reeb graph, very used in computational topology.

The Morse theory was written by Marston Morse in 1934 when he published the first article with the main ideas of this theory [110]. Some years after that, in 1963, Milnor extends the Morse article finishing with the theory that can be called *Morse theory* [111]. Morse introduced techniques in the study of a variety topology through the analysis of differentiable functions defined over the variety. This was a revolution at that time, allowing that the nature of the critical points of differentiable functions over the variety was reflecting the topological complexity of the underlying space. During the 60's decade, there were lots of advances in the study of algebraic topology, incorporating computational methods and introducing novel techniques to the study of old problems.

We will describe the basis of the classic Morse theory in order to understand its application in topological data analysis and our work.

Manifolds

Generally speaking, manifolds are a kind of topological spaces more manageable than other structures that provide with geometric intuition for many methods in computational topology. The formal definition should be accompanied of the definition of the coordinates and charts.

Definition 3.2.16 *Let M be a topological space. Given an open set $U \subseteq M$, we say that a chart is an homeomorphism $\theta : U \to V$, where V is an open subset of $\mathbb{R}^n$. The inverse θ^{-1} equips U with a coordinate system.*

A chart is normally noted as (U, θ). An *atlas* of M is then a collection of charts $\{(U_i, \theta_i), i \in I\}$ which covers M, i.e. $\cup_{i \in I} U_i = M$.

Definition 3.2.17 *An n-dimensional topological manifold M is a second-countable (i.e. has a countable base) Hausdorff topological space with an atlas where the charts are all subsets of $\mathbb{R}^n$.*

Intuitively, an n-manifold M is a topological Hausdorff space such that around every point there is a neighborhood that is topologically equivalent to the open unit ball B^n in $\mathbb{R}^n$ (defining an homeomorphism $\theta_i : U_i \to B^n$, for every open cover $\{U_i\}_{i \in I}$ of M).

For two charts $(U_\alpha, \theta_\alpha)$ and (U_β, θ_β) of a manifold M such that $U_\alpha \cap U_\beta \neq \varnothing$, the composites

$$\theta_\alpha \theta_\beta^{-1} : \theta_\beta(U_\alpha \cap U_\beta) \to \theta_\alpha(U_\alpha \cap U_\beta)$$

are referred as transition functions or transition maps for comparing two charts of an atlas and explain the changes of coordinates. For example, two overlapping intervals of $\mathbb{R}$ would be two overlapping charts. Each chart has a coordinate system and transition functions connect these coordinates on the overlaps.

For example, the torus can be also considered a manifold with charts being overlapping squares.

Morse Theory Details

Morse theory comes from the aim of describing the geometric structure of the manifold. It starts by considering a compact manifold M and a differentiable function $f : M \to \mathbb{R}$. The idea is to study the topological information about M encoded in the inverse images $h^{-1}(k)$, $\forall k \in \mathbb{R}$. The planes where the inverse images change are called the *critical points* of the function and is precisely what Morse theory characterizes: a space can be determined by the (finite) critical points of suitable continuous functions from $M \to \mathbb{R}$ with null value for those critical points. From the critical points we will define the level complexes, and both give us information about the homotopy type of M.

Definition 3.2.18 *Let $f : M \to \mathbb{R}$ a differentiable function. A point $p \in M$ is called a critical point of f if the differentiable application of f in p is null. In this case, the real value $f(p)$ is called the critical value of f.*

Definition 3.2.19 *Let $p \in M$ a critical point of f. Fixed $(x_1, ..., x_n)$ a local coordinates system in an environment of p small enough, it is said that the critical point*

p is non-degenerate if the Hessian matrix

$$H_{f,p} = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)(p)$$

is non-singular.

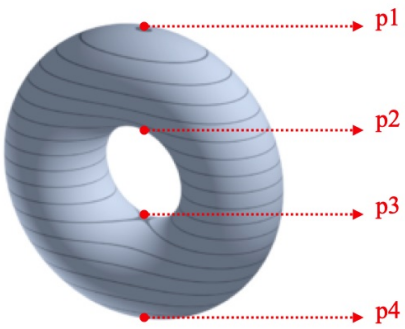
The non-degenerative condition is independent of the coordinates system used, as the non-singularity property is equivalent to the non-null determinant.

Definition 3.2.20 *A differentiable function $f : M \to \mathbb{R}$ is called a Morse function if all its critical points are non-degenerate.*

The meaning of a non-degenerate critical point p is basically the possession of multiple directions in p over f.

As an example, the height function $f : T \to \mathbb{R}$ over the vertical 2-torus T is a Morse function with four critical points $\{p_1, p_2, p_3, p_4\}$.

Figure 3.5: Critical points of the torus.



We could define now the level sets, that are the basic objects introduced by the Morse theory, and are related to the CW-complex definition [112]:

Definition 3.2.21 *A CW-complex is a topological space X inductively constructed as follows:*

1. *The 0-skeleton $X^{(0)}$ is a finite set of 0-cells (or points).*
2. *Given the $(n-1)$-skeleton $X^{(n-1)}$, we construct the n-skeleton $X^{(n)}$ joining the n-cells to $X^{(n-1)}$ through its boundaries. We obtain, therefore, the quotient space of $X^{(n)}$ with the associated quotient topology.*

3. *Repeating the process a finite number of times, we have* $X = X^{(n)}$ *for any* $n < \infty$.

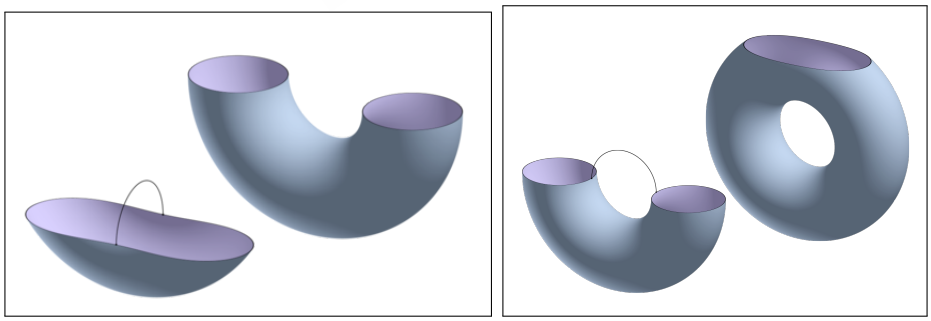
A subcomplex of a CW complex X is a subspace $A \subset X$ which is a union of cells of X, such that the closure of each cell in A is contained in A.

Definition 3.2.22 *Let* $f : M \to \mathbb{R}$ *be a differentiable function over a variety* M. *For each* $a \in \mathbb{R}$, *the level set* M_a *is defined as the closed subset of* M *given by* $M_a := f^{-1}(-\infty, a] = \{p \in M | f(p) \leq a\}$.

If we analyze the 2-torus, we can clear differentiate the level sets contained on it (first example given in [111]). For each $a \in \mathbb{R}$, we have:

1. if $a < f(p_4)$ the level set M_a is empty.
2. For every $f(p_4) < a < f(p_3)$, M_a is homeomorphic to a disc (2-cell), which is homotopically equivalent to a point, i.e., a 0-cell.
3. For every $f(p_3) < a < f(p_2)$, the level set M_a is homeomorphic to a cylinder, which is homotopically equivalent to a disc with a 1-cell attached.
4. For every $f(p_2) < a < f(p_1)$, M_a is the result of extracting a disc to a torus, which is homotopically equivalent to a cylinder with a 1-cell attached.
5. Finally, for every $a > f(p_1)$, M_a is the complete torus, homotopically equivalent to stuck to the previous level complex a 2-cell.

Figure 3.6: Steps 2-3 and 3-4 specified above.



Therefore, we can construct the torus with the level sets derived from its critical points. We conclude, generally, that the critical points of a Morse function allow us to characterize the homotopy type of a variety, characterizing the variety with the structure of a cell complex. Please note that a Morse function defined on a

compact manifold admits only a finite number of critical points, and therefore the manifold can be described by a finite number of elements.

We could then state a link between the number of critical points of a function f and the global topology of the manifold M through the Euler characteristic [113]. The *Euler characteristic* is a combinatorial topological invariant defined as:

$$\chi(M) = \sum_{i=0}^{n} (-1)^i \beta_i(M),$$

where β_i are the Betti numbers of M found at dimension at most n. We have then that

$$\chi(M) = \sum_{\lambda=0}^{n} (-1)^\lambda \mu_\lambda,$$

where $\mu_\lambda(M)$ are the number of cells of dimension λ at most n. Please note that this is equivalent to define the Euler characteristic as the alternating sum of λ-simplices for a finite simplicial complex with simplices of dimension at most n.

The decomposition of M into cells previously described characterizes M by means of the structure of a cell complex. For 2-manifolds embedded in $\mathbb{R}^3$, the sequence of indices of the critical points of a function does not code the way the manifold is embedded into the space. Therefore, we need consider other topological structures, such as Reeb graphs.

The Reeb Graph

The *Reeb graph* and the *Reeb space* are derived from the main ideas of the Morse theory and define the construction of manageable analytical structures [114]. Defined by the George Reeb in 1946 [115], the Reeb graph was not introduced into computer science until 1991 [116]. It is defined as follows:

Definition 3.2.23 *We define an equivalence relation on the topological space X with a Morse function $f : M \to \mathbb{R}^n$, by stipulating that $p \cong q$ if $p, q \in f^{-1}(k)$ for some k and moreover that p and q are in the same path component. The Reeb space of M is the quotient of M by this equivalence relation.*

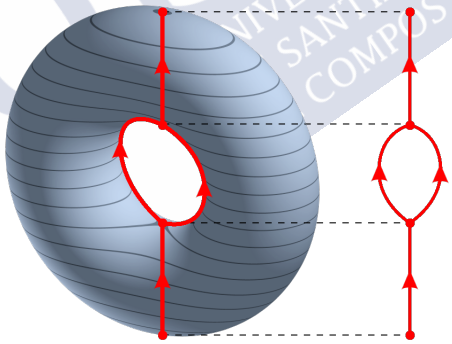
The term *Reeb graph* is used to identify the simplicial complex associated with the quotient space defined by Reeb. It means, if $n = 1$ in the previous definition, we obtain a graph construction referred as the *Reeb graph* (please see figure 3.7 below). The vertices of the Reeb graph correspond to components of the level sets, with edges connecting components as k varies. For orientable, closed 2-manifolds, the

number of loops (defined as the first Betti number) in the Reeb graph corresponds to the genus (number of holes) of the manifold. The generalizations to dimensions greater than two became complicated but we will basically work in $\mathbb{R}$.

The Reeb graph provides, therefore, with the switch from a topological space to a real valued function on that space derived from the Morse level sets, that acts as a high-level shape descriptor from a topological and a geometrical perspective. It is therefore really used for topological data analysis as it allows us to extract easily the underlying data space characteristics in the form of a manageable graph element. As described in the prior section 3.1, the detection of the "shape" of a data cloud is incredibly important for many data analysis applications, where the size and complexity of the data make difficult to visualize it and extract prior information about its distribution. The selection of the function f could be adapted depending on the study aims.

The TDA algorithm Mapper could be defined as an approximation to the Reeb graph with a clustering procedure, as we will explain in the section 3.4.

Figure 3.7: The Reeb graph of a 2-torus with the height function is given by the Reeb space when $n = 1$.



3.3 Persistent Homology

Persistent homology is a technique that has been mainly developed and applied over the last 20 years, but initial ideas were already presented in the early 1990's and the idea of persistence was introduced in 1999 [117]. Initially, persistent homology was used to obtain a large scale geometric understanding of complex datasets, encoded as finite metric spaces. Many different applications of persistent homology were developed in the last years in different areas as evasion problems, viral evolution, cancer research, or neural networks [119]. There are still several fields of ongoing study, mainly based on the interpretation and comparison of the PH outcomes, sensitivity to noise, and the application of persistent homology to complexes produced by random models.

Filtration

Persistent homology is based on the idea of a filtration over a simplicial complex, as described in the following definition.

Definition 3.3.1 *If K is a finite simplicial complex, a filtration of K is a chain of subcomplexes $K_0 \subseteq K_1 \subseteq ... \subseteq K_n = K$.*

For each $k \geq 0$ associated to any filtration $K_{()}$ of K, we will have a sequence of vector spaces

$$H_k(K_0) \to H_k(K_1) \to ... \to H_k(K_n)$$

applying the homology functor to the chain of simplicial complexes defined by the filtration. So, every filtered system of simplicial complexes produces a sequence of vector spaces. Please note that we denote $H_k(K_i)$ or $H_k(K_i; \mathbb{F})$ to the homology group in dimension k instinctively.

Generally, given $K_i \subseteq K$ an element of the filtration, and given a homology class $\alpha \in H_k(K_i)$ we will say that:

1. α is born in i if it is not in the image of the application $H_k(K_{i-q}) \to H_k(K_i)$, for $0 < q < i$.
2. α dies at $l > i$ if it becomes zero in $H_k(K_l)$, or its image in $H_k(K_l)$ coincides with the image of another class that was born before.

All the information in the sequence of vector spaces generated by the homology groups can be encoded as *persistence barcodes* (PB), which are a collection of non-empty intervals where we track the birth and death (and so lifespan) of a

topological feature found by the filtration. We can also think on a barcode in $\mathbb{R}^2$, defined as a collection of points plotted by their appearance (x axis) and disappearance (y axis) and that is called a *persistence diagram* (PD).

The Vietoris-Rips complex and the Čech complex produce natural examples of filtered systems of simplicial complexes from the data of a finite metric space varying the scale of ϵ. We will mainly use the Vietoris-Rips complex filtration for our analyses, as will be described in subsection 3.3.2.

Therefore, given (X, d_X) a finite metric space (that represents our data) and a sequence of real numbers $\epsilon_0 < \epsilon_1 < ... < \epsilon_m$, we have a filtration of the complex $VR_{\epsilon_m}(X, d_X)$:

$$VR_{\epsilon_1}(X, d_X) \to \cdots \to VR_{\epsilon_m}(X, d_X)$$

and therefore we have a sequence of vector spaces

$$H_k(VR_{\epsilon_1}(X, d_X)) \to \cdots \to H_k(VR_{\epsilon_m}(X, d_X))$$

On a homology level then, the functor $H_k(VR_{(-)}(X, d_X))$ provides a means to compare between the homology of the complexes as ϵ varies. For example, an element $\gamma \in H_k(VR_{\epsilon_i}(X, d_X))$ is a k-dimensional feature at scale ϵ_i. We determine the significance and stability of γ by finding the maximum $j > i$ such that the image of γ under the group homomorphism:

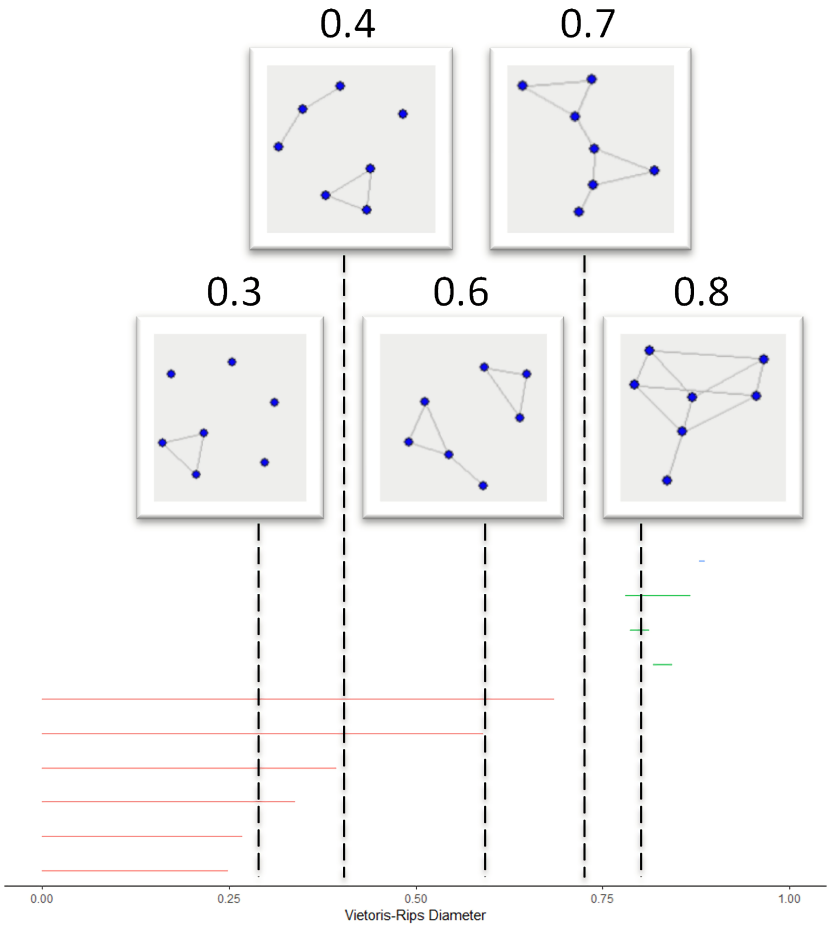
$$\theta_{ij} : H_k(VR_{\epsilon_i}(X, d_X)) \to H_k(VR_{\epsilon_j}(X, d_X))$$

is non-zero. We could then say that an element $\gamma \in H_k(VR_{\epsilon_i}(X, d_X))$ represents a k-dimensional hole in the geometric realization of the Vietoris-Rips complex at ϵ_i. We say that the feature was "born" at ϵ_i if γ does not exist for $\epsilon^{'} < \epsilon_i$. If $\theta_{ij}(\gamma) = 0$, we know that the corresponding hole is filled and therefore we say that γ "dies" at j. This calculation of birth and death times lead us to calculate also the lifespan of a feature as a measure of its persistence.

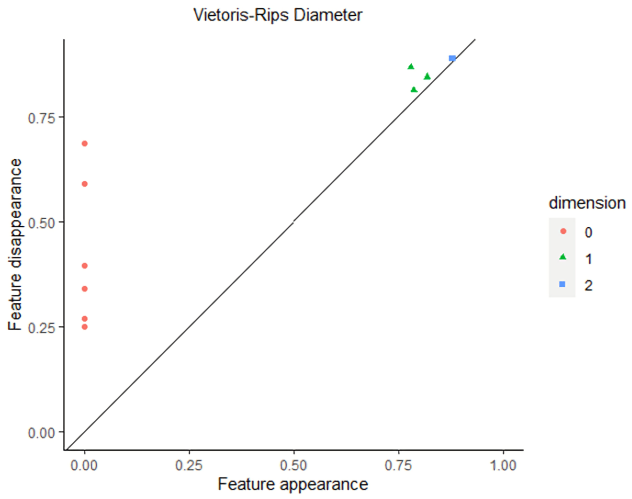
In figure 3.8 we present an example of a PB and a related PD for a simplicial complex with seven vertices or CpGs. Please note that this figure is just an example of the persistence barcodes and diagrams layout, and the way of computing the PH with the simplicial complex will be detailed in the subsection 3.3.2. Each pair of nodes is joined by the inverse of their absolute Pearson correlation. The filtration is based on the construction of the Vietoris-Rips complex, so the complete subgraph of each element is generated as each step (i.e. the clique complexes). Results are created with the R package *TDAstats* [118], which produces one less zero-dimensional feature than expected (to avoid redundancy).

Figure 3.8: Example of persistent homology results.

(a) Persistence barcode.



(b) Persistence diagram.



3.3.1 Stability of Persistent Homology under Perturbation

If we are going to use topological invariants to describe our data, we would need to know how those invariants may change with data perturbations. One of the advantages of persistent homology is that the set of barcodes forms a metric space, i.e., the distance between the barcodes allows us to measure changes in the related TDA output associated with the underlying data.

The definition of a metric in the input data metric space permits to define stability theorems of persistent homology that relates perturbations of the input data with the input metric with perturbations in the output barcodes with an output metric [120]. We will describe input and output metrics, together with the persistence theorems, below.

Input metrics

The *Gromov-Hausdorff* input metric allows to define and proof stability theorems. It is defined as:

Definition 3.3.2 *Let A and B be non-empty subsets of a metric space (X, d_X). Then we define the Hausdorff distance between A and B to be*

$$d_H(A,B) = max\left(\underset{a\in A}{sup}\,\underset{b\in B}{inf} d_X(a,b), \underset{b\in B}{sup}\,\underset{a\in A}{inf} d_X(a,b)\right),$$

or equivalently

$$d_H(A,B) = \underset{\epsilon>0}{inf}\{B \subseteq A_\epsilon, A \subseteq B_\epsilon\},$$

where A_ϵ and B_ϵ denotes the sets of all points within distance ϵ of A and B, respectively.

With other words, the Hausdorff distance is determined by the point in A with the largest distance to the closest point in B (and in the other way around).

Lemma 3.3.1 *The Hausdorff distance imposes a metric on the set of non-empty subsets of a metric space (X, d_X).*

In order to generalize the definition above, Gromov proposed to consider the infimum of the Hausdorff distance over all isometric embeddings of the two metric spaces into a larger ambient metric space.

An isometric embedding $\phi : (X, d_X) \to (Y, d_Y)$ is an injective map $X \to Y$ such that $d_X(x_1, x_2) = d_Y(\phi(x_1), \phi(x_2))$. So an isometric embedding identifies X with a submetric space of Y.

Definition 3.3.3 *Let* (X_1, d_{X_1}) *and* (X_2, d_{X_2}) *be compact metric spaces. The Gromov-Hausdorff distance between* X_1 *and* X_2 *is defined to be*

$$d_{GH}((X_1, d_{X_1}), (X_2, d_{X_2})) = \inf_{\theta_1: X_1 \to Z, \theta_2: X_2 \to Z} d_H(X_1, X_2)$$

where θ_1 *and* θ_2 *are isometric embeddings of* (X_1, d_{X_1}) *and* (X_2, d_{X_2}) *in* (Z, d_Z) *respectively.*

We say that two metric spaces are *isometric* if there exists an isomorphism $f : X \to Y$ that preserves all distances. This leads to the following theorem [120]:

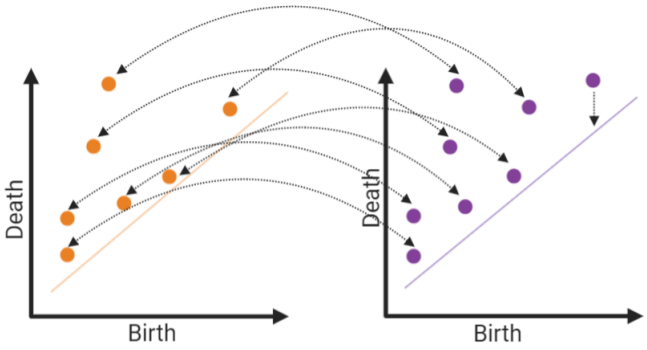
Theorem 3.3.2 *The Gromov-Hausdorff distance is a metric on the set of isometry classes of compact metric spaces.*

In practice, the Gromov-Hausdorff distance measures the maximum distortion in the best matching between two metric spaces. This distance is a suitable means for capturing perturbations of data sets.

Output metrics

We can compare different barcodes with several barcodes distances as the *bottleneck distance*, that measures the distance between two persistence diagrams P and Q by the maximum distance between two points in a matching from P to Q; or the *Wasserstein distance*, that considers the total distance between the matched pair of points. Of course, the optimal measures of those distances need to be assessed within each particular study comparing with, for example, permutation or random tests.

Figure 3.9: Example of matching points to calculate the distance between two persistence diagrams.



Before presenting the formal definitions of those distances, we define generally a distance between two intervals $[a_1, b_1)$ and $[a_2, b_2)$ as:

$$d_\infty([a_1, b_1), [a_2, b_2)) = max(|a_1 - a_2|, |b_1 - b_2|)$$

This leads to have $d_\infty([a, b), \varnothing) = \frac{|b-a|}{2}$.

Now, we have the barcodes B_1 and B_2 that are sets of intervals. We assume that $|B_1| < |B_2|$ and define a bijection $\phi : A_1 \to A_2$, where A_1 is a multi-subset of B_1 and A_2 is a multi-subset of B_2. We also add $\varnothing$ to B_1 and B_2, and consider the elements of B_1/A_1 and B_2/A_2 as matched with $\varnothing$. Then, we can define:

Definition 3.3.4 *Let B_1 and B_2 be barcodes. The bottleneck distance is defined to be*

$$d_B(B1, B2) = \inf_{\phi} \sup_{Z \in B_1} d_\infty(Z, \phi(Z))$$

where ϕ varies over all matching between B_1 and B_2 and the supremum is taken over bars in B_1.

This means, the bottleneck distance measures the worst discrepancy in the best matching between the two barcodes. Please note that this distance could also be defined in terms of persistence diagrams.

Definition 3.3.5 *Let B_1 and B_2 be barcodes. For $p > 0$, the p-Wasserstein distance is defined to be*

$$d_{W_p}(B_1, B_2) = \left(\inf_{\phi} \sum_{Z \in B_1} d_\infty(Z, \phi(Z))^p \right)^{1/p}$$

Normally, we will take $p = 1$, so d_{W_1}.

Link Between Input/Output Metrics

We can now establish the link between the bottleneck distance and the Gromov-Hausdorff distance with the following theorem [120]:

Theorem 3.3.3 (Stability theorem) *Let (X, d_X) and (Y, d_Y) be finite metric spaces. Then for all $k \geq 0$,*

$$d_B(PH_k(VR(X, d_X)), PH_k(VR(Y, d_Y))) \leq d_{GH}((X, d_X), (Y, d_Y))$$

where PH_k denotes the persistent diagrams resulting from applying persistent homology up to the homology dimension k to the Vietoris-Rips complex.

Please note that we could obtain similar results using the Čech complex or the Wasserstein metric.

There are several extensions of the persistent homology theory, as the *zigzag persistence* (that takes into account filtrations in different directions) or the *multidimensional persistence* (that filters by more than one function). However, we will not present them here as they are not of interest for our work.

3.3.2 Persistent Homology with Networks

Of particular interest for us is the application of PH to complex network analysis, particularly to correlation networks that will be derived from our DNA methylation data cloud. During the last years, PH was largely applied for the study of complex networks as a multiscale feature extractor. The main applications of PH to networks are with brain and protein-protein interaction (PPI) networks [121], or to study the three-dimensional genome structure identifying loops [122]. But PH is also applied to non-biological data as social structures or the financial context. Interestingly, [123] proposes the use of PH for the selection of molecular therapies analyzing the relationship between PPI networks and cancer survival. They in fact suggested a promising future application of those methodologies:

> "[...] We hope that the use of advanced mathematics in medicine will provide timely information about the best drug combination for patients, and avoid the expense associated with an unsuccessful clinical trial, where drug(s) did not show a survival benefit [...]".

PH is normally applied to a defined simplicial complex or network but can also be applied to study the differences among networks extracting their topological features to compare them. Many studies have used the bottleneck or Wasserstein distances to measure the differences among barcodes obtained for different networks. In addition, different statistical classifications techniques (community detection, k-means clustering, etc.) are normally applied to the PH results in order to differentiate them among sample groups [124, 125].

Different filtration functions were applied to weighted and undirected networks, basically creating subgraphs obtained via a sequence of edge weights [119]. Other solutions were developed for unweighted networks, as assigning edge weights based on some network property, such as edge-betweenness centrality, or the use of discrete Morse theory to obtain the filtration [126]. In our case, we use the Vietoris-Rips filtration for the weighted networks we create. Before defining it, we need to formally introduce the definition of a *clique complex*:

Definition 3.3.6 *Let $G = (V, E)$ be an undirected graph. The clique complex $Cl(G)$ is the simplicial complex with all complete subgraphs of G as its faces.*

The 1-skeleton of $Cl(G)$ is G itself.

Definition 3.3.7 (Vietoris-Rips (VR) filtration) *Let $G = (V, E)$ be an undirected weighted graph with the weight function $W : V \times V \to \mathbb{R}$ defined on E. For any $\epsilon \in \mathbb{R}$, the 1-skeleton $G_\epsilon = (V_\epsilon, E_\epsilon) \subseteq G$ is defined as the subgraph of G where $V_\epsilon = V$ and its edge set $E_\epsilon \subseteq E$ only includes the edges whose weight is less than or equal than ϵ. Then, for any $\epsilon \in \mathbb{R}$, we define the Vietoris-Rips complex as the clique complex of the 1-skeleton G_ϵ, $Cl(G_\epsilon)$, and the Vietoris-Rips filtration is then defined as $\{Cl(G_\epsilon) \to Cl(G_{\epsilon'})\}_{0 \leq \epsilon \leq \epsilon'}$.*

With this filtration, we apply PH for two main objectives:

1. To study, in the first place, the topology of the correlation structure of specific groups of CpG sites (as CpG islands). We create graphs or simplicial complexes where each node is a CpG site and they are joined by their absolute Pearson correlation: the graph $G = (V, E)$ is formed by $V = \{CG_1, ..., CG_n\}$, where n is the number of selected CpG sites; and

$$e_{CG_i, CG_j} = 1 - |\rho(CG_i, CG_j)|, \quad \forall\, i, j \in 1, ..., n \tag{3.1}$$

   To generate a system of filtered graphs from a starting correlation graph $G = (V, E)$, we need to subsequently modify the correlation matrix C associated to G that defines the weights of the set of edges E. We define then the altered matrix $\bar{C} = 1 - |C|$, and a set of thresholds $\delta = \{0, 0.1, ..., 1\}$ to be used for the filtration. For each threshold δ, we generate a graph $G_\delta = (V, E_\delta) \subseteq G$ with a set of nodes V and a set of edges $E_\delta \subseteq E$, that derives from the weighted adjacency matrix $\bar{C}$, such tat:

$$\text{if} \quad \bar{C}_{ij} > \delta \quad \text{then} \quad \bar{C}_{ij} = 0 \tag{3.2}$$

   Please note that, as we are using the inverse of the correlation, the first filtered graphs correspond to the higher correlation coefficients.

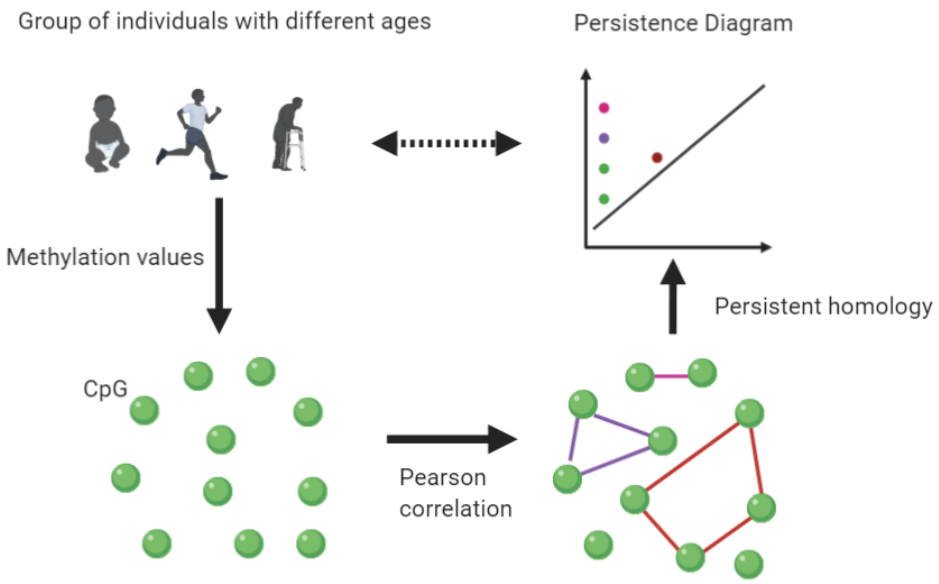
   We will apply persistent homology with the VR-filtration to the described graphs. The same analysis will be done for separated groups of age that are afterwards compared (see figure 3.10 below as an analysis reference). This analysis will be presented in Part II of the present document.

2. To interpret, in a second step, the topological features of the simplicial complexes obtained with MultiNet (that will be described in the part III). In this case, each node is a cluster of CpG sites, so the graph $G = (V, E)$ is formed by

$$V = \{V_1, ..., V_n\} = \{\{CG_{1,1}, ..., CG_{1,n_1}\}, ..., \{CG_{n,1}, ..., CG_{n,n_l}\}\},$$

where n is the number of MultiNet nodes and n_i is the number of CpG sites within each node; and $e_{V_i,V_j} = 1 - mean(|C|)$, $\forall\, i, j \in 1, ..., n$, where C is the correlation matrix of the CpGs contained on V_i and V_j. This analysis will be presented in Part III.

Figure 3.10: Illustration of the study of the epigenetic changes with the aging process using persistent homology.



Of course, the efficiency of the PH algorithm depends on the number of initial simplices and the homology dimension used. Usually, the study of the homology up to dimension H_2 is enough to track interesting topological features in a data analysis study because it allows to already detect structural information.

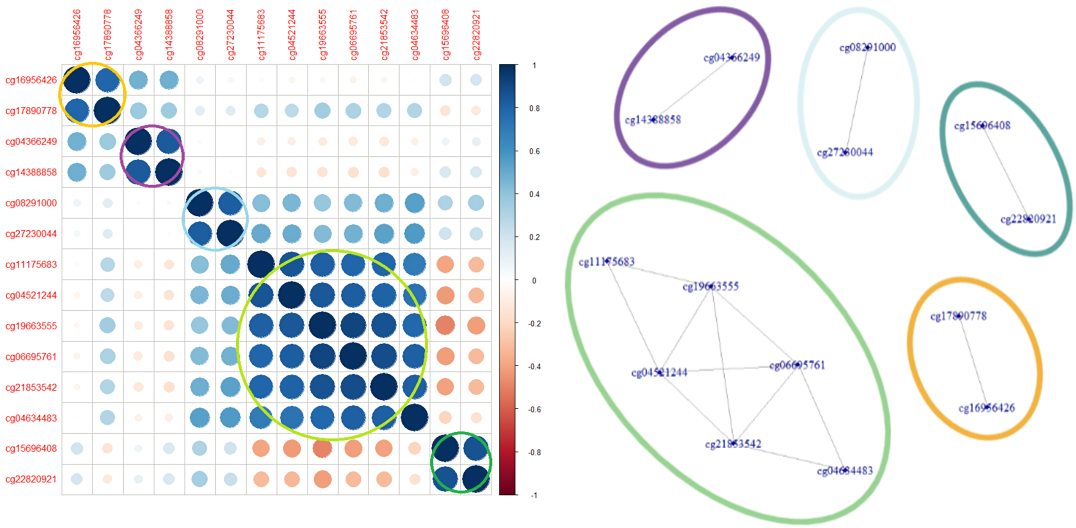
3.3.3 Persistent Homology with Methylation Data

Persistent homology applied to the weighted correlation network extracts the main topological elements that define their structure. Due to the computational limitation of PH, we restrict this study to the analysis of known CpG islands. Despite the main results will be presented in the next part II, we will also describe

here an illustrative example of how the VR-filtration and PH work over the correlation network.

For this purpose, we use a 450K methylation dataset (GSE40279) from a healthy adult population (656 subjects, 19-101 years, whole blood) and select the CpG sites associated with the CpG island chr7:94284858-94286527, the biggest island of chromosome 7. We then generate a network of CpG sites (nodes) joined by the modification of the Pearson correlation coefficient described in equation (3.1). The correlation plot below (figure 3.11 at left) represents the correlation matrix with colors indicating higher (dark blue or dark red) or lower correlation coefficients. The higher correlated clusters circled correspond to the edge-connected subgraphs or modules of the network at right. Some of the CpG sites are also highly positively correlated with age (with coefficients over 0.6).

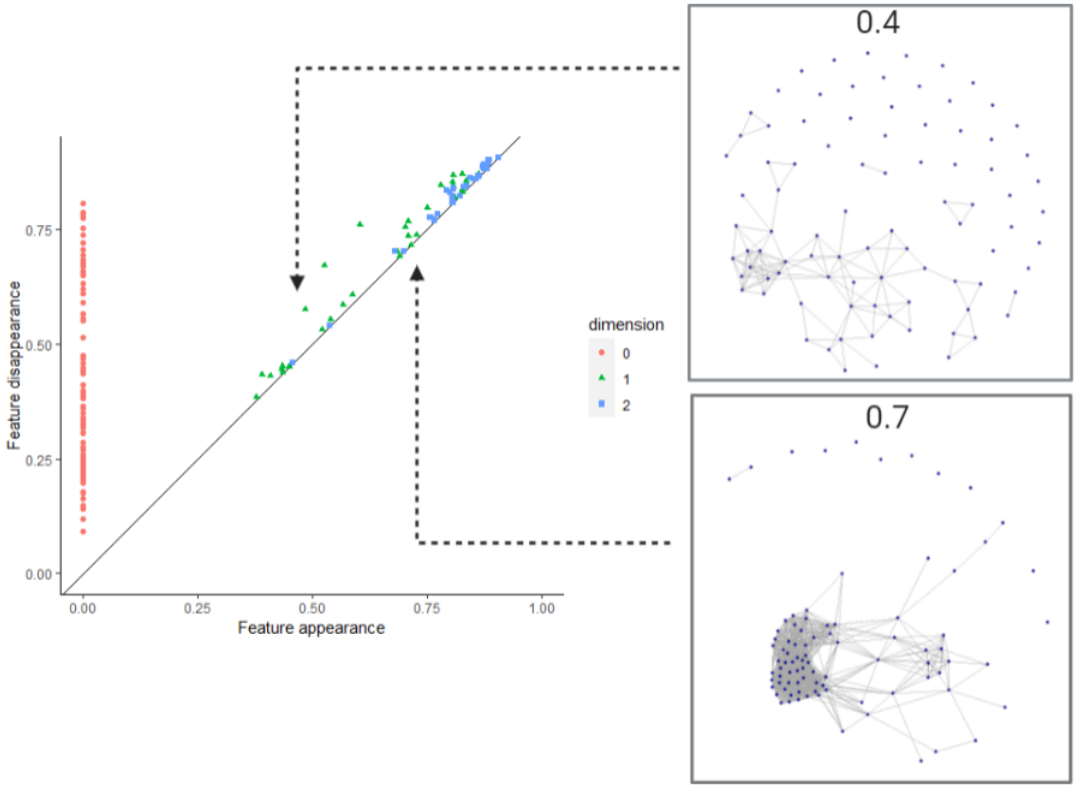
Figure 3.11: Correlation matrix and related weighted filtered graph at a threshold $\delta = 0.2$.



The bigger green circle describes a stronger network of interactions among the CpG sites included. Respectively, the related subgraph has more nodes and edges. The topological features formed by this green module would be then describing the correlation structures formed due to a higher correlation concentration, which may unravel interesting biological conclusions.

Consequently, the application of PH to the weighted network points out different topological features in homology dimensions 0, 1, and 2 observed in the persistent diagram of figure 3.12.

Figure 3.12: Persistence diagram and related filtered correlation graphs at filtration thresholds 0.4 and 0.7.



In the first steps of the filtration, that correspond to high levels of correlation, we observe basically topological elements of dimension 0 because of the strict correlation filter. Those elements are derived from the non-joined points or the joining of 2 or 3 CpGs that form a 1-simplex or a 2-simplex. The first topological features of dimension 1 (holes) appear from 0.4, having a higher incidence with higher weights (lower correlation coefficients) due to the increase of the edge-density. The same applies to dimension 2. We see how the subgraphs that started with a weight of 0.4 continue to grow in edge-density with increasing VR-weights, presenting a clearer module-separation. We pay special attention to the features appearing from dimension 1 with lower weights because they are showing a structured high correlation among the corresponding CpG sites.

In the part II of the present work, we will study deeply the local behavior of the correlation in CpG islands and its differentiation among age groups. The sample characteristics the modulated design of the associated correlation networks will be key to understand the birth, death and lifespan of the related topological features.

3.3.4 Challenges of Persistent Homology

Persistent homology is a successful technique to detect and visualize the underlying data topological features that may describe better the data structure than routine techniques. However, this methodology has also several constraints. The networks used on the main PH applications until now are mostly small networks with less than 1,000 vertices due to the increase of the computational time if we have more nodes. Besides, even if the computational time is still reasonable or feasible, the results of PH may be uninterpreted if we apply it to huge networks, where the birth or death of topological features is not comparable anymore with the network visualization.

As the homology dimensions calculated with reduced computational times are basically H_0 and H_1, we may be loosing higher-dimensional topological features that could bias the conclusions of the analysis.

In addition, despite there are some basic stability parameters, PH results may not be robust when dealing with noisy sampling data, and slight alterations on the underlying distribution may derive in very different barcodes.

The implementation of programming functions related to PH needs to be refined, having only a couple of packages in R (*TDA*, *TDAstats*). Despite they work fine, there are only a few functions inside to operate. One may find a slightly higher level of sophistication with Matlab or Phyton.

For us, the extrapolation of the analysis just presented to the entire 450k array would not be an easy task due to the high computational time and the complexity to interpret the results. For that reason, we decided to complete our local correlation analysis with a global one inspired in another TDA technique called *Mapper*.

3.4 Mapper

An alternative topological data analysis technique arises with power, with the aim of analyzing complex data structures with a clustering procedure. It is called *Mapper*, and it is mainly based on Morse theory concepts.

First implemented in 2007 [127], Mapper is a computational algorithm that extracts descriptors of high-dimensional data sets represented as simplicial complexes, which retain topological and geometric information. Mapper does not act directly over the data, but instead, it selects a filter or combination of filters that act as a data mapping to a metric space. Then, cluster analysis is applied to the point cloud with overlapped intervals defined by the filter functions to construct the network joining the clusters (or nodes) with common points.

Mapper is an alternative to the common hypothesized models of data description, as it does not assume any data properties. It combines then topological basis and machine learning tools into an integrated data analysis methodology. The dimension of Mapper is determined by the number of filter functions used to map the information contained in the data cloud.

Starting from a point cloud X on a finite metric space (X, d_X), the idea of Mapper is to study the topology of the sub-level sets of a function $f : X \to \mathbb{R}^n$, for any $n > 0$, defined on the point cloud. The generic version of the one-dimensional Mapper algorithm on X computed with the filter function f can be then summarized as:

1. Cover the range of values $Y = f(X)$ with a set of consecutive intervals $\{I_s\}_{(1 \leq s \leq S)}$ with a defined percentage of overlap.
2. Apply a clustering algorithm to each inverse image $f^{-1}(I_s)$, $s \in \{1, \ldots, S\}$. This defines a pullback cover $C = \{C_{(1,1)}, \ldots, C_{(1,k_1)}, \ldots, C_{(S,1)}, \ldots, C_{(S,k_S)}\}$, normally overlapped, of the point cloud X, where $C_{(s,k)}$ denotes the kth cluster of $f^{-1}(I_s)$.
3. The Mapper is then the nerve (as defined in 3.2.14) of C. Each vertex $v_{(s,k)}$ of the Mapper corresponds to one element $C_{(s,k)}$, and two vertices $v_{(s,k)}$ and $v_{(s',k')}$ are connected if and only if $C_{(s,k)} \cap C_{(s',k')} \neq \varnothing$, i.e. they have common points.

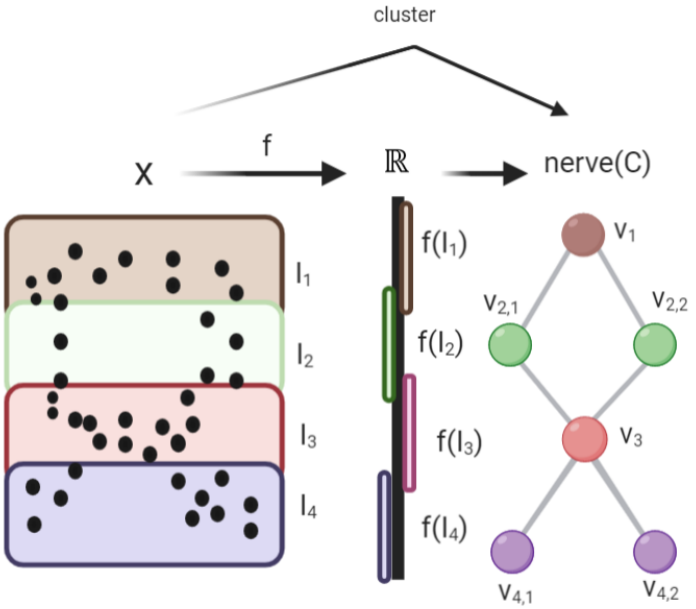
Optionally, we could assign a color to each vertex in the graph based on a function of interest to see the distribution of that function over the graph.

It is obvious that the process described above and its results are mostly dependent on the filter function f selected and the metric d. Depending on what we would

like to see, we would define our filter function(s). Moreover, the filter functions can be specifically selected by the user or defined from the data (using coordinates from PCA analysis, for instance). Apart from the filter function, the number of clusters that you can specify as part of the cluster analysis might depend on the data or might be a fixed number. Similarly, the overlapping percentage or the cluster analysis method (despite Mapper uses normally hierarchical clustering) are other of the parameters that can be adjusted or automatically selected.

The fact that the user has to select those parameters may seem a disadvantage at a first step but, if it is correctly managed, it is really providing with a big flexibility to apply Mapper to different types of datasets, samples, and aims. However, wrong parameter selection may lead to unexpected results. There are different ways to deal with the Mapper parameter selection and sensitivity, some more technical [128] and others more experimental. A common check when applying Mapper is to use a different choice for the set of parameters in order to observe the changes of the network obtained and how the clusters are organized.

Figure 3.13: Example of 1D Mapper functioning over a point cloud X.



Mapper is really useful to visualize the structure of the data, which may lead to find clinically relevant results or statistically differentiated features. From the shape of the network obtained with Mapper, we can already extract substantial information, but also from the subnetworks contained on it. A post-processing step (defined by the user) is normally done to understand the underlying data characteristics.

There are multiple examples of the success of Mapper over routine techniques of data analysis and visualization, that emerged above all during the last 5-10 years. One of the most famous applications of Mapper was published in 2011 and applied to a breast cancer dataset, where the gene expression pattern was analyzed in a correlation metric space [107]. They identified a unique subgroup of Estrogen Receptor-positive (ER+) breast cancers from Mapper output that was not identified with common clustering. This study suggests that Mapper design may be useful to differentiate case and control samples but also to differentiate different subtypes or alterations from the same disease. Other examples of recent biological Mapper application are described in [108, 129, 130]. None of them, that we know, was applied to study specific epigenetic modifications as DNA methylation. Additionally, Mapper was applied to many fields as economics or sociology.

We take the Mapper design as a basis for developing an alternative computational algorithm that aims to analyze high-dimensional epigenetic datasets or any other high-dimensional data. This robust and powerful generalization that we call *MultiNet* will be explained deeply in part III, together with the results and a computational implementation guide.

3.4.1 A Toy Example of Mapper Application

In order to fully understand the Mapper functioning, we will present some simple examples of its application and comparison with typical methods as PCA.

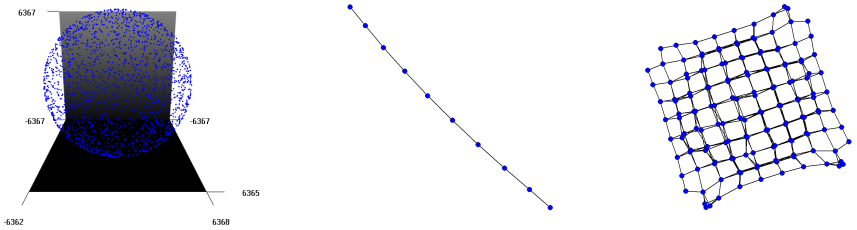
Imagine that we would like to see how Mapper describes a 3D sphere. Using a common euclidean distance as the metric behind, we could use the coordinates of the sphere as filter functions of Mapper. First question would be: which Mapper dimensionality should we choose? If we apply one-dimensional Mapper to the sphere (using the x axis coordinates), results are not substantial and the algorithm is not correctly describing the "shape" of the data (figure 3.14). However, with the two-dimensional Mapper (with x and y axis coordinates), results are closer to what we would have expected and with PCA outcome.

So, with this example we can already imagine how important is to choose the correct dimension of Mapper in order to describe correctly our set of points. The rest of the parameters used were the same for both: 10 intervals (or 10×10), 50% overlap, and 10 clusters per interval. In case we select less intervals or less number of clusters per interval, results would be really similar but with less vertices. Filter functions could be also selected based on the PCA analysis coordinates but results are not varying dramatically.

However, if we do a similar analysis with the torus instead of the sphere, we do

Figure 3.14: Example of a sphere generated with 2,000 points and the corresponding 1D and 2D Mapper results.

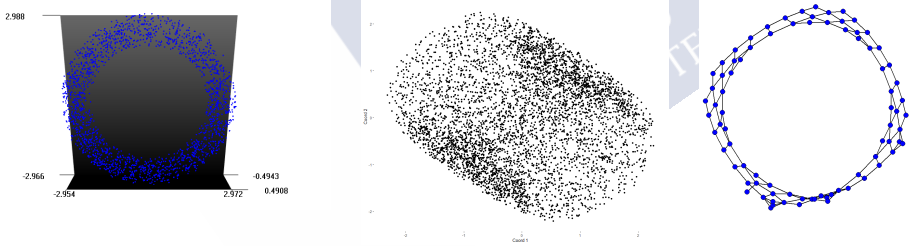
(a) Sphere point cloud. (b) Mapper 1D. (c) Mapper 2D.



see differences between PCA and Mapper results (figure 3.15). In this case, PCA does not describe efficiently all the topological features of the torus and only the 2D Mapper with the two first coordinates of the torus as filter functions detects the "hole" in the middle.

Figure 3.15: Example of a torus generated with 2,000 points and the corresponding PCA analysis and 2D Mapper result.

(a) Torus point cloud. (b) PCA result. (c) Mapper 2D.



Hence, TDA provides with an alternative approach to capture all the information hidden in our data.

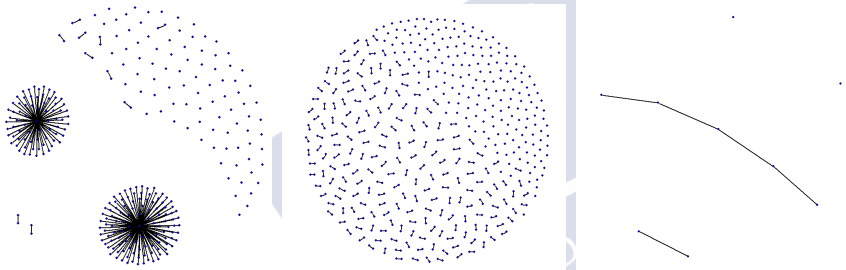
3.4.2 Mapper with Methylation Data

We have presented a toy example to see how Mapper is working with a simple data structure and its sensitivity to the parameters selected, including the dimension used. The next step would be to apply this algorithm to our methylation 450k Illumina dataset and see if it can differentiate the distinct methylation patterns for each sample group, or if Mapper works better than PCA or MDS with this data.

We would like to apply Mapper to visualize how a sample with different ages is organized based on their methylation values. For this purpose, we have used the two different 450k methylation datasets, from a healthy pediatric population

(GSE36054, 192 subjects, 0-17 years, peripheral blood leukocytes) and a healthy adult population (GSE4027, 200 subjects, 19-101 years, human whole blood). We analyze the different methylation patterns among both groups applying PCA, MDS and Mapper. In this case, PCA operates slowly with the entire array and provides with a clear group differentiation, similarly to MDS. Mapper results come faster and obtain a differentiation for one-dimensional Mapper with the median methylation per subject as filter function and the Pearson correlation as a distance measure for clustering (10 intervals, 50% overlap and 3 clusters). It successfully differentiates the sample groups but results are altered dramatically if we change the number of clusters per interval to 2 or 5, as figure 3.16 shows. A similar event occurs with the 2D Mapper, demonstrating its "dangerous" sensitivity to the parameter selection.

Figure 3.16: 1D Mapper results with 3, 2, and 5 clusters.



3.4.3 Challenges of Mapper

We have seen that Mapper is a very useful tool of data analysis applicable to many different fields. A wider application from different perspectives and an improvement of the parameter selection are topics still to be developed and implemented.

Mapper works fine with reduced data sizes but it starts to be computationally slow with higher ones. This is mainly due to the use of hierarchical clustering.

The available programming packages (as the *TDAmapper* package in R) could be improved including more functionalities and a better network representation. Moreover, they do not provide with a post-processing step after creating the graphs that allow a deeper analysis of the data distribution, or even the interpretation of the layout of the graphs from a topological perspective, analyzing the forms of the simplicial complex.

These constrains lead us to develop an alternative computational algorithm (Multi-Net), inspired on the idea of Mapper but with improved functionalities. Despite we will explain it in the part III, the main improved points are explained in the next section.

3.5 From Mapper to MultiNet

Most of the Mapper applications done until now present several disadvantages to apply it to our methylation dataset, that we would like to solve creating the new algorithm MultiNet. Some of those improvement points are:

- Usually Mapper is applied to a data cloud where each point represents a subject, with the objective of differentiating groups. Thus, the groups separation is normally clear but we cannot know the underlying reasons with simply analyzing the graph. For instance, in the previous example with age sample groups, we do not know which are the CpG sites with a highest group differentiation and neither the methylation trend with age. Moreover, the distribution of the beta values in the 450k methylation array is particularly dense in the extremes, not Gaussian, with a standard deviation greatly compressed in the low (between 0 and 0.2) and high (between 0.8 and 1) ranges, potentially biasing the calculation of the mean values per patient. This type of distribution is caused by the fact that the CpGs contained in the 450k Illumina methylation array are not taken uniformly along the genome, but they are spread out around different genomic regions.

- If we only join nodes with common points, we may be loosing interesting information about other common features among nodes. In order to describe the complex correlation structure of the epigenome, we may need to develop an alternative way of joining nodes with common characteristics as for example those with a high mean correlation coefficient.

- Computational times are low with reduced sample sizes. Once the observed sample size increases, the computational times increases dramatically due to the use of hierarchical clustering. Alternative methods could be used.

- Mapper suffers a great instability derived from different parameter selection. There is not a clear guideline for it and most applications are based on prior exploratory analysis. Improvements may be done in terms of a higher stability and selection guidance.

All those observed issues lead us to develop MultiNet, with the same topology basis but a different data analysis approach, that is described in part III. MultiNet is then presented as an improvement of Mapper, that adds several new functionalities of data analysis especially designed to work efficiently with high-dimensional (epigenetic) data.

3.6 TDA Applied to Genomics

Persistent homology and Mapper have been already implemented in several papers to study the topology of genomic data, to describe or characterize biological networks (proteins interactions, gene expression interactions, or neuronal system functioning). However, the application of topological approaches to the study the complex stratified structure of the DNA is still in the early stages, and the expectation is that they grow importantly in the next coming years. Some of the published topological studies with genomic data are:

- Topological study of time series analysis: there are a lot of studies showing, for example, time dependent gene expression profiles. Several mathematical approaches have been followed to study those biological mechanisms, from spectral methods to stochastic processes. TDA in terms of persistent homology can also be applied to this type of data [131].

- Topological study of neuroscience: neurons in the brain and their interactions or connections between different regions of the brain have several properties that can be captured in a network. Researches are currently exploring the use of topological techniques to characterize the molecular, neuronal, and architectural properties of the brain [108, 129].

- Topological study of biomedical imaging: solid tumors appear as masses in various imaging technologies. Tumors have a shape and volume, and these can be modeled as the topological and geometric properties of a three-dimensional object. Rough metrics on these masses are used as a standard prognosis and to evaluate therapeutic efficacy. This work suggested an interesting approach of combining omic data with imaging to better characterize the mechanisms of initiation and progressing of tumor growth [132].

- Topological study of the spreading of infectious diseases: epidemiology has widely used networks to capture information about the spread of an infectious disease. TDA was used to study the mathematical structure of these graphs, their dimensionality, and their topological and geometric properties [133]. An interesting application would be, for instance, the study of the Coronavirus (Covid-19) pandemic spread that affected the entire world in 2019 and 2020.

- Topological study of the cancer behavior, studying molecular data to understand the transitional similarity between tumors in different patients; or to understand the progression of the disease. In addition, the study of the responses to different drug therapies and the heterogeneity of responses among different samples can also be studied with topological analysis techniques [107].

- Topological study of the three-dimensional structure of the DNA: topological data analysis seems to fit for the description of the chromatin structure and the loops identification. Persistent homology could be applied to the contact matrices derived from Hi-C data, for example [134].

Other investigations could be also done, as the topological study of the common genetic variants across diseases can be also included in a network where each node is represented by a specific disease related to other one. The topological aspects of those huge interactions are to be studied deeply in order to understand the hidden genomic features that act in common for related illnesses. As mentioned before, there is a long way to walk for TDA.

3.6.1 TDA Applied to Epigenetics

Our proposal is to apply TDA ideas and techniques to unstudied areas as epigenetics, and to create novel analytical tools more flexible and efficient. Particularly, we mainly apply them to high-dimensional data to obtain substantial information about correlation patterns. Firstly, we would need to introduce some basic items in order to start to use TDA for epigenetic data analysis:

- The finite data cloud X that needs to be analyzed. For epigenetics data, we use a 450K Illumina Methylation array, that is a finite collection of around 485,000 CpG sites with their corresponding DNA methylation Beta level in a selected sample size.
- The finite metric space with a given metric (X, d_X). In our case, the metric is based on the Pearson correlation ρ [135] in order to group features with similar correlation coefficients:

$$\forall x, y \in \mathbb{R}^n,\ d_\rho(x, y) = \sqrt{1 - \rho(x, y)^2},$$

- The associated topological space to recover its invariants, that could be defined as a graph $G = (V, E)$ composed by a set of vertices V and a set of edges E. It is defined from the metric space where the vertices $\{V_1, ..., V_n\}$ are the points of X and two points are linked if $d_X(V_i, V_j) \leq \epsilon$, for any $\epsilon > 0$. The vertices V are the CpG sites or clusters of them.
- The linked simplicial complex that can be defined as the clique complex $Cl(G)$ of an undirected graph $G = (V, E)$ where the vertices of G are its vertices and each k-clique, i.e. the complete sub-graphs with k vertices, in G corresponds to a $(k-1)$-simplex in $Cl(G)$.

With this information, we will present our novel analytical contributions in the next part II and part III.

Part II

A Model for the Local Correlation

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 4

Description of the Local Correlation Structure

How is the correlation structure of the DNA methylation? What are their local characteristics within a CpG island? How does it change over the years? The DNA methylation correlation structure will be firstly analyzed from a local perspective to do it afterwards more globally. In this chapter, we will describe the study of the correlation behavior within CpG islands with graph theory and persistent homology, representing the correlation matrices as weighted graphs. We will analyze its non-random spatial structure and the effect of short-range and long-range correlations. All this study will lead to the development of a graph model described in the next chapter.

4.1 Correlation of a CpG Island

As explained in the first chapter, a lot of the current methylation analysis methods are based on the detection of regional patterns assuming that there is a higher correlation for short-range genomic regions. Additionally, we presented in the chapter 2 the presence of long-range inter-chromosomal significant correlation that could be related to the loops found in the chromatin structure, as described in the section 1.3.

We have also presented that the chromatin structure is altered with the aging process. Despite the unknowns, this modification seems to reduce the long-range interactions due to a less-condensed chromatin design, and this could have an impact on the methylation correlation structure over the years.

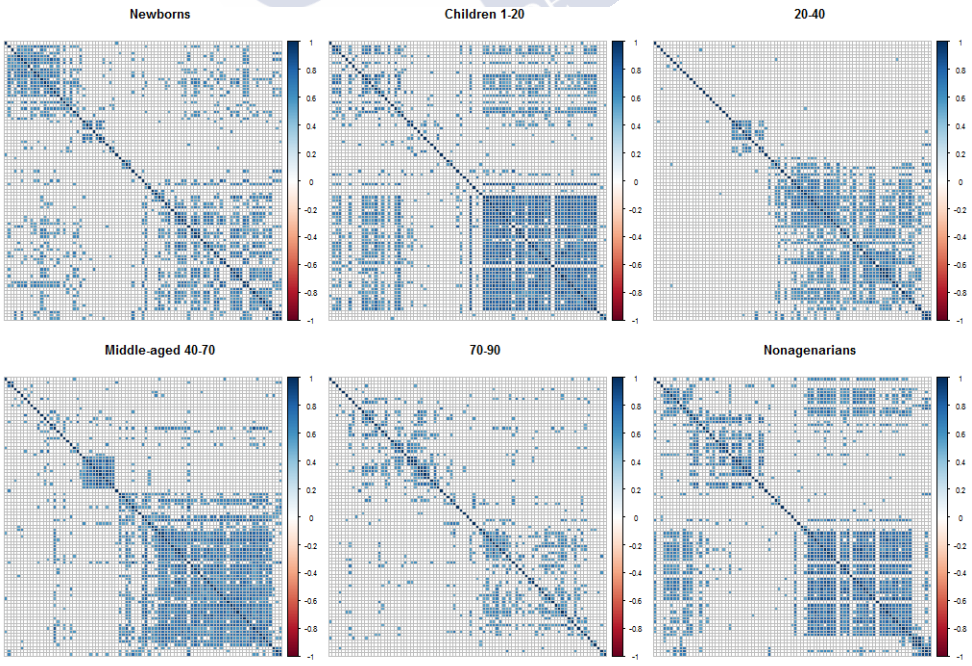
Having those points into account, our hypothesis is that the methylation marks present indeed a structural spatial design and are less coordinated for older ages. Therefore the correlation would decrease with age, especially for the global or long-

range structure. We then aim to describe precisely this correlation design through the study of the related weighted correlation graph, proposing a graph model that accounts for the hypothesis generated based on the study of the CpG islands. Please note then that, in this case, the short-range and medium/long-range positions are all within an island. In addition, the extension of this analysis to other genomic regions will be also presented.

For this analysis, we will not take into account sites related to SNPs and sex chromosomes, as they were linked to a higher methylation variability.

We start with the observation of correlation matrices of a CpG island (ordered by genomic position) for distinct age groups of around 20 individuals, from newborns to nonagenarians: newborns (GSE30870, cord blood leukocytes), children between 1-16 years (GSE36064, peripheral blood mononuclear), middle-age subjects (GSE33233, whole blood, 39-72 years), nonagenarian people (GSE30870, peripheral blood leukocytes), and 20-40, 70-90 samples using GSE40279 of whole blood. The data used is not longitudinal, but different subjects are measured at different ages. In addition, the data sources are different despite all come from blood samples. The first studied island is the biggest on chromosome 7 and it is in the promoter region of genes PEG10 [136] and SGCE.

Figure 4.1: Correlation matrices of the CpG island chr7:94284858-94286527 ordered by genomic position and ascending age groups. Only absolute coefficients greater or equal than 0.5 are presented.



Observing the correlation matrices, presented in figure 4.1, we could already detect a structured spatial behavior similar to the expected, with heavier correlation clusters near to the diagonal (related to short-range positions). Despite we use different datasets, we detect the existence of a structure based on correlation clusters that seems to be altered over the years. Especially interesting is the case of the nonagenarian group, where the distribution of the correlation is more similar to the children group than to the prior age group between 70 and 90 years old. Please note that, in line with literature, the oldest groups have a higher inter-subjects variability, while the younger groups present higher intra-subject variability.

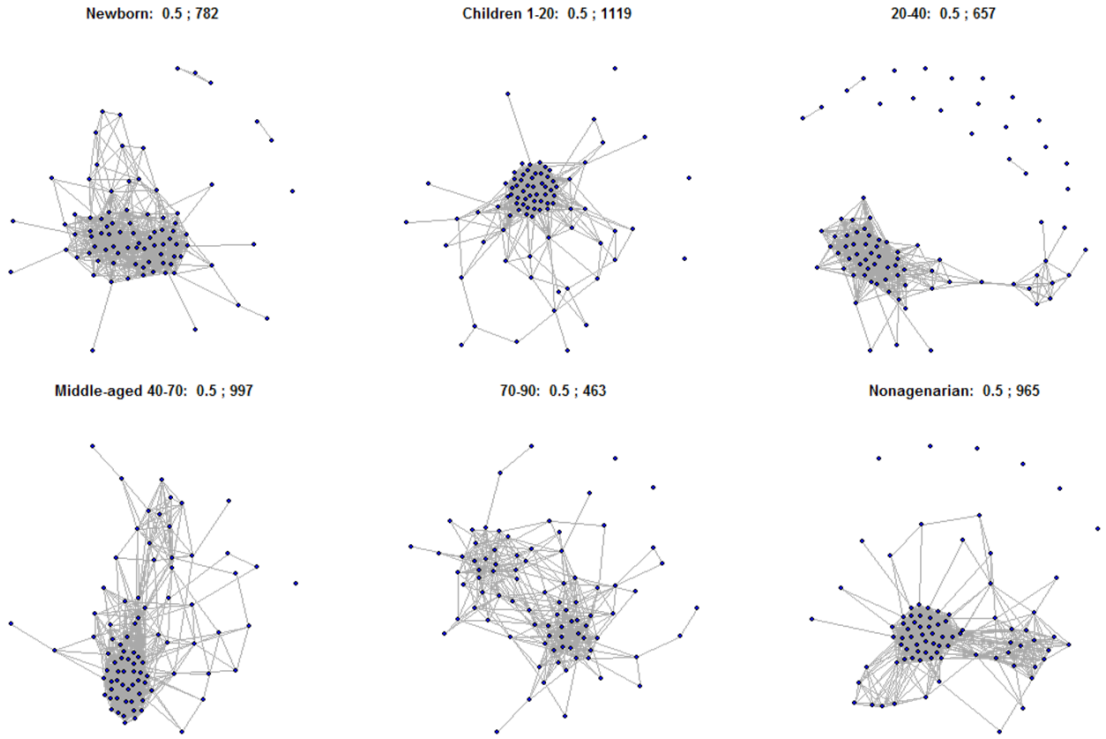
This structural design is repeated for different CpG islands, and it is more clearly seen for those with a higher dimension. Clusters around the diagonal and in long-range positions are present. Generally, the correlation seems to be higher for the children group: selecting the biggest 110 CpG islands (with at least 50 CpGs), we obtain a greater average absolute correlation for children compared to the rest of the groups. If we limit the study to long-range correlation, the children's group is also the winner.

The detection of structural information with the simple observation of the correlation matrix is normally not feasible (for bigger dimensions) or potentially not accurate. We propose then to analyze the matrices by the properties of their associated weighted correlation graphs. We will analyze the characteristics of the graph with classical graph theory, and its topology using persistent homology, to take into account all the information layers. Hence, the observed correlation matrices can also be represented as weighted graphs (see figure 4.2) where each node is a CpG site of the island and they are connected as per their Pearson correlation coefficient in absolute value (in order to take into account negative coefficients).

As we observe at a first step, networks are mainly separated into modules, that correspond to the correlation matrix diagonal clusters, which makes that some CpGs have a higher degree (number of edges of a node). Our analysis will be based on describing and modeling this community design, and other graph properties as the edge-density (ratio of the number of edges with respect to the maximum possible edges) or the degree distribution. Moreover, we will link those properties with the homology study of the networks. This approach will allow us to establish an analytical measure of matrices distinction in a "spatiotemporal" way, translating the topological features into parameters of statistical differentiation.

Please note then that the correlation graphs presented here have a special property: there exists a (genomic) distance between the nodes that has an effect on

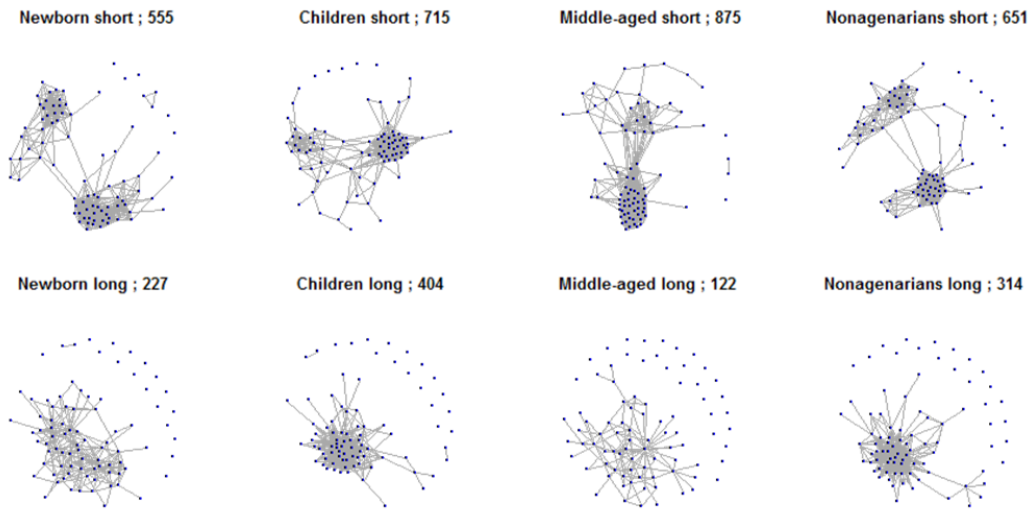
Figure 4.2: Correlation networks of the CpG island chr7:94284858-94286527 by age group with coefficients greater or equal than 0.5. In the title we present the age group, the coefficients filtration threshold, and number of edges of each network.



the own graph characteristics (as modularity). The genomic distance determines a different correlation structure for sites located in near positions and sites located in non-near positions. Indeed, if we separate the sites in short and long distances (below and above the median of the position distance matrix), we observe differences in the topology of the graphs of figure 4.3, that are created for some of the age groups. Additionally, those differences indicate that a strong local or short-range correlation structure may not be accompanied by a similar long-range one (as it happens with the middle-age group).

There are different published models that take into account the modularity design of the networks (as the Stochastic Block Models [137]), but we have not found any that accounts for the distance between the nodes as a feature that influences the topology of the network. Here, we have two different network descriptors: the weighted adjacency matrix that collects the network connections based on nodes correlation, and the distance matrix that collects the genomic distance among the nodes. This spatial design is the key of our study and allows us to generate novel models of random graphs.

Figure 4.3: Some of the related correlation networks separated for short and long-range correlations.



Despite we have a biological motivation for this work, we also aim to present general graph models that suit with other types of spatial data and that could successfully describe their structure. Social networks with a distance component among the individuals may be an example to study population dynamics within and between community structures as families, towns, cities or countries.

4.1.1 Modularity and Clustering Coefficient

Network modularity is an important measure of the structure of graphs that is vital in our work, as the modularity level and the module's singularities of our studied networks will allow us to characterize the graphs and to establish a relationship with results from persistent homology.

As described by Newman in [138], "[...]the modularity is a measure of the extent to which like is connected to like in a network[...]". Modularity measures the level of division of a network into modules (groups, clusters, or communities) through the difference between the actual and the expected number of edges at a random process. Therefore, networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.

It is technically defined as:

Definition 4.1.1 *Let $G = (V, E)$ be a network of n nodes and m edges with adja-*

cency matrix A and suppose we have divided the nodes into C clusters. Then

$$Q = \frac{1}{2m} \sum_{i,j} \left(a_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j),$$

where k_i is the degree of node i and $\delta(C_i, C_j)$ is the Kronecker delta of the community C_i, C_j of vertices i and j.

So the higher Q, always lower than 1, the stronger the community structure of the network. Q takes negative values if there are less edges between vertices of the same type than we would expect by chance. Sometimes, negative values of Q are all considered as zero for implementation purposes. We will obtain networks' modularity using the R function *cluster_fast_greedy*, which works fast for huge networks and detects modules optimizing the modularity score [139].

Biological or social networks, for example, present a high degree of modularity that has substantial importance for the dynamics of the network. Understanding and modeling the community design is crucial to determine any future network behavior. For instance, a closely connected social community will imply a faster rate of transmission of information or a higher rate of disease transmission than a loosely connected community.

Another important network property is *transitivity*. A transitive relation in a network refers to three vertices connected transitively: if vertex u is connected to vertex v, and v is connected to w, then u is also connected to w. So networks showing this property are themselves said to be *transitive*. Perfect transitivity only occurs in networks where each component is a fully connected subgraph or clique, i.e., a subgraph in which all vertices are connected to all others. We can quantify the level of transitivity in a network with the *clustering coefficient*, that was defined in several ways by different authors.

Newman [138] proposed a global way to measure the clustering coefficient, defined as the fraction of paths of length two (edges between two nodes) that are closed (i.e. the triangles formed by three nodes and three edges). That is, we count all paths of length two, and we count how many of them are closed, and we divide the second number by the first to get a clustering coefficient C that lies in the range from zero to one:

$$C = \frac{3t}{|P_2|} = \frac{3|C_3|}{|P_2|}$$

where $t = |C_3|$ is the total number of triangles, and $|P_2|$ is the number of paths of length two in the network. The 3 in the numerator comes from the fact that a triangle consists of three closed triplets (i.e. three nodes connected by three edges). A higher C implies a better transitivity.

A different proposal was done by Watts and Strogatz in 1998 [140], whose work establishes a global coefficient from the average of the local ones. The local clustering coefficient of node i is defined by:

$$C_i = \frac{\text{number of connected triangles including node i}}{\text{total number of possible edges of node i}}$$

If t_i designates the number of triangles attached to node i of degree k_i, then

$$C_i = \frac{t_i}{k_i(k_i-1)/2} = \frac{2t_i}{k_i(k_i-1)}$$

Please note that $0 \leq C_i \leq 1$. Thus, the global average clustering coefficient of the entire network is:

$$\bar{C} = \frac{1}{n}\sum_i C_i$$

Generally, the Watts-Strogatz index quantifies how clustered a network is locally, while the Newman index indicates how clustered the network is as a whole. Indeed, they are normally correlated.

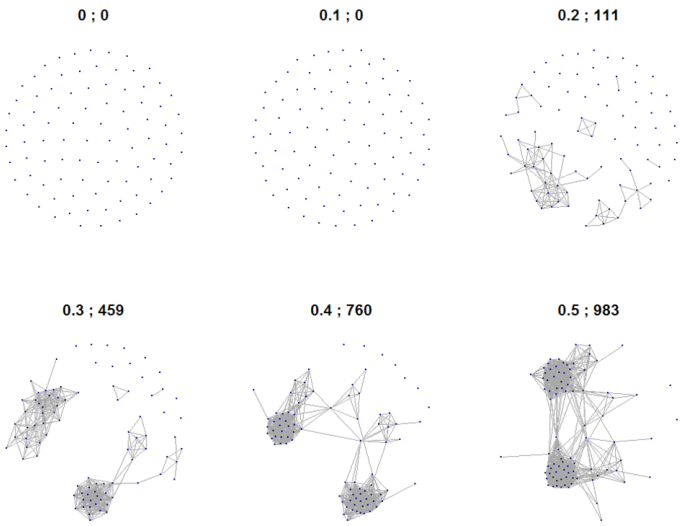
In the next sections we will use the presented concepts to describe the characteristics of our correlation graphs.

4.2 Characteristics of the Correlation Network

Taking into account the design of the correlation matrices observed in the previous section, a main objective would be to analyze the evolution of the heavily short/long-range correlated clusters. In terms of networks, those clusters form network modules with special graph properties. Therefore, if the clusters are maintained during different age periods, then the topology of the related networks should be very similar. If we study those changes from a non-topological perspective, however, we may obtain high differences that do not account for clusters topology and are not informative of the correlation evolution (measuring, for example, the distance between the matrices). For this reason, it is especially important to observe the topology of the correlation networks, more than measure their analytical difference.

Before proceeding with the analysis of the correlation evolution over the years, we will focus on analyzing the structural behavior of the CpG island correlation, studying the associated graph properties and homology results. The modulated design that we mentioned in the first section of this chapter is clearly observed when generating filtered graphs, as the ones in figure 4.4 related to the same island chr7:94284858-94286527. Please note that we removed the CpG site cg10312334 located in this island as it was considered an "outlier" following its genomic coordinates.

Figure 4.4: Filtered correlation graphs of the island chr7:94284858-94286527 associated to the children group. For each graph, we specify the threshold of the filtering and the number of edges.



To generate a system of filtered graphs from a starting correlation graph $G = (V, E)$, we need to subsequently modify the correlation matrix C associated to G that defines the weights of the set of edges E. We define then the altered matrix $\bar{C} = 1 - |C|$, and a set of thresholds $\delta = \{0, 0.1, ..., 1\}$ to be used for the filtration. For each threshold δ, we generate a graph $G_\delta = (V, E_\delta)$ with a set of nodes V and a set of edges $E_\delta \subseteq E$, that derives from the weighted adjacency matrix $\bar{C}$, such tat:

$$\text{if} \quad \bar{C}_{ij} > \delta \quad \text{then} \quad \bar{C}_{ij} = 0 \tag{4.1}$$

Please note that, as we are using the inverse of the correlation, the first filtered graphs correspond to the higher correlation coefficients. The number of edges of figure 4.4 is different from those in figure 4.2 as we are using now all the children sample (around 80 individuals).

In order to study generally the graph properties and compare them with the ones from a random graph, we have designed the following procedure.

4.2.1 Random Test with Persistent Homology

One may think that the modules found in the graph of figure 4.4, and so the island clustered correlation structure that they represent, appear randomly. In order to test this randomness hypothesis, we define a random process simulating a correlation matrix and studying the results of persistent homology. The idea is to simulate a matrix as similar as possible to the original one and see how different are the obtained graphs and persistence barcodes.

Given the correlation matrix C with elements $\rho_{i,j}$ and dimension $n \times n$, we simulate a truncated (from -1 to 1) normal distribution per row given the mean and variance of that row:

Algorithm 2: Random algorithm

Parameters: C original correlation matrix $n \times n$;
initialization;
while *simulation* $s = 1$ **do**
  $\forall i \in \{1, ..., n\}$ rows of C, determine μ_i and σ_i;
  $\forall i \in \{1, ..., n\}$, $\hat{C}_i \sim \mathcal{N}_{(-1,1)}(\mu_i, \sigma_i^2)$;
  Apply PH to $\hat{C}$ and get the Betti numbers for H_1 and H_2: β_{1_1} and β_{1_2};
end
Repeat simulation $s = 2, ..., 100$;
Result: Calculate $mean_s(\beta_{s_1})$ and $mean_s(\beta_{s_2})$ $\forall s \in \{1, ..., 100\}$.

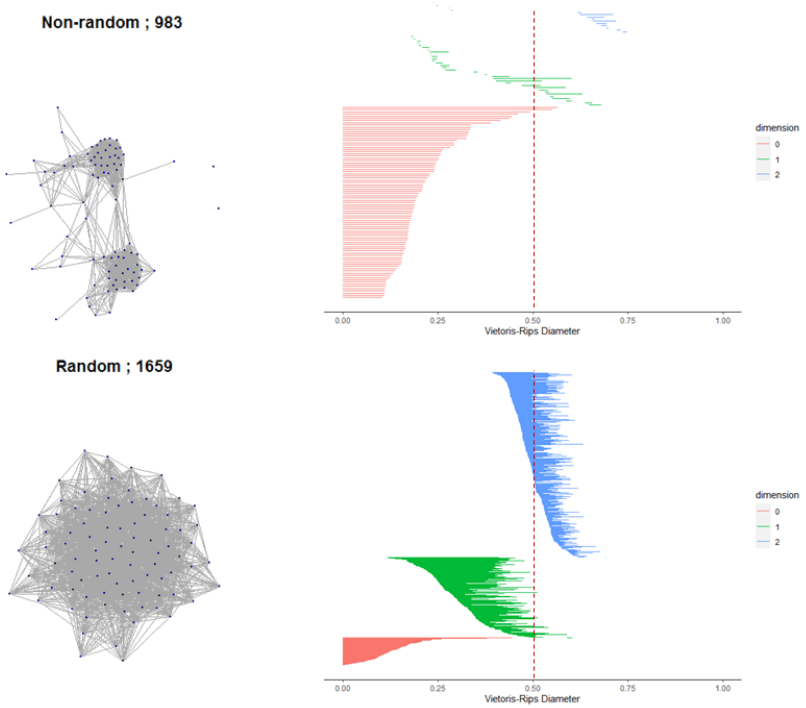
The obtained average Betti numbers are then compared with the ones obtained

from the original matrix C.

In order to test the process on a determined dataset, we select the children population between 1-16 years old of the dataset GSE36064 (peripheral blood mononuclear, around 80 individuals) as representative of the two types of correlations that we may have (among near/non-near genomic positions), and use the CpG island of the chromosome 7 as reference. With it, the original and simulated networks are totally different (please refer to figure 4.5) in terms of classical graph theory characteristics. The random one has a higher edge incidence and density, lower modularity, and lower degree variability.

Following the process described in the algorithm 2, applying a PH analysis with the VR-filtration over the modified correlation matrices ($1-|C|$ and $1-|\hat{C}|$), we obtain indeed substantially different results in terms of Betti numbers up to homology dimension 2 for the non-random and random networks (33 vs. 256 for H_1, and 13 vs. 581 for H_2, respectively). The distribution of the lifespan times is also different (higher for the random graph), with means of 0.03 and 0.1 for non-random and random networks in H_1, and 0.02 vs. 0.05 in H_2. None of the simulated random graphs obtained lower Betti numbers that the original one.

Figure 4.5: Non-random vs. random graphs (at a threshold of 0.5) and persistence barcodes. The red dotted line marks the threshold corresponding to the presented networks.



The significant difference of PH results demonstrates the non-randomness of our initial network, i.e., there are some hidden features ruling the correlation of the CpG island. The reason of this differentiation is based on the graph modularity that is also related, at the same time, with the degree distribution of the network. The fact that the random network has a higher edge density but a lower modularity score tells us that the edge density of the modules is lower than in the non-random network, and so presents a less variable node degree distribution. This fact is strictly related to the higher incidence of Betti numbers in the random graphs, especially for dimensions greater than 0. As explained in [141, 142], clique complexes and random VR-complexes present an interesting Betti numbers distribution. In particular, since the Betti number β_0 of a simplicial complex is equal to the number of connected components, its value is always decreasing with an increasing edge density p of the network submodules. The higher Betti numbers β_j, with $j > 0$, have however an uni-modal distribution with the density p of the network modules. This implies, for instance, that the Betti number β_1 initially grows with increasing values of p but starts to decrease as the density of links p increases. For that reason, the higher modularity and edge-density within the modules in the real network produces a decrease in the associated Betti numbers, as the topological features are "killed" very quickly.

Therefore, we could say that the main differentiated point of the non-random network is its ability to create communities, improving the general modularity of the network and altering the variability of the degree distribution. Those communities arise from the spatial distribution of the CpGs, as we describe below.

4.2.2 Detection of Short-range/Long-range Correlations

Once we have seen that the correlation structure is not random, we analyze which are the factors that influence this design. One important feature is the genomic location of the CpG sites on the island, which determines a relationship between the nodes. The correlation graphs properties are related to the distribution of the highest correlated clusters of CpGs, that can be nearly located or not. We need then to analyze the distinct topology of the correlation structure for short-range and long-range interactions, that are key to characterize the overall correlation design and its evolution with the aging process.

To do it, we could propose several methods but we will focus on the study of short-range and long-range correlations through the definition of different filtration functions that generate distinct persistence diagrams to be compared.

Two Vietoris-Rips Filtration Functions: Short-range and Long-range

To take into account the spatial feature we define a double filtration that takes into account the correlation coefficient and the genomic distance among CpGs. To do it, we define two VR-filtrations, one for short-range positions, and another one for long-range ones. Those two filtrations will be then based on the following two functions W and D.

Let $G = (V, E)$ be the weighted undirected correlation graph with the weight function $W : E \to \mathbb{R}$ and the distance function $D : V \times V \to \mathbb{R}$. Those functions are defined as:

$$W(e_{u,v}) = 1 - |\rho(u, v)|, \, \forall \, u, v \in V, \, e \in E$$
$$D(u, v) = d(u, v) = |u - v|, \, \forall \, u, v \in V,$$

being d the distance matrix of the genomic coordinates of the nodes and ρ the Pearson correlation. Then, we establish the following filtration functions:

Definition 4.2.1 *(Vietoris-Rips-ShortDistance filtration (VRSD)) For any $\delta \in \mathbb{R}$, the 1-skeleton $G_\delta = (V_\delta, E_\delta) \subseteq G$ is defined as the subgraph of G where $V_\delta = V$ and its edge set $E_\delta \subseteq E$ only includes edges that fulfill the following conditions:*

- $W(e_{u,v}) \leq \delta, \quad \forall \, u, v \in V, \, e \in E, \, and$
- $D(u, v) \leq median(d), \quad \forall \, u, v \in V$

Then, for any $\delta \in \mathbb{R}$, we define the Vietoris-Rips-ShortDistance complex as the clique complex of the 1-skeleton G_δ, and the Vietoris-Rips-ShortDistance filtration is then defined as $\{Cl(G_\delta) \to Cl(G_{\delta'})\}_{0 \leq \delta \leq \delta'}$.

Definition 4.2.2 *(Vietoris-Rips-LongDistance filtration (VRLD)) For any $\delta \in \mathbb{R}$, the 1-skeleton $G_\delta = (V_\delta, E_\delta) \subseteq G$ is defined as the subgraph of G where $V_\delta = V$ and its edge set $E_\delta \subseteq E$ only includes edges that fulfill the following conditions:*

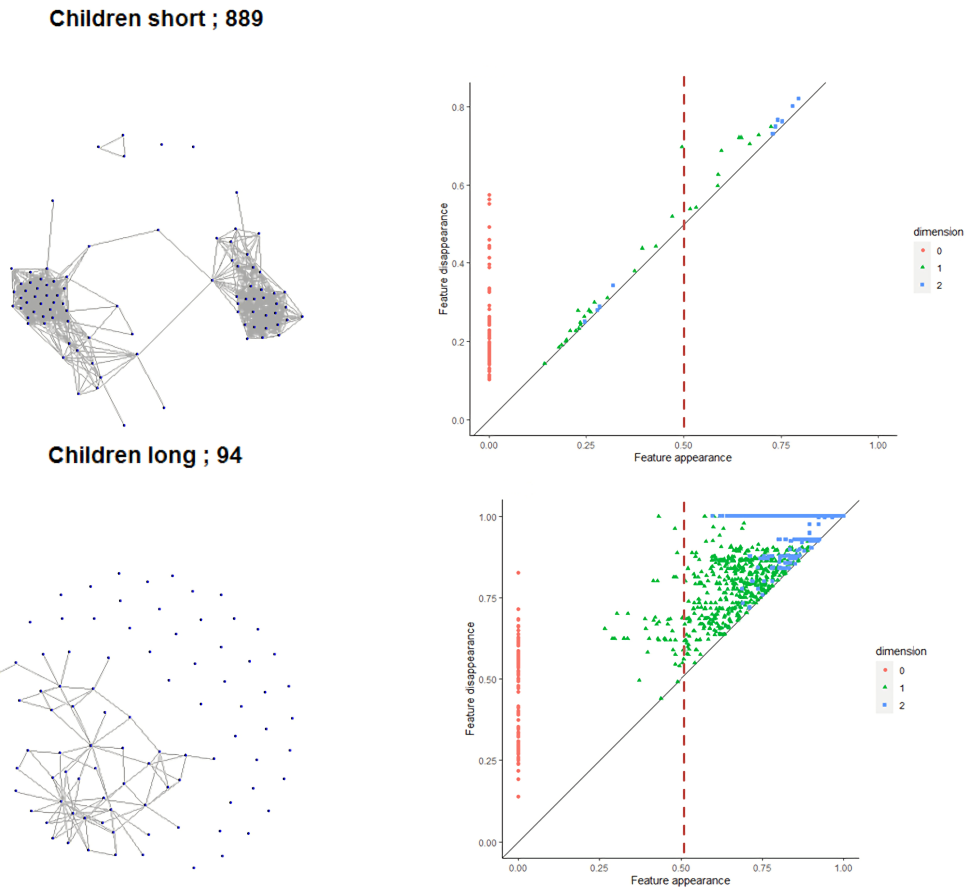
- $W(e_{u,v}) \leq \delta, \quad \forall \, u, v \in V, \, e \in E, \, and$
- $D(u, v) > median(d), \quad \forall \, u, v \in V$

Then, for any $\delta \in \mathbb{R}$, we define the Vietoris-Rips-LongDistance complex as the clique complex of the 1-skeleton G_δ, and the Vietoris-Rips-LongDistance filtration is then defined as $\{Cl(G_\delta) \to Cl(G_{\delta'})\}_{0 \leq \delta \leq \delta'}$.

The resulting networks and PH results from applying both filtrations to the same children correlation matrix of the island in chromosome 7 are very different in terms

of Betti numbers and average lifespan times, as we observe in the persistence diagrams of figure 4.6 below. Short-range graph presents a higher edge-density and modularity, improving the network transitivity. The reason of those differences is the genomic distance itself, as the correlation is higher in nearer regions and the number of potential edges per node changes depending on the node position. Please note that the density of the CpGs may not be equally distributed in distinct regions as per array design. For example, a CpG situated at the beginning or the end of the CpG island (with ordered increasing positions) will have a higher number of potential edges in the long-range network than in the short one. On the other side, a CpG situated in the middle will have a similar number of potential edges in both.

Figure 4.6: Short-range and long-range correlation networks at a threshold of 0.5, and persistence diagrams.

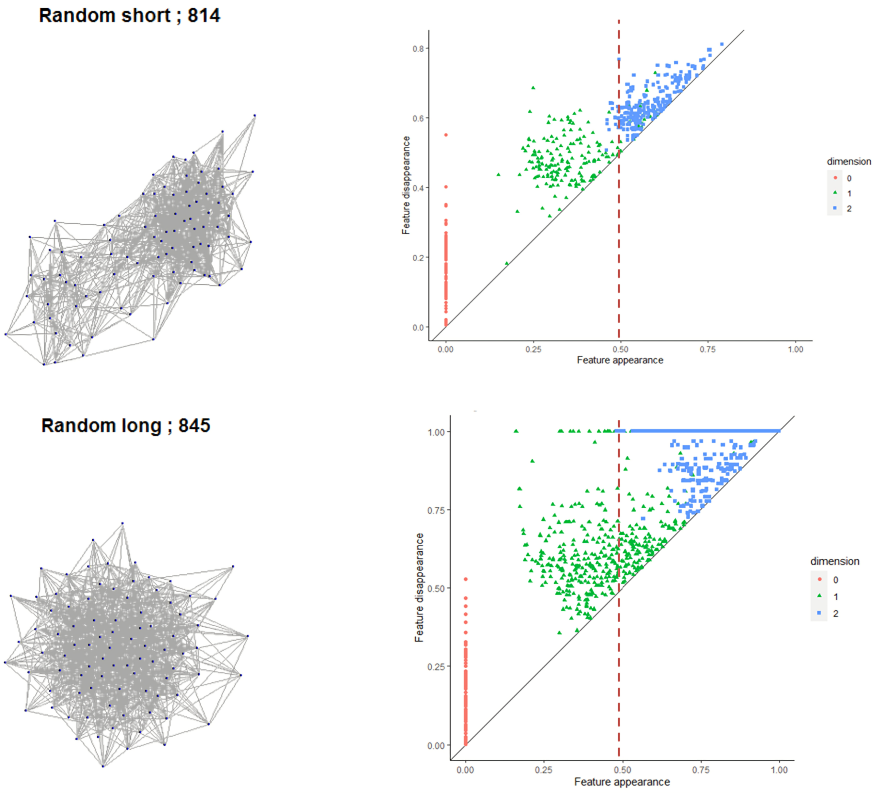


As the design of the short/long networks is completely different, the associated persistent homology results are related to this layout, as we have explained in the previous subsection: the incidence and lifespan of the topological features found in the long-range network are higher than in the short one due to a lower modularity

(0.15 vs. 0.3) and within-modules edge-density. Moreover, the long-range related features born generally after the short-range ones, indicating a lower long-range correlation or a distinct long-range network design that difficult the earlier creation of features.

If we create the short and the long-range networks with the random process described in the algorithm 2, results are indeed different to the ones just presented in figure 4.6, especially for the short-ranges, but they also present some similarities (please see figure 4.7). While the mean absolute values of the random short and long-range matrices is almost the same, the modularity of the short-range duplicates the one from the long-range and their topology is different, demonstrated by the differences between the graphs and the homology results. Similarly to what we observed in the figure 4.6, the short design presents less Betti numbers and a lower average lifespan. The short-range random design presents more differences with the non-random one, indicating its stronger structure.

Figure 4.7: Short-range and long-range random correlation networks at 0.5, and persistence diagrams.



We conclude then that the distance between the nodes characterizes the overall network and its modularity. The study of the topology of the correlation matrix

as a weighted graph gives us more information than the calculation of the matrix mean, and its use increases in power for high-dimensional matrices where the visualization is difficult. Persistent homology detects not only if the correlation is random/non-random or higher/lower, but also if it follows a short or long-range design. Thus, the study of the distance between the obtained PH barcodes would be a way to analyze the differences in the network's structure. Advanced mathematical studies could be developed in this sense as post-investigations to the present work.

After a comprehensive characterization of the local correlation structure of DNA methylation, we will present in the next chapter a model to describe and predict this structure.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 5

A Model for the Local Correlation Structure

We have demonstrated that the observed local correlation structure is not random and has a specific design based on the distance between the nodes. How could this observation be generalized and modeled? We propose in this chapter a novel random network model that is able to reproduce the behavior of the correlation graphs. In addition, we analyze the evolution of the correlation structure over increasing age ranges based on the designed model.

5.1 A Stochastic Block Model with Distance

Once we have determined which are the features that have an influence on the correlation structure of a CpG island and so on the correlation graphs studied, we propose a model for its behavior. We could define it in terms of a stochastic block model (SBM) [137, 143] design, with the novelty of including the distance between nodes as an additional parameter. This parameter will be key to estimate the blocks and to take into account indirectly the correlation structure (i.e., the weights of the graph). We will call it a *stochastic block model with distance* (SBM-D), which is constructed as follows.

Let $G = (V, E)$ be a graph, where V is the node set of dimension n and E is the set of edges. Let K be the number of groups of nodes defined taking into account the distance D between the nodes, i.e., nodes are grouped initially based on their genomic distance on a partition of K consecutive intervals. Each node belongs to one of those groups defining a belonging matrix Z of dimension $n \times K$, such that each row $Z_i = 0$ except exactly once that takes the value $Z_i = 1$ (it represents the

group whose the node belongs).

$$Z = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \dots & \dots & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & 0 & 1 \\ 0 & \dots & 0 & 0 & 1 \end{pmatrix}$$

Group sizes are then derived from Z as the sum of their columns. Please note that the rows of the matrix Z are ordered by their genomic location and the groups of the columns represent the contiguous partitions of their genomic coordinates.

In addition, we could define a degree matrix H of dimension $K \times K$ that can be derived from Z and the adjacency matrix A: H_{ij} represents the number of edges between the groups i and j.

$$H = \begin{pmatrix} \frac{1}{2}\sum_{i,j\in k_1} A_{ij} & \sum_{i\in k_1, j\in k_2} A_{ij} & \dots & \sum_{i\in k_1, j\in k_K} A_{ij} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i\in k_K, j\in k_1} A_{ij} & \sum_{i\in k_K, j\in k_2} A_{ij} & \dots & \frac{1}{2}\sum_{i,j\in k_K} A_{ij} \end{pmatrix}$$

where $k_1, ..., k_K$ represent the K node groups. Please note that the adjacency matrix A is in principle non-weighted (has only values 1 and 0) and symmetric, and the diagonal of A is zero (which corresponds to a non-weighted undirected non-loop graph design).

In order to describe the generation of edges of G according to the groups, a $K \times K$ block matrix, denoted by C, is introduced. Each element of the block matrix C_{ij} represents the probability of occurrence of an edge between a node in group k_i and

a node in group k_j. C will be symmetric in our case as we treat with undirected graphs, and the sum of its rows or columns is not necessarily 1.

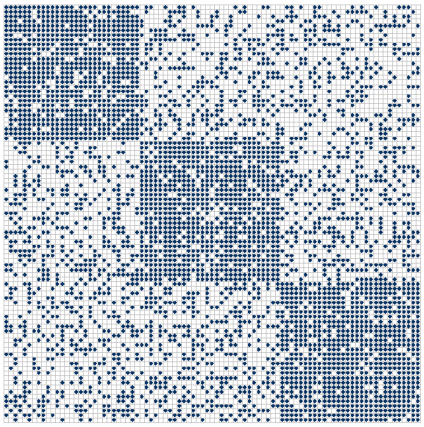
$$C = \begin{pmatrix} p_{k_1,k_1} & \cdots & p_{k_1,k_K} \\ \vdots & \ddots & \vdots \\ p_{k_K,k_1} & \cdots & p_{k_K,k_K} \end{pmatrix}$$

We could then describe the adjacency matrix A_{ij} as a Bernoulli distribution with success probability $Z_i C Z_j^T$: $A_{ij} \sim Be(Z_i C Z_j^T)$, where Z_i is a row of matrix Z. Therefore, the total number of edges between any two blocks k_i and k_j is a Binomial distributed random variable with mean equal to the product of C_{ij} and the number potential edges $n_i n_j/2$, where n_i, n_j are the sizes of the groups k_i and k_j: $H_{ij} \sim B(n_i n_j/2, C_{ij})$.

Then, knowing the composition of the groups determined by the matrix Z and their link probabilities determined by C, we are able to generate an adjacency matrix for our type of graphs. A first example of an easy SBM-D design could be selecting $K = 3$ blocks or clusters with a similar size, where the probability of connecting two nodes in the same cluster is 0.8, and the probability of connecting two nodes in different clusters is 0.2 (presented in figure 5.1):

$$C = \begin{pmatrix} 0.8 & 0.2 & 0.2 \\ 0.2 & 0.8 & 0.2 \\ 0.2 & 0.2 & 0.8 \end{pmatrix}$$

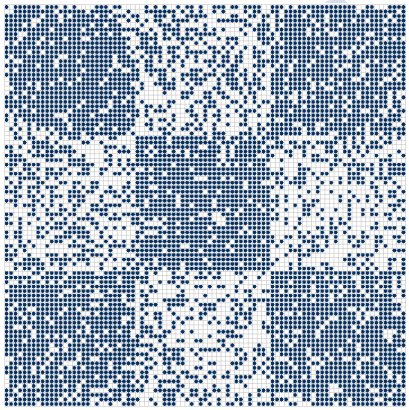
Figure 5.1: Example of the described adjacency matrix generated with a SBM-D.



A different adjacency matrix could be generated specifying in C high probabilities out of the diagonal too, as showed in figure 5.2:

$$C = \begin{pmatrix} 0.8 & 0.2 & 0.8 \\ 0.2 & 0.8 & 0.2 \\ 0.8 & 0.2 & 0.8 \end{pmatrix}$$

Figure 5.2: Second example of the described adjacency matrix generated with a SBM-D.



The stochastic process is then determined by the assumption that the edge probabilities between a pair of nodes would depend on the nodes membership determined by Z.

However, in reality, the matrices Z (so the number of groups K) and C are unknown and we need to infer them. There are many published ways of estimating those parameters. Here, focusing on the correlation structure of CpG islands described in the previous chapter 4, we will use the DNA methylation data itself to do it.

5.1.1 Parameters Inference

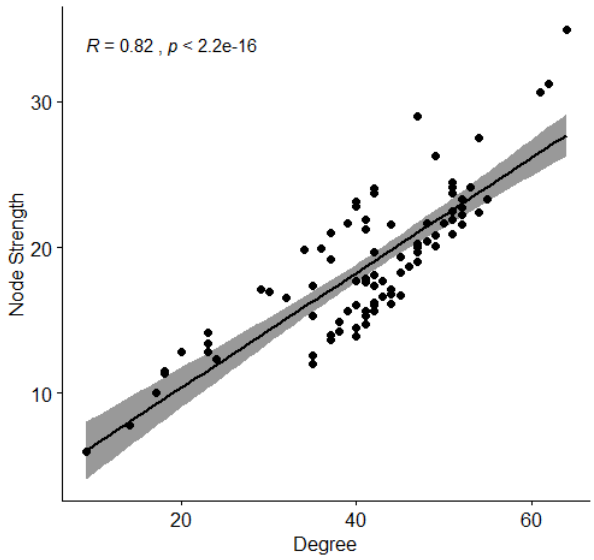
We will adjust the matrices Z and C of the SBM-D model taking into account the studied properties of the correlation graphs generated from the correlation matrix of a CpG island described in the chapter 4. Particularly, the clusters that form the matrix Z will be estimated with a k-means clustering procedure, and the joining probabilities within/between clusters of matrix C will be determined with a sampling process.

Hub Correlated Nodes

In chapter 4, we presented that the modules of the correlation graphs were related to the correlation matrix clusters. Additionally, we saw that those modules/clusters are determined by the distance between the nodes. From these facts, we derive the following rationale to obtain the parameters Z and C.

First, the degree distribution of the studied correlation graphs is highly related to the overall correlation of each node. Applying graph filtration up to a threshold of 0.5 (please refer to the filtration process described in 4.1), the Pearson correlation coefficient between the degree of the networks and the node strength (defined as the sum of each adjacency matrix row) is extremely high around 0.9. If we increase the filtration threshold, i.e., introduce lower levels of correlation in the network, then this relationship gets worse due to the noise included (being the Pearson coefficient still greater than 0.8 for a threshold of 0.7, as presented in figure 5.3).

Figure 5.3: Correlation between the node degree and node strength up to a threshold of 0.7.



The fact that the degree distribution of our networks is greatly related to the node strength, means that it is hugely related to the level of correlation of a node with the rest of the nodes. Consequently, the degree of the nodes is a valid measure to detect highly correlated sites or clusters of sites. We will call *hub nodes* to the ones with higher strength and so a higher degree.

Secondly, the degree distribution is, at the same time, ruled by other variables. Let d be the degree distribution of our correlation graph. We consider X_1, X_2, and

X_3 three variables determining the position of the node i in the island, the mean of the methylation levels for node i and the related variance. If we apply a linear regression model to d with X_1, X_2, and X_3 as independent parameters:

$$d = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

we obtain that the position and the mean methylation are significantly (p-value<0.05) associated with the degree distribution of the correlation networks filtered at low thresholds (i.e., containing highly correlated nodes). Therefore, X_1 and X_2 will be selected as degree estimators in the next process.

Computation of Z and C

We use then the two significant variables X_1 and X_2 (position and mean methylation) to infer the SBM-D parameters with the following process:

1. A k-means cluster analysis with the variables X_1 and X_2. The optimal number of clusters is selected following the Elbow method, which aims to stabilize the intra-cluster variation, i.e., the within-cluster sum of squares over a variety of potential clusters:

   $$\sum_{k=1}^{K} \sum_{i \in n_k} z_{ik} (x_i - \bar{x}_k)^2$$

   where n_k is the set of observations in the kth cluster, $\bar{x}_k$ is the center for the kth cluster, and z_{ik} is an indicator function that takes the value 1 if x_i is in the cluster k and takes the value 0 if x_i is not in the cluster k. The optimal cluster K is selected such that the sum of squares is stabilized in a low value. If needed, please refer to the subsection 6.2.1 for a detailed description of the clustering technique.

   Please note that the k clusters will contain contiguous CpG sites because the genomic position variable X_1 has a big weight on their distribution. If it would not be the case, then we could consider as different groups those that contain non-contiguous CpG sites.

2. Over each differentiated cluster k, we define adjacency matrices A^k whose values are generated following a Bernoulli distribution with probability p_{int_k} (i.e. a Binomial distribution over one experiment):

   $$A^k_{ij} \sim B(1, p_{int_k}), \quad \forall\, k = 1, \ldots, K,\ i, j = 1, ..., n_k$$

   We will explain afterwards how we could calculate p_{int_k}.

3. The diagonal of the overall adjacency matrix A is therefore formed from all A^k:

   $$A = A^1 \oplus ... \oplus A^K$$

4. Two sites belonging to different clusters k_l, k_m are joined following a Bernoulli distribution with probability $p_{ext_{k_l k_m}}$ (i.e. a Binomial distribution over one experiment):

$$A_{ij}^{k_l k_m} \sim B(1, p_{ext_{k_l k_m}}), \quad \forall\, l, m = 1, ..., K,\ i = 1, ..., n_{k_l},\ j = 1, ..., n_{k_m}$$

We will explain below how we could calculate $p_{ext_{k_i k_j}}$.

The probabilities p_{int} and p_{ext} are expected to met $p_{int} \geq p_{ext}$, following our distance-based modulated design. Those two values could be fixed, as $p_{int} = 0.8$ and $p_{ext} = 0.2$ for example, or could be estimated with a sampling process:

- For each k cluster, we select randomly a couple of CpGs v_1, v_2 and calculate their correlation coefficient $\rho(v_1, v_2)$. We repeat the process N times and obtain the cluster probability p_{int_k} as:

$$p_{int_k} = \frac{\#(|\rho(v_1, v_2)| > \delta_{int})}{N}$$

The threshold δ_{int} could be adjusted as desired, as $\delta_{int} = 0.5$, depending on the level of exigency that we set for the correlation. The number of simulations N could also be adapted depending on the adjacency matrix size. The necessary sample size to estimate successfully the population mean would be:

$$N = z^2 \frac{\sigma^2}{\alpha^2}$$

where z corresponds to the z-score for a confidence level of 95% ($z = 1.96$), $\alpha = 0.05$ is the error margin, and σ is the standard deviation of the population (calculated as the mean of the standard deviation of each row of the correlation matrix). For example, for the case of the island on the chromosome 7, we obtain $N = 142.9$, that we round to $N = 200$. The used N allows to be 95% sure that our estimated p_{int} is accurate with a margin error of 5%.

- For each couple of clusters k_i, k_j, we select randomly one CpG per cluster, $v \in k_i$ and $w \in k_j$, and calculate their correlation coefficient. We repeat the process N times and obtain the probability $p_{ext_{k_i k_j}}$ as:

$$p_{ext_{k_i k_j}} = \frac{\#(|\rho(v, w)| > \delta_{ext})}{N}, \quad \forall\, i, j = 1, ..., K$$

The threshold δ_{ext} could be adjusted as desired, as $\delta_{ext} = 0.5$ or lower.

Please note that this sampling process was generally established for the case of dealing with huge correlation matrices, where a sampling process is quicker than a exact calculation. However, for the cases of our CpG islands, we could directly calculate p_{int_k} and $p_{ext_{k_i k_j}}$ probabilities over the calculated correlation matrix.

With this process, that is fairly quick to compute, we obtain a different internal probability for each cluster, and a different external one for each pair of clusters. So the block matrix would be (for K clusters):

$$C = \begin{pmatrix} p_{int_{k_1}} & \cdots & p_{ext_{k_1 k_K}} \\ \vdots & \ddots & \vdots \\ p_{ext_{k_K k_1}} & \cdots & p_{int_{k_K}} \end{pmatrix}$$

The resulting adjacency matrix represents a non-weighted network with a degree distribution similar to the original one. If the matrix Z is generated randomly, then the obtained adjacency matrix is not representative of the original one, as it loses the design in groups.

The original and simulated adjacency matrices for the island chr7:94284858-94286527 are really similar reproducing the diagonal highly correlated blocks that we are interested in, as observed in figure 5.4. Indeed, the hub nodes, or nodes with higher degrees (greater than the mean of the degrees), match in both in about an 80%. Generally, selecting the 10 biggest CpG islands, the average percentage of matching is about 70%, as reflected in figure 5.5. Results are reproduced in other datasets, as using GSE40279 with more than 600 individuals.

Figure 5.4: Original adjacency matrix of island chr7:94284858-94286527(with coefficients with an absolute value greater or equal than 0.5) at the left and the related modeled adjacency matrix at right, both ordered by genomic position.

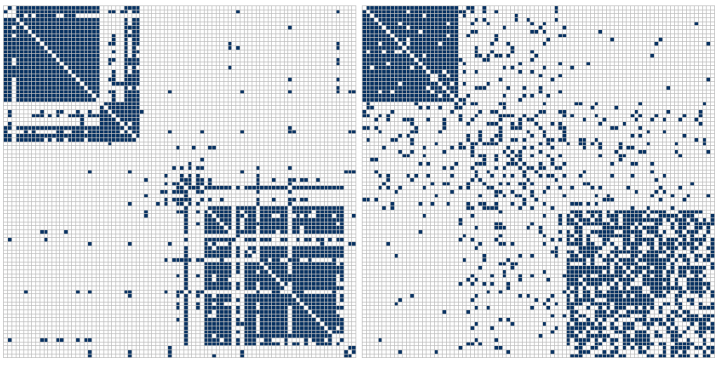
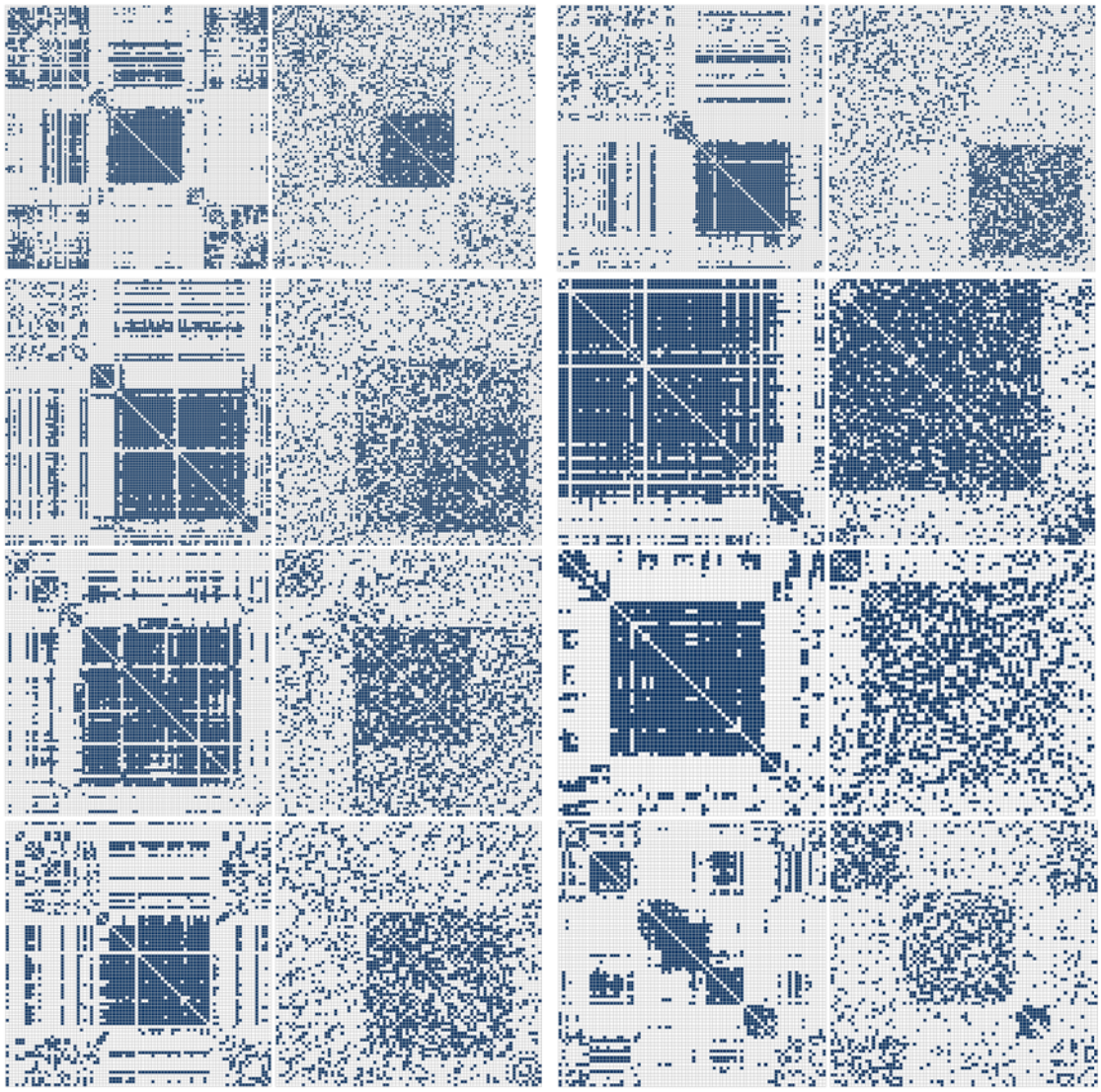


Figure 5.5: Original adjacency matrices of other of the biggest CpG islands (with coefficients with an absolute value greater or equal than 0.5) at the left and the related modeled adjacency matrices at right, both ordered by genomic position.



5.2 Application of the SBM-D Model beyond CpG Islands

The study done within CpG islands could be extended to other genomic regions, aiming to detect as well highly correlated CpG sites or clusters of CpG sites (regions). We present here the results of applying the described model to a chromosome and a couple of gene families.

5.2.1 Intra-chromosomal Interactions

We could try to identify with our model interacting zones within the same chromosome, as the chromosome 22. We select all the CpG islands contained in that chromosome and compute the matrices Z and C with the algorithm described in the previous section. Using $\delta_{int} = \delta_{ext} = 0.5$, we identify highly locally correlated sites related to different CpG islands that belong to clusters with probabilities $p_{int} > 0.5$.

A lot of those sites are related to genes linked to the 22q11.2 deletion syndrome [144, 145]. Genes as TBX1, CDC45, RANBP1, ZDHHC8, DGCR6L, DGCR8, SNAP29, or CLDN5 are associated with this genetic alteration, that seems to have an epigenetic related modification too. The deletion of a small piece of chromosome 22 may cause problems in children, including developmental delays or intellectual disability. Despite the epigenetic alteration related to this deletion is not completely known, an alteration of the coordination/correlation among the detected CpG sites could be an insight for future research. The sites found have overall a great mean absolute correlation of ~ 0.7.

If we apply a similar process to other chromosomes, as the chromosome 17, we indeed find a significant number of genes related to the 17q12 deletion syndrome. This analysis could be then extended to multiple chromosomes or genomic regions of interest, as an automatic and rapid way of detecting highly correlated sites.

5.2.2 Inter-chromosomal Interactions

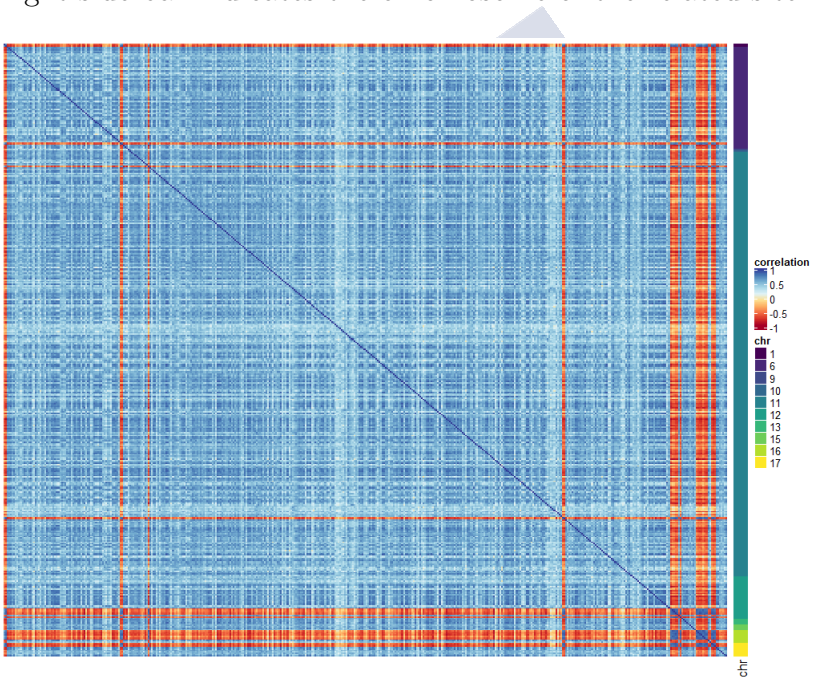
Additionally, we could think about detecting interacting CpG sites that belong to different chromosomes. For instance, we could apply our model to CpG sites related to the same gene family.

The olfactory receptor (OR) gene family has a special interest for us, as we knew it is greatly variable and correlated. Being one of the biggest gene families of the genome, the CpG sites associated to the OR family are located in different chromosomes. Multiple genes of this network were associated with many different diseases,

as we will also see in the part III of the present work.

With more than 1,000 related genes, the interactions of the OR gene family are not easy to visualize within a correlation matrix, and the detection of correlated clusters should be done analytically. Generally, the sites related to this gene family present a high variance on their methylation levels and also a high correlation. The application of our method to the CpG sites associated with those genes allows us to detect the highest correlated CpG sites among this big gene network. Those sites belong to different chromosomes, above all to chromosome 11, 6, and 12, as observed in the correlation heatmap below.

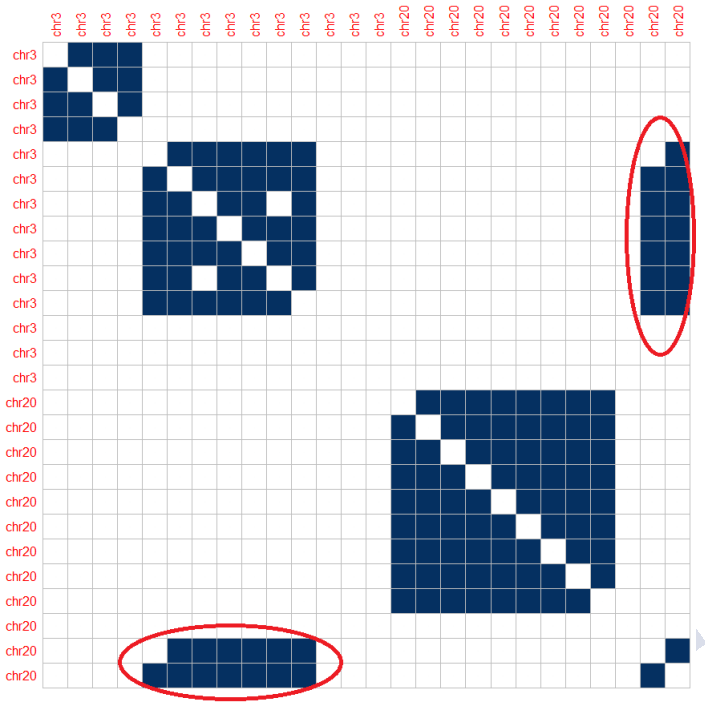
Figure 5.6: Correlation matrix ordered by genomic position for the OR-related sites found. The right-side bar indicates the chromosome of the related site.



We also applied the algorithm to other associated genes, as the OXT gene and its related OXTR gene (located in chromosomes 20 and 3, respectively). The model detects successfully the interactions between intra and inter-chromosomal zones observed in the original adjacency matrix of figure 5.7. Those two genes were indeed related in literature (as stated in the NCBI web page), sharing multiple functionalities, belonging to the same biological pathway "Oxytocin signaling pathway", and being commonly related to several disorders [146].

Similarly, we could apply the model to study interactions between CpG islands in the same or distinct chromosomes using the appropriate parameters.

Figure 5.7: Adjacency matrix ordered by genomic position for OXTR, OXT related sites in chromosomes 3 and 20 respectively. Only correlation coefficients greater than 0.5 in absolute value are taken into account.



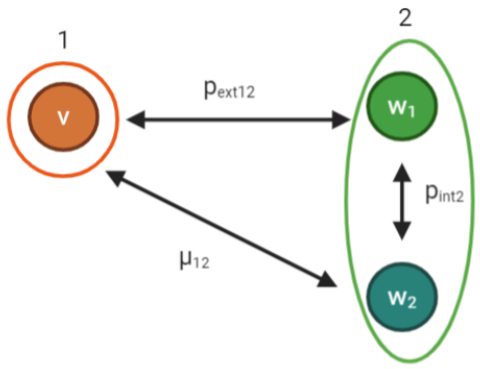
So, despite CpG islands present a very accentuated spatial correlation design, there are other genome zones that could also present highly correlated clusters. CpGs related to the same gene family but located in different chromosomes, or different genes located in the same chromosome, are candidates for this design. The study of the associated networks and the prediction of their interactions with our proposed SBM-D provides with an alternative and fast way to detect those hidden correlations and further investigate their biological meaning.

5.3 Reduction of the Long-range Noise

The proposed previous model generates successfully the diagonal highly correlated clusters. However, it presents a spread distribution for the non-diagonal or long-range positions. This noise present for long-range positions could be better adjusted with lower levels of δ_{ext} or with a specific probability model for those areas taking the distance into account. For example, if we have a node $v \in k_i$ and a set of nodes $w_1, ..., w_l \in k_j$ nearly located, we could join v and w_1 with probability $p_{ext_{k_i k_j}}$ but augment the probability (with a new parameter $\mu_{k_i k_j}$) of having edges among the rest of the nodes $(v, w_2), ..., (v, w_l)$ due to the fact that $w_1, ..., w_l$ are nearly located.

This update would introduce then the new parameter $\mu_{k_i k_j}$ and would take into account the "transitivity rule": if v and w_1 are connected by an edge, and w_1 and w_2 are connected by an edge, then it is likely that v and w_2 are connected by an edge. In our case, the last likelihood increases with a lower distance between w_1 and w_2.

Figure 5.8: Illustration of the new model's idea.



In practical terms, this would be equivalent to define a stochastic process where the time is represented by the genomic distance and the likelihood of joining two nodes belonging to different clusters k_i, k_j, would depend on the probabilities $p_{ext_{k_i k_j}}$ (that join v with w_i) and $p_{int_{k_j}}$ (that join w_i with w_j). The new parameter is defined then as: $\mu_{k_i k_j} = mean(p_{ext_{k_i k_j}}, p_{int_{k_j}})$.

The update with respect to our previous model is then related to the p_{ext} probabilities. We will modify now the generated adjacency matrix A to create $\hat{A}$. In principle $\hat{A} = A$, and we define a genomic distance matrix d_{k_j} between the nodes in the group k_j. We modify the inter-cluster values of $\hat{A}$ as follows:

$$\text{if } \hat{A}_{ij} = 1 \text{ and } d_{k_j}(j, s) < median(d_{k_j}) \text{ then } \hat{A}_{is} = max(A_{is}, B(1, \mu_{k_i k_j})),$$

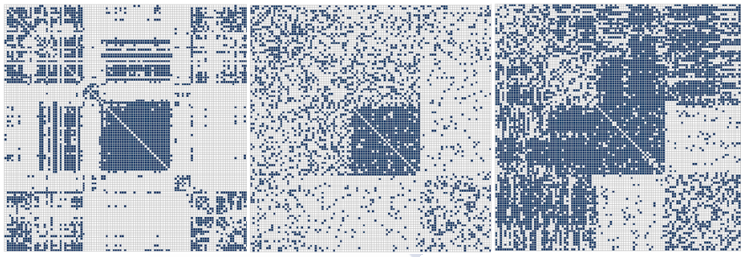
$\forall\, i \in k_i$ and $j, s \in k_j$ being $s \neq j$, and $\mu_{k_i k_j} = mean(p_{ext_{k_i k_j}}, p_{int_{k_j}})$.

We are then altering the environment of j, increasing the probability that nearly located nodes in k_j are connected with starting nodes in k_i. With other words, as normally $p_{ext} < p_{int}$, what we are doing with this modification is to augment the probability that a node is connected with contiguous ending nodes.

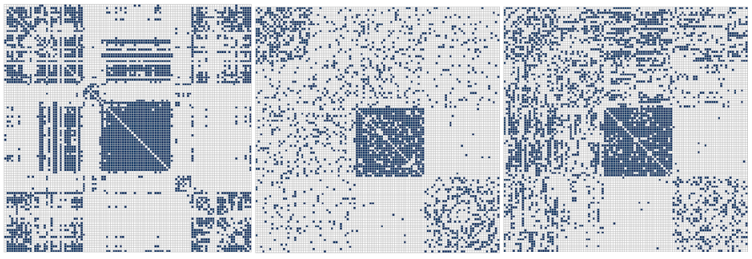
Indeed, the adjacency matrix of one of the biggest islands located in the chromosome 6, that presents high local (short-range) and global (long-range) correlation, is better predicted with this second proposed model that reduces the noise in the long-range zones. As we have defined different thresholds δ_{int} and δ_{ext} for the internal and external probabilities, we could also adjust those to obtain the most accurate representation. The computational time is similar than the initial model, it augments slightly with a high number of CpG sites.

Figure 5.9: Adjacency matrix of the island chr6:31830299-31830948 at left (coefficients with an absolute value greater or equal than 0.5), modeled matrix A in the center, and the related adjusted matrix $\hat{A}$ at right, all ordered by genomic position.

(a) Using $\delta_{int} = \delta_{ext} = 0.5$



(b) Using $\delta_{int} = 0.5$, $\delta_{ext} = 0.6$



Concluding, with the original SBM-D and this improved version, we are able to reduce the high dimension of a correlation matrix to the estimation of just three parameters: K, p_{int}, and p_{ext}. This model provides then with an alternative fast approach to identify highly correlated sites within CpG islands and other genomic regions.

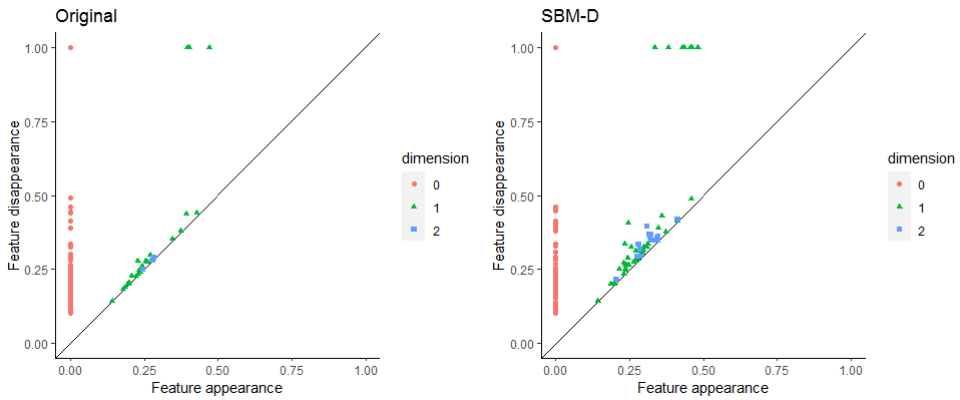
5.4 Comparison with Persistent Homology

More than studying the similarities among the original and estimated graph properties, as the degrees of each node in the original and the estimated adjacency matrix, we are interested in comparing the topological properties related to the structural correlation. Those could be studied from the persistent homology results of the original network and the estimated one and also comparing our random SBM-D model with other known models.

5.4.1 Comparison with the Original Network

With the proposed SBM-D, we obtain non-weighted adjacency matrices. To compare the original network with the generated one, we need to assign edge weights to the adjacency matrix A. As done with the original network, we assign them following the modified Pearson correlation coefficients $(1 - |\rho|)$ between each pair of connected CpGs. A VR-filtration over the weighted adjacency matrix is then applied to obtain the persistence diagrams of figure 5.10.

Figure 5.10: Persistence diagrams of original and estimated networks of the chr7:94284858-94286527 island restricted to coefficients with an absolute value greater or equal than 0.5, using $\delta_{int} = \delta_{ext} = 0.5$.



The obtained persistence diagrams are indeed really similar. Both networks present similar topological elements in dimension 1 than 2, being slightly higher in the modeled network. Despite the edge-density of both is almost the same, the modularity is a little bit greater for the original one, what is increasing the edge-density in the modules and makes difficult the creation of topological elements in H_1 and H_2.

Please note that we compare the SBM-D with the original network at weight 0.5 because the SBM-D was designed to represent the higher correlation clusters of the matrix, avoiding the lower correlation values in order to reduce the non-substantial

information.

To demonstrate that the Wasserstein distance among the two above diagrams P_1 (of the original network) and P_2 (of the modeled network) is not statistically significant, we develop a permutation test over a null hypothesis of a null Wasserstein distance between the diagrams ($H_0 : d_W(P_1, P_2) = 0$). This test permutes the values of the original adjacency matrix several times (100) and calculates the distance among the obtained barcodes, as specified in the algorithm 3 below.

Algorithm 3: Permutation algorithm

Parameters: A_o original weighted adjacency matrix, A_s estimated weighted adjacency matrix from SBM-D, d the Wasserstein distance among the two related persistence diagrams;
initialization;
while *simulation* $s = 1$ **do**
  A_{o_1} is the permuted correlation matrix from A_o after permuting each row values once;
  Apply persistent homology to A_{o_1} and A_s and calculate the related barcodes $P_{A_{o_1}}$ and P_{A_s};
  Calculate $d_1 = d_W(P_{A_{o_1}}, P_{A_s})$;
end
Repeat simulation $s = 2, ..., 100$;
Result: Calculate the p-value of the test as $p = \frac{\#(d_s > d)}{100}, \ \forall s = 1, ..., 100$
if $p < 0.05$ **then**
  We reject the null hypothesis of null distance;
else
  We cannot reject the null hypothesis of null distance;
end

All distances obtained permuting the values of the original matrix were higher than the initial one so the null hypothesis of low distances is not rejected, which indicates that the two persistence diagrams are not significantly different.

If we simulate various SBM-D networks using the same parameters and over the same island, we would obtain very similar Wasserstein distances over the generated networks and the original one. Indeed, those distances would be much more similar using the adjusted model presented in the previous section. This indicates that our SBM-D is stable and robust enough.

5.4.2 Comparison with Other Random Models

The proposed SBM-D generates a random graph that depends on the distance between the nodes and the parameters K, p_{int}, and p_{ext}. There are other models to generate random graphs, despite we have not found any that could be used successfully to reproduce the properties of our correlation graphs.

A famous random model is, for instance, the Erdös–Rényi one (1959) [147]. The Erdös–Rényi (ER) random network is written as $G_{ER}(n, p)$, where p is the probability of adding a link between a pair of nodes (a fixed parameter). The degree distribution of an ER network follows a Poisson distribution and the clustering coefficient is small. Please note that the SBM-D design with a fixed probability $p_{int} = p_{ext} = p$ of joining nodes becomes an ER network.

Alternatively, we find the Barabási-Albert (BA) model (1999) [148], where the probability of finding a node degree k decays as a power law of the degree. This means, the probability of attaching node u to node v is proportional to the degree of v. Then, the principal characteristic of this type of random networks is the power-law degree distribution.

Ultimately, we may talk about the Watts-Strogatz (WS) model (1998) [140], that aim to augment the clustering coefficient of the ER model. First, one form the circulant networks with n nodes connected to k neighbors. Then, we rewire some of its links: each of the original links has a probability p of having one of its end points moved to a new randomly chosen node. If p is high, it approaches the ER model.

Our correlation networks could not be efficiently computed with any of those three models. The reason is mainly the difference in the degree distributions and the modularity. BA and WS models present a higher modularity index, being higher for WS, but the degree follows a power-law distribution (where only some vertices have a high degree) or a Dirac delta function (where all nodes tend to have the same degree), respectively. What we generate, however, is a graph with a high modularity and a normal degree distribution where most of the nodes present a high degree (sometimes bi-modal, depending on the modules size and the long-range design).

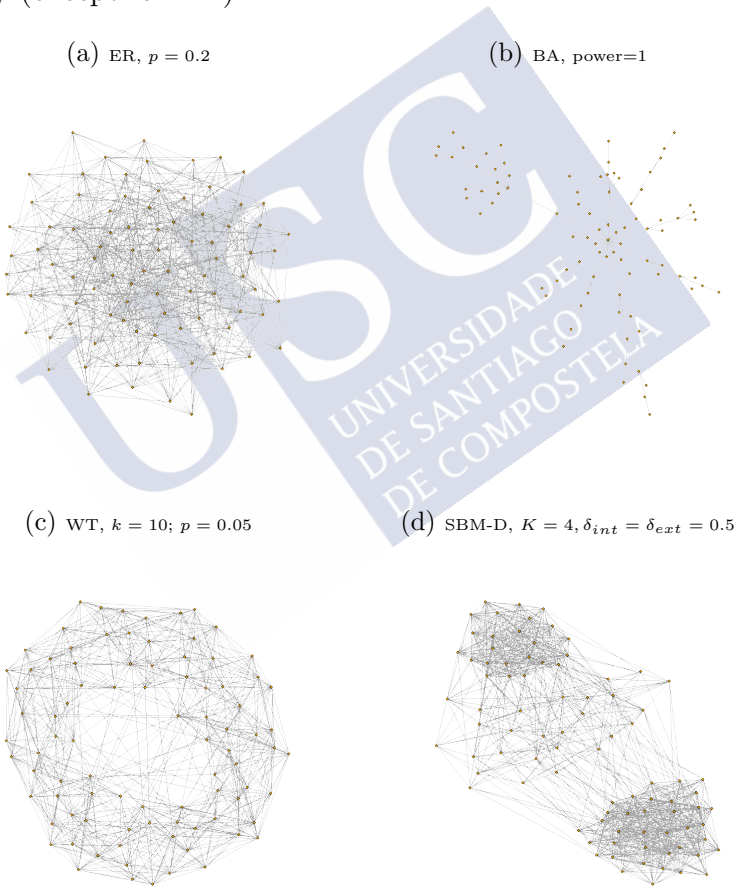
The main difference between the WS model and our SBM-D is that the probability of joining nodes from the same neighborhood is always the same, as well as the probability p of connecting new nodes. That makes WS a worse predictor of hub CpG sites or clusters of sites as per their correlation.

Additionally, the fact that our nodes are ruled by a distance that determines the

creation of the modules is something new that was not taken into account by any of those three designs and we have not explicitly found in literature.

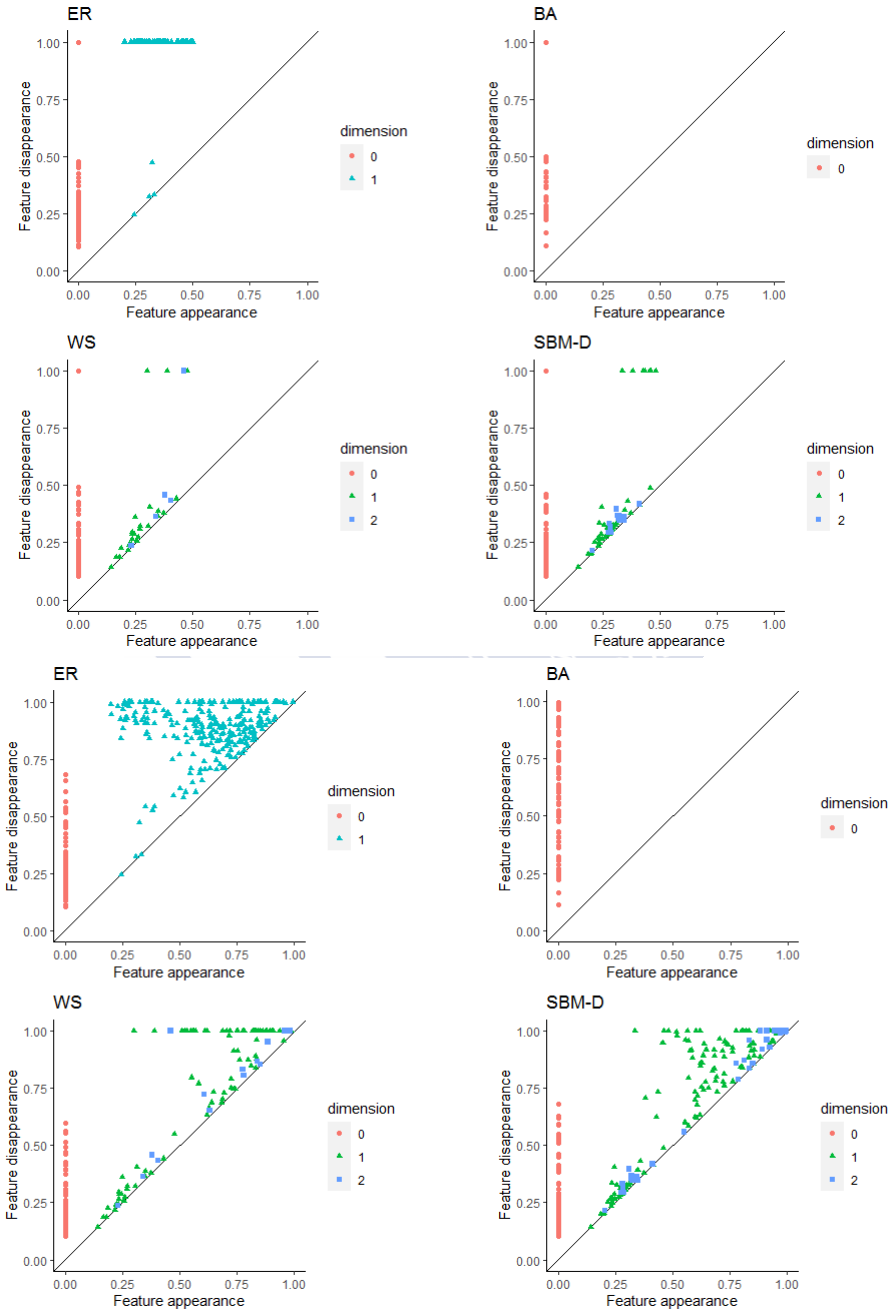
We generate the four graphs with similar edge densities in order to compare their homology results, see figure 5.11. We apply persistent homology to each network to measure and represent their topological differences. Assuming that the nodes represent CpG sites, we assign edge weights based on the Pearson correlation coefficient among two nodes and apply VR-filtration over the modified adjacency matrix $(1 - |A|)$.

Figure 5.11: Generated networks over the island chr7:94284858-94286527 with similar edge density (except for BA).



The graphs generated present some topological differences, as indicated by the differences on the PH results for the four graphs in figure 5.12. Despite the graphs and therefore the persistent diagrams may change with several runs of the random models, we could detect some comparative parameters.

Figure 5.12: Persistence diagrams of the four respective networks. The first four diagrams are restricted to network weights greater or equal than 0.5.



As expected, the BA graph does not generate any feature in dimension greater than zero, resulting from its low edge-density that is unable to create hole structures. The ER persistent diagram has basically elements of dimension 1 (holes) that increase in incidence once we connect more nodes. The lifespan of the features is generally high, thanks mainly to its low modularity.

The most similar ones are WS and our modeled network SBM-D, as both follow a more modulated design. Indeed, the Wasserstein distance between both persistence diagrams is lower than comparing SBM-D with ER. However, they also present differences: the WS network has fewer elements in both dimensions, specially in dimension H_2. Despite both present similar edge-density, transitivity, and modularity values, the degree distribution of the WS network is much less variable, which may difficult the formation of complex elements and destroy the ones that may be formed between modules. The lifespan of the features is generally shorter than ER, despite there are some of them that present a larger lifetime related to less connected subgraphs.

The appearance of the topological features for those two networks occurs earlier than in the ER graph, indicating that ER needs a higher edge-density to complete topological structures with at least four points (as holes).

WS is less sensitive to noise or perturbations than the SBM-D design. The Betti numbers obtained from the application of PH to networks with an increasing edge-density are indeed more altered for the SBM-D model. This derives from a less random design of SBM-D, that increases the edge-density in already densely connected zones.

5.5 Evolution of SBM-D by Age Groups

The evolution of the networks for the different age groups could be measured or modeled in distinct ways [149], and we do it through the SBM-D model.

The age groups presented in section 4.1 belong to different datasets and are not longitudinal. This decreases the power to detect truly modifications of the correlation structure over the years, as the changes may be caused by the different data characteristics. Still, we detect common structural correlation clusters. If we would apply PH to differentiate their network's topology, we would find similar results with lower Betti numbers for the groups with a higher module edge-density (the older ones), as expected.

It is challenging to find a public methylation dataset with such a large age range. Therefore, we restrict the analysis to the dataset GSE40279 dividing the sample in individuals of 20-50 and above 80 years old (with around 100 subjects per group). The absolute mean correlation is higher in the older group.

Taking the island of the chromosome 7, the proposed adjusted model described in section section 5.3 (with $\delta_{int} = \delta_{ext} = 0.3$) predicts successfully the correlation structure (between the 70-80% of hub nodes are detected). Indeed, the parameters generated by the model are different for both age groups, as indicated in the table 5.1 below, indicating a distinct correlation behavior. The number of optimal clusters is the same for both, but the average intra-cluster and inter-cluster probabilities of joining nodes indicate a heavier correlation in the oldest group for short-range and long-range positions.

Table 5.1: SBM-D parameters by age group.

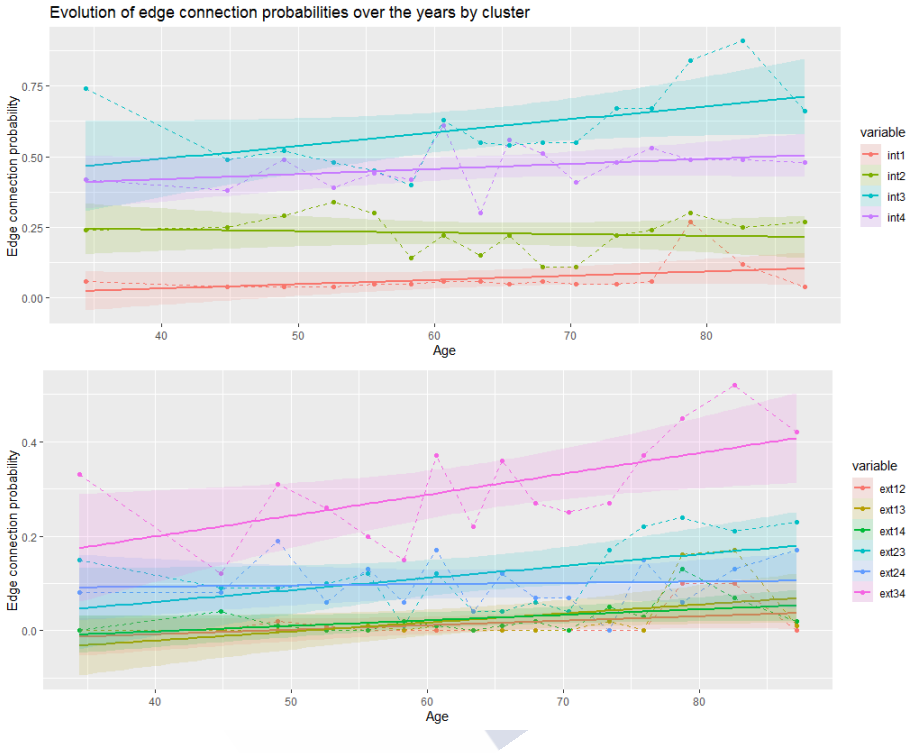
Sample group	K	Average p_{int}	Average p_{ext}
Young: < 50 years (average 41)	4	0.32	0.07
Old: > 80 years (average 86)	4	0.41	0.15

The study above using two age groups can be extended using a wider age range. We take the entire sample of 656 individuals ordered by increasing age and study the model parameters over overlapping increasing windows of 50 individuals (with an overlap of 10 individuals and $\delta_{int} = \delta_{ext} = 0.5$). The optimal number of clusters is always 4 and the results are in line with the ones presented in the previous table, indicating that the probability of joining two nodes is greater for nodes belonging to the same cluster and presents an increasing trend over the years.

However, the evolution picture may be biased by the short age-range (where most

of the individuals have between 40 and 80 years old), as two main groups are missing: children and elderly people.

Figure 5.13: Evolution of edge connection probabilities over the average window age with the related fitted regression lines. The first figure presents the p_{int} probabilities for each cluster, and the second the p_{ext} ones for each pair of clusters.



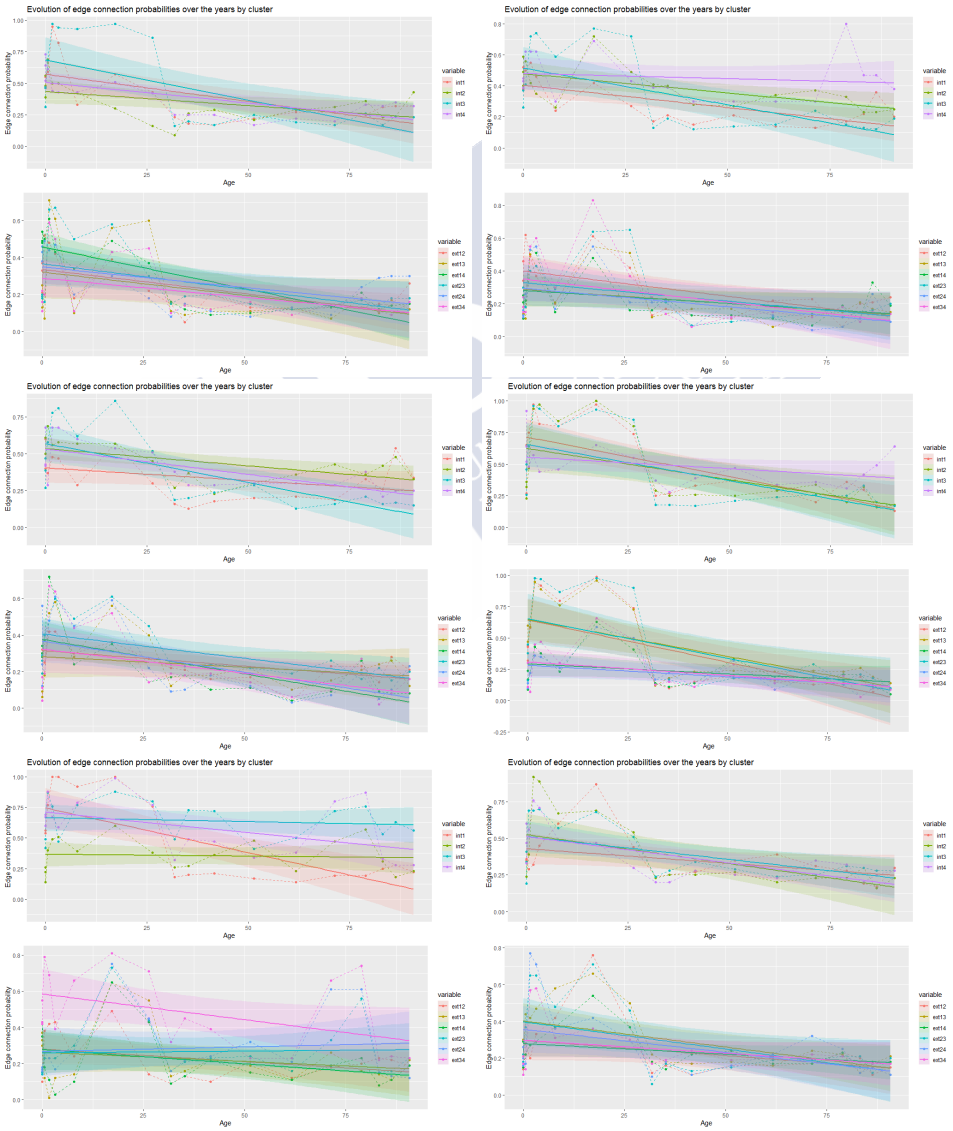
5.5.1 Multiple Age Datasets

The application of the same analysis to a wider age range including data from different arrays (the ones presented in section 4.1), completes the analysis indicating an interesting trend. We include 20 individuals of the following groups: newborns (GSE30870), children 1-16 years (GSE36064), [20,40) years old people (GSE40279), [40,80) years old people (GSE40279), and $\geq$ 80 years old people (GSE40279). We use $\delta_{int} = \delta_{ext} = 0.5$ to take the highest locally correlated sites, and we study the behavior of some of the 10 biggest CpG islands. Results presented in the figure 5.14 re-enforce the idea of the structural local correlation within those regions, as short-range positions present generally higher correlation values. A similar behavior was observed in other islands. Besides, we find a decreasing pattern over increasing years, especially on the long-range regions. Curiously, the correlation seems to be high over short and long-range positions up to 20-25 years old approximately. Then, it starts to decrease up to 70-75 years old when it slightly increases again in

some cases. This observation, that would need a posterior biological confirmation, is then in line with the increasing trend observed above with a shorter age range.

In any case, thanks to the designed SBM-D over the correlation graphs, we are able to measure an evolving time situation using only the ten parameters of inter-cluster and intra-cluster joining probabilities, reducing the dimension of a correlation matrix ($\sim 100 \times 100$) to these ten parameters and their progression.

Figure 5.14: Evolution of p_{int} and p_{ext} probabilities over the average window age for some of the biggest CpG islands.



5.6 Summary

Thanks to the study of the local correlation matrices based on DNA methylation data with graphs and persistent homology, we could take conclusions from their topology. We have demonstrated that their structural behavior is not random and follows a spatial design. Short-range and long-range correlations are distributed forming highly correlation clusters (or modules) associated with near genomic positions and (potentially) with loops, respectively. This understanding allowed us to design a model (SBM-D) of the correlation structure that is able to predict the graph topology and detect heavily local correlated CpG sites within CpG islands.

This model generates a graph trough the estimation of its degree distribution that takes into account the genomic distance between the nodes, and reproduces adequately the structural interactions between methylation marks, even out of CpG islands. It requires mainly three parameters (K, p_{int}, p_{ext}), that can are inferred from the DNA methylation data, and presents different graph properties when compared with other common random network models. Our model is then able to describe and predict a modulated design of a correlation matrix based on those three parameters, providing with a quick analytical tool for analyzing the matrix.

Out of the biological context, the proposed SBM-D may be very useful to measure, for instance, the likelihood of the spread of an infectious disease taking into account the "distance" between the individuals, that can be considered as their localization, the number of common contacts, or any other measure of relatedness.

As an application of the SBM-D, we were able to measure the evolution of the correlation structure with the age, observing a quite stable correlation design over the years in CpG islands. The level and distribution of the correlation in local and global regions are the main differentiating point, presenting generally higher levels of local or short-range correlation. While children and elderly people present high levels of local and global correlation, middle-aged individuals manifest a decrease of the correlation levels above all in the long-range regions. This would validate our initial hypothesis of a decrease in the correlation with age, especially on the long-range interactions for older ages. It is not clear yet how a more or less structured correlation within the islands impacts the genome functioning, but it may augment the gene expression variability producing unexpected results or disease-conditions.

We have not found any explanation for the fact that, in some cases, the oldest individuals present similar levels of correlation as children. Following aging theories associated to DNA methylation, the changes in our methylome are constant and increase over the years, leading to greater variability in older ages. Maybe this variability is causing the high local correlation, or maybe it is a hint about the epigenetic features that could be associated with longevity. Those findings would

need posterior biological research using a wider (ideally longitudinal) age range.

Generally, the study of the local correlation of CpG islands could bring important biological outcomes where a more or less correlation design could be associated to a specific epigenome status and potential disorders. As we have based the successful design of the SBM-D on a local use, this study will be extended to a wider global overview of the correlation in the next part III of the present work with a fast computational algorithm able to manage more CpGs. With both analyses, we aim then to provide with the needed analytical tools to understand better the complex network of interactions inside us.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Part III

An Algorithm for the Global Correlation

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 6

MultiNet

How could we programmatically and efficiently study the correlation of a big-dimensional dataset? How can it be useful for epigenetics? We have developed a computational algorithm, called *MultiNet*, with the main objective of providing with a powerful computational tool of data analysis based on TDA's idea, that is able to extract substantial information from high-dimensional datasets. We mainly apply it to DNA methylation data with different sample cases but its flexibility allows the application to other data sources and research fields. Within this chapter, we will describe the algorithm and its implementation.

6.1 The Big Data Analysis Challenge

From 50s decade, we are living a worldwide revolution in terms of artificial intelligence (AI) that comes basically from three main points: the accessibility to a great variety of big data generated from different sources, the continuous improvement of the computational capacity (Moore's law), and the construction of novel mathematical and computational techniques as *deep learning*. The analysis of huge and complex datasets requires a step forward on the usual way of analyzing networks, as it happens with neuronal networks.

As a result of this new technological design, an enormous part of the mathematical and computational research has been turned into the fast and accurate development of powerful algorithms that can deal with huge complex quantities of data to obtain rigorous results. Big data analysis covers the collection, manipulation, and analyses of massive, diverse data sets that contain a variety of data types, including genomic data [150]. The field of study focused on the development of computer algorithms for transforming data into intelligent actions is known as *machine learning*. Closely related to machine learning, *data mining* comprises the novel analytical techniques that were raised during the last years to explore the data, develop the model or the

algorithm, and identify patterns previously unknown. Currently, machine learning augments, rather than replaces, the analytical power of human brains.

A big part of the traditional statistical methodologies is based on the estimation of a predefined model, while data mining methods try to identify and construct the model based on the evidence without defining a prior one. Generally, the main principle of novel data analysis methods is to "let the data speak" without making assumptions about its prior behavior. There are currently lots of different data mining methods, where the best model is selected based on the data itself. The machine learning process has essentially five linked steps:

1. Data collection: comprises the learning material. In our case, we use DNA methylation datasets or arrays.

2. Data exploration and preparation: establishes the quality of the data collection and prepares it for the learning process. We eliminate, for example, the known altered CpG sites related to SNPs or the sex chromosomes. Moreover, the algorithm is designed to identify other potential biasing features.

3. Model training: the specific machine learning task chosen informs the selection of an appropriate algorithm and the algorithm represents the data in form of a model. For us, the model designed is the own TDA-based algorithm together with several statistical techniques to select CpG sites or regions of interest.

4. Model evaluation: meaning the validation of the results obtained in the previous step. This may comprise the evaluation of the accuracy of the model using a test dataset, or the development of performance measures to the specific study aim. In our case, we use several evaluation methods with different objectives: test the selected epigenetic markers on a different dataset with similar sample characteristics in order to evaluate the prediction accuracy, and test the different algorithm parameters in order to make it as robust and fast as possible.

5. Model improvement: applying different strategies to improve the performance of the algorithm. We may refine the algorithm design based on the biological interpretation of the results. For instance, the selection of the data could be more accurate once we establish the most interesting regions of epigenetic change; or the parameter selection could be programmed to be automatic based on the experience with the same type of data.

These points are the basis of the machine learning design of MultiNet, which is at the same time inspired by TDA. Within the rest of this chapter, we will present the design of MultiNet and its main characteristics, with a detailed description of the different implemented steps of data analysis that compose the algorithm.

After that, in chapter chapter 7, we ill introduce a guideline of parameter selection to understand how MultiNet parameters could be efficiently chosen. The chapter 8 contains the presentation of the main contributions of MultiNet to the analysis of DNA methylation. Additionally, we present a different application of MultiNet to a financial dataset that shows its use with non-biological data. Finally, the chapter 9 contains a guide of MultiNet use in R and the links to the developed programming.

6.2 Introduction to MultiNet

The aim of MultiNet is to extend the local study of the DNA methylation correlation that we did with persistent homology in the part II of this work. As we wanted to work with the entire 450K array without a regional restriction, we needed to develop an alternative model to study the global network of correlation interactions. We have designed a computational algorithm from a TDA-based model that allows us to detect the intra and inter-chromosome correlation structures and differentiate sample groups based on their related methylation levels. This algorithm could be seen as a data analysis tool, that extracts the topology of the data using Morse theory principles. Similarly to Mapper, it is, therefore, not a specific algorithm to study biological or epigenetic data, but a general methodology to detect correlation patterns and sample classification among the hundreds or thousands of variables analyzed. MultiNet is implemented in R.

As mentioned in section 3.4, Mapper is already a very useful tool to detect the shape of the data and visualize it. However, it has some limitations when dealing with huge datasets, either in terms of computational times, parameter selection, or the inability to detect the underlying variables that are causing more impact on the "shape" of the data. Our proposal to treat huge dimensional datasets resides on the following novel approaches:

- Feature modeling: use the variables as observations to construct the topological model (i.e. the simplicial complex). That means, the cluster analysis will be done over a distance matrix among the variables of the dataset. We could then differentiate in the networks or simplicial complexes two levels of information: the characteristics of the variables and how they interact depending on the sample group selected. In case we analyze case/control studies, this method allows us to detect differentiated genomic sites or regions associated with the groups and with different methylation patterns. This provides then with potential epigenetic markers that may help in diagnosis.

- "Divide and conquer" strategy. We divide the dataset into several overlapping windows and apply the algorithm to each one of them. Afterwards, the resulting networks from each window are merged in a unique graph. For this property, we call the algorithm "MultiNet" (from "multiple networks"). This would be similar to do multiple Mappers. The way of selecting and ordering the windows is defined by the user and can be given by the filter functions of the algorithm, so it does not necessarily introduce a new parameter in the model. This method is faster than processing all the windows together and is therefore much faster than Mapper for big data sizes. MultiNet results are less sensitive to the parameter selection than Mapper ones thanks to this window's

design.

- Reduce the computational time using the most relevant information. The algorithm processes the quantity of windows that the user desires, without the need of processing all the dataset information, that still can be done in a reasonable computational time.
- Extend the Mapper design of merging nodes. As our interest is to observe the global correlation structure among the dataset desired, we do not want to reduce the algorithm to join nodes with common points only, as this may limit the network interpretation. Our approach is to join nodes with common and desired characteristics too in order to obtain a more detailed representation of the underlying data structure. In particular, we decided to join nodes with common points or a high mean correlation among the CpG sites contained on them.
- Study the metric from a local and global perspective. That means, we study the differential methylation patterns locally and globally for CpG sites within the same chromosome or in different chromosomes. Both analyses are needed to not overlook the significant behavior of the local structure.
- Define a guide of parameter selection and a more sophisticated machine learning design for MultiNet. We introduce several new parameters not seen in Mapper to have a greater flexibility, as the data windows size or different types of cluster analysis. As a complement, we design a parameter selection guideline with the aim of establishing a parameter stability in terms of results and computational time.

6.2.1 Cluster Analysis

Clustering analysis is the base of MultiNet and its representation. Cluster analysis refers to the grouping of a set of objects in the same group (or cluster), such that the objects in the same group are more "similar" than the objects in other groups. There are many ways of defining this similarity function over the data, and many ways of grouping the objects into clusters. One of the most common methods to do it is the called *k-means clustering* [151]. It aims to form k clusters in which each observation belongs to the cluster with the nearest mean or center such that minimizes the within-cluster variances. Formally, the objective is to find:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2$$

where $(x_1, ..., x_n)$ is a d-dimensional set of observations, $S = \{S_1, ..., S_k\}$ are the sets of observations for each cluster k ($k \leq n$), and μ_i is the mean of points in S_i.

There are several iterative algorithms implemented to solve the prior equation. The most used one is called *Lloyd's algorithm* [152] or *naive k-means*, that assigns each observation to the cluster with the nearest mean (based on the least squared Euclidean distance) and recalculates the means for the observations assigned to each cluster. Other methods, called *initialization methods* (as the Forgy one [153]), randomly choose k observations from the dataset and uses these and the initial means. Hartigan-Wong [154] method provides a variation of k-means which does a local search that iteratively attempts to relocate a sample into a different cluster as long as this process improves the objective function.

The *kmeans* R function uses the Hartigan and Wong method as default, but one could also choose any of the others. For MultiNet, we use the Forgy method as it allows to create empty clusters if the specified k groups are higher than the needed ones.

Other clustering methods as hierarchical clustering one is not directly used in this thesis but indirectly presented the different heatmap plots (programmed with the R function *Heatmap*) that we will specify. The heatmap is a representation of a matrix with colors representing different value ranges. The matrix is ordered by the hierarchical clusters formed over the matrix (either over the rows, the columns, or both). The hierarchical cluster algorithm constructs trees of clusters of objects. The method follows an agglomerative design: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Please refer to [155] for further details about the method.

6.2.2 MultiNet Algorithm

Our starting point is a data cloud X on a finite metric space (X, d_X). In our case of methylation data, we define a matrix B of β values (as defined in the equation (1.1)) of N CpG sites measured over p individuals:

$$B = \begin{pmatrix} \beta_{11} & \dots & \beta_{1n} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \dots & \beta_{pn} \end{pmatrix}$$

Therefore, X is a data cloud with N points in $\mathbb{R}^p$, where each point represents a column of B, i.e., a CpG site.

We define a selection of points of X as a "window" of X, and we will denote it as W_i, $\forall i = 1, ..., D$, where $\{W_1, ..., W_D\}$ could cover all X or could be a subset of

interest. Please note that we will denote W_i indistinctly as a submatrix of B with the selected columns or simply as the subset of the columns selected.

The way of selecting the windows W_i will be determined by a function $h : X \to \mathbb{R}$. For instance, h could be the genomic position of the CpG sites or the median methylation per CpG. We would then obtain a set of overlapping windows $\{W_1, ..., W_D\}$ that contain CpG sites ordered by the values of h (as genomic coordinates or median values). Depending on the study aim, h could be selected differently.

Once we have the windows $W_i \subset X$, we define the filter function $F : W_i \to \mathbb{R}^n$, for any $n > 0$. In our case of methylation data analysis, the filter function will be projected in $\mathbb{R}^2$: $F = (f, g) : W_i \to \mathbb{R}^2$. The functions f and g could be selected as desired depending on the study aim. Could be, for instance, the median and variance of each CpG site.

The metric d could be also defined as needed for the study. In our case of methylation analysis, the metric will be based on the modified Pearson correlation [135]:

$$d = \sqrt{1 - \rho^2_{X,Y}} \tag{6.1}$$

The generic version of the one-dimensional MultiNet algorithm on X computed with the filter function $f : W_i \to \mathbb{R}$ can be then summarized as:

1. Select the function h and the set of ordered and overlapped windows by h: $\{W_i\}_{(1 \leq i \leq D)}$. Each window has a length r and an overlap O. Usually, we will define the same r and O for all the windows analyzed. For instance, we could define h as the variance of each CpG site, select the first 5,000 CpGs more variables and divide them in windows of length $r = 1,000$ with an overlap $O = 500$ (then $D = 10$).

2. For each W_i, cover the range of values $Y_{W_i} = f(W_i)$ with a set of consecutive intervals $\{I_{W_{i_s}}\}_{(1 \leq s \leq S)}$, where S is the number of intervals to cover the window, with length $l_{I_{W_{i_s}}}$ and overlap $o_{I_{W_{i_s}}}$. Usually, the number of intervals S will be the same for all windows analyzed, as well as the length of the intervals l and their overlap o. For instance, we could divide each window of length $r = 1,000$ into $S = 2$ different overlapped intervals of length $l = 500$ and with an overlap $o = 250$ CpG sites.

3. To each window W_i, apply a clustering algorithm to each inverse image $f^{-1}(I_{W_{i_s}}) \subset W_i \subset X$, $s \in 1, \ldots, S$. This defines a pullback cover

$$C = \{C_{(1,1)}, \ldots, C_{(1,k_1)}, \ldots, C_{(S,1)}, \ldots, C_{(S,k_S)}\}$$

of the window W_i, where $C_{(s,k_s)}$ denotes the kth cluster of $f^{-1}(I_{W_{i_s}})$. The number of clusters K per interval is determined by the user and is normally the same for all intervals and windows. The clustering method is normally k-means and the underlying algorithm is selected by the user. We take the Forgy method so in case the number of clusters determined per interval exceeds the optimal number of cluster in that interval, the algorithm does not create the excessive ones. The selection of hierarchical clustering is also possible but slower.

4. For each window W_i we have then a construction similar to the nerve of the cover C_{W_i}, and the corresponding graph M_{W_i}. Each vertex $v_{(s,k)}$ corresponds to one element $C_{(s,k)}$, and two vertices $v_{(s,k)}$ and $v_{(s',k')}$ are connected if at least one of the following conditions is fulfilled:

   (a) The mean absolute correlation of the points of both vertices is equal or greater than a specific level δ:

   $$mean(|H|) \geq \delta$$

   where H is the correlation matrix formed by the elements of $v_{(s,k)}$ and $v_{(s',k')}$, i.e., the correlation matrix of the CpGs contained in both nodes.

   (b) If $v_{(s,k)} \cap v_{(s',k')} \neq \varnothing$ and the quantity of common points is equal or greater than a specific level λ:

   $$\frac{\#\left(v_{(s,k)} \cap v_{(s',k')}\right)}{max\left(\#v_{(s,k)}, \#v_{(s',k')}\right)} \geq \lambda$$

   Please note that the option (a) is checked for vertices belonging to overlapping intervals only (in order to respect the TDA idea of reproducing the "shape" of the data cloud). Normally, we select $\delta = 0.8$ and $\lambda = 0.2$, but the user is free to play with both parameters testing distinct exigency levels for the graph formation.

5. Once all selected windows are measured and all M_{W_i} are created, each sub-network is merged forming the overall simplicial complex M as a union of all the window's simplicial complexes $M = M_{W_1} \cup ... \cup M_{W_D}$. In addition, step 4 is repeated over consecutive windows and two vertices are connected too if any of the two conditions is met. Therefore, MultiNet allows to connect nodes in the network in a bi-directional intra-window and inter-window way. The edges weights of the graph M are assigned based on the absolute mean of the correlation matrix formed by the CpGs in each pair of nodes.

6. A post-processing of the information contained in M is also part of the algorithm and depends on the dataset characteristics. In our case, we do several analysis to obtain substantial biological information and generate colored networks by functions of interest, methylation and correlation heatmaps, statistical results, graphs of biological enrichment analysis and spreadsheets containing all the information. All those steps of post-processing will be also described in the next sections.

6.3 Two Different Perspectives: Local and Global

There are several examples of analyses where different information could be extracted depending on the perspective you use. For example, in our case, is not the same to look at the local correlation structure of the CpG sites located in the same chromosome, that the global correlation structure of the CpG sites located in different chromosomes, as we have already studied in part II of the present work. For that reason, we decided to analyze at once two different approaches in order to have a description of the topological structure of both: local and global. To do it, we adapt the selection of the filter functions as follows:

- Local or intra-chromosomal design: our function h will determine the order of the genomic coordinates. The first filter function f will be also determined by the genomic site, providing with local results in terms of correlation and genes. The second filter function g could be selected as desired but still the results obtained are considered local. We could also select initially only CpG sites located in the same chromosome to guide the local search. This approach is normally faster as chromosomes contain a manageable number of CpG sites.
- Global or inter-chromosomal design: we use different functions h and f that do not depend on the local CpG position in the genome, or any other variable related to the CpG spatial distribution, as the chromosome. We could select, for example, the function h as the difference in the medians between sample groups (case/control) and determine f and g as the median methylation levels and the variance of each CpG.

Thus, depending on the perspective we choose, we will specify a distinct dataset as input for MultiNet and obtain at least 2 MultiNet networks related to local and global analyses.

We could use the results obtained globally to dig into the local approach. In the case of case/control studies, we obtain differentially methylated CpGs with the global algorithm that belongs to different chromosomes with different levels of methylation. Using the local algorithm, we study better those selected CpGs to contextualize the methylation modification within a specific chromosome or genomic region. Of course, this double-design could be also used in many other ways: only local, only global, etc. It provides, in any case, with a solution to detect the different levels of complexity contained in a methylation array.

This is a novel contribution to the DNA methylation analysis, where most of the current methodologies only study the local behavior of the sites, or simply do not normally specify any strategy to deal with these different informative levels.

6.4 MultiNet Implementation over 450k Illumina Methylation Dataset

As previously mentioned, one of the objectives of this present work is to understand the correlation structure of the epigenetic marks based on the study of DNA methylation. Therefore, despite the algorithm could be applied to other types of data, as we will also see afterwards, we will focus from now on the explanation of how we analyze DNA methylation.

Our dataset is a 450k Illumina array, measured on a sample divided into several groups of interest, as control/case for instance. The metric used is the modification of the Pearson correlation specified in the equation (6.1). As a standard pre-process, we normally eliminate from the analysis the CpG sites related to the sex chromosomes X and Y, and CpG sites associated to SNPs, as they may introduce unexpected variability and bias the results [156]. Missing values are not imputed. The normalization or check for data confounders is indirectly done by the algorithm, as we will describe in section 6.7.

The process described in the algorithm 4 below allows to follow the biocomputational algorithm precisely.

The MultiNet resulting graph is a useful element to explore the correlation of the data, as the nodes are created with a correlation metric. Besides, we could extend the analysis perspective (local/global) as described in section 6.3, or apply it to different sample groups to study their differences in terms of correlation and methylation patterns.

Once we obtain the MultiNet graphs (or networks, we use both words instinctively), the analysis of the data should continue extracting more information from the networks which brings interesting biological outcomes. For us, these following steps will be:

1. Networks differentiation. The networks obtained based on sample groups can be differentiated with the processes we will describe in the section 6.5. Once we know they are different, we can take conclusions about the reason of the difference and the interpretation. In addition, extra-information about underlying methylation patterns could be extracted from the networks in form of heatmaps and hierarchical clustering.

2. Detection of significantly differentially methylated sites among sample groups or regions from the MultiNet networks. The statistical significance is tested with the appropriate statistical models, as will be described in the section 6.6.

Algorithm 4: MultiNet algorithm

Data: Methylation matrix: local or global
Result: MultiNet Network
initialization;
while *Parameter Selection* **do**
Select the sample groups and apply the following steps for them;
Select the algorithm dimension n, that is 2 by default;
Select the metric d, that is the modified Pearson correlation by default;
Select h and the D dataset windows W_i, their length r and overlap O. They are $r = 1,000$ and $O = r/2$ by default;
Select the filter functions f and g;
Select the intervals or bins within each window $I_{W_{i_s}}$, their length l and their overlap o. They are $l = r/2$ and $o = l/2$ by default;
Select the cluster method m, that is k-means (Forgy) by default for all windows;
Select the number of clusters $k_{I_{W_{i_s}}}$, that are 3 by default ;
Select the node joining parameters δ and λ, that are 0.8 and 0.2 by default ;
if $W_i \neq \varnothing$ **then**
1. Define a region $R = [min_f, max_f] \times [min_g, max_g]$ so that we cover overlapping $\cup_{i,j} I_{i,j}$;
2. For each i, j select all data points so the values of f and g lies within $I_{i,j}$;
3. Following m, d, and k, apply cluster analysis over those data points and create a 0-dimensional simplices or vertices;
4. For all vertices in the sets $\{I_{i,j}, I_{i+1,j}, I_{i,j+1}, I_{i+1,j+1}\}$, if the intersection of the cluster associated with the vertices is $\geq \lambda$ or have a mean correlation $\geq \delta$, then add a 1-simplex or an edge;
5. The final graph M_{W_i} is then the 1-skeleton of the complex generated;
6. Label each vertex as the variables selected within W_i;
else
Do not produce any cluster or vertex;
end
7. Repeat the process for $\{W_{i+1}, ..., W_D\}$ of length r and overlap O so $W_i \cap W_{i+1} \neq \varnothing$;
8. For all vertices in the complex M_{W_i}, if the intersection with any vertex in $M_{W_{i-1}}$ or $M_{W_{i+1}}$ is $\geq \lambda$ or have a mean absolute correlation $\geq \delta$, then add a 1-simplex or an edge;
9. Create the overall MultiNet graph M joining the M_{W_i} following step 8;
10. Assign edge weights to M as the mean absolute correlation among nodes.
end

3. Validation of results. We will use different datasets, samples and parameters to validate the obtained significant sites. The prediction of the sample group from the differentially methylated sites detected is key to check the model accuracy and for future diagnosis purposes.

4. The biological interpretation of results can be done comparing them with published research. Gene enrichment, gene ontology (GO) enrichment or pathways enrichment are some of the approaches that can be followed in order to obtain more information about the underlying epigenetic or methylation marks detected in the previous steps. The implemented R functions for this aim were used, together with the software STRING.

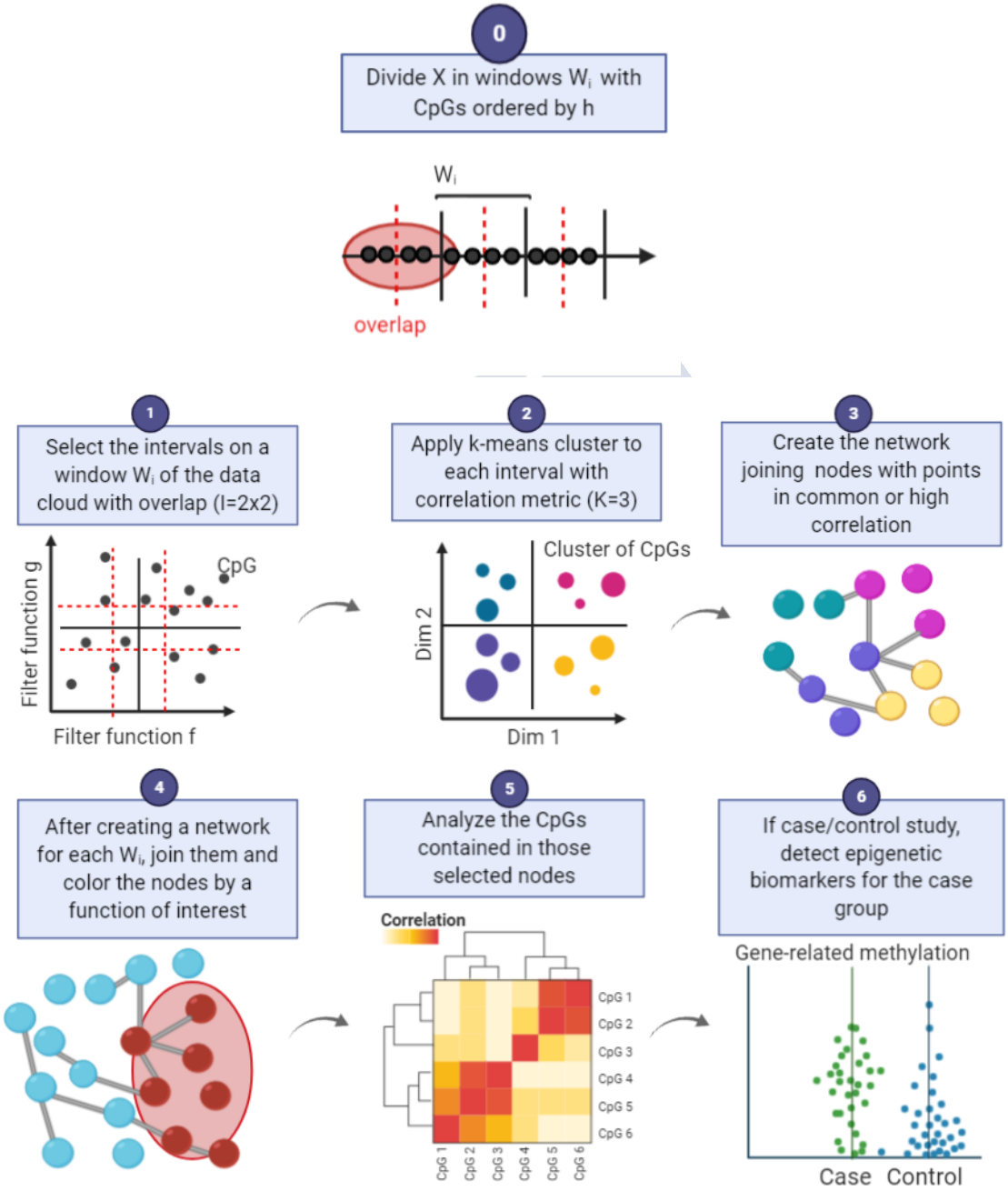
MultiNet is then able to provide information in two levels: 1) variable level: being able to extract the topological structure of the epigenetic marks under study and their correlations; and 2) sample level: being able to differentiate sample groups based on the methylation patterns detected in the first step. This two-level design completes the overall analysis of the data serving as a diagnosis tool.

MultiNet has more parameters than the usual Mapper algorithm. However, their selection is not complex with the adequate guideline, that we will provide in chapter 7. Moreover, they allow a greater flexibility for data analysis with diverse datasets, a higher computationally efficiency, and robustness.

The computational time mainly depends on two parameters (besides the sample size, of course): 1) The number of windows measured (could be the entire array); and 2) the number of intervals that we are specifying for each window. In a bi-dimensional MultiNet, the number of intervals is the product of the number of intervals specified in each direction (x axis and y axis). It means, that for 5×5 intervals, we would need to do 25 cluster analyses and we obtain $25 \times 3 = 75$ clusters or vertices per window (if we determine 3 clusters per interval). That parameter selection would slow down the computational times unnecessarily and produce uninterpretable networks. As we will study afterwards, such a quantity of intervals and clusters may not be needed for a correct application of the algorithm, and with only 2×2 intervals per window the computational time is decreasing almost 10 times. For this reason, a sensitivity analysis of parameter selection is always needed to avoid extra computational time without further substantial results.

With the correct selection, we obtain our MultiNet networks for 10,000 CpGs in a couple of minutes and the entire array (with more than 450,000 variables) in around one hour (using a computer Intel(R) Core(TM) i7-4710MQ CPU @ 2.50 GHz, RAM 24GB, Windows 7).

Figure 6.1: MultiNet main steps.



6.5 Network Differentiation

If we apply MultiNet to different sample groups (as case/control), the algorithm will produce one graph per sample group. Those graphs may have a different number of nodes and edges. As a first step, we need then to demonstrate that the obtained graphs are indeed different and detect the reason of the differences.

During the last years, the increasing study of complex networks led to the development of alternative mathematical methods of graphs analysis. One of the challenges is to differentiate networks. Several techniques were developed recently but most of them are unable to compare networks with different nodes/edges or more than two networks, and require a high computational time for big sizes.

On the other side, usual graph theory characteristics as centrality, modularity, dimension, or transitivity are usually local and mainly based on differences between either node or edge measurements, without considering the network topology. Moreover, they may not describe completely the network complexity and could be unfeasible or confusing with huge networks. Due to the MultiNet design that creates networks per sample with the same parameters and a similar number of edges and nodes, classical graph theory measurements are not very useful to differentiate the graphs created.

Having all those points into account, we decided to apply several exploratory and analytical methods of network differentiation that are defined in the next subsections. They have the objective of studying the underlying distribution of the graphs (please see subsection.6.5.1), or the layout of the graphs that determines their differentiation (please see subsection.6.5.2). Besides, we will also apply exploratory techniques of network visualization that will facilitate the search of nodes of interest, described in subsection.6.5.3.

In principle, all those methods are able to compare more than two groups.

6.5.1 Probability distribution of MultiNet Graphs

The aim of MultiNet networks is not so much to identify a distinguished layout of the graphs, but to be able to establish significant differences of the underlying methylation distribution. The MultiNet networks created with different sample groups are designed to be similar, as we use the same MultiNet parameters to create them. However, they may represent different data distributions coming from the sample distinction. For this reason, we propose a method to measure the level of network dissimilarity based on the difference of the underlying data distribution that they represent.

We estimate the probability distribution of each one of the networks and measure if the distributions obtained are significantly different. Let $G_{case} = (V, E)$ and $G_{control} = (W, F)$ the two graphs generated by MultiNet with two sample groups, where M_1 is the total number of nodes of G_{case} and M_2 is the total number of nodes of $G_{control}$. For each nodes v_i, w_i in each network, we calculate:

$$\mu_{case_{v_i}} = mean(\mu_{CG_j}),$$

$$\sigma^2_{case_{v_i}} = mean(\sigma^2_{CG_j})$$

where μ_{CG_j} is the mean over case samples of each variable CG_j contained in the node v_i, $\forall i = 1, ..., M_1$. Similarly,

$$\mu_{control_{w_i}} = mean(\mu_{CG_j}),$$

$$\sigma^2_{control_{w_i}} = mean(\sigma^2_{CG_j})$$

where μ_{CG_j} is the mean over control samples of each variable CG_j contained in the node w_i, $\forall i = 1, ..., M_2$.

Now, we determine the overall mean and variance of each network as:

$$\mu_{case} = \sum_{i=1}^{M_1} \mu_{case_{v_i}} \text{ and } \sigma^2_{case} = \sum_{i=1}^{M_1} \sigma^2_{case_{v_i}}$$

$$\mu_{control} = \sum_{i=1}^{M_2} \mu_{control_{w_i}} \text{ and } \sigma^2_{control} = \sum_{i=1}^{M_2} \sigma^2_{control_{w_i}}$$

Hence, the distribution of the network is determined by the calculated mean and variance of each group, that is Gaussian by the central limit theorem:

$$M_{case} \sim \mathcal{N}(\mu_{case}, \sigma^2_{case}) = \mathcal{N}(\sum_{i=1}^{M_1} \mu_{case_{v_i}}, \sum_{i=1}^{M_1} \sigma^2_{case_{v_i}})$$

$$M_{control} \sim \mathcal{N}(\mu_{control}, \sigma^2_{control}) = \mathcal{N}(\sum_{i=1}^{M_2} \mu_{control_{w_i}}, \sum_{i=1}^{M_2} \sigma^2_{control_{w_i}})$$

The effect size of the difference between both distributions can be measure with several approaches as the Cliff's Delta. Cliff's delta method is a non-parametric approach developed to improve the bias introduced by parametric methods such as

Cohen's d. Cliff (1996) [157] suggested a novel method that examines the probability that individual observations within one group are likely to be greater than the observations in the other group. That is, the population parameter for which such an effect size is intended is the probability that a randomly selected member of one population has a higher value than a randomly selected member of the second population, minus the reverse probability:

$$\delta = P(x_{i1} > x_{j2}) - P(x_{j1} < x_{j2})$$

where x_{i1} is a member of population one and x_{j2} is a member of population two. The sample estimate of this statistic, Cliff's $\hat{\delta}$, is obtained by comparing each of the scores in one group to each of the scores in the other:

$$\hat{\delta} = \frac{\#(x_1 > x_2) - \#(x_1 < x_2)}{n_1 n_2}$$

where x_1 and x_2 are the observations within group 1 and group 2, and n_1 and n_2 are the group sample sizes. Each observation from one group is compared to each one in the other group, and we calculate the number of times that the observations from one group are higher or lower than in the other. The non-parametric nature of Cliff's δ reduces the influence of such characteristics as distribution shape, differences in dispersion and extreme values. The statistic relies on the *dominance analysis*, a concept referring to the degree to which one sample overlaps another: the greater the overlap, the less the difference between the groups. An effect size of 1 or -1 indicates the absence of overlap between the two groups whereas a 0 indicates overlap and equivalent group distributions.

Cliff's method may fail when two distributions differ in spread, but not in central tendency (two distributions centered in zero but with very different variances will have a $\hat{\delta}$ near to 0). In those cases, we are interested in obtaining a low $\hat{\delta}$, as it warns us that both sample groups have similar mean methylation levels but there is an extra confounder that affects their variability and would need a specific investigation.

This method of network differentiation is computationally fast and is able to obtain significant differences of the networks not obtained by randomly selecting the sample groups. The obtained differences between the networks are therefore explaining the different distribution of the CpGs based on their correlation levels and their methylation profiles. Other methods as the detection of neighbor CpG sites may also be feasible for network's distinction but computationally slow.

6.5.2 Persistent Homology over MultiNet Graphs

The layout of the simplicial complexes obtained with MultiNet is an extra source of information about the distribution of the underlying data based on the metric

used (in this case, the modified correlation). For instance, structural holes in the networks can give important information about CpG sites interactions. In low/mid-sized networks, those differentiated topological elements could be more identifiable.

We could then study the topology of the graphs with persistent homology, and study the differences of the graphs generated with distinct sample groups via the distinct Betti numbers or lifespan times. This analysis is possible for low/mid-size graphs, as it becomes difficult to compute and interpret for huge graphs.

As mentioned in the subsection 3.3.2, persistent homology will be applied then to a graph generated by MultiNet, say $G = (V, E)$, that is formed by

$$V = \{V_1, ..., V_n\} = \{\{CG_{1,1}, ..., CG_{1,n_1}\}, ..., \{CG_{n,1}, ..., CG_{n,n_l}\}\},$$

where n is the number of MultiNet nodes and n_i is the number of CpG sites within each node; and $e_{V_i,V_j} = 1 - mean(|C|)$, $\forall i, j \in 1, ..., n$, for C the correlation matrix of the CpGs contained on V_i and V_j.

The interpretation of the topological elements would have here an additional level of complexity, as they may be formed by vertices that are clusters of CpGs.

6.5.3 Colored Nodes

The use of colored nodes is a common exploratory technique used in Mapper to detect specific patterns of information defined by the user. In our case, we color each node by different functions based on the CpG sites contained on each node. It allows to see directly in the graph how those functions are distributed and potentially find the biological interpretation:

- In case we define case and control groups from the sample (or any other type of sample groups), nodes could be colored by the levels of the difference of the median methylation of the CpG sites in a node between both sample groups. The differentially methylated CpG sites are then selected as the CpGs contained in the nodes with a higher case-control difference of medians. This method allows us to extract the **differentially methylated sites (DMSs or DMCs)** or also the **differentially methylated regions (DMRs)** if we do it locally. The statistical significance of the DMSs will be tested with the techniques explained in the next section 6.6.
- In addition, we can color the network based on the median methylation levels for the CpGs contained in each node, so we have an overview of the distribution of the hyper and hypomethylated sites. It allows us to detect **similarly methylated sites (SMSs)** that share common patterns.

- In order to observe the modifications in the correlation structure based on the different methylation levels, we can color each node by its median absolute correlation. It is useful to detect **highly correlated sites (HCSs)** and obtain the biological interpretation of this interaction.
- Coloring nodes by their median variance can also give an idea of the associated data variability.

For instance, the figure 6.2 represents the median and variance distributions per node. We can already detect some sample characteristics from them: the case sample presents a higher overall variance per node and the median levels higher than in the control group.

Figure 6.2: Density plot with histogram overlay of the median and variance per node.

The user is free to select more functions of interest to present with node colors. The ranges of those distributions could be also defined by the user or determined by their quartiles in order to observe the natural data classification. Therefore, the natural ranges would be:

1. Highest values (red nodes): values greater than the third quartile Q_3 of the overall distribution of the function of interest (taking all nodes into account).
2. Medium values (pink nodes): values between Q_1 and Q_3 (both included).
3. Lowest values (blue nodes): values lower than the first quartile Q_1.

Please note that in the figures we will present in the next chapter the quartile Q_1 will be referred as q_2 and Q_3 will be referred as q_4 as the values derive from the R function *quantile* that returns five values (minimum, Q_1, Q_2, Q_3, maximum).

6.6 Diagnosis

Once we generate the MultiNet graphs, analyze their layout, and compute their differences, we may want to extract substantial information about the underlying data characteristics. To do this, we use several statistical methods that allow us to select the most relevant significantly differentiated variables over the sample groups. From this, we extract the CpG sites that perform better on sample prediction and are potential epigenetic biomarkers of a sample feature. We will mainly use logistic regression and random forest algorithm, apart from other common statistical procedures that will be explained when needed.

6.6.1 Logistic Regression

Once we have detected the DMSs with the colored graphs, we need to test their significant difference over the selected sample groups. Our null hypothesis is then

$$H_{0_i} = \text{CpG site } i \text{ is equally expressed in groups 1 and 2, } \forall i = 1, ..., n$$

where n is the number of detected DMSs. To do it, we apply a logistic regression per CpG, with an outcome variable Y indicating the dichotomous sample group classification and an independent variable x_i that indicates the methylation Beta-values of each CpG site.

$$E(Y|x_i) = \beta_0 + \beta_1 X_i, \quad \forall i = 1, ..., n$$

We could denote $\pi(x_i) = E(Y|x_i)$. The specific form of the logistic regression model is:

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

The logit transformation of $\pi(x_i)$ that will be analyzed with a linear regression model is defined as:

$$g(x_i) = ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_i$$

The likelihood function of the model which leads to the maximum likelihoods estimators is, for every feature x_i and observed class y_i:

$$l(\beta) = \prod_{j=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

After estimating the coefficients of the model, we test their significance using a Wald test. The statistical null hypothesis of $\beta_1 = 0$ is tested with the Wald ratio that follows a standard normal distribution:

$$W = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$$

In order to detect the type I errors (i.e., false positives) when testing so many hypothesis, the obtained p-values p_i of each logistic regression model are adjusted with a false discovery rate (FDR) method called *Benjamini–Hochberg* (BH). The FDR is $E[Q]$, where Q is the proportion of false discoveries among the discoveries. The BH method controls the FDR at level α (usually we take $\alpha = 0.05$) listing the p-values in ascending order and denoted by $\{p_{(1)} \dots p_{(n)}\}$. The BH procedure has then two steps:

1. For a given α, find the largest k such that $p_{(k)} \leq \frac{k}{n}\alpha$
2. Reject the null hypothesis for all $H_{(i)}$ for $i = 1, ..., k$

As we test hundreds or thousands of null hypothesis related to the selected CpG sites, we may have a problem with adjusting methods as FDR, that decrease in power and lead to a small number of rejected null hypothesis (where "power" means the proportion of non-true nulls which are rejected). We would need to have very low initial p-values in order to obtain significance with FDR in a great number of tests. For this reason, when FDR methods fail, we will use other newer multiple comparisons adjustments, as the *SGoF* method (from sequential goodness-of-fit) [158]. If k are the number of rejections after performing n tests individually at level γ then, having into account the null hypothesis that n nulls are true, the expected number of rejections is $n \times \gamma$. Therefore, we can compare k with the expectation in order to conclude about its significance. This goodness-of-fit metatest is defined as an exact binomial test with significance level α. Normally we will take $\gamma = \alpha = 0.05$.

The R package *SGoF* implements however the *Conservative SGoF*, which is an asymptotic version (for a larger number of tests) of the Binomial SGoF procedure, where the Binomial quantiles are approximated by the normal ones. Besides, the variance of the number of p-values below gamma is estimated without assuming that all the null hypotheses are true, which typically results in a more conservative decision. When the number of tests is large, Conservative SGoF and Binomial SGoF report approximately the same result. This method has a considerably greater power than BH one when dealing with thousands of tests.

As a general approach, we will first take FDR and only use SGoF when the number of tests increases and results are not as expected.

Additionally, with low sample sizes we may have the problem of *complete separation* [159], very typical using genomic data. This problem is originated when a collection of the covariates completely separates the outcome groups, i.e., the covariates discriminate perfectly. The likelihood function is then monotone and does not

reach the maximum. The problem is manifested by large estimated standard errors or large estimated coefficients. This issue may be solved by different ways, and we have selected the penalized logistic regression model manner [160]. This method ensures the obtaining of the maximum likelihood estimate modifying the likelihood function with a penalty:

$$l(\beta) = l(\beta) \times |\mathbf{I}(\beta)|^{0.5},$$

where $|\mathbf{I}(\beta)|^{0.5}$ is the square root of the determinant of the information matrix with elements defined by the partial derivatives of the log-likelihood function. The R function *logistf* provides with an implementation of this model using Firth's bias reduction method.

Generally we will use the common logistic regression and change to penalized logistic regression when the first one fails and shows unexpected coefficients estimates.

For the cases when we deal with more than two sample groups we could perform a multinomial logistic regression. Please refer to [161] for a detailed description of the logistic regression model and applications.

6.6.2 Random Forest

On the other side, in order to detect which of the significant sites are more relevant to differentiate the sample groups, we apply random forest [162, 163] to the set of DMSs. This model performs normally better than other similar ones as C5.0, and it is already implemented in R (*randomForest* package, or *ranger* function). The selection of this method as a complement to the logistic regression was based on the need of a "relevance" measure to identify anyhow the most "interesting" CpG sites (i.e., those which would lead to a better prediction). In genomics, the number of key biomarkers related to a disease or condition is aimed to be reduced maximally in order to increase the diagnosis efficiency and viability. Random forest is able to quantify the level of relevance of a variable with the Gini index.

This method uses an ensemble of classification tree with recursive partition, that starts searching for all distinct values of each variable to find the variable and the cut-off value that divides the data into two parts minimizing the sum of squares of the residuals. That means, the partition allows us to explain a greater proportion of the total data variability. This technique is fast and easy to use for high-dimensional data, allows multi-class classification, and is generally a good "predictor".

The Gini index [164] measures the purity degree or homogeneity of the tree nodes: a node is pure when it only contains elements of the same class. The node is purer

if the index value is smaller. Generally, we could define it as:

$$Gini = 1 - \sum_{i=1}^{n} (p_i)^2$$

where p_i is the probability of an object being classified to a particular class.

The importance measure consists on the reduction of the Gini index, so the most important variables are those with a higher reduction. With $p_k = n_k/n$ being the fraction of the n_k samples from class $k = \{0, 1\}$ (or dependent variable of sample classification) out of the total of n samples at node v, the Gini impurity $i(v)$ is calculated as

$$i(r) = 1 - p_1^2 - p_0^2$$

Its decrease Δi, that results from splitting and sending the samples to two sub-nodes v_l and v_r (with respective sample fractions p_l and p_r) by a threshold t_θ (a value of our CpGs) on variable θ (our CpGs), is defined as

$$\Delta i(r) = i(r) - p_l i(r_l) - p_r i(r_r)$$

In an exhaustive search over all variables θ available at the node and over all possible thresholds t_θ, the pair $\{\theta, t_\theta\}$ leading to a maximal Δi is determined. The decrease in Gini impurity resulting from this optimal split $\Delta_{i_\theta}(v, T)$ is recorded and accumulated for all nodes v in all trees T in the forest, individually for all variables θ:

$$I_G(\theta) = \sum_T \sum_r \Delta i_\theta(r, T)$$

The Gini importance I_G finally indicates how often a particular feature θ was selected for a split, and how large its overall discriminate value was for the classification problem under study.

As a summary, we will consider a CpG site significantly differentiated among the selected sample groups if the logistic regression is significant (adjusted p-value $<$ 0.05) and the Gini importance is non-zero. The selected sites/genes with a higher Gini importance (that have, at the same time, a lower logistic adjusted p-value) will be presented as the most relevant associated to the sample differentiation. The relevant CpG sites obtained will be tested on distinct datasets to ensure its ability to predict the sample grouping without excessive over-fitting.

6.7 Statistical Considerations of MultiNet

When working on data analysis, the focus should not be only on selecting the appropriate analytical tools, but also on detecting potential data confounders or artifacts. The lack of standardized data or the use of incomparable data may lead to unexpected results that sometimes are considered as "discoveries". This problem is exponential when dealing with huge datasets or high-dimensional data, especially when talking about genomics where there are still a lot of unknowns. For this reason, it is important to re-think the results to take into account all the potential biasing factors.

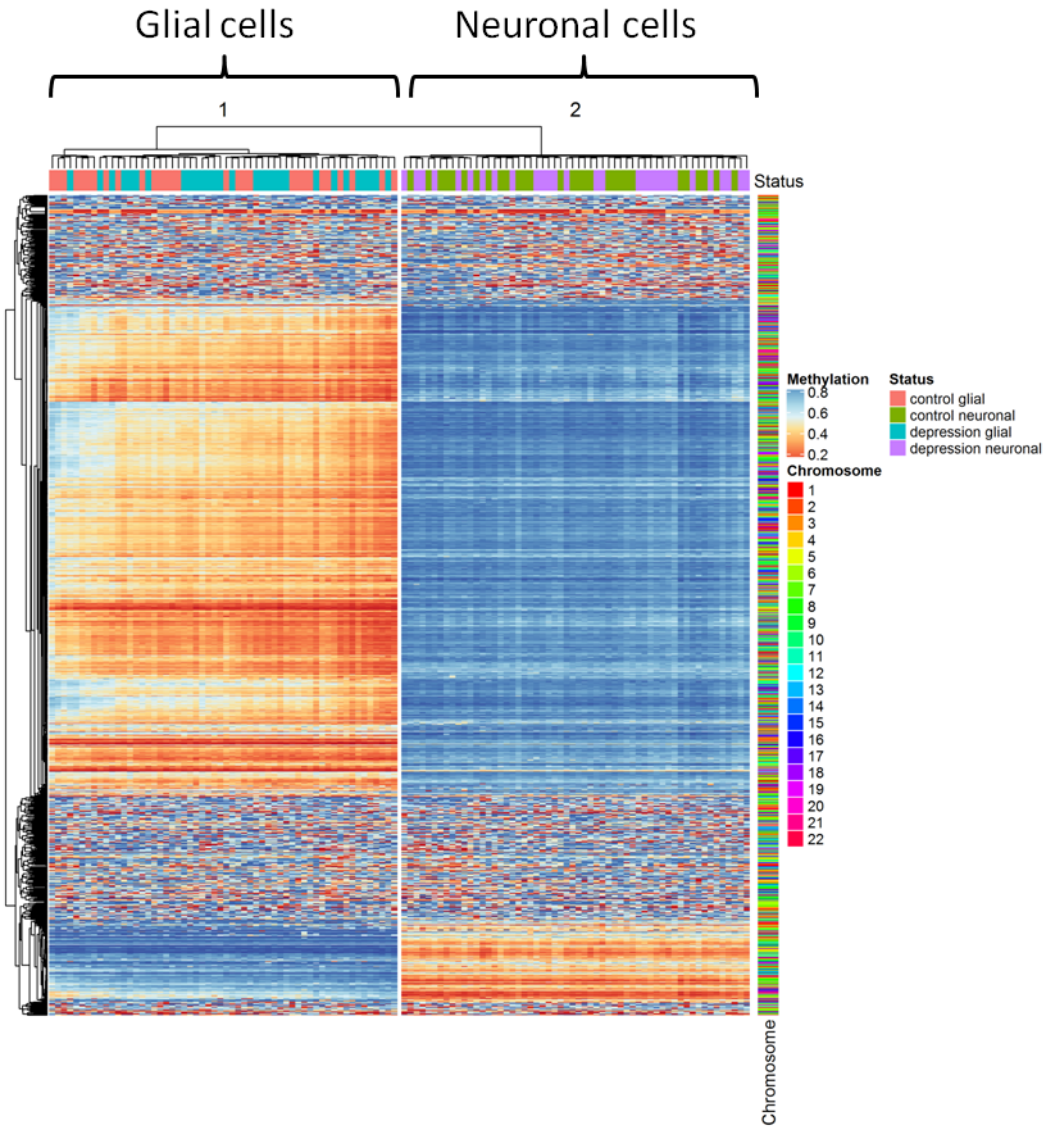
When using MultiNet, we advise to do it with each application, despite the own algorithm have implemented several check-steps to avoid an incorrect data interpretation, as we describe within this section.

Apart from the power of reducing the data dimension through the creation of simplicial complexes, MultiNet has other important statistical properties to mention. One of them is its auto-normalization function. Taking into account a metric based on the Pearson correlation to group clusters of features, the normalization of the data is indirectly done. In this sense, we could say that MultiNet acts as a "data smoother". Despite the Pearson correlation works better with normal data, the MultiNet results are not significantly altered if we use Spearman, for instance. Nevertheless, we could normalize the dataset before its input into MultiNet.

On the other side, the cross-validation of the algorithm can be done with the parameter selection sensitivity analysis, that we recommend and will be described deeper in the next chapter 7; and with the prediction on different datasets to confirm the results. Despite sometimes datasets with great sample sizes are not available, we should always be careful with the interpretations in low/mid sample sizes.

Indeed, MultiNet detects generators of variability different than the sample grouping itself. For instance, it is perfectly able to detect that there is an underlying confounder in the brain methylation data GSE41826, where the difference of the depression and control MultiNet networks is not significant (low Cliff's delta) but the heatmap 6.3 clearly distinguish another factor of differential methylation status. Those results are in line with the associated publication [165], where they identify cell-type-specific DNA methylation differences between neurons and glia that are independent of the psychiatric phenotype.

Figure 6.3: Heatmap of DMSs over depression/control samples and cell types. The columns represent the individuals and the rows the CpG sites. The right-sided vertical bar indicates the chromosome of the CpG site, and the upper horizontal bar indicates the sample group or "status". They are ordered by the results of the hierarchical clustering on rows and columns.



USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 7

Parameter Selection for MultiNet

We have shown that MultiNet precises of a specific parameter selection that increases its flexibility. How can we decide which is the optimal parameter election of the model? In this chapter, we will present a guidance to answer this question.

7.1 The Parameter Selection Challenge

Being based in TDA and Mapper, MultiNet could be described already as an unsupervised clustering machine learning algorithm because it makes minimal assumptions about the data analyzed. On the other side, as we are also using the results of MultiNet to predict epigenetic changes and classify them by sample groups, it can be also viewed as a supervised algorithm of epigenetic classification with a pattern detection aim. The use of random forest, which is considered unsupervised and supervised, completes this meta-learning design.

It is difficult to compare MultiNet with other methods of data grouping or prediction, as this algorithm is a compound of several statistical, computational and topological data analysis techniques. In addition, MultiNet is able to produce results about the genes and biological pathways enriched, together with other substantial biological information. As it happens with many other data mining techniques, there is not an "optimal method" but the user may select it depending on the data characteristics and study aims. In any case, we think that MultiNet provides with a complete and fast approach. Despite it requires more parameters, results are robust and the computational time is low to deal with high-dimensional datasets.

We have seen that MultiNet is quicker than Mapper as it goes with a "divide and conquer" strategy of calculating the correlation matrix by overlapped data windows, please refer to the table 7.1 to see a detailed comparison of computational times. Moreover, the hierarchical clustering method used by Mapper increases the compu-

tational time generally. For the same quantity of CpGs, Mapper could take 10 times more time to complete the process (or even more with higher data sizes). We avoided using computational techniques as parallelization in order to maintain a practical use of MultiNet with limited/normal computational resources. Besides, MultiNet is more stable, as the several window's division removes naturally the big part of the sample noise or bias, resulting in more solid and precise networks representing the underlying selected data. The join of all those networks is, therefore, resembling more accurately the data properties.

Table 7.1: Computational times (minutes) of Mapper and MultiNet (window of 500 CpGs, 2 bins, 2 clusters, 50% of overlap), and the correlation matrix calculation. The biological additions of MultiNet were not included for this calculation.

Variables	MultiNet	Mapper	Correlation matrix
1,000 CpGs	0.15	0.4	0.002
5,000 CpGs	0.7	7	0.05
10,000 CpGs	1.4	>20	0.2
50,000 CpGs	7	>20 (19 Gb)	8 (19 Gb)
100,000 CpGs	14	>20 (75 Gb)	>20 (75 Gb)
400,000 CpGs	∼50	no result	no result

Usually, researchers select Mapper parameters based on a previous data exploratory step. However, the optimal parameter selection of the initial Mapper design is a mathematical challenge and a research topic of the last years, where several investigators tried to present an automatic approach based on topology theory [166, 167] or programmatic techniques [168] improving the selection of the clusters or the bins. Especially interesting and practical is the last article published about Mapper parameter selection [169], where it was presented that some general selections as a low data cloud size, a high number of bins or intervals, and a low percentage of overlap may lead to incorrect, unexpected and very unstable Mapper results. They used the example of the torus to show how a wrong parameter selection may lead to simplicial complexes very far from the torus layout. This article recommendations are in principle in line with our MultiNet parameter selection, where we normally have a high-dimensional data cloud, and select a low number of intervals and a high percentage of overlap.

In the next section, we describe in detail the reasons of our parameter selection.

7.2 Sensitivity Analysis of Parameter Selection

Finding the optimal set of parameters for MultiNet is crucial in order to complete its successful application. Of course, the parameter selection should also take into account the computational time, to find a right balance between stability and quickness. The MultiNet algorithm efficiency depends mainly on the number of windows of the data cloud, and the number of intervals per window.

The parameters selected for the applications of MultiNet (that will be presented in the next chapter) are based on the previous knowledge of the data cloud, the type of data (DNA methylation), the exploratory analyses done with the dataset, and the computational efficiency of the method. Some of those exploratory analyses include the visualization of the filter functions distribution or sampling analyses to select the optimal number of clusters to be included in the k-means model, for instance.

We may need to test several sets of MultiNet parameters before selecting the one that describes better our data. Please note that, following our approach, the parameter selection would only need to be done once, as we decided to keep the same parameters for all the windows selected. That means, all the windows have the same length and overlap, are separated into the same number of intervals and are clustered into the same number of clusters per interval. We decided to follow this design in order to obtain interpretable results and reduce computational times. However, the user is free to re-design this strategy based on data needs.

The following explanation per parameter aims to provide a user guide for their selection.

7.2.1 Filter Functions and Metric

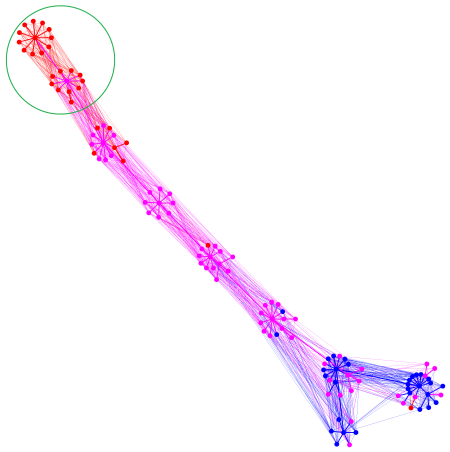
The filter functions and the metric are of course defined based on the previous knowledge of the data type and the aim of our research. In this case, as we are studying the epigenetic biomarkers that lead to specific disease diagnosis or group differentiation, the use of measurements as the difference of medians among groups, median, or variance is reasonable. In addition, as we are interested in studying the correlation structure of the DNA methylation, a metric based on the Pearson correlation was also an adequate selection.

It is important to study the distribution of the filter functions before selecting them for the algorithm, in order to interpret better the resulting graphs and to detect outliers or subsets of data that may be interesting to analyze separately. On the other side, we already saw that the filter functions can be automatically defined

by a previous process, as for example taking the coordinates of principal component analysis.

The use of adequate node coloring functions is also helpful to detect interesting data characteristics in the networks.

Figure 7.1: Example of a MultiNet graph colored by variance levels per node. The green circled region presents higher levels of variance and may deserve a special investigation.



7.2.2 Window Length

The selection of the windows should be based on the aim of the study, selecting windows ordered by genomic position for a local perspective, for instance. The length and number of windows would be based on the computational time needed to calculate the related metric (correlation matrix), and the amount of data we would like to include. Of course, if we select a wider window we would also need more time to calculate the correlation matrix but fewer windows to cover the entire array dimension. Generally, the selection of smaller windows is better as the algorithm takes more minutes on running a window length of 2,000, for example, that running 2 windows of length 1,000 CpGs. The maximum number of windows could be defined by the aim of the study or could be the entire array. For instance, we could select the most differentiated CpGs by their median distributions if we want to detect DMSs.

Furthermore, the window should contain enough data points to do cluster analysis and obtain robust results. In this sense, the final selection of the window would be determined by the number of intervals and clusters that we will detail below. As we decrease the window length, and so the number of windows to cover the dataset length increases, we also increase the number of nodes in the network and it may become non-interpretable.

The percentage of overlap between consecutive windows is of 50% by default. It was selected as a standard measure as low percentages did not reproduce successfully the distribution of the data cloud.

7.2.3 Number of Intervals

Bin length is also another of the parameters we should introduce in the algorithm. The graphs have of course more nodes and edges if we increase the number of intervals. As we mentioned before, this increase does not need to discover any hidden data structure but will make the algorithm 10 times slower if we use 5 intervals instead of 2, for instance.

Similarly to the window overlap, the interval overlap is defined as 50% per default, as it leads to stable results. The user is free to adapt it to other data needs.

Several applications of k-means cluster analysis (with different number of clusters) over different interval lengths selected randomly may give us an indication of the percentage of variance explained and therefore the optimal length.

7.2.4 Number of Clusters and Cluster Method

The number of clusters should be selected adequately to group correctly the data features and represent accurately the underlying data cloud. If we increase it considerably, we may obtain huge networks with many nodes that do not provide with substantial information but only deforms the network. If we decrease the number of clusters too much, we may be grouping features with different characteristics.

The number of optimal clusters per interval can be calculated of many different ways, as using the Elbow method [155]. In fact, R has different packages and functions that apply and compare the methods to select the optimal number (as the package *NbClust*). We could select randomly different CpG sites and test how many clusters would those methods propose. This sampling process simulated several times may give us a good indicator of the data needs. We could develop an alternative machine learning design of the MultiNet algorithm that calculates automatically the number of clusters and intervals needed per window. We would need to decide then on two parameters less. However, the computational time would increase and results should be interpreted and compared carefully, as we may have distinct networks because we have used different underlying parameters. We decided to keep these parameters as user choices to allow a greater flexibility of use.

With regards to the k-means method, MultiNet is prepared to choose the one de-

fined by the user. The *kmeans* function in R has different options: *Hartigan-Wong*, *Lloyd*, *Forgy* (both names for the same method *Lloyd-Forgy*), and *MacQueen*. The algorithm of Hartigan and Wong is used by default. Except for the Lloyd–Forgy method, k clusters would always be returned if a number is specified. That means, they are forced to create k clusters even when there are no points to cover all of them. In this case, R produces an error and the algorithm stops. As we are dealing with high-dimensional datasets where we have not visibility over all the data windows included in the model, we are interested in allowing empty clusters (returning less than k clusters) in the needed cases. For that reason, our selection was Lloyd-Forgy. Of course, there could be a lot of other methods to initialize the k-means algorithm and select the centers, but we restrict it to the most known ones with programmatic implementation. Despite they include a random process, it does not alter significantly the results due to the simplicial complex design. The hierarchical clustering is also an option of MultiNet but we do not recommend it due to its slow computation.

The two main parameters for joining nodes with common points or high correlation (λ and δ), should be selected depending on the exigency level that we want for MultiNet.

We could define the described parameter selection analytically with automatic processes and mathematical analyses on the behavior of MultiNet with different parameters, but it could be an extensive research that would be out of the main scope of this work. It is definitively a topic that would deserve a post-investigation.

Chapter 8

Contributions of MultiNet to Data Analysis

What are the contributions of MultiNet? Could it serve as a diagnostic tool? Once we have described the functionalities of MultiNet we would like to present the contributions of MultiNet to the epigenetics study. Within this chapter, we will show MultiNet use possibilities, from the study of the methylation changes with aging, to the discovery of methylation markers associated to diseases like cancer. We aim to reflect that MultiNet is a powerful tool to detect epigenetic biomarkers in line with published results adding novel discoveries related to the methylation patterns and the correlations of thousands of variables. In addition, we will show how MultiNet can be applied to any other non-biological data type, as stock market data.

Generally, MultiNet has two main parts. A standard one that is based on the algorithm described in subsection 6.2.2, and a specific part that depends on the data type and the data interpretations we want to extract. In our case, we have programmed a complete MultiNet to deal with genomic data (following the steps specified in section 6.4), that produces all the figures presented in this work plus a spreadsheet with the DMSs and the corresponding genomic characteristics: hyper or hypomethylated state, p-values from statistical models, random forest Gini importance, related genes, chromosomes, genomic region, etc. The algorithm just requires the input parameters (described in chapter 7) and the classification of the sample groups (case vs. control or newborns vs. nonagenarians, for instance), if any. If we do not specify sample groups, the algorithm uses all the sample.

In the next sections, will show the potentiality of MultiNet presenting the contributions of the algorithm designed to the study of the epigenetic modifications with the aging process and cancer. Results will consist on the MultiNet networks and

all the biological information that we can extract from them, i.e. their associated differentially methylated sites and regions, and the study of those sites and regions. Moreover, the underlying correlation structure that defines the network's layout will be also studied.

In addition to those, we will present its potential use with distinct diseases (section 8.5), and in a different research field (section 8.6) to show its flexibility.

8.1 MultiNet by Age Sample Groups

As we described previously, there is an epigenetic modification as we age. We aim to use MultiNet to detect the methylation patterns associated to the aging process based on different methylation levels related to distinct age ranges. Lot of research was published during the last years in this regard, with the common basis of what they call "the epigenetic clock" or "biological age", but also with a great heterogeneity of results depending on the data, tissue or type of analysis [170, 171].

The use of persistent homology helped us to understand better the local evolution of the methylation marks interactions in the part II of the present work with the SBM-D design. Now, MultiNet provides with a wider overview of the methylation patterns potentially behind aging and longevity.

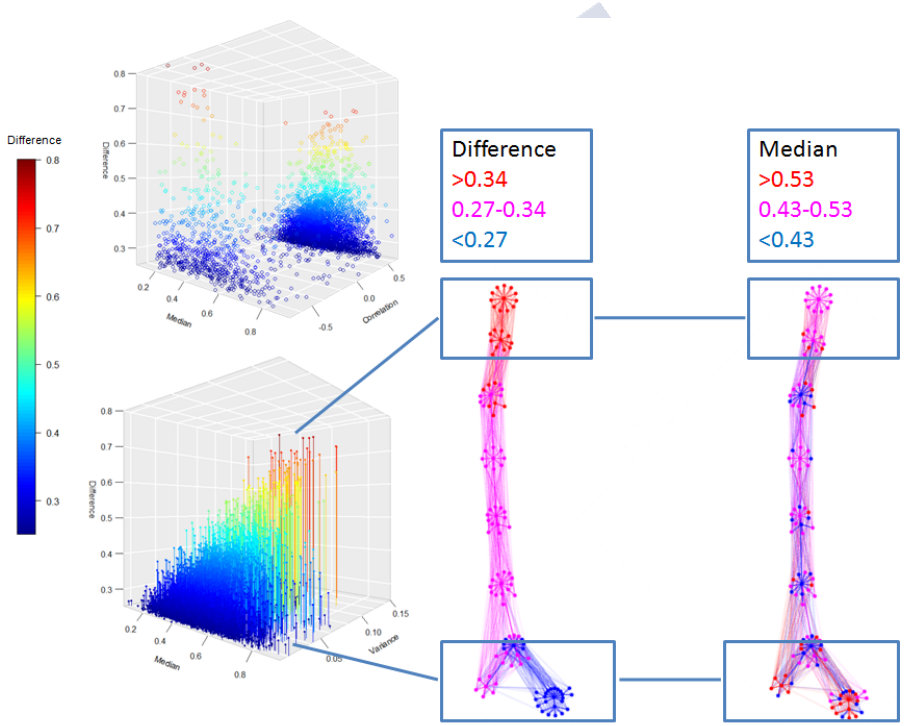
We use several blood arrays to detect the DNA methylation modifications associated with the aging process: GSE30870 containing methylation data from newborns (or case group) and nonagenarian (or control group) people [172], GSE36064 including children sampling (1-16 years old), and GSE33233 that contains methylation data from a third group of middle age subjects (39-72 years, average 59.8). For validation purposes, we use GSE40279 containing data from whole blood of 656 human individuals, aged from 19 to 101 [173], and GSE34639 as an alternative dataset of newborns with distinct cell types. In addition, we test two other datasets, GSE41826 and GSE66351, based on methylation levels taken from multicellular and unicellular brain tissue. Please remind that all those datasets can be downloaded from the NCBI web page [6].

If we detect the epigenetic modifications related to age, we are nearer to a better understanding of the aging mechanisms. In addition, we should also be able to identify and predict those epigenetic marks associated to age-related diseases [174]. For that reason, we also test our algorithm on diseases for which age is a main risk factor: cancer (prostate and colorectal). Please note that the comparison among the results obtained with MultiNet and the CpG sites considered as "CpG clock sites" should be done carefully as those studies were not always using Illumina 450K methylation arrays and had a different study aim that was the heavy correlation with the chronological age [175].

As a first application, we analyze the results obtained with MultiNet using a subset of 5,000 CpGs from the array GSE30870 separated in newborns and nonagenarians sample groups. In particular, we select the first 5,000 CpG sites more differentiated (higher absolute difference of medians) among both sample groups. The first outcome we obtain from MultiNet is a graph for the total population (please see

the figure 8.1), where each node is a cluster of CpG sites grouped based on their correlation levels. The filter functions are the median methylation by CpG and the difference of the median methylation between newborns and nonagenarians. We have also included in the same figure a 3D plot showing the data cloud based on the filter functions and the median correlation (for the first 3D plot) and the filter functions plus the variance (in the second 3D plot). The first step is to link the "shape" of the data cloud to the "shape" of the MultiNet graph obtained.

Figure 8.1: 3D representation of the selected filter functions plus the median correlation and the variance, and the associated MultiNet overall graph colored first by the difference of the methylation medians for the CpGs in each node and then by the median methylation of each node.



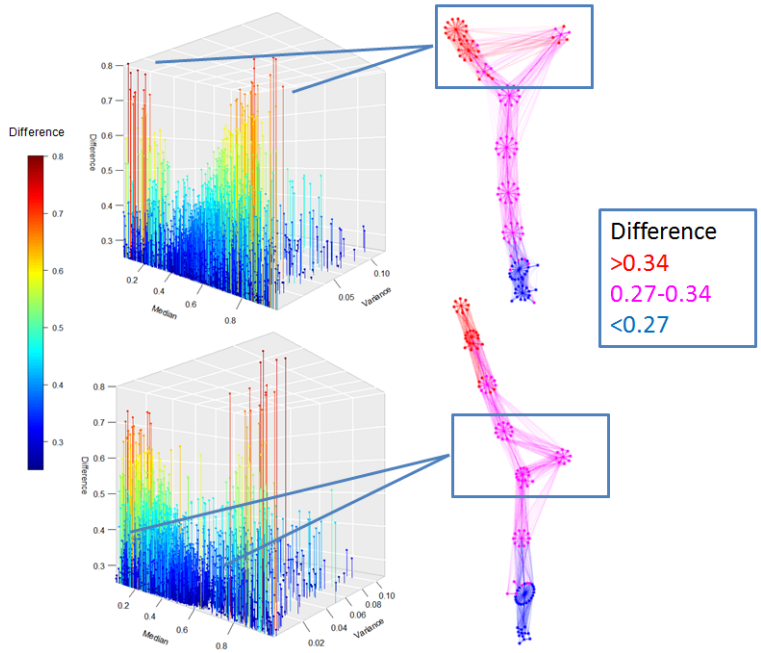
The first 3D plot in 8.1 allows us to see how the data is highly positively and also negatively correlated within the window selected. The distribution of the median values is centered at 0.5 with differences among newborns and nonagenarians up to 0.8. This distribution is indeed reflected in the associated MultiNet graphs of the same figure 8.1, where the algorithm creates two inferior "arms" correspondent to the methylation levels of the less differentiated sites. From the node colors of the two graphs we can indeed see that those with a higher difference among samples are those with a medium methylation median, while the lower differences present a higher heterogeneity of methylation levels. The network describes then correctly the data cloud and its characteristics.

The graph in 8.1 has 120 nodes (10 windows ×4 intervals ×3 clusters per interval) and around 1,200 edges. Please note that the main parameters used were: a windows size of 1,000 CpGs with 50% overlap, 2 intervals per window with 50% overlap, 3 clusters per interval, and "Forgy" as the method for the k-means cluster. The node joining parameters are $\delta = 0.8$ (for the correlation) and $\lambda = 0.2$ (for the points in common).

The calculus of the correlation matrix of those 5,000 CpG would be computationally possible, obtaining a greater mean absolute correlation for the younger group. However, MultiNet provides with a wider data analysis tool extracting more information about the "shape" of the data cloud and the underlying data distribution.

As we are considering newborns and nonagenarians, the structure just saw in figure 8.1 may be caused by the differences of those populations. Therefore, we also study the distribution over each one of the samples applying MultiNet to each group, as presented in figure 8.2. In this case, we use as filter functions the median and variance levels per CpG site.

Figure 8.2: 3D representation of the filter functions plus the difference of the methylation medians for case and control samples, and the associated MultiNet graphs colored by the difference of methylation medians per node.



The obtained graphs allow us to study the correlation structure within the different sample groups and present indeed different layouts, as we will confirm with the comparison of the persistent homology study presented in the figure 8.4. Networks properties as centrality or modularity are similar, as they are designed to use the same MultiNet parameters in order to be comparable. They have 120 nodes with a different number of edges: 1046 vs. 1037 for newborns and nonagenarians respectively. As we are joining nodes with high correlation values or a high percentage of points in common, this difference could indicate a less sparse or more correlated methylation distribution in the newborns' group.

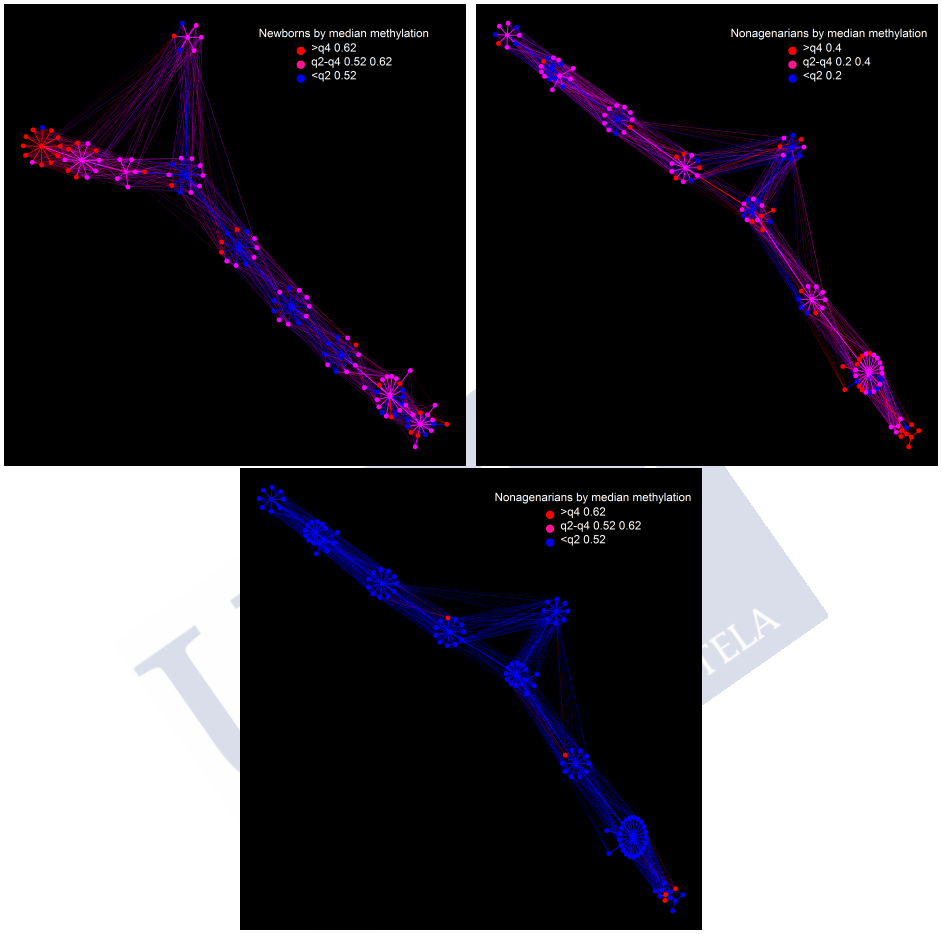
What is indeed different is the distribution that those graphs represent. If we calculate the probability distribution of both networks, as explained in subsection 6.5.1, they result to be significantly different with an effect size calculated by Cliff's delta near to 1. The random selection of the samples results in similar network distributions, with a Cliff's delta effect size near to 0.

Based on the quartile distribution specified within the figures in 8.3, we observe lower and more variable levels of methylation in the nonagenarians group. The most differentiated nodes correspond to higher levels of methylation in the newborns' group and medium/lower levels for nonagenarians. The difference between both methylation distributions is clearer when looking at the second row of figure 8.3, where we have colored the nodes of the nonagenarians graph based on the quartile distribution of the newborns' graph. We can see there that the nonagenarians present clearly lower methylation levels. This is in line with literature and the hypomethylation trend with increasing age.

Therefore, the MultiNet graphs represent a useful tool of dimensionality reduction of the big dataset, and serve as a guide to extract much more information in a low computational time. Other common techniques of dimensionality reduction, as PCA or MDS, would not give us that wide analysis.

The study of the colored graphs based on functions of interest allows us to understand the methylation distribution and to detect different patterns among the sample groups. From those colors, we will extract all the biological information of interest as DMSs, SMSs, HCSs (please refer to section subsection 6.5.3 to remind their definition). This post-processing is described below, after the study of the MultiNet graphs layout with persistent homology.

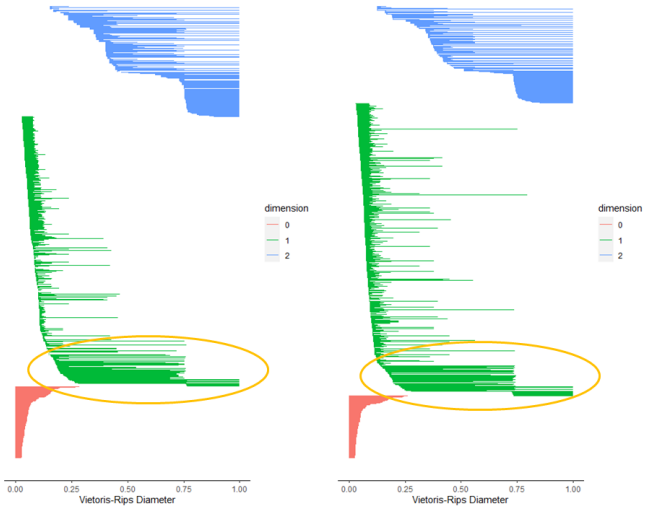
Figure 8.3: Newborns (left) and nonagenarians (right) MultiNet graphs colored by the difference of medians within each node and the median methylation. The graph in the second row indicates the nonagenarians MultiNet graph colored by the quartiles of the newborn's distribution.



8.1.1 MultiNet Topology with Persistent Homology

The weighted graphs presented in 8.3 are really simplicial complexes, so we could also extract information from their topological characteristics. In order to detect interesting topological elements automatically, we apply persistent homology to study the topological features hidden in the MultiNet graphs for their posterior interpretation. The weights of the networks are given by MultiNet, calculated as the mean absolute correlation between the connected nodes. PH with a Vietoris-Rips filtration over the inverse of the weighted adjacency matrix is then applied. We assume then that the nodes that are not joined have a null correlation coefficient. The comparison of the obtained PH results should be done carefully as both MultiNet networks may not have the same number of nodes (due to the MultiNet clustering flexibility).

Figure 8.4: Persistence barcodes of newborn and nonagenarian MultiNet graphs.



In the case of the studied MultiNet networks for newborns and nonagenarians (with 120 nodes each), we observe a difference in the number and lifespan of the topological features found, that are higher for the oldest group in H_1 (445 vs. 557) and similar in H_2. Thus, those differences demonstrate that the graphs are distinct and so also their underlying data distribution. As the newborns' graph presents a higher number of edges and a slightly higher edge-density, there are less topological elements and they generally last a lower time.

There are several common features that are born at a VR-weight of 0.25 and remain up to 0.75 approximately (marked in yellow in the figure above). Those elements represent holes in dimension 1 relatively stable in both networks. Their lifespan is high because we forced the algorithm to join nodes only if they have a high correlation (greater than 0.8). These barcodes represent, therefore, highly correlated clusters of CpG sites.

Generally, the topological elements that we may find in MultiNet graphs would depend on the underlying data distribution and correlation. If we select a lower δ, the number of edges between nodes will be higher, creating more complex topological structures but also reducing the lifespan times of the related homology elements. If we select a restrictive threshold δ for joining nodes only if they have a great mean correlation, then the elements generated with high correlation values would not be broken due to the own MultiNet design. It occurs similarly with the parameter λ.

Therefore, studying the topological features from persistent homology permits to detect sample differences basically based on the correlation levels of the CpG sites selected.

8.1.2 Biological Information from MultiNet Graphs

The MultiNet networks obtained are just the first step of the analysis, and we could extract a lot more information from them. The different functions used to color the MultiNet graphs allow us to identify different types of interesting sites to be explored, as DMSs, HCSs or SMSs.

Only with a window size of 5,000 CpG sites correctly selected by the maximum methylation difference among newborns and nonagenarians, we already obtain interesting results that will be afterwards extended in a wider window. The adequate selection of the filter functions was key to obtain substantial results from data.

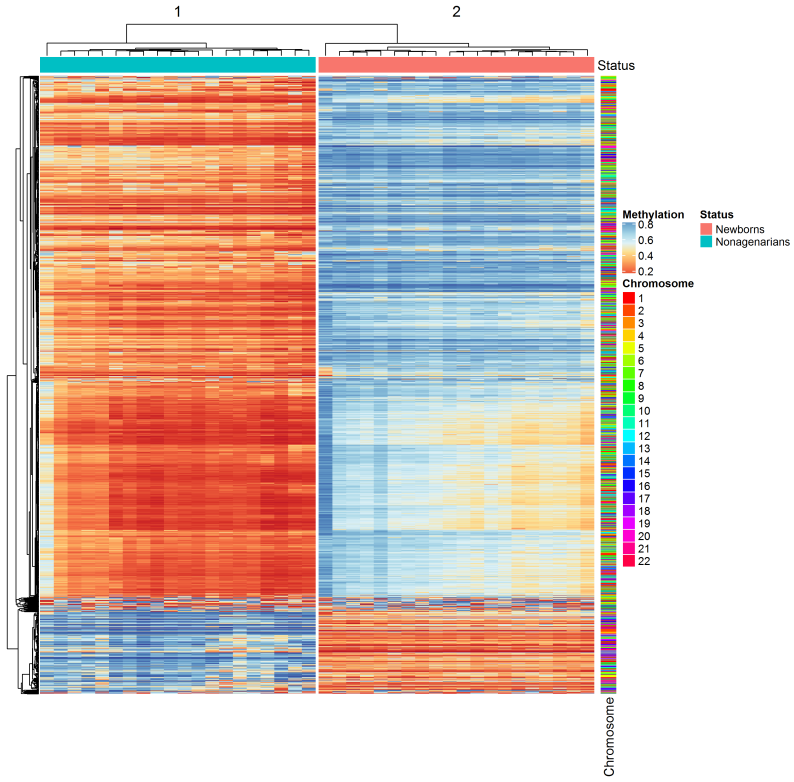
In this case, we find more than 1,800 DMSs contained on the red nodes of the overall network colored by the difference of medians. Please note that those nodes are clustered based on their correlation, so we are indirectly selecting sites very correlated. Most of them belong to "OpenSea" genomic regions, and are located in different chromosomes as 1, 6, 2, 3, 17, 11, or 7. Please note that the 450K array contains also more "OpenSea" regions (they are the 36%), followed by CpG islands (around a 31%). The results from a penalized logistic regression applied to each CpG reduce this number to almost 100 significant DMSs with a SGoF adjustment ($\alpha = 0.05$ and $\gamma = 0.05$).

With the methylation heatmap 8.5, we confirm the hypomethylation trend in the older sample group, except for some sites where the pattern is the opposite, in concordance with literature. The significant hypermethylated sites in nonagenarians mostly belong to CpG islands, while the hypomethylated are in "OpenSea" regions. If we select randomly the sample groups, the results of the logistic regression are not significant as the graphs does not allow us to detect truly differentiated sites.

The 100 significant sites and their related genes are mostly associated with the aging process in literature. Some known age-related methylation genes found are KCNAB3 [176], DDO [177], FHL2 [178], SLC12A9 [179], OXT [180], SNED1, or TTC22. However, we also find other genes without a clear association in literature, as NRP1, PFKP or MGP.

We tested the results on a different dataset with 656 individuals of ages 19-101 (GSE40279) to see how well the selected CpG sites could predict the sample age. We apply a linear regression model over the age variable taking into account the 100 sites as independent estimators and 300 individuals as a training set. Around 20 of the 100 variables were significant (p-value < 0.05) and with those 20 we obtained a correlation greater than 0.8 between the predicted age (for the rest of the sample) and the biological age of those subjects. The average difference among both ages was around 5 years. The selected sites correspond to genes like KCNAB3, FHL2,

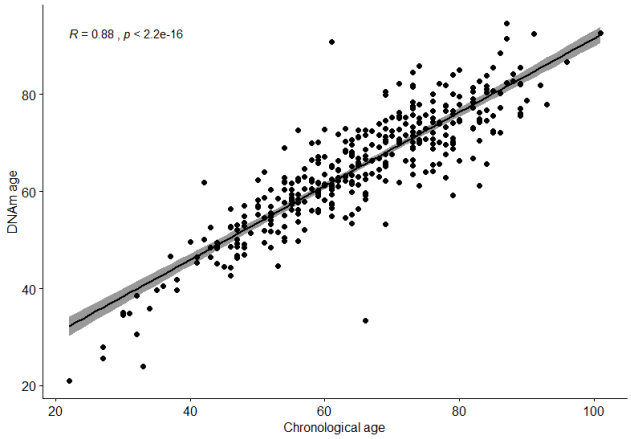
Figure 8.5: Methylation heatmap of the more than 1,800 DMSs. The columns represent the sample and the rows the CpG sites. They are ordered by the results of the hierarchical clustering on rows and columns.



HOXB6, FOXI2, MAS1L, PAK6, SLC12A9, TRIP6, EBF4, PITPNC1, LRFN1, or EDARADD.

We also colored the sample graphs by the absolute median correlation of each node and study the CpGs contained on the nodes with higher values (absolute correlation greater than 0.8). We observed then that the CpGs contained on those nodes are highly correlated, mostly positively correlated, within each age group. The majority of the CpGs were the same for both groups (more than 1,300, mostly in "OpenSea" regions), meaning that the correlation structure of those sites is not significantly altered by the age but they may develop common biological functions. They belong to different chromosomes and could be named highly correlated sites (HCSs). The genes more present as associated to HCSs were CSTA, MS4A3, or GPR109B. The protein network obtained in STRING with those genes is indeed significant, mainly associated with the positive regulation of metabolic process and with KEGG/Reactome pathways associated to the immune system.

Figure 8.6: Correlation between the chronological sample age and the estimated age through the linear regression model over the 20 significant sites.



The CpG sites contained on nodes with similar methylation values are composing the group of similar methylated sites (SMSs). They reflect behavioral patterns of methylation trends. For instance, sites related to HLA, OR, or PCDH gene families were found to be highly methylated in the newborns' sample (some of them significantly differentiated from nonagenarians).

8.1.3 The Addition of the Children and Middle-Aged groups

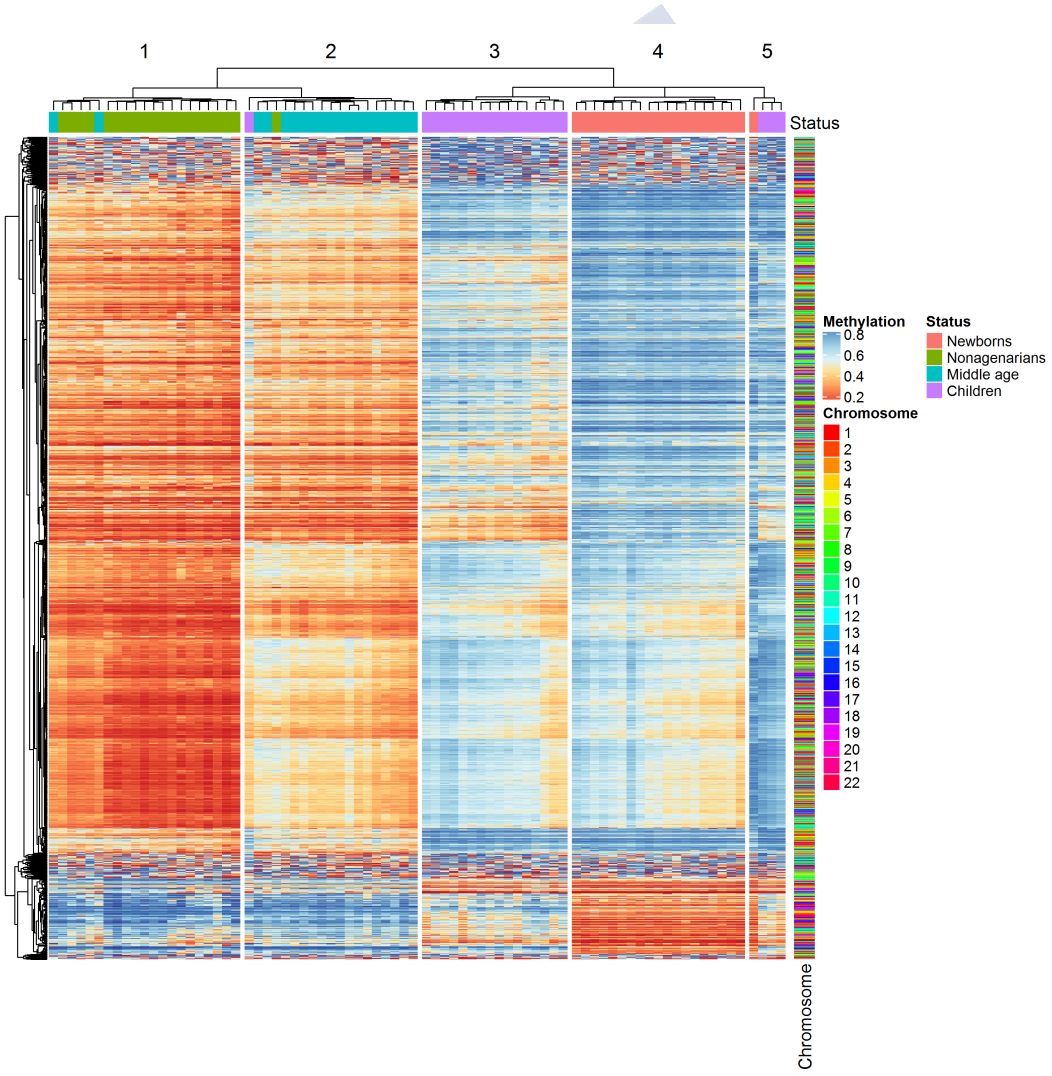
MultiNet allows to separate the sample into more than two groups. However, the computational time and the comparisons may become more difficult to manage. Another option is doing it through the natural differentiation of the methylation patterns establishing only two initial groups, as we will show here.

For example, we could extend our aging analysis above including children between 1-16 years (GSE36064, peripheral blood mononuclear, selection of 20 individuals), and middle-aged individuals (GSE33233, whole blood, 39-72 years). Instead of generating four different MultiNet networks, we group newborns and children versus middle-aged and nonagenarians. It is important to use similar sample sizes in order to be able to interpret correctly the results found in terms of correlation. The algorithm takes the highest differences among the two, and generate the corresponding networks in the same way as before. The layout of the networks changes accordingly, even using the same MultiNet parameters.

We obtain a similar quantity of DMSs (almost 2,000), being around 1,500 significantly different among both sample groups with the same adjusted penalized logistic regression model. Among them, we find the sites detected with only two

groups plus novel ones related to genes as ALDH1A1 or CSRP3. The hierarchical clustering presented in the heatmap 8.7 of DMSs differentiates the four underlying sets with clear distinguished patterns among the youngest and the oldest groups. The hypomethylation trend with age is kept, with a progressive evolution to a less methylation status for older groups. There is also a main distinct area where the newborns present low methylation levels but they increase considerably already in the children group. This allows to distinguish two groups of children, also in line with literature and the variability of the methylation statuses found for children samples. Interestingly, the areas of hypomethylation in the younger groups correspond mostly to CpG islands.

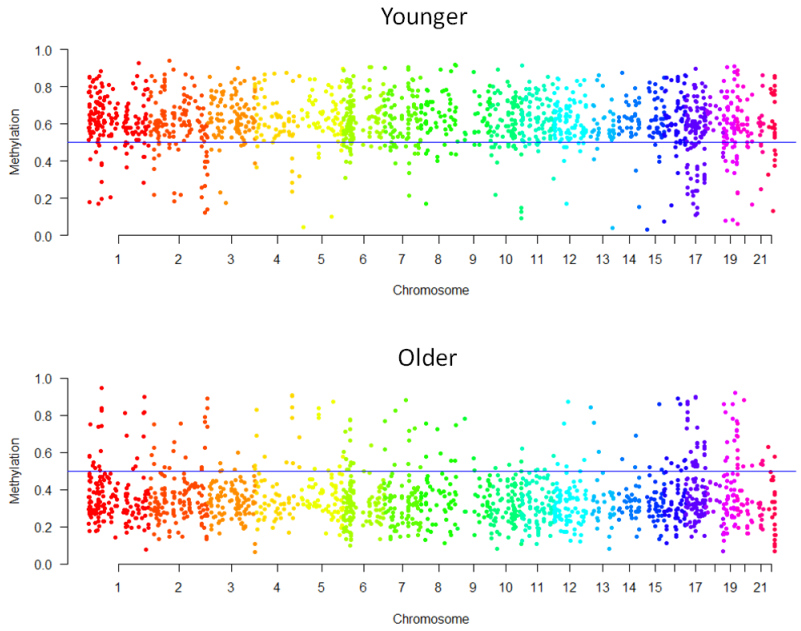
Figure 8.7: Methylation heatmap of the DMSs found by the four age groups.



In the manhattan plot 8.8, we observe better the distribution of the significant

sites by chromosomes.

Figure 8.8: Manhattan plot of methylation levels by chromosome for both sample groups.



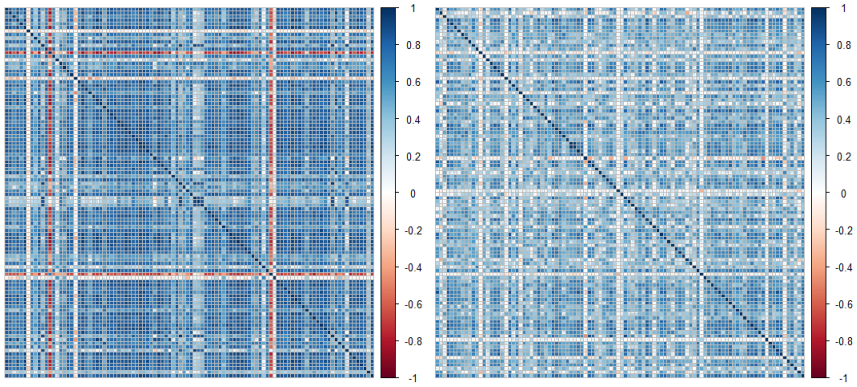
The sites significantly differentiated among both age groups are associated to different genes as the ones presented for the previous study with newborns and nonagenarians. They are also related to regulatory disease pathways associated to aging, as arthritis, cancer or heart problems. The sample classification prediction with the most 100 relevant CpGs (among the significant ones) is successful using a different children population from the same dataset GSE36064 and a new newborns sample (GSE34639, cell type CD4+).

We find indeed 12 sites among the ones selected as age-related by Horvath: cg24550149, cg22242842, cg25105522, cg13072943, cg05451210, cg09655403, cg05922714, cg07593390, cg25456477, cg13120986, cg04098052 and cg10553204.

The relevant sites selected are more correlated in the younger groups, as seen in 8.9 confirming a stronger correlation design for younger individuals.

Some of those relevant sites are also significantly associated to age (linear regression p-value<0.05) with methylation levels taken from brain tissues in multicellular or unicellular designs (GSE41826, GSE66351), despite the correlation between the DNAm age and the chronological age is generally lower than in blood samples (around 0.6). Moreover, the methylation patterns per group are not the same as in

Figure 8.9: Correlation matrix of the most relevant CpGs by younger and older groups.



blood, and therefore the prediction of the groups does not work correctly.

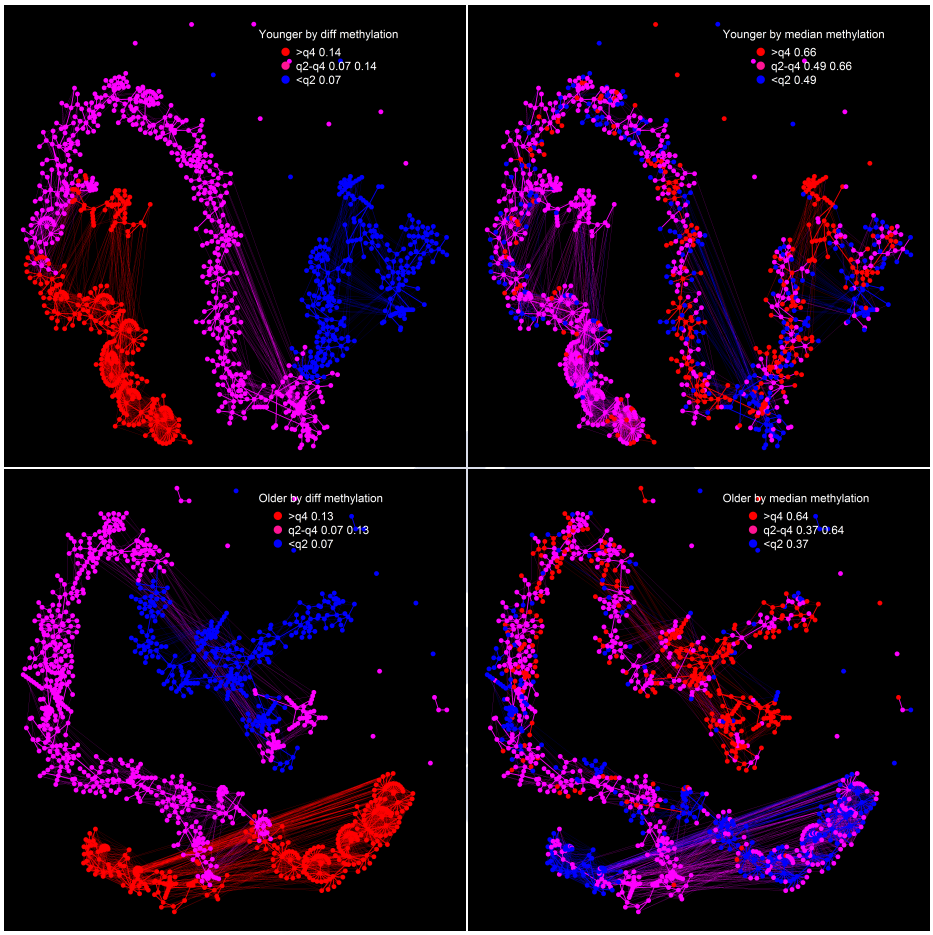
As a conclusion, there is an association among the detected CpG sites methylation levels and the individual's chronological age, but their methylation is not following the same trend for different sample tissues. MultiNet would allow us to extend this insight by applying it to more tissues.

Results with a Wider Window

We extend now the prior analysis with four age groups divided into two to a window of 50,000 CpG sites ordered by the maximum difference among the median methylation levels in both groups. We could use the entire array too, but the filter functions allow us to reduce the computational time and analyze the most interesting sites for sample group differentiation. At this step, the calculation of the raw correlation matrix would be already computationally heavy and uninterpretable, while MultiNet produces the three networks in less than 20 minutes. In this case, both networks (please see 8.10) present a distinct quantity of nodes (1,103 vs. 1,178), and edges (3,004 vs. 3,500) for the young and old groups.

The algorithm detects 13,001 DMSs, with more than 2,000 significantly differentiated (they increase up to almost 6,000 specifying $\gamma = 0.1$ in the SGoF adjustment). As we have seen previously, we observe indeed a general hypomethylation trend over the increasing age groups. Around 90% of the significant sites found are hypermethylated in the younger group and belong to an "OpenSea" genome region, while the hypomethylated sites are located mostly in CpG islands. Besides, most of the sites found are located in gene body regions.

Figure 8.10: MultiNet younger (first row) and older groups (second row) graphs colored by the difference in median methylation at left and by the median methylation of each node at right.



Among the genes found, we have well-known genes related to aging or age-related diseases, some of them also detected with MultiNet over 5,000 CpGs. Indeed, we find more (around a 10%) of the sites specified by Horvath [52] and Hannum [173] (more than 50%). They are, for instance: ELOVL2 [178], FHL2 [182], or OTUD7A [183]. Specifically, some of the CpG sites with higher methylation differences (hyper or hypomethylated in the younger group) are associated to age-related genes as SNED1 (hypo), SLC12A9 (hyper), GNG7 (hyper), CHST3 [184], PDCD1LG2 [185]. Some genes as PLCH1 were also specified in the original data articles or other ones related to aging [186, 187] or neurodegenerative disorders [188]. We also find around 20% of the genes specified in the platform GenAge [189], that collects 308 genes related to the aging process: PTEN, IRS2, CETP, MAP3K5, etc. The same platform has a dataset of longevity-related genes and we also find around 20% of them with our algorithm: some HLA-family genes, IL-family genes, MTHFR or CETP genes are

some of them. Other genes as OSGIN1, CD300LB, or ZPLD1 were not found as related to aging in literature.

Among the hypermethylated sites in the older groups (mainly in CpG islands), we find a lot of related genes of the protocadherins family, PCDHG, which were related to epigenetic dysregulation in multiple ways and have been observed in a variety of different disorders that are directly or indirectly influencing brain structure and function [181].

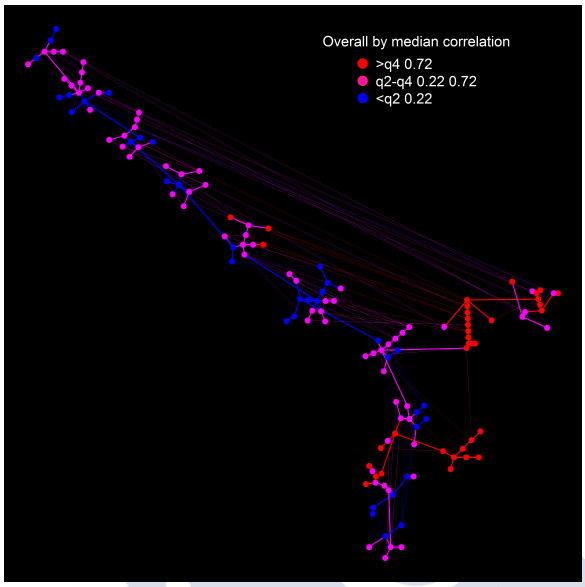
8.1.4 Local MultiNet

Once we have studied the DNA methylation behavior for different age groups globally, we should do it locally to obtain a magnifying glass effect. The identification of differentially methylated sites should be then completed by the identification of differentially methylated regions (DMRs) within chromosomes, that may be overlooked when we analyze the data from a global perspective. Therefore, we use the information obtained from the global MultiNet to detect chromosomes where the methylation variations among both sample groups were higher. We observed that, for example, chromosomes 22 or 7 contains quite a lot of the detected significant DMSs.

Despite the local differences in the median methylation between the sample groups are lower than in the global analysis, some sites were still significantly differentiated. For instance, analyzing the chromosome 22, we obtained that the gene C22orf26 is heavily methylated in the younger group. Please note that in this case the windows were restricted to the chromosome 22 and the filter functions were the genomic position of the sites and their median methylation. In order to validate the local results, we used the *limma* R package over the same chromosome and it found as significant all the ones that we obtained.

Alternatively, we could use the local MultiNet to explore the correlation behavior of a specific sample group and detect HCSs, SMSs, or differentially methylated regions within that group. For example, we could apply MultiNet to the entire children sample and the CpGs from the chromosome 22. If we color the network by the absolute median correlation matrix value of the CpGs per node and select those with a higher correlation (please see figure 8.11), we obtain sites highly locally correlated (mostly hypomethylated) and matching some of the sites found in the subsection 5.2.1, with genes associated with the 22q11.2 deletion syndrome, plus other related genes as MAPK11, MAPK12 and MAPK13. This analysis is very quick, as we obtain the network and the interesting sites in less than 5 minutes.

Figure 8.11: Local MultiNet graph using as filter functions the median methylation level and the CpG sites position, colored by median absolute correlation per node. The sites selected are the ones contained in the red nodes.



8.1.5 Summary

We have summarized the biological information that comes out from MultiNet. Nevertheless, a deeper analysis of those results may deserve an ad-hoc investigation that is not part of this work aim.

As a conclusion, MultiNet is able to reduce the dimension of the dataset and quickly extract relevant biological information from the graphs produced. We detected epigenetic marks globally and locally that may collaborate in the aging and longevity processes and their associated DNA methylation patterns, with a clear trend to the hypomethylation as we age in line with published investigations. The fact that this trend is altered for CpG sites located in CpG islands (presenting higher methylation levels for the oldest groups) leads to consider those regions as interesting epigenetic traits related to the aging process that could be further investigated. The increase of methylation on those sites could be altering then the related gene expression leading to a potentially non-expected gene silencing.

In the next sections, we apply MultiNet to a prostate cancer dataset and a colorectal cancer dataset in order to see how the algorithm is able to provide substantial results for different types of data. Moreover, as both are age-related diseases, we analyze the common epigenetic marks found among the aging study and the cancer study.

8.2 Prostate Cancer

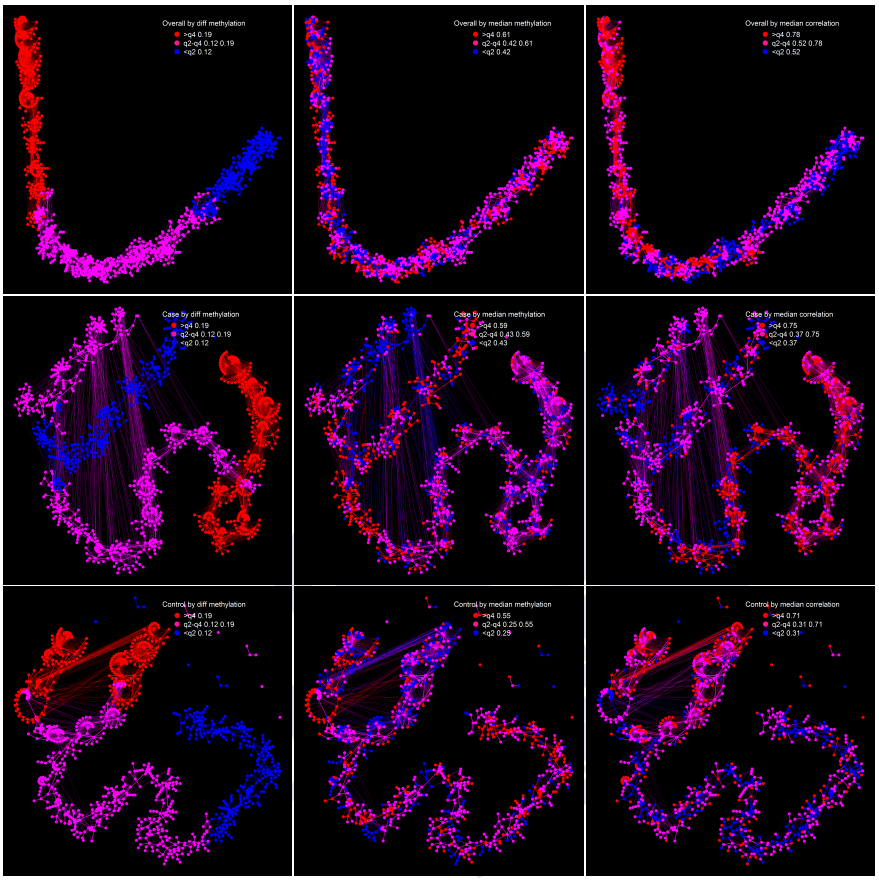
As we described in the first chapter of this thesis, some of the genetic and epigenetic alterations in cancer subjects are known, despite the reasons are not totally understood. A lot of the ongoing cancer research is then focused on finding the genetic and epigenetic biomarkers that make possible a premature cancer diagnosis from a blood sample (or other tissues), in order to be able to prevent the disease even before it is manifested. We focused the research in two different cancer types: prostate cancer and colorectal cancer. Of course, the flexibility of MultiNet would allow to apply it to other cancer types too. We tried to use recent public methylation datasets with a reasonable high sample size using Illumina Human Methylation450 BeadChip (450k). Additionally, an analysis of the common epigenetic biomarkers among those two cancer types and age-related ones will be presented in section 8.4.

Methylation biomarkers have been identified in prostate cancer previously but their efficiency can be improved and extended. The results we are about to show provide with a set of epigenetic biomarkers that may be helpful for this aim. We will use the GSE76938 dataset [190] for the prostate cancer study. It contains data from 73 prostate cancer tissue and 63 healthy tissue, 52 of which are patient-matched. In this article published in 2017, they concluded that DNA methylation patterns are very altered in prostate cancer tissue in comparison to benign-adjacent tissue. They developed a mixed model linear regression analysis and identified 226,235 CpGs with significantly different methylation levels in cancer tissues: $\sim$ 67% had increased methylation (mostly located in CpG islands) and $\sim$ 33% had decreased methylation. We use this dataset as an application of MultiNet to a specific disease, obtaining results in line with current publications and providing with additional findings.

Similarly to the aging study, we select here the 50,000 most differentiated CpGs and use the same filter functions and parameters than before. As observed in the graphs 8.12, the number of nodes and edges are now considerably high, which provides with more underlying information but, at the same time, makes more difficult the graph interpretation (case graph has 1171 nodes and 4270 edges, while control graph has 972 nodes and 3166 edges). The distribution of the control group seems to be less sparse and with lower methylation levels. Indeed, the case graph presented in 8.13 is much more "pink" when we use the quartiles of the control distribution as reference.

We find 13,000 DMSs clearly separated in the two sample groups by their methylation status, almost all of them statistically significant. Where control tissues are generally hypomethylated, the prostate tissue presents a trend to the hypermethylation, specifically for a group of patients. On the other side, where control is

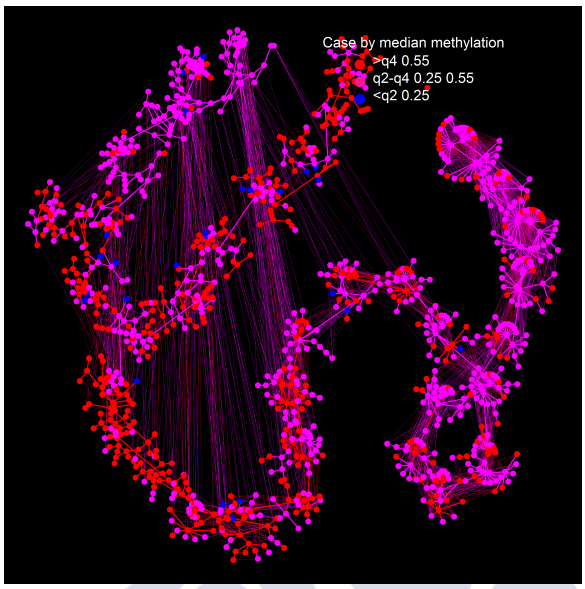
Figure 8.12: Overall (first row), case (second row) and control (third row) MultiNet graphs colored by the difference of median methylation levels (left), median methylation (middle) and median correlation of each node (right).



hypermethylated, the prostate group presents the opposite trend of lower methylation levels. Particularly, we find about a 70% of sites hypermethylated, really in line with the original data article [190]. Generally, the case group is more variable, with some subjects classified as controls, and two differentiated clusters 2 and 3 (see figure 8.14 (a)). This may be an insight on the distinction of case subgroups depending on their disease status or progression. Indeed, a common Mapper analysis over the same dataset with mean and variance as filter functions also differentiates the three clusters, with some case/control tissues sharing a more similar pattern. Among the significant DMSs, there is a higher percentage of hypermethylated case sites in promoter regions and CpG islands, as you could observe in 8.14 (c) and (d).

The genes obtained are in line with the ones found in the data publication in about a 50%. Some of the most relevant ones selected by the random forest algorithm are presented in the table 8.1 and the figure 8.14 (b) and (e). Particularly interesting are the matches with some of the genes from the HLA family, as HLA-F, HLA-G,

Figure 8.13: Case MultiNet graph colored by the median methylation per node with the quartiles of the control distribution.

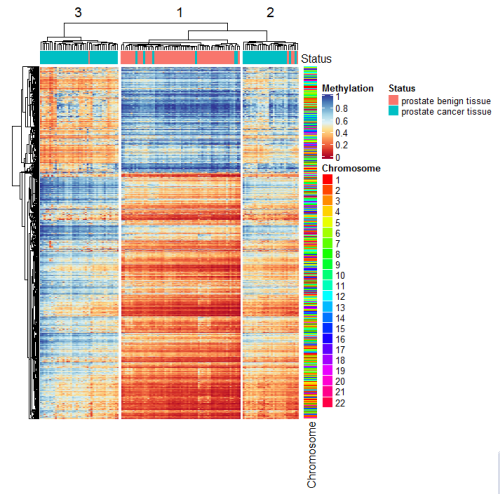


HLA-H, HLA-J, or HLA-L. We also find other relevant genes related to prostate cancer in different articles, as the family of RPS genes (like RPS15), SNORA10, WNT3 (currently associated with prostate cancer or generally with cancer [195]), or OXGR1 [196]. In addition, lot of genes from the olfactory receptor (OR) gene family are hypomethylated in the cancer group [197]. Most of the genes presented in [198] (published in 2019) were detected by MultiNet too as prognostic biomarkers for prostate cancer: DOCK2, FBXO30, GRASP, HIF3A, PFKP, and TPM4. In addition, some genes also match with the ones presented in the IntOGen web page that was released by beginning of 2020 [199]: APC (12 CpGs, 2 hypomethylated and 10 hypermethylated), LRP1B (3 hypo), ZFHX3 (3 hypo), FAT3 (1 hyper and 1 hypo), CDKN1B (1 hyper), FOXP1 (3 hyper), ALK (1 hyper), EHD2 (1 hyper), FAM174B (1 hypo), LZTR1 (1 hyper), FAT1 (2 hyper and 1 hypo), JAK1 (1 hyper). Those interactions indicate potential genes involved in genetic-epigenetic alterations.

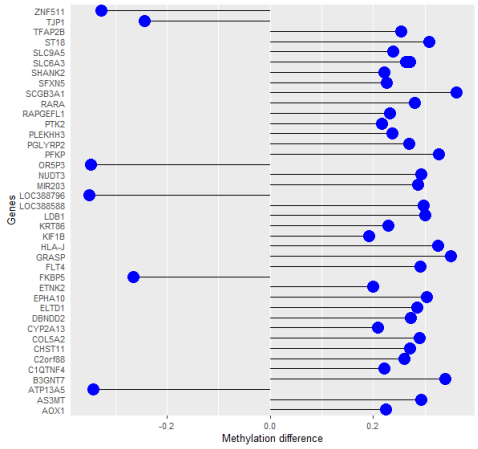
Doing an enrichment analysis on the set of significant DMSs to find common biological pathways and diseases for the related genes, we find quite a huge quantity of KEGG pathways commonly related to cancer as (see figure 8.15): Calcium signaling pathway, axon guidance, Wnt signaling pathway, PI3K-Akt signaling pathway, ECM-receptor interaction, Proteoglycans in cancer, Prostate cancer, Ras signaling pathway, Rap1 signaling pathway, etc. One may be surprised about the appearance of the human papillomavirus infection pathway, but this is fact in line with current literature [201]. On the other side, we find several diseases that share common genes

Figure 8.14: Some of the plots generated by MultiNet.

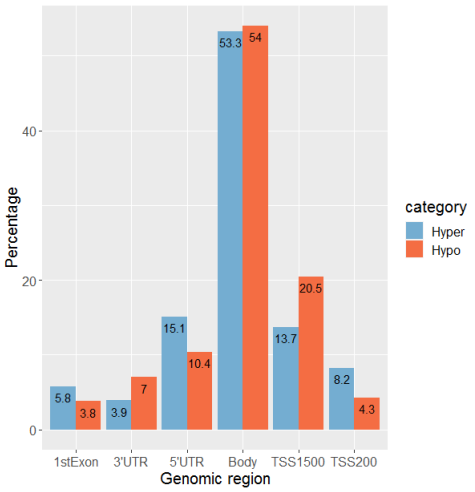
(a) Heatmap of DMSs.



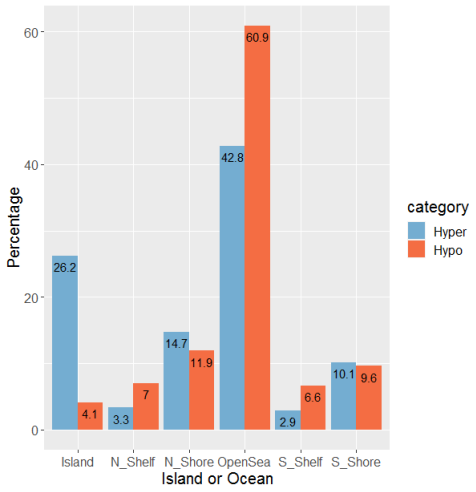
(b) Most relevant genes.



(c) Bar plot of genomic region.



(d) Bar plot of CpG island relatedness.



(e) Boxplot of some relevant genes.

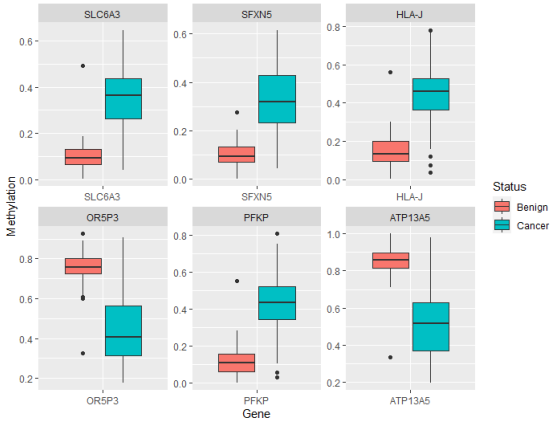
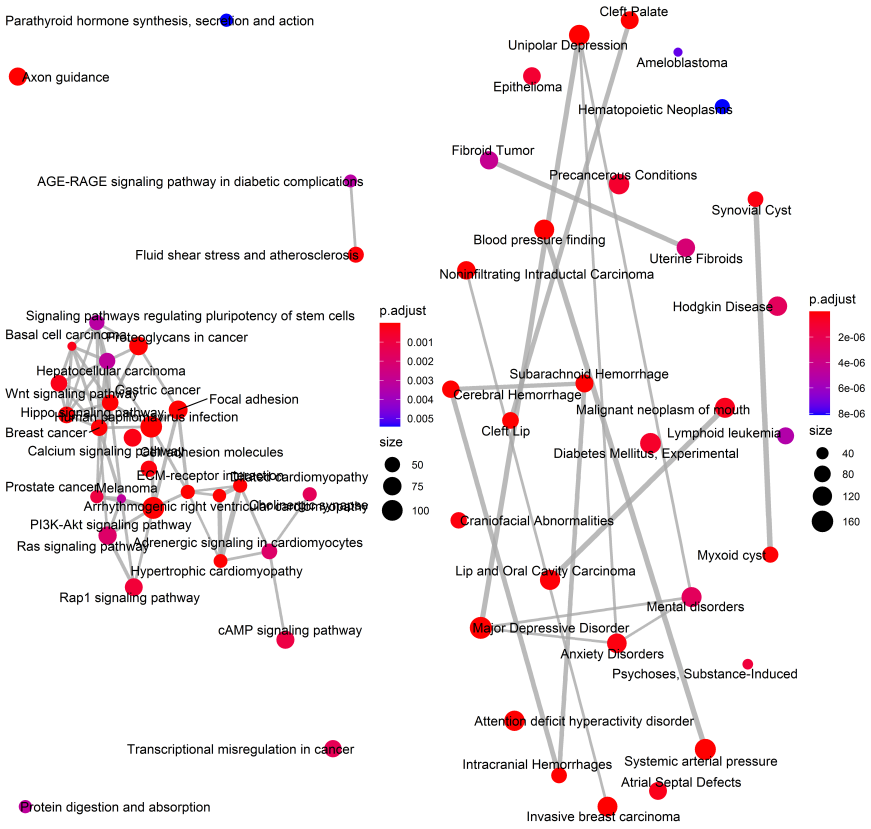


Figure 8.15: KEGG and disease enrichment networks.



with prostate cancer, as other cancer types, blood pressure problems, or some types of mental disorders. Figure 8.15 was created using the functions *emapplot* from the R package *enrichplot*, that uses the two enrichment functions *enrichKEGG* and *enrichDGN* from the packages *clusterProfiler, DOSE* [200] of R. This enrichment map organizes enriched terms into a network with edges connecting overlapping gene sets and colored by the adjusted p-values derived from an over representation test and adjusted for multiplicity [200].

The application of local MultiNet on chromosome 17 (containing more than 14,000 CpGs) detects new genes that were hidden with the global design, as PT53 related to the site cg07760161, that is one of the hub genes associated to cancer.

We then use the DNA methylation array coded as GSE84749 that contains 24 samples of prostate cancer subjects [192], whose are correctly predicted (more than 80% of success) as cancerous status using ten of the most relevant CpG sites classified by the random forest algorithm and included in table 8.1. Other relevant genes as SLC6A3, C1QTNF4, B3GNT7, ETNK2, KRT86, or RARA are also good

Table 8.1: 10 of the most relevant CpG sites selected.

CpG ID	Gene; Region	Recent Publication
cg07198194 (hyper)	PFKP; TSS1500	2019 [191]
cg21523564 (hyper)	<NA>	<NA>
cg23396786 (hyper)	SFXN5; TSS200	2017 [192]
cg16107322 (hyper)	<NA>	<NA>
cg06092265 (hyper)	<NA>	<NA>
cg01748263 (hypo)	ATP13A5; TSS1500	2017 [193]
cg15726260 (hyper); cg16794576 (hyper)	HLA-J; Body	2017 [190]
cg09729613 (hyper)	AOX1; TSS200	2018 [194]
cg04178787 (hyper)	<NA>	<NA>

sample predictors. Indeed, we share about 50% of the genes found in [192] where they analyze whole genome methylation and expression profiling. Please note that the published references specified in this table may not be related exclusively to DNA methylation studies, but could contain other investigations where those genes appeared as associated to prostate cancer.

8.2.1 Summary

Overall, results indicate an epigenetic modification between adjacent tissues that could be valid for disease diagnosis. Indeed, the accuracy of sample prediction is high in different cancer cohorts using only 10 of the detected biomarkers. We detected a hypermethylation trend in the cancer samples, expect for specific sites where we found the opposite pattern. Among cases, we were able to detect a different methylation behavior, potentially caused by a different disease status.

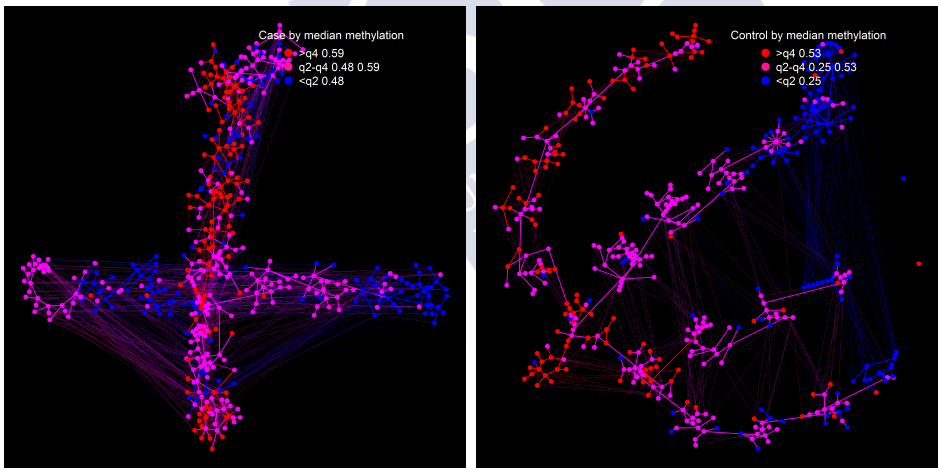
Those results are in line with the original publication, extending it with a wider analysis of correlation and from a global/local perspective. We have found site-related genes in line with literature but also other non-studied ones that may deserve an ad-hoc investigation with greater sample sizes. For instance, the fact that a lot of olfactory receptor genes presented lower levels of methylation in the cancer samples might be pointing out the importance of this regulatory gene network in prostate cancer. In addition, the huge correlation found among several CpG sites could be an indication of epigenetic networks related to cancer epigenetic pathways.

8.3 Colorectal Cancer

As a validation method, we also use MultiNet with a different DNA methylation dataset from colorectal cancer (CRC) and healthy patients. CRC is a complex disease dominated by genetic and epigenetic alterations. We will use the GSE48684 dataset [202], published in 2014 and containing 41 normal colon samples, 42 colon adenomas, and 64 colorectal cancers. The original article presented three classes of cancers and two classes of adenomas based on their DNA methylation patterns. This suggested a variability in the pathogenesis of colorectal cancer from the adenoma step.

Using MultiNet, we obtain results in line with the original publication and other papers related to CRC. Please note that we consider as case samples the subjects with colorectal cancer, while controls are the two types of normal samples (64 vs. 41).

Figure 8.16: Case (left) and control (right) MultiNet graphs colored by median methylation per node.

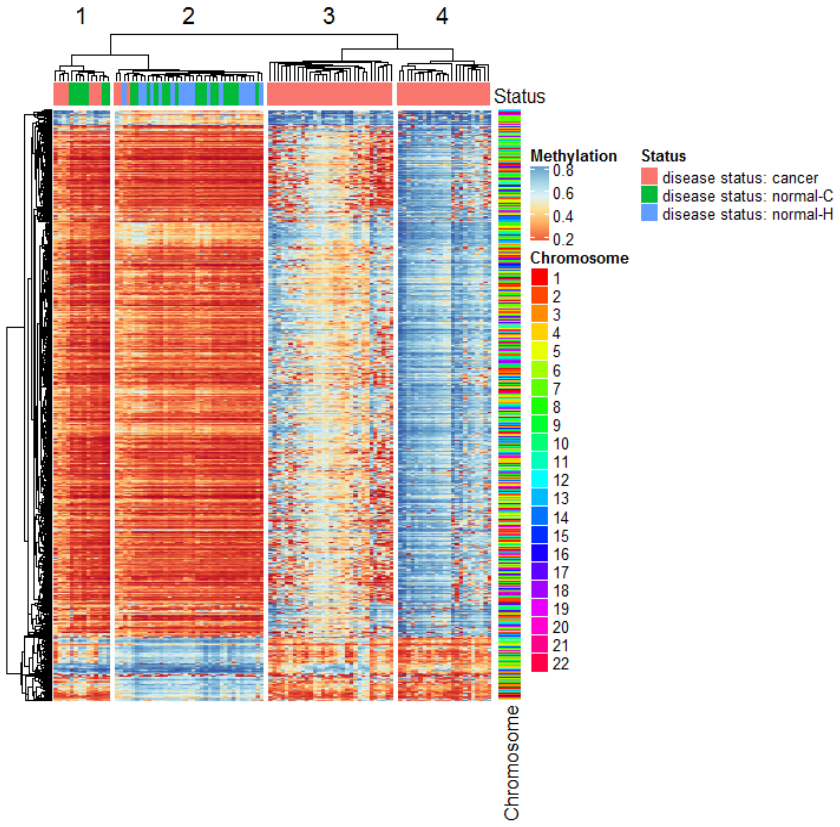


We select the 20,000 most differentiated CpGs among both sample groups for the analysis and obtain around 5,500 DMSs (presented in figure 8.17), where almost all of them are significant and greatly correlated. The case samples are mostly hypermethylated with some CpG sites values hypomethylated too. In addition, we find three different types of case samples: cancer samples with mostly low methylation levels that are hypermethylated in some sites too (this group has a similar behavior to the control samples), cancer samples with medium methylation levels that are also hypomethylated in some sites, and cancer samples with high methylation levels. With regard to the two control groups, the differences are less clear.

Most of the differentiated CpG sites that are hypermethylated in the case group belong to CpG islands. If we add adenoma samples to the analysis, we obtain a

bi-directional prediction, with some (most) adenomas predicted as case samples and others as control samples. It may indicate a progressive change to the tumor status in methylation levels terms for some patients.

Figure 8.17: Methylation heatmap of DMSs by sample groups.



We find sites related to genes specified in the original publication and in other papers, as: SND1, OPLAH, C1orf70, GATA2, ADAMTS5, FOXD3, FOXF2, GNAO1, SLIT3, MAD1L1, or GRIA4 [203, 204, 205, 206, 207, 208]. Specifically, OPLAH cg26256223 hypermethylation is associated with reduced gene expression in the CRC. On the other side, is interesting to see which are the most frequent genes altered for hypo or hypermethylated CpG sites. For instance, the family of ZNF genes has a higher incidence for hypermethylated sites in CRC, while the OR gene family is more present for the hypomethylated trend. Both are in line with previous publications [209, 210]. Some genes were also in line with the repository IntOGen [199]: FAT4, PCBP1, NBEA.

We use the data from GSE101764 [217], selecting the 112 tumor samples, to validate the results. The random forest algorithm predicts the CRC class correctly for 100% of the subjects. If we select only some of the most relevant CpG

Table 8.2: 14 of the most relevant CpG sites.

CpG ID	Gene; Region	Recent Publication
cg06319475 (hyper)	<NA>	<NA>
cg10526659 (hyper); cg06952671 (hyper)	ITGA4; 5'UTR,1stExon	2015 [211]
cg01350077 (hyper); cg27200446 (hyper)	MDFI; 5'UTR,1stExon	2017 [212]
cg13596497 (hyper)	COL4A1; TSS1500	2018 [213]
cg04806177 (hypo)	SEPX1; Body	2012 [214]
cg17494199 (hypo)	<NA>	<NA>
cg17698295 (hyper)	OPLAH; Body	2020 [215]
cg16601494 (hyper); cg08738570 (hyper); cg15487867 (hyper)	C1orf70; 5'UTR,1stExon	2013 [141]
cg18533201 (hyper)	GDF6; Body	2018 [216]
cg25223771 (hyper)	<NA>	<NA>

sites specified in the table 8.2, the prediction success is 99%. The differentially methylated sites found allow us, therefore, to predict the diagnosis of CRC patients based on their DNA methylation levels for those sites. Other relevant genes as EN2, YPEL3 or SLCO3A1 also lead to a good prediction rate. If we use the adjacent mucosa samples of the same array, however, they are predicted as normal samples.

Please note that the publications specified in the table 8.2 may not be exclusively studies of DNA methylation data. Apart from the genes specified in that table, we shared more genes of those publications, as BMP2 or PITX2.

Doing an enrichment analysis on the set of DMSs, we find quite a huge quantity of KEGG pathways commonly related to cancer as: Calcium signaling pathway, axon guidance, Wnt signaling pathway, PI3K-Akt signaling pathway, ECM-receptor interaction, etc. On the other side, we find several diseases that share common genes with colorectal cancer, as different mental disorders, autism, blood pressure problems, weight gain, or craniofacial abnormalities. The presence of different mental disorders as pathways associated with CRC genes may be surprising, but different recent studies published show that indeed CRC patients have a high risk of developing mental disorders even after the disease ends [218, 219]. Those main articles highlight that CRC survivors are at increased risk for mental health disorders in

the short-term and long-term. Survivors who develop mental health disorders also experience decreased survival. A deeper investigation about this link would be an interesting path.

8.3.1 Summary

Similarly to what we presented in the prostate cancer study, MultiNet is able to detect the methylation differences between CRC patients and control groups. We observed a clear hypermethylation trend for the cancer samples, starting in some cases from adenomas. In addition, the cancer samples' behavior was not homogeneous, but it presented a three-class design that would deserve an official clinical classification. We have identified 14 CpG sites which are good cancer predictors and may help with future diagnosis.

The appearance of genes related to mental disorders pathways would be an interesting research path. In addition, some olfactory family genes presented a common hypomethylated status in the cancer samples, as we observed with prostate samples. This finding may be also important for understanding overall cancer functioning. The presence of hypermethylated sites located in CpG islands is also an important observation, as occurs in CRC and prostate cancer but also in the oldest sample groups studied previously.

Additional studies could include a greater sample size, more cancer or disease status types, and the ability to extract similar information from blood samples. The effect of potential confounders as the gender or the colon region of the sample selected could be also studied.

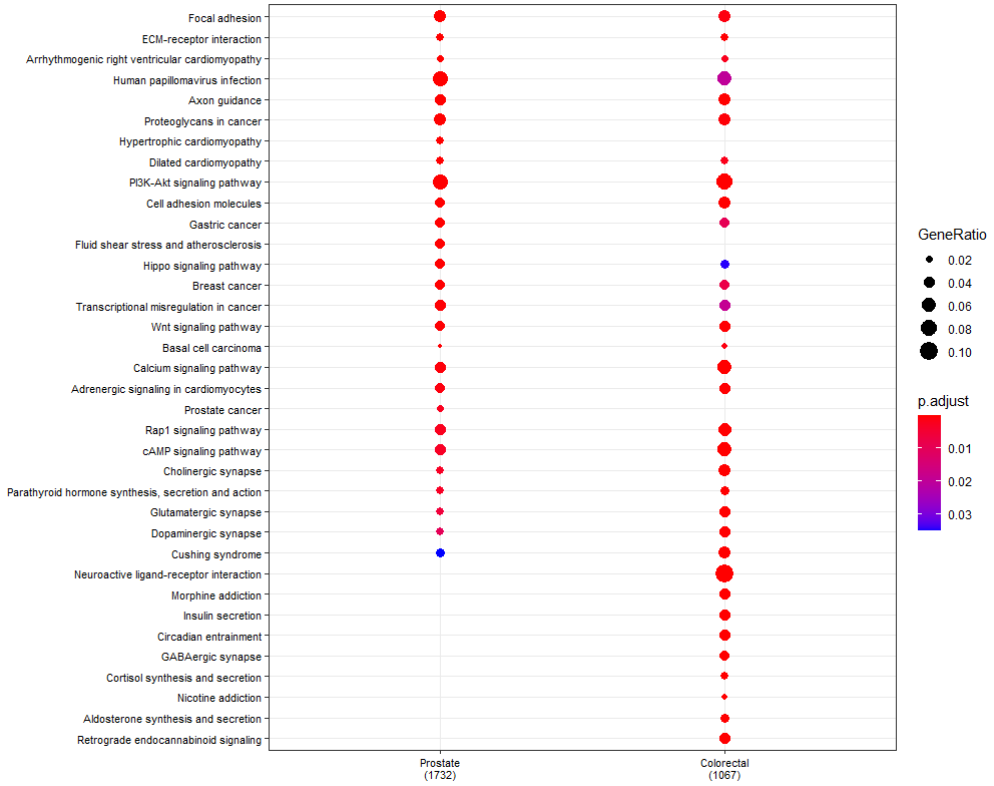
Generally, the study presented for CRC and prostate cancer could be extended to other cancer types.

8.4 Common Genes for Aging and Cancer

As cancer is an age-related disease, is not surprising that the three studies presented above share common genes and regulatory pathways. On the other side, is has been observed how different cancer types share common genes, that may explain or help to understand the tumorigenesis process. Here, we analyze the common sites and genes obtained with the three studies above. We use for that the significant DMSs obtained from running MultiNet over each dataset with 50,000 CpGs (and the same parameters).

As it was expected, both cancer studies share sites, genes and common KEGG pathways as we could observe in figure 8.18. This figure was created with the function *compareCluster* from the R package *clusterProfiler*, that uses the function *enrichKEGG*. The pathways found are colored by the adjusted p-values derived from an over representation test and adjusted for multiplicity, and the GeneRatio is the rate of genes included that are related to the pathway.

Figure 8.18: KEGG pathways for prostate and colorectal studies.



For instance, calcium signaling pathway, axon guidance, transcriptional misreg-

ulation in cancer, PI3K-Akt signaling pathway, ECM-receptor interaction, Wnt signaling pathway, cell adhesion molecules, or proteoglycans in cancer. Prostate cancer pathway is only present for the prostate study. Neuroactive ligand-receptor interaction or other addiction-related pathways are more present in the colorectal cancer group. Some OR genes are also common to both. Future studies that are able to detect more common epigenetic markers among different cancer types will be helpful to understand better the underlying genome alteration.

Additionally, some genes are shared among the age study and the other two. Specifically, CSTA, SULT1E1, MGP genes and PCDH gene family are common with prostate results. On the other side, ECEL1P2 [179], PI3 or SPTA1 are present in aging and colorectal results. Those findings indicate target CpG sites and genes that could be further studied as epigenetic biomarkers of accelerated biological age and aging-related cancer.

8.5 MultiNet with Other Diseases

There are other diseases where the case/control differentiation is not so clear and therefore the task of finding prognostic genes through DNA methylation is not straightforward. This is, above all, the case with neurological disorders, as Autism Spectrum Disorder (ASD), Alzheimer or Depression. Probably, most of them should be studied somehow together to discover the underlying biological dysfunction, as started in [220]. On the other side, the epigenetic inheritance derived from pregnancy or parent's life is not very studied and may be crucial for those disorders. The early diagnosis of those is normally key to ensure a better treatment and quality of life, but is also difficult to achieve due to the disease condition itself. For this reason, the automation of diagnosis tools using machine learning algorithms is extremely important for neurological disorders. Unfortunately, the lack of public standard datasets with enough sample size makes very difficult to create and validate trustful results.

We have tested MultiNet on neurological disorders datasets, that generally contain a high inter/intra-subject variability. The results obtained should be carefully interpreted. The algorithm detected groups of DMSs and known genes significantly differentiated, but also other non-published genes. Indeed, publications are, at the same time, very heterogeneous, demonstrating the sensitivity of current analysis methods to noise.

From our ASD study, using the dataset GSE109905 and the recent publication as reference [221], we detect the same epigenetic biomarker (cg20793532 related to the gene PPP2R2C) plus the 50 CpGs identified as DMSs. In addition, we find other significantly DMSs non specified in the article and associated to autism as DIP2C, XKR3, some HLA family genes, OR family genes, PCDH family genes, etc. The algorithm detects two types of ASD patients, most of them with higher levels of methylation but also some others with a methylation status similar to the control group. Those sites and genes would deserve a separate investigation.

8.6 Other Applications of MultiNet

As we mentioned at the beginning of the present document, MultiNet is a flexible algorithm that is able to analyze the correlation structure from any type of high-dimensional data. Despite we have applied it mainly to DNA methylation data, we would like to present here how the algorithm works with a very different non-biological dataset.

Generally, it is not an easy task to obtain public datasets with thousands of variables out of the biological or genetic world. For that reason, we decided to generate high-dimensional data using economic time series. We analyze the longitudinal data of the stock market (IBEX35) from 1st of January 2005 up to 1st of October 2020, where each daily price evolution (close-open) corresponds to a different variable for each one of the biggest 35 companies in Spain (our sample). As the strongest 35 companies were changing since 2005, we select the ones that stayed all that time. We work then with a matrix of 29 companies and more than 4,000 variables. Missing values were imputed to zero.

The main idea is to apply MultiNet to explore the correlation of the data and find periodic trends of market behavior and subgroups of companies with a similar progression. In this case, as there is not a sample differentiation, we do not specify sample subgroups as case or control. Instead, we include the overall data matrix with all companies and their price evolution per day with the objective of exploring its correlation structure.

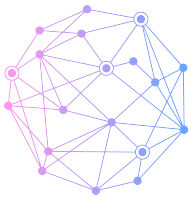
Figure 8.19: Evolution of BBVA stock prices during the selected period. The red stars are days with heavy decreases that we obtain with MultiNet.

The parameters were changed slightly in order to be in concordance with the data used. For example, data windows were selected by 30 days, the filter functions will the median and the variance of the price evolution per day, and the windows of 30 days were selected by decreasing median values. The metric used for the cluster analysis was still the metric associated to the Pearson correlation.

The algorithm took less than one minute to run over the timespan and obtain the networks. Indeed, we were able to distinguish different key periods of huge ups and downs thanks to the correlation among the variables that is represented in the MultiNet networks. Among the nodes with a lower median (with low closing values compared to open prices), we obtain the "black day" dates related to the biggest decreases, as occurred in 2007-02-27, 2007-04-24, 2007-06-06, 2008-01-21, 2008-09-29, 2008-10-22, 2008-11-06, 2010-02-04, 2010-05-14, or 2020-03-12 (being the last one the worst ever). Similar days are obtained by selecting the nodes with higher levels of variance or correlation.

The correlation values among the different entities did not provide with any new information but just confirm that companies from the same area are highly positively correlated historically. A special high correlation is presented among Santander and BBVA banks, for example. On the other side, those two banks have higher degrees of correlation with almost all other entities, suggesting how they are representative of the IBEX behavior. This is another important feature to be taken into account for prediction.

We could of course extend the study of this data considerably: increasing the observation period, using time values instead of daily values, or studying deeply the network layouts and trends before the marked periods of increases and decreases in order to make accurate predictions. However, that would be an extensive work out of the aim of this thesis and the objective of this section of showing another application for MultiNet.



Chapter 9

MultiNet in R

How could we apply MultiNet to our research? We have developed MultiNet in R to be useful and easy to use. The main function needs the input dataset plus the filter functions, and it only asks for a group of parameters before creating the networks. In addition, we have developed several post-processing functions of biological interpretation that program all the graphics presented in this document (from the colored networks to the heatmaps, gene frequencies or enrichment networks), plus output documents as spreadsheets with the results. The implementation of those functions would depend on the data used, as they are mainly focused on epigenetic data. Despite MultiNet is not yet an official R package, in this chapter we present the developed programming and a guide of use.

9.1 Guide of MultiNet Use

MultiNet R function is a compound of several R functions, as indicated in the figure 6.1, some of them already implemented in R and others programmed by us. It is prepared to deal with high-dimensional data and has the main objectives:

1. Prepare the data, select the parameters, and generate MultiNet networks for each sample group (normally divided in overall/control/case), including colored graphs.

2. Differentiation of MultiNet networks according to the method described in subsection 6.5.1.

3. Select substantial information from the networks, as DMSs, HCSs, or SMSs.

4. Do hierarchical cluster analysis over those DMSs and represent them in a heatmap for sample differentiation from distinct methylation patterns.

5. Select significantly differentiated sites over case/control groups with logistic regression. Select the most relevant CpG markers through random forest.

6. Validate the significant sites found in the previous step in a different dataset (selected by the user). Select the most relevant CpG markers for sample prediction.

7. Study the correlation of the detected DMSs.

8. Analyze the regions, genes and biological pathways associated to the significant DMSs.

9. Apply local MultiNet over the chromosomes of interest, as those that contain a greater number of significant DMSs.

10. Record main results in a spreadsheet.

The first step would be to download the selected array and the genome annotated database. The main packages needed are: *IlluminaHumanMethylation450kanno.ilmn12.hg19, IlluminaHumanMethylation450kmanifest, minfi, GEOquery, igraph, png, grid, gridExtra, gapminder, tidyverse, lme4, ggpubr, np, logistf, sgof, qvalue, multtest, randomForest, circlize, BioCircos, circular, OmicCircos, migest, ReactomePA, enrichplot, clusterProfiler, org.Hs.eg.db, DOSE, qqman, calibrate, netbiov, plot3D, ggplot2, gtools, effsize, scales, ComplexHeatmap.*

```
> gse1 <- getGEO(''GSE76938'')
Found 1 file(s)
GSE76938_series_matrix.txt.gz
trying URL
'https://ftp.ncbi.nlm.nih.gov/geo/series/GSE76nnn/GSE76938/
matrix/GSE76938_series_matrix.txt.gz'
Content type 'application/x-gzip' length 240733165 bytes (229.6 MB)
downloaded 16.5 MB
[...]

> gse1
$GSE76938_series_matrix.txt.gz
ExpressionSet (storageMode: lockedEnvironment)
assayData: 347898 features, 136 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: GSM2041110 GSM2041111 ... GSM2041245 (136 total)
```

```
  varLabels: title geo_accession ... tissue:ch1 (34 total)
  varMetadata: labelDescription
featureData
  featureNames: cg00000029 cg00000109 ... cg27666123 (347898 total)
  fvarLabels: ID Name ... SPOT_ID (37 total)
  fvarMetadata: Column Description labelDescription
experimentData: use 'experimentData(object)'
  pubMedIds: 28412973
Annotation: GPL13534

> ann450k <- getAnnotation(IlluminaHumanMethylation450kanno.ilmn12.hg19)
> dim(ann450k)
[1] 485512     33
> annotation0 <- ann450k
```

Secondly, we convert the downloaded dataset into a manageable data frame containing the DNA methylation Beta levels and the metadata. In addition, we select the sample groups case/control from the subjects metadata. We rename the variable that contains the group differentiation as "disease".

```
> all.dat0 <- exprs(gse1[[1]])
> dim(all.dat0)
[1] 347898    136

> all.meta1 <- pData(phenoData(gse1[[1]]))
> dim(all.meta1)
[1] 136  34

> names(all.meta1)[34] <- "disease"
> case <- all.meta1[all.meta1$disease %in% c("prostate cancer tissue"),]
> dim(case)
[1] 73 34
> rcase <- rownames(case)
> control <- all.meta1[all.meta1$disease %in% c("prostate benign tissue"),]
> dim(control)
[1] 63 34
> rcontrol <- rownames(control)
```

We then eliminate the CpG sites related to SNPs or to the sex chromosomes, as they may be an unexpected source of variability:

```
> newdata <- intersect(rownames(all.dat0), rownames(annotation0))
```

```
> all.dat<-all.dat0[newdata,]
> dim(all.dat)
[1] 347898    136

> annotation <- annotation0[rownames(all.dat),]
> dim(annotation)
[1] 347898     33

> ##Remove chromosomes X and Y
> nox <- annotation[annotation$chr!="chrX" & annotation$chr!="chrY",]
> noxr <- rownames(nox)

> ##Remove SNPs
> snps <- annotation[is.na(annotation$Probe_rs),]
> snpsr <- rownames(snps)
>
> bothr <- intersect(noxr,snpsr)
> length(bothr)
[1] 278061

> all.dat1 <- na.omit(all.dat[bothr,])
> dim(all.dat1)
[1] 247252    136
```

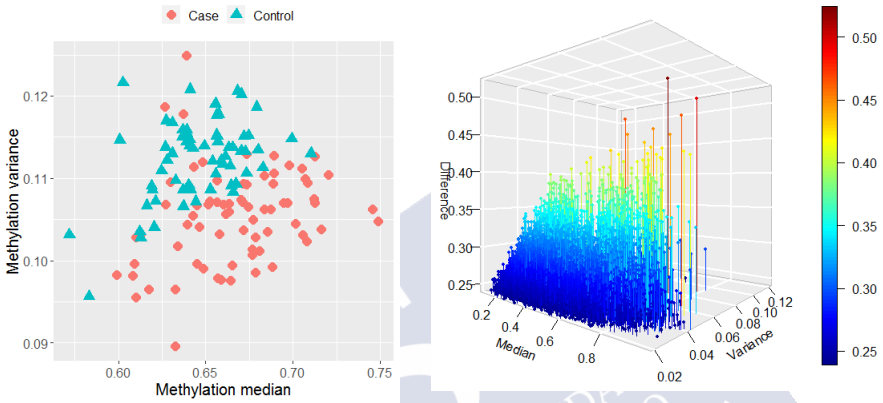
We prepare now the filter functions of MultiNet:

```
> medi_global <- apply(all.dat1,1,median)
> medi_global_case <- apply(all.dat1[,rcase],1,median)
> medi_global_control <- apply(all.dat1[,rcontrol],1,median)

> var <- apply(na.omit(all.dat1),1,var)
> var_case <- na.omit(apply(all.dat1[,rcase],1,var))
> var_control <- na.omit(apply(all.dat1[,rcontrol],1,var))

> diff_global <- na.omit(abs(medi_global_case-medi_global_control))
> diff_global1 <- sort(diff_global,decreasing=TRUE)

> ##Dataset for MultiNet ordered by maximum difference of medians
> all.dat_ord_global <- na.omit(all.dat1[names(diff_global1),c(rcase,rcontrol)])
> dim(all.dat_ord_global)
[1] 247252    136
```

We could visualize the filter functions or do any other kind of descriptive analysis in order to observe how the data is mapped and understand better the MultiNet outcome. In addition, it would help us to identify missing data or outliers that may be imputed or removed. This step is open for the user (i.e. not determined by MultiNet) and one can do multiple exploratory analyses as PCA, MDS, sampling correlation or clustering, etc.

Figure 9.1: Descriptive plots of the data cloud.

(a) Subject's distribution. (b) Filter functions.



After the correct pre-processing, we are then prepared to run the global algorithm selecting the appropriate parameters. Please note that the overall network with all subjects is always created. The input variables are the overall dataset and the filter functions for all sample groups. Moreover, the requested parameters are:

1. The creation of biological results: 1=Yes; 0=No
2. The creation of case: 1=Yes; 0=No
3. The creation of control: 1=Yes; 0=No
4. The initial CpG to be selected in the input array
5. The last CpG to be selected in the input array
6. The window size
7. The window's overlap in percentage
8. The number of intervals per window
9. The interval's overlap in percentage

10. The number of cluster per interval
11. The edge-joining parameter δ
12. The edge-joining parameter λ
13. The k-means methodology

If the first parameter of creation of biological results is zero, then the algorithm will only create the networks without all the post-processing. This was implemented for the cases where the biological interpretation is not needed, as the IBEX35 analysis.

The algorithm starts then the implementation based on those parameters. It indicates the phases of the execution as follows:

```
> multinet(all.dat_ord_global=all.dat_ord_global,
filter1_global=medi_global,filter2_global=diff_global1,
filter1_control=medi_global_control,filter2_control=var_s1_control,
filter1_case=medi_global_case,filter2_case=var_s1_case)

[1] "Please include MultiNet parameters"
Biological results: 1
Case option: 1
Control option: 1
Initial position: 1
Last position: 5000
Window size: 1000
Percent window overlap: 50
Number of intervals: 2
Percent interval overlap: 50
Number of clusters: 3
Points in common among nodes: 0.2
Median correlation among nodes: 0.8
Cluster method: Forgy

[1] "MultiNet per window - start"
[1] 1001
First window: 9.27 sec elapsed
[1] 1501
Next window: 8.22 sec elapsed
[1] 2001
Next window: 8.07 sec elapsed
[1] 2501
```

```
Next window: 8.34 sec elapsed
[1] 3001
Next window: 8.05 sec elapsed
[1] 3501
Next window: 9.47 sec elapsed
[1] 4001
Next window: 11.58 sec elapsed
[1] 4501
Next window: 8.39 sec elapsed
[1] 5001
Next window: 8.55 sec elapsed
[1] 5501
Next window: 8.45 sec elapsed
[1] "MultiNet per window - end"

[1] "MultiNet: joining of window networks - start"
[1] "MultiNet: joining of window networks - end"

[1] "MultiNet running time"
Time difference of 2.6141 mins

[1] "MultiNet: colored networks - start"
[1] "MultiNet: colored networks - end"

[1] "MultiNet: node histograms - start"
[1] "MultiNet: node histograms - end"

[1] "MultiNet differences between case/control networks - start"
[1] "Cliffs delta"
delta estimate: 1 (large)
[1] "MultiNet: differences between case/control networks - end"

[1] "MultiNet: DMS identification - start"
[1] 1791
[1] "MultiNet: DMS identification - end"

[1] "MultiNet: logistic regression"
[1] "Complete separation warning, proceed with penalized regression & SGoF"

[1] "MultiNet: random forest"

[1] "MultiNet: correlation heatmaps - start"
```

```
[1] "MultiNet: correlation heatmaps - end"

[1] "MultiNet: DMSs regions and trends - start"
[1] "MultiNet: DMSs regions and trends - end"

[1] "MultiNet: DMSs pathways - start"
[1] "MultiNet: DMSs pathways - end"

[1] "MultiNet: create spreadsheet - start"
[1] "MultiNet: create spreadsheet - end"

[1] "SUMMARY"
[1] "TOTAL DMSs: "
[1] 1791
[1] "TOTAL SIGNIFICANT DMSs: "
[1] 1748
[1] "TOTAL SIGNIFICANT DMSs HYPERMETHYLATED: "
[1] 986
[1] "TOTAL SIGNIFICANT DMSs HYPOMETHYLATED: "
[1] 762

[1] "MultiNet TOTAL RUNNING TIME"
Time difference of 7.320919 mins
```

The output of the algorithm consists then on the MultiNet networks plus all the post-processing presented in the chapter 8. Please refer to [230] to access the R code of MultiNet and all its auxiliary functions.

Part IV

Last Considerations

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Chapter 10

Discussion and Open Research

With this research project, we aimed to overcome the initial challenges of our brain machinery and develop novel mathematical tools with a biological motivation, which are able to process big quantities of information with a complex design. Particularly, we had the main objective of detecting hidden correlation patterns in high-dimensional epigenetic datasets trough the main study of their topology, that could be related with the three-dimensional structure of the genome.

The first mathematical hypothesis of a non-random (structured) network of short-range and long-range epigenetic correlations was validated leading to further investigate their local and global structure. The use of topology-based mathematical techniques was key to success in this investigation, as topological data analysis methods present a high flexibility to work with huge and complex data as the genomic one [227]. Inspired by this methodology, we developed our own processes, models, and applications.

The stochastic block model with distance of the local correlation structure and the development of a computational algorithm as a powerful analytical tool were the two main novel outcomes of this research. Both mathematical procedures covered the deep study of the DNA methylation correlations and their modifications with different conditions or sample characteristics. Moreover, both analyses would be possible in a non-biological context with similar data properties and a high-dimensional data space. This is an important fact to be taken into account to generalize the results and open future investigation paths. Being a transversal work, the conclusions and open research are also cross-sectional. From a topological, computational, statistical or even biological perspective, we could continue developing the research done obtaining promising results.

In this chapter, we will present the investigation summary based on the mentioned aims, containing the developed work and results, together with the related open research.

10.1 A Model for the Local Correlation

Due to the known complexity of the epigenetic processes and the three-dimensional design of the chromatin, we decided to face the analysis from two main perspectives: a local intra-chromosomal behavior especially focused on CpG islands, and a global inter-chromosomal one. The results of both studies are of course in line but they provide different information layers. This differentiation provides a wider vision non-present in other similar studies and that could be useful with other high-dimensional data, where substantial information could be extracted with distinct perspectives of analysis.

The first local analysis was accompanied by the study of the local correlation evolution with age. The representation of the correlation matrices as weighted correlated graphs with a (genomic) distance between the nodes allowed us to study the correlation structure based on the graph's properties, and their topological characteristics with persistent homology. This process established already a novel way of studying epigenetic modifications and their correlations. We understood the structured design of the correlation, mainly affected by the spatial disposition of the CpG sites along the genome and the distribution of the short-range and long-range correlations, far away from any random design.

Persistent homology studies confirmed the non-random design of the local correlation and pointed out several topological features associated with specific graph properties, as the modules edge-density and its relationship with the Betti numbers. Moreover, persistent homology differentiated the design of short and long-range networks and opened the door to establish mechanisms of graph distinction through the distance of the related persistence diagrams.

The spatial design determined a special kind of graph, with a high modularity and whose degree distribution was highly related to the correlation clusters of the associated matrix. Those facts allowed us to develop a model of prediction of this degree distribution. Having into account the genomic position and the methylation levels of each CpG site, we developed a clustering-based model (called SBM-D) to estimate the adjacency matrix of the CpG island graph (or other regions, as gene networks-related sites) with three main parameters K, p_{int}, and p_{ext} estimated from the data. Those parameters measure the modules or clusters of CpG sites based on their correlation, and the distribution of short-range and long-range correlations for the clusters determined. The dimension of the initial correlation matrix was then reduced to the estimation of three indexes.

This model succeeded on finding the hub CpG sites and regions with a higher

level of correlations. This prediction is crucial to detect multiple epigenetic interactions that may have an impact on biological functioning and related-diseases. The application of persistent homology to study the topological properties of the SBM-D graph compared with other random graphs models was once again helpful, providing with an analytical way of tracking the observed differences. Persistent homology gave us, in addition, an idea of the topological elements that we would be able to generate with each random model.

The SBM-D model is able to detect an evolution of the correlation structure in CpG islands with different age ranges. It confirms the stronger short-range correlation generally and a overall decreasing trend of correlation as we age. However, despite the initial trend is a loss of short-range and especially long-range correlation with age, the pattern seems to change for very elderly people. They indeed present similar levels of correlation as observed in childhood. This finding may be explaining some of the longevity mechanisms but should be confirmed with post investigations.

The study of complex real networks is itself a very broad and complicated research field, so advanced graph mathematical investigation of the behavior of complex epigenetic (including also negative correlation as a feature of interest) or genetic-epigenetic networks would be an interesting research way and could be useful for the study of complex diseases. The study of the chromatin folding design could be also extended with this methodology, trying to establish a solid link between the genome structure and its functionality. In addition, the evolution of the studied networks with age could continue by defining specific aging models and generalizing the results found on more datasets, ideally with longitudinal data.

The same study could be done with non-biological data, as the SBM-D model is prepared to describe any type of information with a spatial and modulated design. The application on social sciences seems to be one potential path.

10.2 An Algorithm for the Global Correlation

Attending to the aim of developing novel useful data analysis tools, we presented MultiNet. We created a machine learning-based algorithm to detect local/global correlation trends on high-dimensional data and different sample group patterns. Inspired in Mapper and the idea of representing the "shape" of the data, reducing the dimension of the data through the creation of simplicial complexes, MultiNet provides with a wider data analysis tool than current techniques. Indeed, it is a compound of several additional methods of networks' differentiation and diagnosis models, considering the effect of data confounders. We developed our own method of network's distinction and use efficiently logistic regression and random forest models. MultiNet is designed to be manageable and understandable by distinct research professionals.

As we decided to use MultiNet to contribute to the epigenetics study and the understanding of DNA methylation patterns, we obtained substantial results on that area. Nevertheless, the algorithm was designed to study the form and global correlation structure of other data types.

As the main contributions to the study of the DNA modifications with the age, MultiNet was able to identify the overall increasing hypomethylation trend in line with published studies. Interestingly, the CpG sites with an opposite trend of hypermethylation for older sample groups belong mostly to CpG islands. This indicates a strong importance of those CpG islands to regulate aging pathways with the potential power of accelerating or decelerating the process altering the expression of the corresponding genes.

The algorithm was then perfectly able to distinguish among the selected sample groups, detecting a differentiated methylation trend and correlation design. Moreover, it allowed us to distinguish separated children groups based on those trends. We have found significant epigenetic markers (significant DMSs), a lot of them among the "clock CpGs", with a heavier correlation in younger groups and a group of genes with a powerful sample prediction.

Also, we have found important cancer-related genes and significant DMSs for prostate and colorectal cancer that may be used as epigenetic markers of disease diagnosis. Despite the overall hypermethylation for cancer samples related to some genes found in literature, validating our results, we also found novel sites and genes that could be important to study and even present an opposite trend. Both types of cancer shared genes and biological pathways, suggesting an interesting research way to study some of their alterations together.

Some of the genes and related biological KEGG pathways were indeed common for the aging study and the cancer ones, in line with the fact that the age is a cancer risk factor. For instance, the fact that lots of aging-related genes were associated with the protocadherins family may suggest a potential link with age-related disorders as cancer. Indeed, it was observed that interactions among PCDHs represent a long-range epigenetic silencing by hypermethylation related to this disease [222]. The appearance of the olfactory gene family as related to several diseases or age-related status is something that would be interesting to investigate, as we know that the olfactory sense alteration was also related to neurological disorders [223, 224] or to the lack of social skills [225, 226].

Of course, the genetic and epigenetic results provided by MultiNet could be studied by a team of biological experts that can also find other interesting applications of the same algorithm. The extension of the biological analysis could be also done, studying the chromatin states of the DMSs found or linking our methylation findings with gene expression profiles. The use of the novel methylation array with more than 850,000 CpG sites could be a step forward in this study.

As an example of the use of MultiNet with non-biological data, we presented how it is able to detect correlation patterns with a stock market dataset. Despite we just presented an initial approach of this analysis, it could be extended in many ways using the information provided by MultiNet to design prediction models or to investigate further the cycled patterns.

MultiNet is constructed to be the seed of something greater, a commercial package designed with advanced techniques of computer science, with more functionalities and with an automatic learning of data needs. The successful applications of this algorithm, even out of the biological world, spot a promising future for the real use of MultiNet. This methodology is designed to be understandable and usable by different research groups, asking for minimal data assumptions and producing results with reasonable computational resources and time. The increase of the power of MultiNet could bring a bright future as a great contributor to the data analysis in many research fields.

10.3 Conclusion

Concluding, we were able to complete our research aims, combining the novel approach and design of new topological data analysis methods specially focused on the managing of high-dimensional correlation structures, with a higher understanding of epigenetic mechanisms. The initial challenges were covered and the consequent generated hypotheses were successfully tested.

We would like to emphasize the transversal design of the present work, applying mathematical procedures to genomic data from different perspectives: theoretical (using topological and statistical knowledge); computational (designing MultiNet); and biological (using mathematics to describe observed phenomena). This design is then in line with current data analysis research needs, where the quantity and heterogeneity of big data from different fields need to be analyzed.

This whole analysis embraces the multidisciplinary design of generating biological hypothesis from observation to be solved analytically with advanced mathematical techniques, spotting a great evolution in both fields which complement each other.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Appendix A

Resumo (Galego)

A capacidade do ser humano actual para xerar e almacenar grandes cantidades de datos en varias áreas provoca unha necesidade crecente de analizalos, o que desperta o entusiasmo dun exército de matemáticos. Dependemos máis que as nosas xeracións anteriores dos datos que analizamos e, polo tanto, das predicións que xeramos. A nova era da intelixencia artificial apóiase na dispoñibilidade de novos algoritmos computacionais baseados, idealmente, en teorías matemáticas fundadas.

Entre esta crecente oferta, hai unha fonte de datos específica que ten un interese especial para nós pola súa importancia na evolución humana: a xenómica (ou os estudos "omics" en xeral). En particular, estúdase cada vez máis o conxunto das modificacións xenómicas producidas pola interacción entre os suxeitos e o seu contorno, chamado *epixenética*. A epixenética ofrece unha perspectiva diferente á vida humana, xa que apoia a idea de que as nosas decisións sobre o estilo de vida xogan un papel importante en nós mesmos e nas xeracións futuras [1]. Tamén abriu a porta a un xeito diferente de facer medicina, xa que os trastornos epixenéticos poden reverterse.

Ademais, a relación entre a epixenética e o deseño espacial do ADN dentro do núcleo da célula (o que normalmente se denomina "estrutura tridimensional" do xenoma) estase estudando cada vez máis, engadindo un nivel máis de complexidade a todo o sistema e apuntando novos descubrimentos. Este deseño espacial é capaz de "unir" rexións xenómicas distantes que poden acabar tendo un funcionamento común ou unha alteración xenómica.

Por tanto, o estudo das modificacións epixenéticas e o seu funcionamento xeral dentro do xenoma non é unha tarefa fácil. Ten moitos niveis de complexidade diferentes e quizais algúns deles sexan aínda descoñecidos. O seu éxito depende cada vez máis do desenvolvemento de modelos matemáticos e algoritmos computacionais eficientes que sexan capaces de analizar con precisión esa grande recompilación

de datos para a súa posterior interpretación. De feito, o campo de investigación chamado "bioloxía matemática" céntrase no uso de ferramentas matemáticas para estudar sistemas biolóxicos e investigar a súa estrutura, desenvolvemento e comportamento. Desde as primeiras aplicacións máis sinxelas da serie de Fibonacci ou a proporción áurea, o uso da realización matemática complexa aplicada ou motivada por problemas biolóxicos está crecendo sobre todo a partir da secuenciación do xenoma (coa "bioloxía de sistemas"). Estamos nos albores dun novo xeito revolucionario de entender a nosa bioloxía e a futura medicina. É necesario construir pontes intelectuais entre a formulación matemática abstracta e os problemas reais para analizar todos os niveis de complexidade e facer o que os matemáticos saben facer: viaxar ás profundidades da cova do razoamento.

Ante esta situación, esta tese céntrase en proporcionar ferramentas alternativas de análise de datos epixenéticos desde unha perspectiva topolóxica. Pero, que é realmente a "análise de datos"? e a "topoloxía"?

Un cerebro é unha máquina perfecta de análise de datos. Desde os primeiros instantes de vida, recibe información a través do sistema sensorial e analízaa dun xeito cada vez máis sofisticado. A primeira fase deste proceso de pensamento analítico que se produce de forma natural adoita ser a segmentación de datos. Os bebés humanos, por exemplo, diferencian o sabor dun limón do sabor do leite, pero non saben o motivo da diferenza. A próxima vez que proben un alimento cítrico, este "irá" directamente ao "lado do limón" (memoria sensorial). Este tipo de aprendizaxe automática, que vai antes do uso da linguaxe e da comunicación, podería considerarse unha "aprendizaxe non supervisada". Non obstante, unha vez que comezamos a ser máis conscientes de nós mesmos e do noso contorno, tendemos a etiquetar a información aprendida e así os grupos que detectamos anteriormente. Isto pódese considerar entón unha "aprendizaxe supervisada". De feito, hai unha fase da infancia moi típica, onde continuamente preguntamos por nomes e razóns (¿que é iso?, ¿por que iso é así?). Cada vez que aprendemos algo novo, os datos pasan directamente pola cadea de procesamento, que os etiqueta e almacena rapidamente. Por exemplo, un bebé de tres anos é capaz de distinguir unha xirafa dun golfiño e incluso é capaz de distinguir os animais terrestres dos animais acuáticos. Ás veces, esta etiqueta é incorrecta, pero a propia maquinaria de aprendizaxe probablemente o solucione. Tendo en conta todo isto, pódese facilmente imaxinar como a cadea do procesamento e a etiquetaxe se ven afectados por unha alteración do sistema sensorial, relacionada con moitos trastornos neurolóxicos.

Curiosamente, nesta segunda fase do desenvolvemento cognitivo, o sistema sensorial tamén é capaz de adoptar unha perspectiva xeométrica, organizando o espazo segundo as relacións elementais de semellanza, proximidade, separación, ou continuidade. Esta primeira forma de detectar estruturas matemáticas altamente

abstractas está presente en nenos de baixa idade, que poden estar indirectamente debuxando estruturas xeométricas desde unha perspectiva topolóxica a través dos seus principais invariantes topolóxicos. Por exemplo, é posible que non debuxen un círculo moi diferente dun triángulo, pero debuxarán con bastante precisión o feito de que dúas figuras non se intersecan. De feito, a topoloxía é o campo da matemática que estuda as propiedades dos obxectos xeométricos non afectados por continuas transformacións de forma ou tamaño das figuras. A capacidade de deformar unha rosquilla nunha taza ou de definir o aspecto diferenciado da letra "D" e a letra "C" (buratos / non buratos). As deformacións topolóxicas parecen ser acuáticas e naturais, e describen o mundo desde unha perspectiva moi particular. Estudada por siglos, a topoloxía ten aplicacións en moitos campos distintos como a bioloxía, a física ou incluso a arte.

O desenvolvemento tecnolóxico e dixital engade máis información á cadea de procesamento, pero a súa etiquetaxe nun grupo complícase. Ás veces, podemos adestrar o noso cerebro para poder facer unha etiquetaxe aproximada (con formación analítica, por exemplo), pero a maioría de veces, a cantidade ou o nivel de complexidade dos datos lévanos a pedir axuda. A análise de datos provén deste fracaso natural e intenta imitar a cadea de procesamento do noso cerebro. Así, inicialmente pasamos por unha primeira fase exploratoria sen supervisión para segmentar os datos e detectar grupos que despois se utilizarán para clasificar a nova información nunha fase supervisada. Desta maneira nace a análise matemática de datos.

Actualmente hai falta de ferramentas analíticas estándar e eficientes para tratar as grandes cantidades e variedades de datos de alta dimensión (tamén chamados datos de "alta dimensión, baixo tamaño da mostra" (HDLSS, das súas siglas en inglés), o que significa que temos moitas máis variables que observacións independentes) como os xenéticos. Particularmente, a análise das estruturas de correlación de grandes dimensións é un tema pendente no campo da epixenética. Na nosa opinión, coñecer a estrutura de correlación epixenética é crucial para comprender completamente as interaccións biolóxicas que poden ser desreguladas e manifestadas por un fenotipo específico (como unha enfermidade), e poderían asociarse coa arquitectura tridimensional do xenoma. Como amosamos nesta tese, a análise topolóxica das grandes estruturas de correlación contribúe enormemente á súa comprensión e interpretación.

Xeralmente, a aplicación da topoloxía alxebrica na análise de datos mediante a análise de datos topolóxicos (TDA, das súas siglas en inglés) proporciona unha perspectiva diferente e tremendamente útil, xa que o estudo da "forma" dos datos é clave para extraer as características subxacentes facendo mínimos supostos previos sobre a súa distribución . Ademais, a topoloxía pode extraer os invariantes topolóxicos do propio espazo de datos, o que axuda ao deseño preciso de modelos

matemáticos para describilo e predilo. Por outro lado, a redución de dimensionalidade é clave para tratar conxuntos de datos de alta dimensión con centos ou miles de variables. Varios elementos da topoloxía alxebrica permiten reducir a dimensión dos datos extraendo os seus principais elementos ou características críticas. A aplicación dos principais aspectos da teoría de Morse ou da teoría de grupos de homoloxía permite reducir a dimensionalidade e extraer esas características mediante a creación de complexos simpliciais. Estos complexos son unha representación da nube de puntos que contén moita información sobre os propios datos. Como un poderoso método de topoloxía computacional, a metodoloxía TDA ten un deseño de aprendizaxe automática. TDA é a inspiración técnica do presente traballo.

Motivados pola necesidade biolóxica de comprender mellor as complexas interaccións epixenéticas, desenvolvemos un traballo de investigación onde o estudo da correlación destes conxuntos de datos tan enormes é o principal obxectivo. Pretendemos extraer os principais elementos topolóxicos da estrutura de correlación das modificacións epixenéticas (particularmente da metilación do ADN) para comprender e modelar o deseño de correlación. Desenvolvemos novos enfoques matemáticos e ferramentas analíticas para axudar a comprender mellor os patróns de correlación de metilación que serven tamén para outras aplicacións non biolóxicas.

Usando a idea de análise de datos topolóxicos, a nosa proposta principal é estudar as estruturas de correlación a través das propiedades topolóxicas das redes de correlación asociadas, o que representa un novo método para describir e modelar estas estruturas epixenéticas. Esta análise fíxose a nivel local e global, deseñando diferentes metodoloxías para cada obxectivo. Xeramos un modelo para describir a estrutura de correlación local e desenvolvemos un algoritmo computacional para estudar a correlación global e as alteracións da metilación do ADN con diferentes condicións de mostra (como o envellecemento ou o estado da enfermidade). A idea principal do algoritmo deseñado é facer análise de datos de gran tamaño dun xeito sinxelo e rápido, automatizando o proceso de tratar con miles de variables. Ambos métodos son novas formas de estudar a paisaxe epixenética e sinalan novos e prometedores descubrimentos biolóxicos neste campo.

Esta tese ten, polo tanto, unha elevada transversalidade, onde se empregaron diferentes técnicas matemáticas para abordar problemas biolóxicos. Para desenvolver de xeito eficiente as análises feitas, utilizáronse aspectos principais de topoloxía alxebrica, teoría de grafos, estatística e informática. Ademais, podería considerarse un traballo multidisciplinar xa que analizamos profundamente estruturas biolóxicas que se están a investigar actualmente. Foi necesaria unha comprensión completa das estruturas xenómicas complexas para propoñer unha hipótese e un modelo matemáticos razoables. Non obstante, as nosas ideas e propostas técnicas poderían aplicarse a calquera outro campo onde se poidan recompilar datos de alta dimensión.

Derivada desta transversalidade, esta tese abre a porta a diferentes camiños de investigación, desde un camiño máis teórico centrado na descrición das redes de correlación e as súas propiedades topolóxicas; a unha dirección máis aplicable centrada na evolución e contribucións do algoritmo que se creou e as interpretacións biolóxicas correspondentes.

O presente traballo divídese en catro partes principais. A primeira parte é unha introdución ó planteamento do problema e á metodoloxía. A segunda e terceira parte presentan o noso traballo e, finalmente, a última parte inclúe unha descrición completa das conclusións e investigacións abertas. En concreto, esta tese contén:

1. A primeira parte é unha introdución aos desafíos biolóxicos e matemáticos que hai que resolver. Divídese en tres capítulos que explican o contexto biolóxico, a hipótese matemática relacionada e as principais técnicas matemáticas empregadas.

   (a) O primeiro capítulo contén unha descrición dos principais conceptos biolóxicos que empregamos e que son necesarios para comprender o enfoque do problema e a relevancia do traballo. Describimos extensamente a regulación epixenética, centrándonos na metilación do ADN e as súas alteracións con varias características como o proceso de envellecemento ou as enfermidades como o cancro. Aquí explicamos tamén as características dos conxuntos de datos de metilación que empregamos. Teñan en conta que se utilizaron sempre datos públicos recollidos en humanos como base do presente traballo.

   (b) O segundo capítulo inclúe a hipótese matemática derivada do problema biolóxico e a súa argumentación. Ademais, especificamos as primeiras análises exploratoriasque se fixeron baseadas nesta hipótese. Introducimos os desafíos matemáticos que as técnicas actuais non son capaces de resolver (como a análise de estruturas de gran correlación), propoñendo a necesidade de novas ferramentas analíticas inspiradas nun novo enfoque de análise de datos topolóxicos.

   (c) O capítulo 3 desta primeira parte é unha descrición profunda da metodoloxía TDA, na que nos inspiramos, e da teoría matemática relacionada. Tamén incluímos unha descrición extensa de dúas das principais técnicas de TDA, Mapper e homoloxía persistente, incluíndo exemplos no campo epixenético e a introdución á nosa proposta analítica.

2. A segunda parte deste traballo contén dous capítulos e consiste na análise completa da estrutura de correlación da metilación do ADN desde unha perspectiva local. Estudamos a distribución espacial da correlación dentro das illas CpG e as súas principais propiedades topolóxicas. Para facelo, representamos as

matrices de correlación como grafos de correlación que teñen un deseño modulado especial baseado nesta distribución espacial dos sitios CpG. O capítulo 4 comprende un estudo extenso das características dos grafos baseado nos principios da teoría de grafos e na homoloxía persistente. O capítulo 5 describe un modelo da estrutura de correlación das illas CpG baseado nas propiedades estudadas. Este modelo estima con éxito as interaccións entre sitios CpG cunha alta correlación. Ademais, o modelo é capaz de distinguir comportamentos de correlación distintos entre os diferentes grupos de idade da mostra, polo que somos capaces de medir a evolución potencial do deseño de correlación co proceso de envellecemento.

3. A terceira parte inclúe catro capítulos para introducir o algoritmo computacional desenvolvido, chamado *MultiNet*. Presentamos a súa descrición principal e a súa implementación no capítulo 6. O o capítulo 7 contén unha guía para a selección de parámetros necesarios e presentamos as contribucións do algoritmo ao estudo epixenético no capítulo 8. Propoñemos aquí un modelo para detectar a estrutura de correlación a nivel global que mellora os algoritmos actuais empregados para analizar conxuntos de datos de alta dimensión. Isto podería considerarse como unha extensión da parte II, usando a mesma perspectiva topolóxica pero cun deseño máis amplo. O algoritmo detecta, ademais, tendencias de metilación con diferentes grupos de mostra de interese (como estudos de casos/controles) incluíndo modelos de predición para detectar biomarcadores epixenéticos que poden ser útiles no diagnóstico de enfermidades. Usamos o algoritmo para estudar as alteracións epixenéticas co proceso de envellecemento e co cancro, obtendo resultados acordes coa información pública máis novos descubrimentos prometedores. Ademais, tamén demostramos o seu potencial con datos non biolóxicos. Computacionalmente, foi deseñado para ser rápido e sinxelo de usar por diversos profesionais. O algoritmo desenvolveuse en R, como se indica no capítulo 9.

4. A última parte está composta polas conclusións do traballo. Resumimos as novas contribucións e a interpretación dos resultados no seu conxunto. Tamén destacamos os diferentes campos de investigación que se poderían abrir a partir de esta tese.

Nas seguintes seccións detallamos os resultados obtidos, as conclusións, e as novas liñas de investigación que poderían seguirse.

Un modelo matemático para a correlación local

Debido á complexidade coñecida dos procesos epixenéticos e ao deseño tridimensional da cromatina, decidimos afrontar a análise desde dúas perspectivas principais: un comportamento intra-cromosómico local enfocado especialmente nas illas CpG e outro inter-cromosómico. Os resultados de ambos estudos están por suposto en liña, pero proporcionan diferentes capas de información. Esta diferenciación é unha novidade comparada con outras técnicas actuales, e podería ser útil con outros datos de alta dimensión, onde se podería extraer información substancial con distintas perspectivas de análise.

A primeira análise local acompañouse do estudo da evolución da correlación local coa idade. A representación das matrices de correlación como grafos correlacionados ponderados cunha distancia (xenómica) entre os nodos permitiunos estudar a estrutura de correlación en función das propiedades do grafo e as súas características topolóxicas con homoloxía persistente. Este proceso estableceu xa un xeito novidoso de estudar as modificacións epixenéticas e as súas correlacións. Entendemos o deseño estruturado da correlación, afectado principalmente pola disposición espacial dos sitios CpG ao longo do xenoma e a distribución das correlacións de curto e longo alcance, moi lonxe de calquera deseño aleatorio.

Os estudos de homoloxía persistente confirmaron o deseño non aleatorio da correlación local e sinalaron varias características topolóxicas asociadas a propiedades específicas dos grafos, como a densidade de eixes dos módulos e a súa relación cos números de Betti. Ademais, a homoloxía persistente diferenciou o deseño de redes de curto e longo alcance e abriu a porta a establecer mecanismos de distinción de grafos a través da distancia dos diagramas de persistencia relacionados.

O deseño espacial determinou un tipo especial de grafo, cunha alta modularidade e cuxa distribución de graos estaba moi relacionada cos grupos de correlación da matriz asociada. Estes feitos permitíronnos desenvolver un modelo de predición desta distribución de graos. Tendo en conta a posición xenómica e os niveis de metilación de cada sitio CpG, desenvolvemos un modelo baseado en agrupacións (chamado SBM-D, das súas siglas en inglés) para estimar a matriz de adxacencia do grafo da illa CpG (ou doutras rexións, como sitios relacionados con redes xénicas) con tres parámetros principais K, p_{int} e p_{ext} estimados a partir dos datos. Eses parámetros miden os módulos ou clusters de sitios CpG en función da súa correlación determinando a distribución de correlacións de curto e longo alcance para os clusters. A dimensión da matriz de correlación inicial reduciuse entón á estimación destes tres índices.

Este modelo conseguiu atopar os sitios e rexións CpG cun maior nivel de correlación. Esta predición é crucial para detectar múltiples interaccións epixenéticas que poden ter un impacto no funcionamento biolóxico e enfermidades relacionadas. A aplicación da homoloxía persistente para estudar as propiedades topolóxicas do grafo SBM-D en comparación con outros modelos de redes aleatorios volveu ser útil, proporcionando un xeito analítico de rastrexar as diferenzas observadas. A homoloxía persistente deunos, ademais, unha idea dos elementos topolóxicos que seriamos capaces de xerar con cada modelo aleatorio.

O modelo SBM-D é capaz de detectar unha evolución da estrutura de correlación das illas CpG con diferentes rangos de idade. Confirma a correlación de curto alcance máis forte en xeral e unha tendencia de correlación decrecente a medida que envellecemos. Non obstante, a pesar de que a tendencia inicial é a perda da correlación de curto alcance e especialmente de longo alcance coa idade, o patrón parece cambiar para persoas moi anciás. Efectivamente presentan niveis de correlación similares aos observados na infancia. Este achado pode explicar algúns dos mecanismos de lonxevidade, pero debería confirmarse con investigacións posteriores.

O estudo de redes reais complexas é un estudo de investigación moi amplo e complicado, polo que a investigación matemática avanzada das redes epixenéticas (incluíndo tamén a correlación negativa como característica de interese) ou xenético-epixenéticas sería unha investigación interesante con posibles aplicacións no campo das emfermidades complexas. O estudo do deseño da cromatina tamén podería ampliarse con esta metodoloxía para establecer unha relación sólida entre estrutura e funcionalidade. Ademais, a evolución das redes estudadas coa idade tamén podería ampliarse definindo modelos de envellecemento específicos e xeneralizando os resultados atopados en máis conxuntos de datos, idealmente con datos lonxitudinais.

O mesmo estudo podería facerse con datos non biolóxicos, xa que o modelo SBM-D está preparado para describir calquera tipo de información cun deseño espacial e modulado. A aplicación en ciencias sociais parece ser un camiño potencial.

Un algoritmo para a correlación global

Atendendo ao obxectivo de desenvolver novas ferramentas útiles de análise de datos, presentamos MultiNet. Creamos un algoritmo baseado na aprendizaxe automática para detectar tendencias de correlación local / global en datos de alta dimensión e diferentes patróns de grupos de mostra. Inspirado en Mapper e na idea de representar a "forma" dos datos reducindo a súa dimensión mediante a creación de complexos simpliciais, MultiNet ofrece unha ferramenta de análise de datos máis ampla que as técnicas actuais. De feito, é un composto de varios métodos adicionais de diferenciación de redes e modelos de diagnóstico. Presentamos una nova forma de diferenciar redes a través da súa distribución, e una maneira eficiente de utilizar a regresión loxística e os modelos de "random forest" para establecer modelos de predición. Está deseñado para ser manexable e comprensible por distintos profesionais da investigación.

Como decidimos usar MultiNet para contribuír ao estudo da epixenética e á comprensión dos patróns de metilación do ADN, obtivemos resultados substanciais nesa área. Non obstante, o algoritmo foi deseñado para estudar a forma e a estrutura de correlación global doutros tipos de datos.

Como principais contribucións ao estudo das modificacións do ADN coa idade, MultiNet foi capaz de identificar a tendencia global de hipometilación crecente en liña cos estudos publicados. Curiosamente, os sitios de CpG cunha tendencia oposta de hipermetilación para grupos de mostra maiores pertencen principalmente ás illas CpG. Isto indica unha forte importancia desas illas CpG para regular as vías de envellecemento co poder potencial de acelerar ou desacelerar o proceso alterando a expresión dos xenes correspondentes.

O algoritmo foi entón perfectamente capaz de distinguir entre os grupos de mostra seleccionados, detectando unha tendencia de metilación diferenciada e un deseño de correlación. Ademais, permitiunos distinguir grupos de nenos separados en función desas tendencias. Atopamos marcadores epixenéticos significativos, moitos deles entre os chamados "clock CpGs", cunha correlación máis forte en grupos máis novos e un grupo de xenes cunha forte predición de mostra.

Ademais, atopamos importantes xenes e sitios diferencialmente metilados (DMS, das súas siglas en inglés) relacionados co cancro para o cancro de próstata e colorrectal que se poderían empregar como marcadores epixenéticos no diagnóstico destas enfermidades. A pesar da hipermetilación xeral para mostras de cancro relacionadas con algúns xenes atopados na literatura, validando os nosos resultados, tamén atopamos novos sitios e xenes que poderían ser importantes para estudar e

incluso presentan unha tendencia contraria. Ambos tipos de cancro presentan tamén xenes e vías moleculares comúns, indicando un camiño para estudiar as súas alteracións en conxunto.

Algúns dos xenes e vías moleculares atopados con KEGG foron comúns para o estudo do envellecemento e o cancro, en liña co feito de que a idade é un factor de risco no cancro. Por exemplo, o feito de que moitos xenes relacionados co envellecemento estivesen asociados á familia das protocadherinas pode suxerir un potencial vínculo cos trastornos relacionados coa idade como o cancro. De feito, observouse que as interaccións entre os PCDH representan un silenciamento epixenético de longo alcance por hipermetilación relacionado con esta enfermidade [222]. A aparición da familia dos xenes olfativos relacionada con varias enfermidades ou estado relacionado coa idade é algo que sería interesante investigar, xa que sabemos que a alteración do sentido olfatorio tamén estaba relacionada con trastornos neurolóxicos [223, 224] ou co falta de habilidades sociais [225, 226].

Por suposto, os resultados xenéticos e epixenéticos proporcionados por MultiNet poderían ser estudados por un equipo de expertos en bioloxía que tamén pode atopar outras aplicacións interesantes do mesmo algoritmo. Tamén se podería extender o análise estudando os estados da cromatina dos DMS atopados ou ligando os nosos achados de metilación cos perfís de expresión xénica. O uso da nova matriz de metilación con máis de 850.000 sitios CpG podería ser un paso adiante neste estudo.

Como exemplo do uso de MultiNet con datos non biolóxicos, presentamos como é capaz de detectar patróns de correlación cun conxunto de datos da bolsa. A pesar de que simplemente presentamos un enfoque inicial desta análise, podería ampliarse de moitas maneiras utilizando a información proporcionada por MultiNet para deseñar modelos de predición ou investigar máis a fondo os patróns cíclicos.

MultiNet está construído para ser a semente de algo maior, un paquete comercial con máis funcionalidades e cunha aprendizaxe automática en función das necesidades dos datos. As aplicacións exitosas deste algoritmo, incluso fóra do mundo biolóxico, teñen un futuro prometedor para o uso real de MultiNet. Esta metodoloxía está deseñada para ser comprensible e utilizable por diferentes grupos de investigación, pedindo supostos mínimos dos datos e producindo resultados con recursos e tempo de cálculo razoables. O aumento da potencia de MultiNet podería traer un futuro brillante como un gran contribuínte á análise de datos en moitos campos de investigación.

Conclusión

Con este proxecto de investigación, pretendíamos superar os retos iniciais da nosa maquinaria cerebral e desenvolver novas ferramentas matemáticas capaces de procesar grandes cantidades de información cun deseño complexo. Particularmente, tivemos como obxectivo principal detectar patróns de correlación ocultos en conxuntos de datos epixenéticos de alta dimensión a través do estudo principal da súa topoloxía, que puideran estar relacionados coa estrutura tridimensional da cromatina.

A primeira hipótese matemática dunha rede non aleatoria (estruturada) de correlacións epixenéticas de curto e longo alcance validouse levando a investigar máis a fondo a súa estrutura local e global. O uso de técnicas matemáticas baseadas na topoloxía foi clave nesta investigación, xa que os métodos de análise de datos topolóxicos presentan unha alta flexibilidade para traballar con datos de gran tamaño e complexidade como os xenómicos [227]. Inspirados por esta metodoloxía, desenvolvemos os nosos propios procesos, modelos e aplicacións.

Para concluír, puidemos completar os nosos obxectivos de investigación, combinando un enfoque novidoso e o deseño de novos métodos de análise de datos topolóxicos enfocados especialmente ao manexo de estruturas de correlación de alta dimensión, cunha maior comprensión dos mecanismos epixenéticos. Cubríronse os retos iniciais e comprobáronse con éxito as consecuentes hipóteses xeradas da observación biolóxica.

Queremos resaltar o deseño transversal do presente traballo, aplicando procedementos matemáticos a datos xenómicos desde diferentes perspectivas: teórica (utilizando coñecementos topolóxicos e estatísticos); e computacional (deseñando de MultiNet). Este deseño está en consonancia coas necesidades actuais de investigación de análise de datos, onde téñense que analizar gran cantidade e heteroxeneidade de datos de diferentes campos.

Este traballo abraza o deseño multidisciplinar de xerar hipóteses biolóxicas da observación para resolver analíticamente con técnicas matemáticas avanzadas, facendo posible unha gran evolución nos dous campos que se complementan mutuamente.

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

Bibliography

[1] Egger, G., Liang, G., Aparicio, A. et al. "Epigenetics in human disease and prospects for epigenetic therapy". Nature 429, 457–463 (2004).

[2] Adam Purcell. "Basic Biology: An Introduction", New Zealand ISBN Agency, National Library of New Zealand (2018).

[3] Mills, M.C., Rahal, C. "A scientometric review of genome-wide association studies". Commun Biol 2, 9 (2019).

[4] Xie, Nina et al. "Novel Epigenetic Techniques Provided by the CRISPR/Cas9 System." Stem cells international vol.7834175 (2018)

[5] https://www.encodeproject.org/

[6] https://www.ncbi.nlm.nih.gov/

[7] Emmert-Streib F., Dehmer M., Haibe-Kains B. "Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks". Front Cell Dev. Biol. 2:38 (2014).

[8] Deshpande R., VanderSluis B., Myers CL. "Comparison of Profile Similarity Measures for Genetic Interaction Networks". PLoS ONE 8(7): e68664 (2013).

[9] Fortney, K., Xie, W., Kotlyar, M., Griesman, J., Kotseruba, Y., and Jurisica, I. "NetwoRx: connecting drugs to networks and phenotypes in Saccharomyces cerevisiae". Nucleic Acids Res. 41, D720–D727 (2013).

[10] https://www.genome.jp/kegg/pathway.html

[11] Lind, M.I., Spagopoulou, F. "Evolutionary consequences of epigenetic inheritance". Heredity 121, 205–209 (2018).

[12] Weinhold, Bob. "Epigenetics: the science of change." Environmental health perspectives vol. 114,3 (2006).

[13] Nessa Carey. "The Epigenetics Revolution". Icon Books (2012).

[14] Holliday, R. "Epigenetics: A Historical Overview". Epigenetics, 1:2 76-80 (2006).

[15] C. H. Waddington. "The Epigenotype (1942)", International Journal of Epidemiology, Volume 41, Issue 1, Pages 10–13 (2012).

[16] Alegría-Torres, Jorge Alejandro et al. "Epigenetics and lifestyle." Epigenomics vol. 3,3 267-77 (2011).

[17] Mario F. Fraga et al. "Epigenetic differences arise during the lifetime of monozygotic twins". Proceedings of the National Academy of Sciences 102 (30) 10604-10609 (2005).

[18] Francine E. Garrett-Bakelman. "The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight". SCIENCE (2019).

[19] Morey C., Avner P. "The Demoiselle of X-Inactivation: 50 Years Old and As Trendy and Mesmerising As Ever". PLoS Genet 7(7): e1002212 (2011).

[20] Julia Romanowska, Anagha Joshi. "From Genotype to Phenotype: Through Chromatin", Genes, 10(2), 76 (2019).

[21] Mengwei Li et al. "EWAS Atlas: a curated knowledgebase of epigenome-wide association studies". Nucleic Acids Research, Volume 47, Issue D1, Pages D983–D988 (2019).

[22] Zhang, Bo et al. "Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm." Genome research vol. 23,9: 1522-40 (2013).

[23] Mary Beth Terry, Lissette Delgado-Cruzata, Neomi Vin-Raviv, Hui Chen Wu & Regina M. Santella. "DNA methylation in white blood cells". Epigenetics, 6:7, 828-837 (2011).

[24] Fraser HB, Lam LL, Neumann SM, Kobor MS. "Population-specificity of human DNA methylation". Genome Biol. 13(2):R8 (2012).

[25] Zhang, Fang Fang et al. "Significant differences in global genomic DNA methylation by gender and race/ethnicity in peripheral blood." Epigenetics vol. 6,5 623-9 (2011).

[26] Kurdyukov, Sergey, and Martyn Bullock. "DNA Methylation Analysis: Choosing the Right Method." Biology vol. 5,1 3. (2016).

[27] Du, P., Zhang, X., Huang, C. et al. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis". BMC Bioinformatics 11, 587 (2010).

[28] Robinson, Mark D. et al. “Statistical methods for detecting differentially methylated loci and regions.” Frontiers in genetics vol. 5 324. (2014).

[29] Dao-Peng Chen, Ying-Chao Lin, Cathy S. J. Fann. “Methods for identifying differentially methylated regions for sequence- and array-based data”. Briefings in Functional Genomics, Volume 15, Issue 6, Pages 485–490 (2016).

[30] Ritchie, ME., Phipson, B., Wu, D., Hu, Y., Law, CW., Shi, W., and Smyth, GK. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” Nucleic Acids Research 43(7), e47 (2015).

[31] Izaskun Mallona, Susanna Aussó, Anna Díez-Villanueva, Víctor Moreno, Miguel A. Peinado. “Modular dynamics of DNA co-methylation networks exposes the functional organization of colon cancer cells genome”. bioRxiv 428730 (2018).

[32] Martin TC., Yet I., Tsai PC., Bell JT. “coMET: visualisation of regional epigenome-wide association scan results and DNA co-methylation patterns”. BMC Bioinformatics. 16(1):131 (2015).

[33] Sun, L., Sun, S. “Within-sample co-methylation patterns in normal tissues”. BioData Mining 12, 9 (2019).

[34] Sofer T. et al. “A-clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure”. Bioinformatics, 29, 2884–2891 (2013).

[35] Jaffe AE., Murakami P., Lee H., Leek JT., Fallin DM., Feinberg AP., Irizarry RA. “Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies.” International journal of epidemiology, 41(1), 200–209 (2012).

[36] Butcher L.M., Beck S. “Probe Lasso: a novel method to rope in differentially methylated regions with 450k DNA methylation data”. Methods, 72, 21–28 (2015).

[37] Haohan Wang, Benjamin J. Lengerich, Bryon Aragam, Eric P. Xing, “Precision Lasso: accounting for correlations and linear dependencies in high-dimensional genomic data”, Bioinformatics, Volume 35, Issue 7, 01, Pages 1181–1187 (2019).

[38] Wang, Tao et al. “Estimating DNA methylation levels by joint modeling of multiple methylation profiles from microarray data.” Biometrics vol. 72,2: 354-63 (2016).

[39] Peters T.J. et al. “De novo identification of differentially methylated regions in the human genome”. Epigenet. Chromatin, 8, 1–16 (2015).

[40] Ruiz-Arenas C., González J.R. "Redundancy analysis allows improved detection of methylation changes in large genomic regions". BMC Bioinformatics 18, 553 (2017).

[41] Hodges E. et al. "High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing". Genome Res., 19, 1593–1605 (2009).

[42] Guy Leonard Kouemou. "History and Theoretical Basics of Hidden Markov Models". EADS Deutschland GmbH, Germany.

[43] Lawrence R. Rabiner. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". Proceedings of the IEEE, Vol. 77, No. 2 (1989).

[44] Gary A. Churchill. "Stochastic Models for Heterogeneous DNA Sequences." Pergamon Press plc, Society for Mathematical Biology. Bulletin of Mathematical Biology Vol. 51, No 1, pp. 79-94 (1989).

[45] Rafael A. Irizarry et al. "Redefning CpG islands using hidden Markov models". Biostatistics, 11, 3, pp. 499-514 (2010).

[46] Larson, J., Yuan, G. "A hidden Markov model for detecting multi-gene chromatin domains". BMC Bioinformatics 11, O5 (2010).

[47] Linghao Shen, Jun Zhu, Shuo-Yen Robert Li, Xiaodan Fan. "Detect differentially methylated regions using non-homogeneous hidden Markov model for methylation array data". Bioinformatics, Volume 33, Issue 23, Pages 3701–3708 (2017).

[48] Farhad Shokoohi et al. "A Hidden Markov Model for Identifying Differentially Methylated Sites in Bisulfite Sequencing Data". Biometrics. 75(1):210-221 (2019).

[49] Aryee MJ., Jaffe AE., Corrada-Bravo H., Ladd-Acosta C., Feinberg AP., Hansen KD., Irizarry RA. "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays." Bioinformatics, 30(10), 1363–1369 (2014).

[50] Pal Sangita, and Jessica K. Tyler. "Epigenetics and aging." Science advances vol. 2,7 e1600584. (2016).

[51] Karen Hodgson et al. "Epigenetic Age Acceleration Assessed with Human White-Matter Images". Journal of Neuroscience 37 (18) 4735-4743 (2017).

[52] Horvath, Steve. "DNA methylation age of human tissues and cell types." Genome biology vol. 14,10 R115 (2013).

[53] Vidaki A, Kayser M. "From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence". Genome Biol. 18:238 (2017).

[54] Shi, Yuan Yuan et al. "Epigenetic modification of gene expression in honey bees by heterospecific gland secretions." PloS one vol. 7,8 e43727 (2012).

[55] Sato, Takahiro et al. "Transcriptional Selectivity of Epigenetic Therapy in Cancer". Cancer research vol. 77,2 470-481 (2017).

[56] Sharma, Shikhar et al. "Epigenetics in cancer." Carcinogenesis vol. 31,1 27-36 (2010).

[57] Joanna Lewandowska, Agnieszka Bartoszek. "DNA methylation in cancer development, diagnosis and therapy—multiple opportunities for genotoxic agents to act as methylome disruptors or remediators". Mutagenesis, Volume 26, Issue 4 Pages 475–487 (2011).

[58] Aref-Eshghi, Erfan et al. "Genomic DNA Methylation-Derived Algorithm Enables Accurate Detection of Malignant Prostate Tissues." Frontiers in oncology vol. 8 100 (2018).

[59] Eshraghi AA., Liu G, Kay SS., et al. "Epigenetics and Autism Spectrum Disorder: Is There a Correlation?". Front Cell Neurosci. 12:78 (2018).

[60] Moosavi A. Motevalizadeh Ardekani A. "Role of Epigenetics in Biology and Human Diseases". Iran Biomed J. 20(5):246–258 (2016).

[61] Dekker, J., Belmont, A., Guttman, M. et al. "The 4D nucleome project". Nature 549, 219–226 (2017).

[62] https://www.youtube.com/watch?v=4Z4KwuUfh0A

[63] Jacobson, Elsie et al. "A potential role for genome structure in the translation of mechanical force during immune cell development." Nucleus (Austin, Tex.) vol. 7,5: 462-475 (2016).

[64] Williams A., Spilianakis CG., Flavell RA. "Interchromosomal association and gene regulation in trans". Trends Genet. 26(4):188-197 (2010).

[65] Maass PG., Barutcu AR., Rinn JL. "Interchromosomal interactions: A genomic love story of kissing chromosomes". J Cell Biol. 218(1):27-38 (2019).

[66] Flavahan WA., Drier Y., Liau BB., et al. "Insulator dysfunction and oncogene activation in IDH mutant gliomas". Nature. 529(7584):110-114 (2016).

[67] Szalaj, P., Plewczynski, D. “Three-dimensional organization and dynamics of the genome”. Cell Biol Toxicol 34, 381–404 (2018).

[68] Jérôme D. Robin, Frédérique Magdinier. “High-order chromatin organization in diseases: from chromosomal position effect to phenotype variegation”. Handbook of Epigenetics (Second edition). The New Molecular and Medical Genetics Pages 73-92 (2017).

[69] Fortin, J., Hansen, K.D. “Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data”. Genome Biol 16, 180 (2015).

[70] Nothjunge, S., Nührenberg, T.G., Grüning, B.A. et al. “DNA methylation signatures follow preformed chromatin compartments in cardiac myocytes”. Nat Commun 8, 1667 (2017).

[71] Zhang, Ling et al. “DNA Methylation Landscape Reflects the Spatial Organization of Chromatin in Different Cells”. Biophysical Journal Volume 113, Issue 7, Pages 1395-1404 (2017).

[72] Joanna Achinger-Kawecka & Susan J Clark. “Disruption of the 3D cancer genome blueprint”, EPIGENOMICS VOL. 9, NO. 1 (2016).

[73] Valton, Anne-Laure, and Job Dekker. “TAD disruption as oncogenic driver”, Current opinion in genetics & development vol. 36: 34-40 (2016).

[74] Achinger-Kawecka, J., Valdes-Mora, F., Luu, P. et al. “Epigenetic reprogramming at estrogen-receptor binding sites alters 3D chromatin landscape in endocrine-resistant breast cancer”. Nat Commun 11, 320 (2020).

[75] Ling, J., Hoffman, A. “Epigenetics of Long-Range Chromatin Interactions”. Pediatr Res 61, 11–16 (2007).

[76] Jeong, Mira et al. “Large conserved domains of low DNA methylation maintained by Dnmt3a.” Nature genetics vol. 46,1: 17-23 (2014).

[77] Dekker, J., Marti-Renom, M. & Mirny, L. “Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data”. Nat Rev Genet 14, 390–403 (2013).

[78] Sarnataro S., Chiariello AM., Esposito A., Prisco A., Nicodemi M. “Structure of the human chromosome interaction network”. PLoS ONE 12(11) (2017).

[79] Cremer T. et al., “Chromosome territories, interchromatin domain compartment, and nuclear matrix: an integrated view of the functional nuclear architecture”. Crit. Rev. Eukaryot Gene Expr. 10: 179–212 (2000).

[80] Kioussis D. "Gene regulation: kissing chromosomes". Nature 435: 579–580 (2005).

[81] Kaufmann S. et al., "Inter-Chromosomal Contact Networks Provide Insights into Mammalian Chromatin Organization". PLoS ONE 10(5): e0126125 (2015).

[82] Ornella Affinito et al. "Nucleotide distance influences co-methylation between nearby CpG sites". Genomics 112 144-150 (2020).

[83] Evan Gtev et al., "CoMeBack: DNA Methylation Array Data Analysis for Co-Methylated Regions". Oxford University Press (2020).

[84] Zhang X., Jeong M., Huang X., et al. "Large DNA Methylation Nadirs Anchor Chromatin Loops Maintaining Hematopoietic Stem Cell Identity". Mol. Cell 78(3):506-521.e6 (2020).

[85] Sun, Luyang et al. "Chromatin Architectural Changes during Cellular Senescence and Aging." Genes vol. 9,4 211. (2018).

[86] Koestler, Devin C et al. "Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL)." BMC bioinformatics vol. 17 120 (2016).

[87] Salas, L.A., Koestler, D.C., Butler, R.A. et al. "An optimized library for reference-based deconvolution of whole-blood biospecimens assayed using the Illumina HumanMethylationEPIC BeadArray". Genome Biol 19, 64 (2018).

[88] Lehne, B., Drong, A.W., Loh, M. et al. "A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies". Genome Biol 16, 37 (2015).

[89] Fisher, R. A. "Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population". Biometrika. 10 (4): 507–521 (1915).

[90] Fisher, R. A. "On the 'probable error' of a coefficient of correlation deduced from a small sample". Metron. 1: 3–32 (1921).

[91] Saelens, W., Cannoodt, R. & Saeys, Y. "A comprehensive evaluation of module detection methods for gene expression data". Nat Commun 9, 1090 (2018).

[92] Yu, D., Zhang, Z., Glass, K. et al. "New Statistical Methods for Constructing Robust Differential Correlation Networks to characterize the interactions among microRNAs". Sci Rep 9, 3499 (2019).

[93] Fern, Xiaoli & Brodley, Carla & Friedl, Mark. "Correlation Clustering for Learning Mixtures of Canonical Correlation Models", (2005).

[94] Darst, B.F., Malecki, K.C. & Engelman, C.D. "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data". BMC Genet 19, 65 (2018).

[95] Frédéric Chazal and Bertrand Michel. "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists". arXiv:1710.04019 [math.ST] (2017).

[96] Edelsbrunner et al. "Topological persistence and simplification". Discrete Comput. Geom. 28:511-533 (2002).

[97] Carlsson, G. "Topology and data", AMS Bulletin, 46(2):255-308 (2009).

[98] Maria C., Boissonnat JD., Glisse M., Yvinec M. "The Gudhi Library: Simplicial Complexes and Persistent Homology". Lecture Notes in Computer Science, vol 8592. Springer, Berlin, Heidelberg (2014).

[99] Fasy et al. "Introduction to the R package TDA". arXiv:1411.1830 [cs.MS] (2014).

[100] https://github.com/paultpearson/TDAmapper/

[101] Ann E. Sizemore et al. "The importance of the whole: Topological data analysis for the network neuroscientist". Network Neuroscience 3:3, 656-673 (2019).

[102] Lum, P., Singh, G., Lehman, A. et al. "Extracting insights from the shape of complex data using topology". Sci. Rep. 3, 1236 (2013).

[103] https://www.ayasdi.com/

[104] https://www.businesswire.com/news/home/20190924005402/en/Symphony-AyasdiAI-Launches-Next-Generation-AI-Solution-Anti-Money

[105] https://www.youtube.com/watch?v=8nUBqawu41k

[106] Cámara, Pablo G. "Topological methods for genomics: present and future directions." Current opinion in systems biology vol. 1 95-101 (2017).

[107] Nicolau, Monica et al. "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival." Proceedings of the National Academy of Sciences of the United States of America vol. 108,17: 7265-70 (2011).

[108] Saggar, M., Sporns, O., Gonzalez-Castillo, J. et al. "Towards a new approach to reveal dynamical organization of the brain using topological data analysis.", Nat Commun 9, 1399 (2018).

[109] James R. Munkres. “Elements of Algebraic Topology”. Addison-Wesley, Menlo Park, CA, 1984.

[110] Morse, M. “The calculus of the variations in the large”. Amer. Math. Soc. Colloquium Publication, vol. 18 (1934).

[111] Milnor, J. “Morse Theory”. Ann. of Math. Studies, vol. 51, Princeton University Press (1963).

[112] Hatcher, A. “Algebraic Topology”. Cambridge University Press, Cambridge (2002).

[113] S. Biasotti et al. “Reeb graphs for shape analysis and applications”, Theoretical Computer Science Volume 392, Issues 1–3, 28 Pages 5-22 (2008).

[114] de Silva, V., Munch, E. & Patel, A. “Categorified Reeb Graphs”. Discrete Comput Geom 55, 854–906 (2016).

[115] G. Reeb, “Sur les points singuliers d’une forme de Pfaff completement integrable ou d’une fonction numerique”, Comptes Rendus Hebdomadaires des S’eances de l’Académie des Sciences 222 847–849 (1946).

[116] Y. Shinagawa, T.L. Kunii, Y.L. Kergosien. “Surface coding based on Morse theory”, IEEE Computer Graphics and Applications 11 66–78 (1991).

[117] V. Robins. “Towards computing homology from finite approximations”. Proceedings of the 14th Summer Conference on General Topology and its Applications, Topology Proc. 24, 503-532 (1999).

[118] Wadhwa RR, Williamson DFK, Dhawan A, Scott JG.“TDAstats: R pipeline for computing persistent homology in topological data analysis.” Journal of Open Source Software, 3(28), 860 (2018).

[119] Aktas, M.E., Akbas, E. & Fatmaoui, A.E. “Persistence homology of networks: methods and applications”. Appl Netw Sci 4, 61 (2019).

[120] David Cohen-Steiner, Herbert Edelsbrunner, John Harer. “Stability of Persistence Diagrams”. Discrete Comput Geom 37:103–120 (2007).

[121] Benzekry, S., Tuszynski, J.A., Rietman, E.A. et al. “Design principles for cancer therapy guided by changes in complexity of protein-protein interaction networks”. Biol Direct 10, 32 (2015).

[122] Kevin Emmett, Benjamin Schweinhart, Raul Rabadan. “Multiscale Topology of Chromatin Folding”, arXiv:1511.01426 [q-bio.GN] (2015).

[123] Benzekry, S., Tuszynski, J.A., Rietman, E.A. et al. "Design principles for cancer therapy guided by changes in complexity of protein-protein interaction networks". Biol Direct 10, 32 (2015).

[124] Ali Nabi Duman and Harun Pirim. "Gene Coexpression Network Comparison via Persistent Homology". International Journal of Genomics 10.1155/7329576 (2018).

[125] Meng, Z., Anand, D.V., Lu, Y. et al. "Weighted persistent homology for biomolecular data analysis". Sci Rep 10, 2079 (2020).

[126] Kannan, H., Saucan, E., Roy, I. et al. "Persistent homology of unweighted complex networks via discrete Morse theory". Sci Rep 9, 13817 (2019).

[127] Gurjeet Singh , Facundo Mémoli and Gunnar Carlsson. "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition". Eurographics Symposium on Point-Based Graphics (2007).

[128] Mathieu Carrière et al. "Statistical analysis and parameter selection of Mapper". Journal of Machine Learning Research 19 1-39 (2018).

[129] Caleb Geniesse, Olaf Sporns, Giovanni Petri, and Manish Saggar. "Generating dynamical neuroimaging spatiotemporal representations (DyNeuSR) using topological data analysis". Network Neuroscience 3:3, 763-778 (2019).

[130] Rachel Jeitziner, Mathieu Carrière, Jacques Rougemont, Steve Oudot, Kathryn Hess, Cathrin Brisken. "Two-Tier Mapper, an unbiased topology-based clustering method for enhanced global gene expression analysis". Bioinformatics, Volume 35, Issue 18, 15, Pages 3339–3347 (2019).

[131] Yu-Min Chung, William Cruse, Austin Lawson. "A Persistent Homology Approach to Time Series Classification". arXiv:2003.06462 [stat.ME] (2020).

[132] Oyama A, Hiraoka Y, Obayashi I, et al. "Hepatic tumor classification using texture and topology analysis of non-contrast-enhanced three-dimensional T1-weighted MR images with a radiomics approach". Sci. Rep. 9(1):8764. (2019).

[133] Glass RJ., Glass LM., Beyeler WE., et al. "Targeted Social Distancing Designs for Pandemic Influenza. Emerging Infectious Diseases". 12(11):1671-1681 (2006).

[134] Norton HK., Emerson DJ., Huang H., Kim J., Titus KR., Gu S., Bassett DS., Phillips-Cremins JE. "Detecting hierarchical genome folding with network modularity". Nat Methods; 15(2):119-122 (2018).

[135] Stijn van Dongen, Anton J. Enright. “Metric distances derived from cosine similarity and Pearson and Spearman correlations”. arXiv:1208.3145 [stat.ME] (2012).

[136] Xie, T., Pan, S., Zheng, H. et al. “PEG10 as an oncogene: expression regulatory mechanisms and role in tumor progression”. Cancer Cell Int 18, 112 (2018).

[137] Clement Lee, Darren J Wilkinson. “A Review of Stochastic Block Models and Extensions for Graph Clustering”. Applied Network Science, Springer Science and Business Media LLC, vol. 4 2364-8228 (2019).

[138] M. E. J. Newman. “Networks—An Introduction”. Oxford: Oxford University Press (2010).

[139] Aaron Clauset, M. E. J. Newman, Cristopher Moore. “Finding community structure in very large networks”. arXiv:cond-mat/0408187 [cond-mat.stat-mech] (2004).

[140] Watts, D. J., and Strogatz, S. H. “Collective dynamics of “small-world” networks”. Nature 393, 440–442 (1998).

[141] Alexander P.Kartun-Giles, Ginestra Bianconi. “Beyond the clustering coefficient: A topological analysis of node neighbourhoods in complex networks”. Chaos, Solitons & Fractals: X Volume 1, 100004 (2019).

[142] Matthew Kahle. “Topology of random clique complexes”. Discrete Mathematics Volume 309, Issue 6, Pages 1658-1671 (2009).

[143] Funke T, Becker T. “Stochastic block models: A comparison of variants and inference methods”. PLoS ONE 14(4): e0215296 (2019).

[144] Morrow BE, McDonald-McGinn DM, Emanuel BS, Vermeesch JR, Scambler PJ. “Molecular genetics of 22q11.2 deletion syndrome”. Am J Med Genet A. 176(10):2070-2081 (2018).

[145] Qiumei Du, M. Teresa de la Morena and Nicolai S. C. van Oers. “The Genetics and Epigenetics of 22q11.2 Deletion Syndrome”. Front. Genet., 06 February (2020).

[146] Harricharran T, Ogunwobi OO. “Oxytocin and oxytocin receptor alterations, decreased survival, and increased chemoresistance in patients with pancreatic cancer”. Hepatobiliary Pancreat Dis Int. 19(2):175-180 (2020).

[147] Erdös, P.; Rényi, A. “On Random Graphs. I.”. Publicationes Mathematicae 6: 290–297 (1959)

[148] Albert-László Barabási & Réka Albert. "Emergence of scaling in random networks". Science 286: 509-512 (1999).

[149] Yook SH, Jeong H, Barabási AL, Tu Y. "Weighted evolving networks". Phys Rev Lett. 86(25):5835-5838 (2001).

[150] He, Karen Y et al. "Big Data Analytics for Genomic Medicine." International journal of molecular sciences vol. 18,2 412. (2017).

[151] MacQueen, J. B. "Some Methods for classification and Analysis of Multivariate Observations". University of California Press. pp. 281–297. MR 0214227 (1967).

[152] Lloyd, S. P. "Least squares quantization in PCM. Technical Note, Bell Laboratories". Published in 1982 in IEEE Transactions on Information Theory, 28, 128–137 (1957, 1982).

[153] Forgy, E. W. "Cluster analysis of multivariate data: efficiency vs interpretability of classifications". Biometrics, 21, 768–769 (1965).

[154] Hartigan, J. A. and Wong, M. A. "Algorithm AS 136: A K-means clustering algorithm". Applied Statistics, 28, 100–108 (1979).

[155] G.A.F. Seber, "Multivariate Observations", Wiley, (1984).

[156] Wang Z., Wu X., Wang Y. "A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip". BMC Bioinformatics. 19(Suppl 5):115 (2018).

[157] Norman Cliff. "Ordinal methods for behavioral data analysis". Routledge (1996).

[158] Carvajal-Rodríguez, A., de Uña-Alvarez, J. & Rolán-Alvarez, E. "A new multitest correction (SGoF) that increases its statistical power when increasing the number of tests". BMC Bioinformatics 10, 209 (2009).

[159] Albert, A. and J. A. Anderson. "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models." Biometrika 71: 1-10 (1984).

[160] Firth, D. "Bias Reduction of Maximum Likelihood Estimates." Biometrika 80: 27-38 (1993).

[161] David W. Hosmer Jr, Stanley Lemeshow, Rodney X. Sturdivant, "Applied Logistic Regression", Wiley, (2013).

[162] Breiman, L. "Random Forests". Machine Learning 45, 5–32 (2001).

[163] Díaz-Uriarte, R., Alvarez de Andrés, S. "Gene selection and classification of microarray data using random forest". BMC Bioinformatics 7, 3 (2006).

[164] Menze, B.H., Kelm, B.M., Masuch, R. et al. "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data". BMC Bioinformatics 10, 213 (2009).

[165] Guintivano J, Aryee MJ, Kaminsky ZA. "A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression". Epigenetics. 8(3):290-302 (2013).

[166] Mathieu Carriere, Bertrand Michel, Steve Y. Oudot. "Statistical analysis and parameter selection for Mapper". Journal of Machine Learning Research, Microtome Publishing (2018).

[167] Tamal K. Dey and Facundo Memoli and Yusu Wang. "Topological Analysis of Nerves, Reeb Spaces, Mappers, and Multiscale Mappers". arXiv:1703.07387 [cs.CG] (2017).

[168] Paweł Dłotko. "Ball mapper: a shape summary for topological data analysis". arXiv:1901.07410 [math.AT] (2019).

[169] Francisco Belchí and Jacek Brodzki and Matthew Burfitt and Mahesan Niranjan. "A numerical measure of the instability of Mapper-type algorithms". arXiv math.AT 1906.01507 (2019).

[170] Dhingra R., Kwee LC., Diaz-Sanchez D., Devlin RB., Cascio W., Hauser ER., et al. "Evaluating DNA methylation age on the Illumina MethylationEPIC Bead Chip". PLoS ONE 14(4): e0207834 (2019).

[171] Xiao F-H., Wang H-T. and Kong Q-P. "Dynamic DNA Methylation During Aging: A Prophet of Age-Related Outcomes". Front. Genet. 10:107 (2019).

[172] Heyn, Holger et al. "Distinct DNA methylomes of newborns and centenarians." Proceedings of the National Academy of Sciences of the United States of America vol. 109,26 (2012).

[173] Hannum, Gregory et al. "Genome-wide methylation profiles reveal quantitative views of human aging rates." Molecular cell vol. 49,2 359-367 (2013).

[174] Salameh Y., Bejaoui Y., El Hajj N. "DNA Methylation Biomarkers in Aging and Age-Related Diseases". Front Genet. 11:171 (2020).

[175] Field AE., Robertson NA., Wang T., Havas A., Ideker T., Adams PD. "DNA Methylation Clocks in Aging: Categories, Causes, and Consequences". Mol Cell; 71(6):882-895 (2018).

[176] Freire-Aradas A., Phillips C., Girón-Santamaría L., et al. "Tracking age-correlated DNA methylation markers in the young". Forensic Sci. Int. Genet. 36:50-59 (2018).

[177] Daniela Punzo et al. "Age-Related Changes in d-Aspartate Oxidase Promoter Methylation Control Extracellular d-Aspartate Levels and Prevent Precocious Cell Death during Brain Aging". Journal of Neuroscience, 36 (10) 3064-3078 (2016).

[178] Paolo Garagnani et al. "Methylation of ELOVL2 gene as a new epigenetic marker of age". Aging Cell (2012).

[179] Salpea P., Russanova VR., Hirai TH., et al. "Postnatal development- and age-related changes in DNA-methylation patterns in the human genome". Nucleic Acids Res. 40(14):6477-6494 (2012).

[180] Ong ML., Holbrook JD. "Novel region discovery method for Infinium 450K DNA methylation data reveals changes associated with aging in muscle and neuronal pathways". Aging Cell 13(1):142-155 (2014).

[181] Nady El Hajj, Marcus Dittrich, Thomas Haaf. "Epigenetic dysregulation of protocadherins in human disease". Seminars in Cell & Developmental Biology Volume 69, Pages 172-182 (2017).

[182] Sang-Eun Jung et al. "DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples", Forensic Science International: Genetics, V38 1-8 (2019).

[183] Sae Rom Hong et al. "DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers". Forensic Science International: Genetics, V29 118-125 (2017).

[184] David Baidoe-Ansah, M. Sadman Sakib, Shaobo Jia, Andre Fischer, Rahul Kaushik, Alexander Dityatev, "Epigenetic mechanism of carbohydrate sulfotransferase 3 (CHST3) downregulation in the aging brain". bioRxiv 741355; (2019).

[185] Marttila, S., Kananen, L., Häyrynen, S. et al. "Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression". BMC Genomics 16, 179 (2015).

[186] Zhang, Y., Wilson, R., Heiss, J. et al. "DNA methylation signatures in peripheral blood strongly predict all-cause mortality". Nat Commun 8, 14617 (2017).

[187] Levine ME., Lu AT., Quach A., et al. "An epigenetic biomarker of aging for lifespan and healthspan". Aging (Albany NY) 10(4):573-591 (2018).

[188] Prasad R., Jho EH. "A concise review of human brain methylome during aging and neurodegenerative diseases". BMB Rep 52(10):577-588 (2019).

[189] https://genomics.senescence.info/genes/index.html

[190] Kirby, Marie K et al. "Genome-wide DNA methylation measurements in prostate tissues uncovers novel prostate cancer diagnostic biomarkers and transcription factor binding patterns." BMC cancer vol. 17,1 273 (2017).

[191] Bjerre MT., Strand SH., Nørgaard M., et al. "Aberrant DOCK2, GRASP, HIF3A and PKFP Hypermethylation has Potential as a Prognostic Biomarker for Prostate Cancer". Int J Mol Sci. 20(5):1173. Published (2019).

[192] Bilir B., Sharma NV., Lee J., et al. "Effects of genistein supplementation on genome wide DNA methylation and gene expression in patients with localized prostate cancer". Int J Oncol 51(1):223-234 (2017).

[193] Zhang S., Xu Y., Hui X., et al. "Improvement in prediction of prostate cancer prognosis with somatic mutational signatures". J Cancer. 8(16):3261-3267 (2017).

[194] Li W., Middha M., Bicak M., et al. "Genome-wide Scan Identifies Role for AOX1 in Prostate Cancer Survival". Eur Urol. 74(6):710-719 (2018).

[195] He, Sha et al. "Wnt3a: functions and implications in cancer." Chinese journal of cancer vol. 34,12 554-62. (2015).

[196] Nikas JB., Nikas EG. "Genome-Wide DNA Methylation Model for the Diagnosis of Prostate Cancer". ACS Omega. 4(12):14895-14901. (2019).

[197] Neuhaus EM., Zhang W., Gelis L., et al. "Activation of an olfactory receptor inhibits proliferation of prostate cancer cells". The Journal of Biological Chemistry. 284(24):16218-16225 (2009).

[198] Xu N., Wu YP., Ke ZB., et al. "Identification of key DNA methylation-driven genes in prostate adenocarcinoma: an integrative analysis of TCGA methylation data". J Transl Med. 17(1):311 (2019).

[199] https://www.intogen.org/search

[200] G. Yu, LG. Wang, GR. Yan, QY. He. "DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis". Bioinformatics, 31(4):608-609 (2015).

[201] Moghoofei M., Keshavarz M., Ghorbani S., et al. "Association between human papillomavirus infection and prostate cancer: A global systematic review and meta-analysis". Asia Pac. J. Clin. Oncol. 15(5):e59e67 (2019).

[202] Luo, Yanxin et al. "Differences in DNA methylation signatures reveal multiple pathways of progression from adenoma to colorectal cancer." Gastroenterology vol. 147,2 418-29.e8 (2014).

[203] Vladimir A. Naumov, Edward V. Generozov, Natalya B. Zaharjevskaya, Darya S. Matushkina, Andrey K. Larin, Stanislav V. Chernyshov, Mikhail V. Alekseev, Yuri A. Shelygin & Vadim M. Govorun. "Genome-scale analysis of DNA methylation in colorectal cancer using Infinium HumanMethylation450 BeadChips", Epigenetics, 8:9, 921-934 (2013).

[204] Ishak, M., Baharudin, R., Rose, I.M., Sagap, I., Mazlan, L., Azman, Z.A.M., Abu, N., Jamal, R., Lee, L.-H., Mutalib, N.S.A. "Genome-Wide Open Chromatin Methylome Profiles in Colorectal Cancer". Biomolecules 10, 719 (2020).

[205] Dickinson RE., Dallol A., Bieche I., et al. "Epigenetic inactivation of SLIT3 and SLIT1 genes in human cancers". Br J Cancer. 91(12):2071-2078 (2004).

[206] Li, J., Liao, Y., Huang, J. et al. "Epigenetic silencing of ADAMTS5 is associated with increased invasiveness and poor survival in patients with colorectal cancer". J Cancer Res Clin Oncol 144, 215–227 (2018).

[207] Zhijin Li et al. "DNA methylation and gene expression profiles characterize epigenetic regulation of lncRNAs in colon adenocarcinoma". Journal of cellular biochemistry, Volume 121, Issue3, Pages 2406-2415 (2020).

[208] Hauptman, N., Jevšinek Skok, D., Spasovska, E. et al. "Genes CEP55, FOXD3, FOXF2, GNAO1, GRIA4, and KCNA5 as potential diagnostic biomarkers in colorectal cancer". BMC Med. Genomics 12, 54 (2019).

[209] Weber L., Al-Refae K., Ebbert J., et al. "Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis". PLoS One. 12(3):e0172491 (2017).

[210] Cassandri, M., Smirnov, A., Novelli, F. et al. "Zinc-finger proteins in health and disease". Cell Death Discov. 3, 17071 (2017).

[211] Gerecke C., Scholtka B., Löwenstein Y., et al. "Hypermethylation of ITGA4, TFPI2 and VIMENTIN promoters is increased in inflamed colon tissue: putative risk markers for colitis-associated cancer". J Cancer Res Clin Oncol. 141(12):2097-2107 (2015).

[212] Li J., Chen C., Bi X., et al. "DNA methylation of CMTM3, SSTR2, and MDFI genes in colorectal cancer". Gene 630:1-7 (2017).

[213] Huang R., Gu W., Sun B., Gao L. "Identification of COL4A1 as a potential gene conferring trastuzumab resistance in gastric cancer based on bioinformatics analysis". Mol Med Rep. 17(5):6387-6396 (2018).

[214] Slattery ML., Lundgreen A., Welbourn B., Corcoran C., Wolff RK. "Genetic variation in selenoprotein genes, lifestyle, and risk of colon and rectal cancer". PLoS One. 7(5):e37312 (2012).

[215] Ishak, M.; Baharudin, R.; Mohamed Rose, I.; Sagap, I.; Mazlan, L.; Mohd Azman, Z.A.; Abu, N.; Jamal, R.; Lee, L.-H.; Ab Mutalib, N.S. "Genome-Wide Open Chromatin Methylome Profiles in Colorectal Cancer". Biomolecules 10, 719 (2020).

[216] Pellatt AJ., Mullany LE., Herrick JS., et al. "The TGFβ-signaling pathway and colorectal cancer: associations between dysregulated genes and miRNAs". J. Transl. Med. 16(1):191 (2018).

[217] Barrow TM., Klett H., Toth R., et al. "Smoking is associated with hypermethylation of the APC 1A promoter in colorectal cancer: the ColoCare Study". J Pathol; 243(3):366-375 (2017).

[218] Peng YN., Huang ML., Kao CH. "Prevalence of Depression and Anxiety in Colorectal Cancer Patients: A Literature Review". Int. J. Environ. Res. Public Health. 16(3):411 (2019).

[219] Lloyd, Shane et al. "Mental Health Disorders are More Common in Colorectal Cancer Survivors and Associated With Decreased Overall Survival". American Journal of Clinical Oncology: Volume 42 - Issue 4 p 355-362 (2019).

[220] Erfan Aref-Eshghi et al. "Evaluation of DNA Methylation Episignatures for Diagnosis and Phenotype Correlations in 42 Mendelian Neurodevelopmental Disorders". The American Journal of Human Genetics, Volume 106, Issue 3, Pages 356-370, ISSN 0002-9297 (2020).

[221] Kimura R., Nakata M., Funabiki Y., et al. "An epigenetic biomarker for adult high-functioning autism spectrum disorder". Sci. Rep. 9(1):13662 (2019).

[222] Vega-Benedetti, A.F., Loi, E., Moi, L. et al. "Clustered protocadherins methylation alterations in cancer". Clin Epigenet 11, 100 (2019).

[223] Olender T., Lancet D., Nebert DW. "Update on the olfactory receptor (OR) gene superfamily". Hum Genomics 3(1):87–97 (2008).

[224] L. Koehler, A. Fournel, K. Albertowski, V. Roessner, J. Gerber, C. Hummel, T. Hummel, M. Bensafi. "Impaired Odor Perception in Autism Spectrum Disorder Is Associated with Decreased Activity in Olfactory Cortex". Chemical Senses, Volume 43, Issue 8, Pages 627–634, (2018).

[225] Rochet, Marion et al. "Depression, Olfaction, and Quality of Life: A Mutual Relationship." Brain sciences vol. 8,5 80. (2018).

[226] Zou, L., Yang, Z., Wang, Y. et al. "What does the nose know? Olfactory function predicts social network size in human". Sci. Rep 6, 25026 (2016).

[227] Raúl Rabadán and Andrew J. Blumberg. "Topological data analysis for genomic and evolution, Topology in Biology". Cambridge University Press (2020).

[228] Trevor Hastie, Robert Tibshirani, Jerome Friedman. "The Elements of Statistical Learning". Springer, (2008).

[229] Alan Agresti. "Categorical Data Analysis". Wiley (2013).

[230] https://github.com/SPRADA1/MultiNet/tree/master

USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA