

ALMA MATER STUDIORUM · UNIVERSITÀ DI BOLOGNA

---

**DOTTORATO DI RICERCA IN**

**FISICA**

**Ciclo 33**

**Settore concorsuale:** 02/D1 - FISICA APPLICATA, DIDATTICA E  
STORIA DELLA FISICA

**Settore scientifico disciplinare:** FIS/07 - FISICA APPLICATA A  
BENI CULTURALI, AMBIENTALI, BIOLOGIA E MEDICINA

**CHARACTERIZATION OF DNA SEQUENCE PROPERTIES  
THROUGH NETWORK AND STATISTICAL  
APPROACHES**

**Presentata da:** Alessandra Merlotti

**Coordinatore Dottorato:**

Prof. Michele Cicoli

**Supervisore:**

Prof. Daniel Remondini

**Esame finale anno 2021**



## Abstract

In this thesis we will see that the DNA sequence is constantly shaped by the interactions with its environment at multiple levels, showing footprints of DNA methylation, of its 3D organization and, in the case of bacteria, of the interaction with the host organisms. In the first chapter, we will see that analyzing the distribution of distances between consecutive dinucleotides of the same type along the sequence, we can detect epigenetic and structural footprints. In particular, we will see that CG distance distribution allows to distinguish among organisms of different biological complexity, depending on how much CG sites are involved in DNA methylation. Moreover, we will see that CG and TA can be described by the same fitting function, suggesting a relationship between the two. We will also provide an interpretation of the observed trend, simulating a positioning process guided by the presence and absence of memory. In the end, we will focus on TA distance distribution, characterizing deviations from the trend predicted by the best fitting function, and identifying specific patterns that might be related to peculiar mechanical properties of the DNA and also to epigenetic and structural processes.

In the second chapter, we will see how we can map the 3D structure of the DNA onto its sequence. In particular, we devised a network-based algorithm that produces a genome assembly starting from its 3D configuration, using as inputs Hi-C contact maps. Specifically, we will see how we can identify the different chromosomes and reconstruct their sequences by exploiting the spectral properties of the Laplacian operator of a network.

In the third chapter, we will see a novel method for source clustering and source attribution, based on a network approach, that allows to identify host-bacteria interaction starting from the detection of Single-Nucleotide Polymorphisms along the sequence of bacterial genomes.

---

## Contents

---

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>From sequence to epigenetic and structural footprints</b>	<b>7</b>
2.1	Dinucleotide distance distribution . . . . .	7
2.2	CG distance distribution . . . . .	8
2.2.1	Fitting method . . . . .	9
2.2.2	Mammal genome analysis . . . . .	10
2.2.3	Genome analysis across different levels of biological complexity . .	10
2.2.4	The role of memory in CG positioning process . . . . .	14
2.2.5	The human genome and a new fitting function . . . . .	18
2.3	From CG to TA distance distribution . . . . .	21
2.3.1	TA random percolation . . . . .	22
2.3.2	TA in primates . . . . .	23
2.3.3	Deviations from the expected trend within human genome . . . .	24
2.4	Conclusions . . . . .	36
<b>3</b>	<b>Unfolding the genome: from chromosome structure to sequence</b>	<b>38</b>
3.1	Hi-C contact maps . . . . .	38
3.2	Hi-C data and genome assembly . . . . .	40
3.3	My contribution in the Genigma project . . . . .	41
3.3.1	The pipeline . . . . .	41
3.4	The Laplacian matrix of a network . . . . .	51
3.4.1	Properties of the Laplacian matrix . . . . .	52
3.4.2	Spectral clustering . . . . .	53
3.4.3	Laplacian embedding . . . . .	54

3.5	Chromosome identification through spectral clustering . . . . .	54
3.5.1	Results on a toy-model . . . . .	55
3.5.2	Results on a real Hi-C contact map . . . . .	56
3.6	Sequence reconstruction through Fiedler vector . . . . .	68
3.6.1	Results on a toy-model . . . . .	68
3.6.2	Results on a real Hi-C contact map . . . . .	68
3.7	Conclusions . . . . .	75
<b>4</b>	<b>From sequence to host-bacteria interactions</b>	<b>76</b>
4.1	The source attribution problem . . . . .	76
4.1.1	The case of <i>Salmonella enterica</i> serovar Typhimuirum and its monophasic variant . . . . .	78
4.2	A Network-based method to identify host-bacteria interactions . . . . .	78
4.2.1	Data set . . . . .	79
4.2.2	Source clustering results . . . . .	80
4.2.3	Source attribution results . . . . .	81
4.3	Validation of the method . . . . .	89
4.3.1	Data set . . . . .	89
4.3.2	Results . . . . .	89
4.4	Conclusions . . . . .	95
<b>5</b>	<b>Conclusions</b>	<b>96</b>
<b>6</b>	<b>Appendix A</b>	<b>99</b>
6.1	Comparison among the 16 dinucleotide distance distributions within hu- man genome . . . . .	99
<b>7</b>	<b>Appendix B</b>	<b>106</b>
7.1	TA distance distribution within mammal genomes . . . . .	106
<b>8</b>	<b>Appendix C</b>	<b>112</b>
<b>9</b>	<b>Appendix D</b>	<b>129</b>
9.1	Deviation at 91 bp from TA distance distribution . . . . .	129

# CHAPTER 1

---

## Introduction

---

In this study we will see how just simply reading the information content of the DNA sequence we can gain knowledge about processes that go beyond the sequence itself, detecting footprints of DNA methylation, of the 3D organization of the genome and, in case of bacteria such as *Salmonella enterica* serovar Typhimurium, of the interaction between the micro-organisms and their hosts. In particular, in this work, we will always refer to the DNA sequence as a 1D object composed by four letters, corresponding to four units composing its chains: A (adenine), C (cytosine), G (guanine), and T (thymine), called *nucleotides*.

In the first chapter, we will analyze the DNA sequence by considering as fundamental units the *dinucleotides*, (i.e. pairs of nucleotides) and focusing on the characterization of their distances along the sequence. In fact, recent studies revealed that dinucleotide distance can be a powerful tool for detecting DNA properties [1, 2], such as the identification of CpG islands [3] and the characterization of epigenomic regulation through methylation [4, 5]. In particular, Paci et al. [4] highlighted a peculiar feature of mammals CG dinucleotides: the distributions of the distances between consecutive CG show an exponential tail, whereas all non-CG distributions are characterized by heavier tails, more similar to a power law. This might be due to the specific role that CGs play inside mammals genomes, since they are the preferential sites of methylation, a fundamental epigenetic mechanism involved in gene regulation [6, 7, 8, 9, 10] and structural conformation of chromatin [11, 12]. Therefore, in the first chapter we will focus on the characterization of the whole CG distributions in a set of mammal model organisms and subsequently we will extend the analysis to a larger set of 4425 genomes, belonging to a wide range of organism categories (bacteria, protozoa, plants, fungi, invertebrates, mammal and non-mammal vertebrates) in order to better understand the heterogeneous

scenario found among non-mammals [4] and to obtain a global picture associated to this particular feature. The results of these analyses are published in [13].

We will then focus on human genome and we will see that CG and TA distance distributions can be described by the same mathematical function, that is a shifted power-law with exponential tail, suggesting a relationship between the two. In particular, we will propose an interpretation for the observed trend, associating respectively the power-law and the exponential trend to a positioning process with and without memory. In the end, we will focus on TA distance distribution, detecting deviations from the trend predicted by the best fitting function, which are characterized by specific sequence patterns that might be related to peculiar mechanical properties of the DNA and also to epigenetic and structural processes.

In the second chapter, we will see how we can use the information about the 3D proximity of DNA fragments within the nucleus to obtain its sequence. In other words, we will see how we can get a genome assembly starting from Hi-C data. In fact, Hi-C (High-throughput Chromosome conformation capture) is a technique that allows to identify chromatin interactions across the entire genome [14, 15], providing information about the spatial proximity between pairs of DNA fragments sampled from a population of millions of cells [16]. In the last 7 years, Hi-C data have gained more and more attention in context of genome assembly [17, 18, 19, 20, 21] since, unlike the usual DNA fragments produced by NGS (Next-Generation sequencing) technique, they have the advantage of carrying information about their spatial proximity within the nucleus, thus allowing to identify chromosomal rearrangements, which are particularly relevant for studying complex pathologies such as cancer [22, 23]. In particular, I had the opportunity to deepen the study of this problem starting from my internship in Prof. Marc A. Marti-Renom's Lab, where I collaborated to GENIGMA (<https://www.genigma.app>), a citizen science project that aims to develop an app that people from all over the world can install on their devices (smartphones or tablets) and play to identify translocations in human cancer cell-lines. The novelty of the project consists in using as inputs Hi-C data sets, that probe the contacts between pairs of genomic regions, and in establishing a consensus over the structural features annotated during the game, thanks to the participation of the players. My contribution to the project consisted in developing a pipeline that allowed to: (1) prepare the input data for the game, (2) mimic the players' activity and (3) analyze the outcome of the game. The first and the third points, in particular, led me to devise an algorithm, based on a network approach, that produces a genome assembly starting from Hi-C contact maps, by exploiting the spectral properties of the Laplacian matrix.

In the third chapter, we will see a novel method, based on a network approach, that allows to detect host-bacteria interactions starting from SNPs (Single-Nucleotide Polymorphism) within the sequence of bacterial genomes. Specifically, this work has been developed within the COMPARE (COllaborative Management Platform for detecion and Analyses of (Re-)emerging and foodborne outbreaks in Europe) EU project, with the aim

of attributing a human case of *Salmonellosis* to the putative source of infection. This procedure, also known as source attribution [24, 25], is in fact an enduring challenge, allowing fast identification of possible cures, the start of an outbreak, the identification and the prioritization of targeted interventions in the food chain, as well as the evaluation of the effectiveness of each intervention.

Many methods have been developed to estimate the relative contribution of different food sources to human foodborne diseases worldwide, including microbial subtyping, comparative exposure assessment, epidemiological analysis of sporadic cases, analysis of data from outbreak investigations, and expert elicitation [24, 26]. Each of these approaches has strengths and limitations, and the usefulness of each depends on the public health questions being addressed [26]. Usually, source attribution studies are conducted by using frequency-matching models like the Dutch and Danish models based on phenotyping data (serotyping, phage-typing, and antimicrobial resistance profiling) [24, 25].

In this chapter, we will see a different approach that exploits the single genetic profiles of the infecting agent, using as input pairwise distance matrices between the bacterial genomes identified in food and human origin samples. In particular, the proposed method represents these pairwise distance matrices as undirected weighted networks where nodes correspond to bacterial isolates and links to a function of genetic distances (i.e. the number of different nucleotides along DNA sequences): the weaker the link, the higher the genetic distance between two isolates. The aim is to extract clusters corresponding to different animal sources (source clustering) and then attributing the human isolates to the putative sources of infection (source attribution). The main idea behind the method is that genomes coming from the same source should show smaller distance values. Therefore, to identify clusters, the algorithm removes all the links whose weight is lower than an optimal threshold value  $t$ , determined by applying a cross validation procedure that maximizes intra-cluster coherence and minimizes the number of isolated nodes. The algorithm has been trained and tested on animal origin samples, providing optimal threshold values that have been validated on an independent data set. Moreover, we will see that the network approach is also useful for investigating which structural features of a data set play a fundamental role in determining the internal coherence of clusters, such as animal sources, the country of origin of the food samples and the year of collection. The major results shown in this chapter are published in [27].



## CHAPTER 2

---

### From sequence to epigenetic and structural footprints

---

In this chapter we will see that the way CG dinucleotides are positioned along the sequence of human genome can be related to the effects that DNA methylation has on it. Moreover, we will see that one of the parameters characterizing the fitting function used to describe CG distance distributions in more than 4000 different organisms, correlates with their biological complexity. We will also see that CG and TA distance distributions, within human genome, can be described by the same mathematical function, suggesting a common process giving rise to the positioning of these two dinucleotides along the sequence. In the end, we will focus on TA distance distribution, detecting deviations from the trend predicted by the best fitting function. In particular, we will see that the parts of the DNA sequence associated with these deviations are characterized by specific patterns that might be related to peculiar mechanical properties of the DNA and also to epigenetic and structural processes.

### 2.1 Dinucleotide distance distribution

Recent studies revealed that dinucleotide distance can be a powerful tool for detecting DNA properties [1, 2], such as the identification of CpG islands [3] and the characterization of epigenomic regulation through methylation [4, 5]. In particular, Paci et al. [4] highlighted a peculiar feature of mammals CG dinucleotides: the distributions of the distances between consecutive CG show an exponential tail, whereas all non-CG distributions are characterized by heavier tails, more similar to a power law. This might be due to the specific role that CGs play inside mammals genomes, since they are the preferential sites of methylation, a fundamental epigenetic mechanism involved in gene

regulation [6, 7, 8, 9, 10] and structural conformation of chromatin [11, 12]. In light of these preliminary observations, we believe that a characterization of the complete CG distribution would provide a better comprehension of their role inside genomes of all organisms, guided by the idea that similar functionalities should share similar statistical properties. Moreover, the identified distribution can be the basis for hypothesizing specific physical models to describe the observed DNA sequence characteristics.

Paci et al. [4] also noticed that the distinction between CG and non-CG distance distributions is less sharp in non-mammal organisms, by considering a set of 21 genomes, belonging to 10 mammal and 11 non-mammal organisms. Therefore, in the first part of this chapter, the study will be extended to CG distance distributions extracted from 4425 genomes, belonging to a wide range of organism categories (bacteria, protozoa, plants, fungi, invertebrates, mammal and non-mammal vertebrates) in order to better understand the heterogeneous scenario found among non-mammals and to obtain a global picture associated to this particular feature.

## 2.2 CG distance distribution

The first step of the analysis consisted in the estimation of the relative frequency distributions  $\hat{p}(\tau)$  of CG distances in a selected set of organisms, namely the DNA sequences of 9 mammal model organisms: *Bos taurus*, *Canis familiaris*, *Equus caballus*, *Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Ornithorhynchus anatinus*, *Pan troglodytes* and *Rattus norvegicus*, chosen since in a previous work [4] they showed very homogeneous characteristics in terms of CG distribution. The data were downloaded from NCBI database [28] and pre-processed by extracting the longest sequence from each genome, except sex chromosomes [4], and by removing the unknown bases, identified with the “N” symbol in the fasta files. This operation did not affect the computation of  $\hat{p}(\tau)$ , because the ratio of N inside the sequences was in general low (see Table 2.1) and they were mainly located contiguously at the centromere and telomere regions, thus producing only a very small number of large distances (that could eventually be easily removed from the analysis). Subsequently we found the positions  $x_j$  of each CG dinucleotide inside the sequence, and we calculated the distance between two consecutive CG as  $\tau_j = x_{j+1} - x_j$ ; finally, for each distance value  $\tau$ , we counted its abundance along the sequence and estimated its relative frequency  $\hat{p}(\tau)$ , as described in eq. 2.1. In this way we obtained a relative frequency distribution that we called CG distance distribution.

$$\hat{p}(\tau) = \frac{\#\{j|\tau_j = \tau\}}{\#\{\tau_j\}} \quad (2.1)$$

### 2.2.1 Fitting method

In order to find a complete characterization of mammal CG distribution, we firstly represented  $\hat{p}(\tau)$  for the 9 mammal model organisms in semilogarithmic scale. In this way, we immediately recognized an exponentially decaying trend in the tails (see Supplementary Materials in [4]), which led us to consider the following functions: exponential and double exponential distributions, which can be associated to physical processes respectively governed by a single and a double characteristic scale (that would correspond to characteristic CG distances along the genome); stretched exponential and gamma distributions, which are related to physical processes involving both a characteristic scale and a power-law trend [29, 30, 31, 32, 33, 34, 35, 36, 37, 38]. We also took into account the q-exponential distribution, as suggested by a recent work [5] that studied CG distance distributions on a small interval of about 2 – 300 dinucleotide distance values for human genome. In our study we consider the whole distance distribution up to about 2000 nucleotides for the same organism, and of the same order of magnitude for the other higher-order organisms of the considered subset. The proposed distributions were fitted to the data by using a non-linear least square method (*fit* function, Mathworks Matlab software).

$$p(\tau) = \log_{10}(ae^{-\tau/b}) \quad (2.2)$$

$$p(\tau) = \log_{10}(ae^{-\tau/b} + ce^{-\tau/d}) \quad (2.3)$$

$$p(\tau) = \log_{10}(ce^{-\tau^a/b}) \quad (2.4)$$

$$p(\tau) = \log_{10}[1 + (1 - a)\tau]^{\frac{-1}{(1-a)}} \quad (2.5)$$

$$p(\tau) = \log_{10}(c\tau^{a-1}e^{-\tau/b}) \quad (2.6)$$

We noticed that the extreme region of the right tail of our CG distributions adversely affected fit results, due to poor sampling, therefore we decided to exclude from the fit procedure all distances beyond the 90th percentile (leaving an interval of distances from 0 up to about 1000 – 2000 bases in all 9 higher-order organisms). The goodness of fit was initially estimated by  $r^2$  parameter (eq. 2.7), defined as:

$$r^2 = 1 - \frac{SSR}{SST} \quad (2.7)$$

where SSR represents the sum of squares of the regression and SST the sum of squares about the mean, also called total sum of squares. Due to the large number of distances fitted for these organisms, any correction for sample size to the goodness of fit estimation

was not relevant. A comparison of  $r^2$  values allowed to discard some distributions with a clear low fitting performance. In order to find the best fitting distribution among the remaining, we considered additionally the mean value of residual distribution (reported in Table 2.2), that allowed a further discrimination.

## 2.2.2 Mammal genome analysis

Goodness-of-fit parameters showed that gamma distribution (eq. 2.6) is the function that best describes CG distance distribution for the 9 mammal subset (see Fig. 2.1 for the case of human genome). In particular, if we look at  $r^2$  values in Table 2.3, we can see that the worst fit results are given by q-exponential distribution, since the corresponding  $r^2$  values are the lowest ones, followed by single exponential distribution. The choice of best fit distribution among the remaining was more difficult, because  $r^2$  values were very similar or even identical. Therefore, we also considered the mean values of residual distribution, that provided a clear distinction among the considered distributions (see Table 2.2), with values around  $10^{-11}$  for gamma fit,  $10^{-8}$  for stretched exponential fit,  $10^{-7}$  for double exponential fit,  $10^{-8}$  for exponential fit and  $10^{-1}$  for q-exponential fit. These values confirmed that q-exponential was the worst fitting distribution, and showed that gamma is the best fit function for mammal CG distance distributions (see Table 2.4 for fit results).

Organism	Sequence	N (%)
Bos taurus	chr1	0.7
Canis familiaris	chr1	0.5
Equus caballus	chr1	1.2
Homo sapiens	chr1	7.4
Macaca mulatta	chr1	6.5
Mus musculus	chr1	7.9
Ornithorhynchus anatinus	chr3	6.4
Pan troglodytes	chr1	2.1
Rattus norvegicus	chr1	5.2

Table 2.1: *Percentage of unknown bases N inside each analyzed sequence of the first set of organisms.*

## 2.2.3 Genome analysis across different levels of biological complexity

Once obtained the best fitting function for the mammal organism set, we applied it to all the 4425 organisms chosen for our analysis (see Table 2.5). The fit parameters associ-

<b>Mammal</b>	<b>Gamma</b>	<b>S. Exp</b>	<b>D. Exp</b>	<b>Exp</b>	<b>Q-exp</b>
Bos taurus	-1.96E-11	-5.19E-6	1.05E-7	-7.65E-12	1.17E-1
Canis familiaris	-8.88E-11	-4.05E-6	3.26E-7	2.47E-8	1.18E-1
Equus caballus	6.53E-10	-5.86E-9	-3.67E-4	6.64E-12	1.51E-1
Homo sapiens	7.69E-10	-1.05E-6	1.21E-7	2.41E-9	1.40E-1
Macaca mulatta	3.13E-11	-2.93E-8	1.26E-7	1.02E-8	1.39E-1
Mus musculus	-2.04E-11	-2.93E-8	2.70E-7	4.00E-8	1.23E-1
Ornithorhynchus anatinus	8.37E-11	-1.25E-7	2.56E-7	3.63E-8	1.10E-1
Pan troglodytes	7.77E-10	-1.79E-6	1.90E-7	2.83E-9	1.39E-1
Rattus norvegicus	7.31E-10	-3.14E-9	1.35E-7	-3.09E-12	1.49E-1

Table 2.2: *Residual mean values of gamma, stretched exponential (S. Exp), double exponential (D. Exp), exponential (Exp) and q-exponential (Q-exp) fit of mammal CG distance distributions.*

ated to the best distribution, together with the goodness-of-fit parameters, were used to describe the analyzed organisms, thus allowing to obtain a global picture from a point of view of organism complexity. In fact, we expected that genomes with similar CG distance distributions would show similar fit parameter values, reflecting similarities in the functional roles of CG dinucleotides in these organisms. Even if for some organism categories the chosen distribution is not optimal as for the initial subset, we hypothesize that organisms with similar distributions (even if not corresponding to the chosen one) should present similar parameters anyway, allowing a global classification with a unified approach. Anyway, to filter out possible fit errors due to bad genome sequence reconstruction, we only considered for our analyses the organisms which goodness-of-fit exceeded a value  $r^2 = 0.9$ . With this filter we discarded on average about 15% of our genomes (from 2% in bacteria to 25% in non-mammal vertebrates), homogeneously distributed along the considered categories, resulting in 3857 genomes left for our analysis. Looking at Fig. 2.2, we notice that  $b$  is the parameter that mainly discriminates between the organism categories while the value  $a$  of the power term in gamma distribution is equally spread across all organisms of all categories (see also Fig. 2.3). Furthermore,  $b$  values seem to increase with the “biological complexity” of the considered categories, being minimum for bacteria and protozoa, and maximum for vertebrates (higher in mammals than in non-mammals) and with an intermediate value for invertebrates. Vertebrate categories have a median value of  $b$  in the range 200 – 300, while it is an order of magnitude lower for bacteria (about 30). We remark that this value is very close to the typical length of DNA enveloped around a histone (146 bp envelope around histone octamer plus a linker region summing up to about 200-220 bp), thus there might be a relation between DNA enveloping around histones and our observation in term of CG distances, even if we cannot provide an explanation for this.

<b>Mammal</b>	<b>Gamma</b>	<b>S. Exp</b>	<b>D. Exp</b>	<b>Exp</b>	<b>Q-exp</b>
Bos taurus	0.982	0.982	0.981	0.961	0.805
Canis familiaris	0.981	0.981	0.977	0.947	0.832
Equus caballus	0.986	0.987	0.775	0.964	0.797
Homo sapiens	0.985	0.985	0.983	0.962	0.799
Macaca mulatta	0.987	0.987	0.986	0.965	0.804
Mus musculus	0.983	0.985	0.983	0.960	0.803
Ornithorhynchus anatinus	0.978	0.981	0.978	0.949	0.831
Pan troglodytes	0.986	0.985	0.984	0.963	0.800
Rattus norvegicus	0.984	0.987	0.985	0.958	0.800

Table 2.3: *R-squared values of gamma, stretched exponential (S. Exp), double exponential (D. Exp), exponential (Exp) and q-exponential (Q-exp) fit of mammal CG distance distributions.*

<b>Mammal</b>	<b>Sequence</b>	<b>a</b>	<b>b</b>	<b>c</b>	<b>r<sup>2</sup></b>
Bos taurus	chr1	0.25 ± 0.03	316 ± 5	0.10 ± 0.02	0.982
Canis familiaris	chr1	0.03 ± 0.03	324 ± 7	0.23 ± 0.04	0.981
Equus caballus	chr1	0.17 ± 0.03	226 ± 4	0.16 ± 0.03	0.986
Homo sapiens	chr1	0.16 ± 0.03	280 ± 5	0.14 ± 0.02	0.985
Macaca mulatta	chr1	0.17 ± 0.03	267 ± 4	0.15 ± 0.02	0.987
Mus musculus	chr1	0.22 ± 0.03	330 ± 6	0.12 ± 0.02	0.983
Ornithorhynchus anatinus	chr3	0.15 ± 0.04	250 ± 6	0.16 ± 0.03	0.978
Pan troglodytes	chr1	0.16 ± 0.03	281 ± 5	0.14 ± 0.02	0.986
Rattus norvegicus	chr1	0.09 ± 0.03	281 ± 5	0.21 ± 0.04	0.984

Table 2.4: *Gamma fit parameter values for the first set of 9 mammals. Errors on parameters are estimated at 95% confidence level and rounded to the first significant digit.*

Since we are considering a large class of organisms, with DNA sequence size differing by several orders of magnitude (from  $10^8$  for mammals to  $10^4 - 10^5$  for bacteria and protozoa), we checked if  $b$  parameter could be associated with the length of the analyzed genomic sequence. This does not seem the case, since the Pearson's correlation coefficient  $r$  between the logarithm of  $b$  and the logarithm of the length of the analyzed genome sequences is very close to zero:  $r = -0.12$ .

In light of these observations, we also tested whether the gamma scale parameter (i.e.,  $b$ ) could depend on CG density inside the sequence (number of CG dinucleotides with respect to sequence length), representing  $b$  as a function of %CG in double logarithmic scale (see Fig. 2.4). In a simple null model, the average distance between dinucleotides should decrease proportionally to the inverse of dinucleotide density inside the sequence, thus with a slope equal to  $-1$  in double logarithmic plot. Therefore, we fitted the  $b$

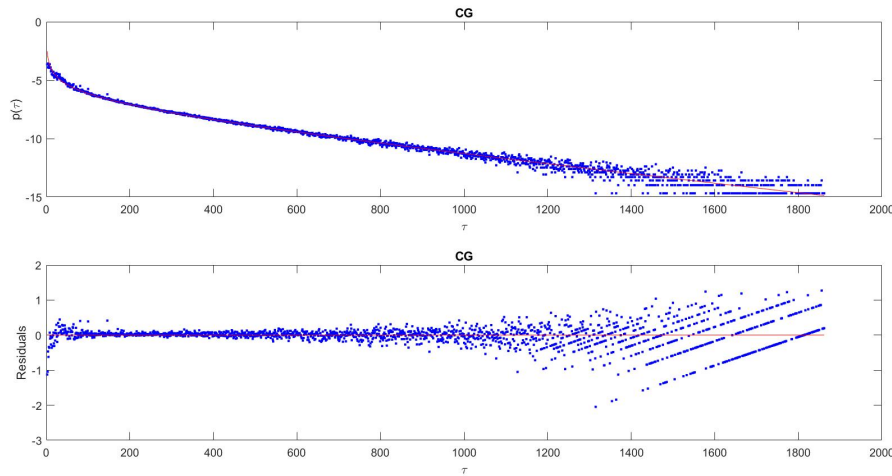


Figure 2.1: *Log-linear plot of CG distance distribution within human chromosome 1, together with gamma distribution as fitting function (upper panel). Fit results are accompanied by residual plots (lower panel).*

vs %CG double logarithmic plot to a straight line using linear least square method, obtaining the results shown in Table 2.6. We observe that the relation between  $b$  and %CG is in general very close to the fitted lines for each organism category, with an average value of Pearson’s coefficient  $\langle r \rangle = -0.65$  (minimum correlation  $r_{MIN} = -0.54$  for invertebrates, maximum correlation  $r_{MAX} = -0.75$  for protozoa). In particular, from this analysis we can identify two groups of organisms, according to the values of the coefficient  $m$ , corresponding to the slope of the line in log-log plot and thus to the exponent of the relation  $b \propto \%CG^m$ : bacteria, plants, fungi, protozoa and invertebrates have an exponent approximately equal to  $-1$ , while mammal vertebrates and non-mammal vertebrates have a smaller exponent in absolute value closer to  $0.5$ , significantly different from the others in terms of 95% confidence interval. Some organism categories thus seem to verify the null model hypothesis, while for vertebrates the significant deviation from the null model suggests a different mechanism for CG dinucleotide placement along the genome rather than a “maximum entropy” process.

A possible biological interpretation of this grouping could be a different role of CG methylation in these two classes of organisms. CG methylation is known to be an important mechanism in higher-order organisms (like vertebrates, that in our analysis show a slope significantly smaller than  $-1$ ), with an active role on gene transcription regulation [39]. For most of the biological categories that showed an exponent close to  $-1$  it is not clear how (or even if) the CG methylation mechanism is used [40, 41, 42], since in some cases different nucleotide sequences are involved in methyl group binding (like the GATC motif in *E. Coli*, or other motifs in plants [43]) and in general is not used

for gene regulation, if not only during embryonic development [44]. We speculate that a characterization of CG distribution parameters for a specific organism could be an index to hypothesize a role of CG methylation at a single organism level, even if we did not go further in the analysis in this direction. In order to extend the range of applications, we think that the method developed in this work can be applied to further repeated genomic sequences (e.g. transcription-factor-binding-site motifs mapped in ENCODE project [45] and repeated sequences associated to transposable elements [46]) in order to gain a deeper insight into DNA properties of single organisms or for comparison between organism categories. Moreover, considering our approach as providing a null model for CG (or other dinucleotide) distribution, we can look for deviations from such null model and study their possible biological meaning (e.g. in relation to CpG islands).

Category	Number of genomes	Size
Vertebrates non-mammals	200	210 Gb
Vertebrates mammals	219	525 Gb
Plants	297	288 Gb
Protozoa	348	17 Gb
Invertebrates	507	168 Gb
Bacteria	1251	5 Gb
Fungi	1603	44 Gb

Table 2.5: *Number and size of genome assemblies downloaded from GenBank database, divided into categories.*

Category	m	q	r <sup>2</sup>
Bacteria	-1.06 ± 0.02	2.23 ± 0.02	0.858
Protozoa	-1.11 ± 0.05	2.30 ± 0.04	0.875
Fungi	-0.87 ± 0.03	2.07 ± 0.02	0.726
Invertebrates	-0.9 ± 0.1	2.30 ± 0.06	0.460
Plants	-1.11 ± 0.09	2.50 ± 0.04	0.707
Vertebrates non-mammals	-0.76 ± 0.08	2.34 ± 0.02	0.704
Vertebrates mammals	-0.51 ± 0.07	2.43 ± 0.01	0.523

Table 2.6: *Linear regression parameters of CG-b relationship, together with r-squared values.*

## 2.2.4 The role of memory in CG positioning process

As shown in the previous sections, mammal CG distance distribution seems to be well described by a gamma distribution, which is characterized by a power-law trend of the



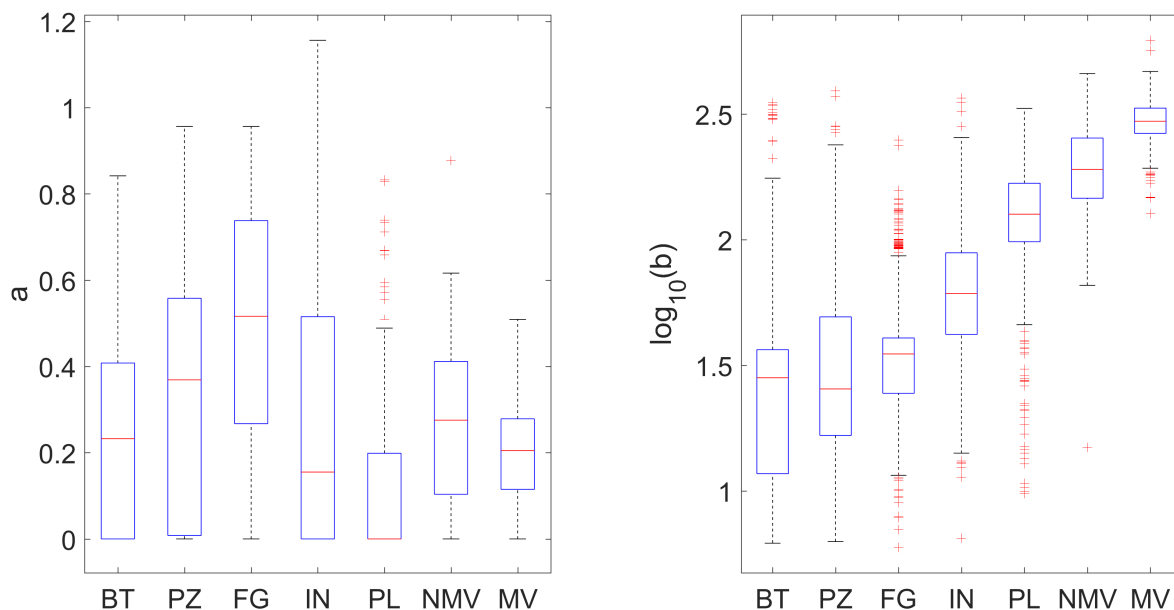


Figure 2.2: *Boxplot of gamma shape parameter  $a$  (left-hand side) and gamma scale parameter  $b$  (right-hand side) for the seven considered categories: bacteria (BT), protozoa (PZ), fungi (FG), invertebrates (IN), plants (PL), non-mammal vertebrates (NMV) and mammal vertebrates (MV).*

type  $\sim x^{-0.8}$  plus an exponential tail of the type  $\sim e^{-x/300}$ . An important question arising at this point is the following: what process can give rise to such a trend? The answer is not trivial, since there is no a priori biological information that can guide us in the modeling. Anyway, we can provide an interpretation for the two trends - power-law and exponential - dominating the distribution respectively at low and high distance values. In fact, we can think of the distance distribution between dinucleotides as the analogous of the waiting time distribution between events. In this parallelism, the DNA sequence corresponds to the time line of the events and the detection of a dinucleotide at a certain position corresponds to the event itself. The waiting time distribution following an exponential trend is typical of memoryless processes [47, 48, 49, 50, 51, 52], such as that observed for a random breaking of a stick; on the other hand, a waiting time distribution following a power-law trend is typical of processes characterized by a long-term memory, such as those observed for a stochastic breaking-stick process with memory constraints, for earthquakes and solar flares [47, 53, 54].

The simplest way to generate such distributions is to represent a chromosome as a segment of finite length, whose points correspond to the different positions that can be occupied by a dinucleotide, getting a power-law or an exponential distance distribution by imposing respectively that:

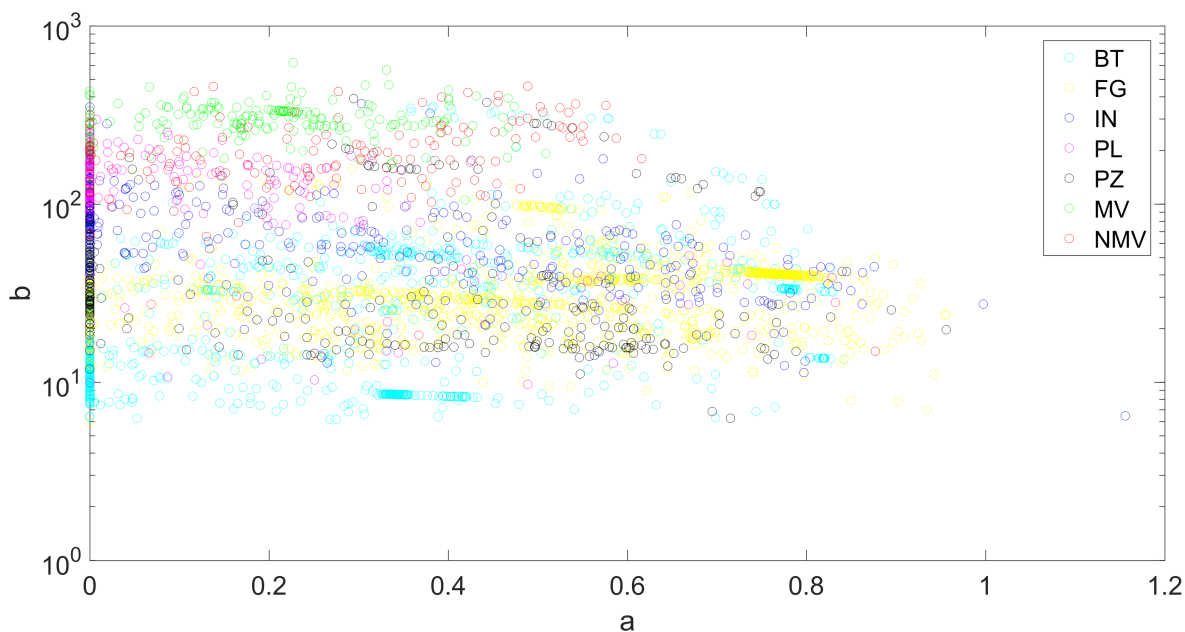


Figure 2.3: *Semilogarithmic plot of gamma scale parameter  $b$  as function of gamma shape parameter  $a$  for the 4425 analyzed genomes, divided into seven categories: bacteria (BT), fungi (FG), invertebrates (IN), plants (PL), protozoa (PZ), mammal vertebrates (MV) and non-mammal vertebrates (NMV).*

1. dinucleotides are not randomly placed along the segment, but with the  $n$ -th dinucleotide having memory of the positions of the previous  $n - 1$ .
2. dinucleotides are randomly placed along the segment, with the  $n$ -th dinucleotide having no memory of the positions of the previous  $n - 1$ .

Specifically, the two distributions can be simulated through a broken-stick process, where the chromosome sequence is represented by a segment corresponding to the interval  $I = [0, 1]$ , and the positions of the dinucleotides are represented by the breaking points. In particular, we can obtain respectively an exponential and a power-law distribution by:

1. extracting  $N$  uniformly distributed random numbers in the interval  $[0, 1]$ .
2. extracting  $N$  numbers through the following steps:
  - Divide the segment into  $M$  sub-segments by extracting a random number from a uniform probability distribution defined in the interval  $[0, 1]$ .
  - Each sub-segment survives with a probability  $1 - p$  and with a probability  $p$  is divided by extracting a random number from a uniform probability distri-

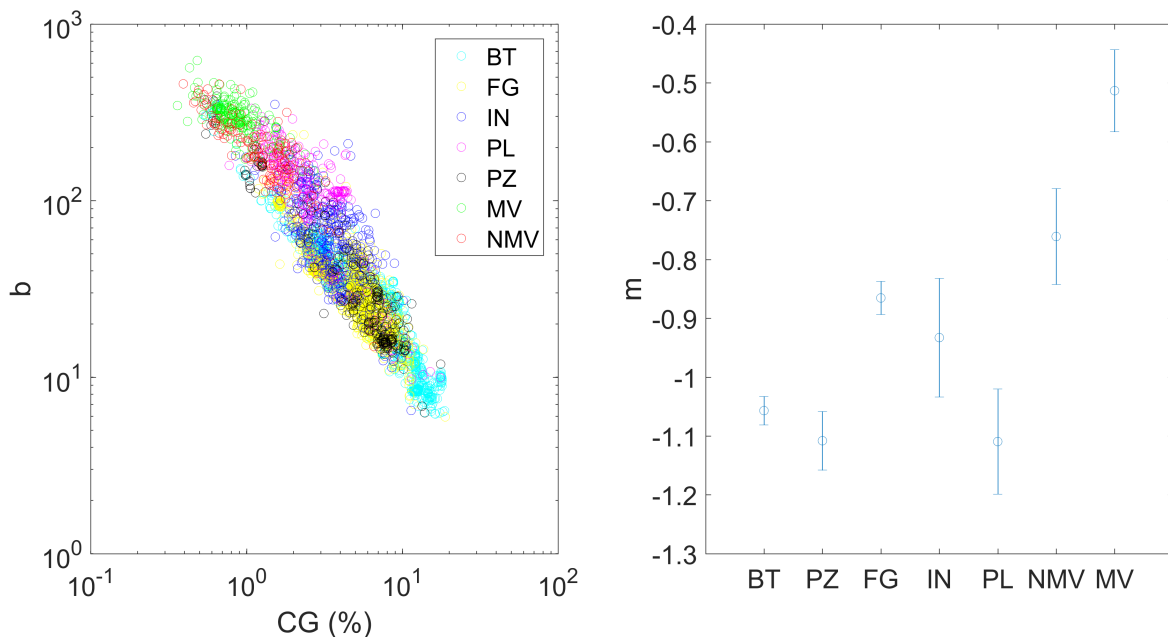


Figure 2.4: *Double logarithmic plot of gamma scale parameter  $b$  as a function of CG percentage for each of the 4425 genomes belonging to the seven considered categories: bacteria (BT), protozoa (PZ), fungi (FG), invertebrates (IN), plants (PL), non-mammal vertebrates (NMV) and mammal vertebrates (MV), (left-hand side). Plot of the angular coefficient  $m$  obtained from linear regression of CG- $b$  relationship, for each considered category (right-hand side)*

bution in the interval  $[a, b]$ , where  $a$  and  $b$  represent the starting and ending points of the broken sub-segment.

- Iterate the previous point  $k$  times, depending on the number  $N$  of dinucleotides that we want to simulate.
- If a sub-segment survives at the  $i$ -th iteration, it will not be broken until the end of the process.

As shown by Fig. 2.5, the two simulations, based respectively on the absence and presence of memory in the positioning process of dinucleotides, give rise to an exponential and a power-law distribution. In terms of CGs, these results suggest the presence of a mechanism of memory preservation giving rise to a non-random positioning when CGs are close to each other along the sequence, and a mechanism of memory loss producing a random positioning when CGs are placed far from each other. It can be hypothesized that this memory preservation concerning CG positioning at low distance values, might be related to the formation and conservation of CpG islands, which are short DNA sequences often associated with gene promoters and defined according to the following

parameters: (1) a minimal length of 200 bp; (2) a minimal CG content of 55%; (3) a minimal observed/expected CG ratio of 0.60 [55]. Another important property differentiating CpG islands from all other CGs, is the low level of DNA methylation [56, 57], which makes them less prone to the spontaneous mutation  $\text{CG} \rightarrow \text{TG}$ , known as deamination, occurring at methylated CG sites, which represent around the 80% of all CGs. Therefore, the memory loss process associated with CGs positioned at high distance values, might be interpreted as an “erosion” of the original positioning, originating from the deamination process occurring at the majority of methylated CGs outside CpG islands.

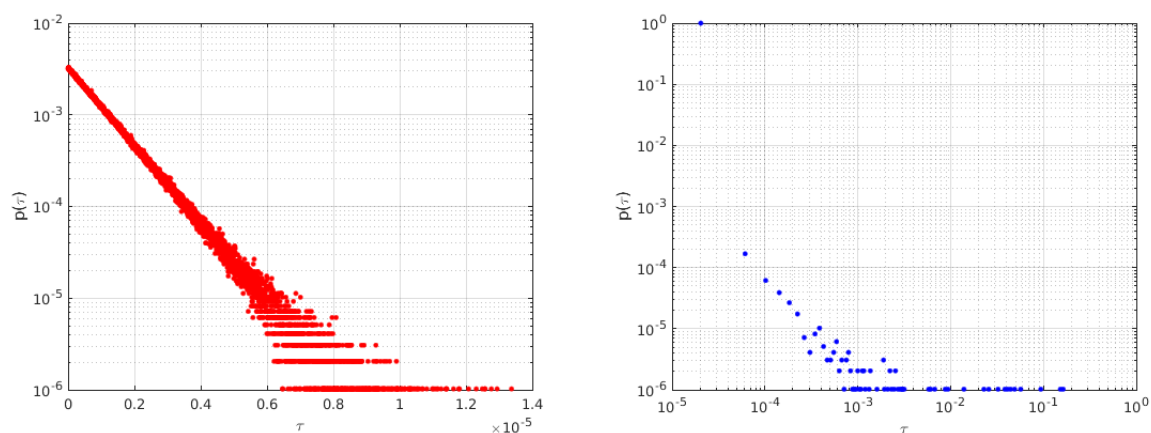


Figure 2.5: Simulation of **memory absence** (left panel) and **presence** (right panel) in dinucleotide positioning, giving rise respectively to an **exponential** and a **power-law** distance distribution, here represented respectively in log-linear scale and in log-log scale, thus showing a linear trend in both cases. If the log-linear plot of the exponential trend in the left panel is compared to the log-linear plots of CG distribution in Fig. 2.6, the exponential tail can immediately be detected. The parameter values used to generate the power-law distribution represented in the right panel are  $M = 10$ ,  $p = 0.8$  and  $k = 25$ .

## 2.2.5 The human genome and a new fitting function

As seen in section 2.2.2, gamma distribution was selected as the best fitting function for describing the trend of CG distance distribution of the 9 mammal model organisms chosen for the study. However, if we focus on human genome, we can clearly see that the fit is characterized by systematic deviations at low distance values (below 50 bp), as shown by the residual plot in Fig. 2.6, suggesting that gamma distribution cannot correctly capture the behavior of CGs when placed at distances lower than  $\sim 50$  bp. Therefore, two variants have been tested for describing CG distance distribution: (1) a shifted gamma (see eq. 2.8) and (2) a shifted power-law with exponential tail (see eq. 2.9). The comparison among the residual plots of the three fitting functions (see

Fig. 2.6) clearly shows that the introduction of a shift parameter removes the systematic deviations at low distance values. Moreover, from an inspection of the plots and the parameter values obtained from the fit (see Table 2.7), we can conclude that the shifted gamma and the shifted power-law with exponential tail are equivalent. In the light of these results and the comments made in section 2.2.4, the latter has been chosen as the best describing function for human CG distance distribution.

$$p(\tau) = \log(c(\tau + d)^{a-1}e^{-\frac{\tau+d}{b}}) \quad (2.8)$$

$$p(\tau) = \log(c(\tau + d)^{-a}e^{-\frac{\tau}{b}}) \quad (2.9)$$

Distribution	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	$r^2$
G	$0.16 \pm 0.03$	$280 \pm 5$	$0.14 \pm 0.02$	–	0.985
SG	$2.9\text{E-}5 \pm 0.07$	$295 \pm 8$	$0.4 \pm 0.2$	$11 \pm 6$	0.985
SPLE	$1.02 \pm 0.08$	$298 \pm 8$	$0.4 \pm 0.2$	$12 \pm 6$	0.985

Table 2.7: *Fit parameter values obtained for CG distance distribution within human chromosome 1, corresponding to the three considered fitting functions: gamma (G), shifted gamma (SG) and shifted power-law with exponential tail (SPLE).*

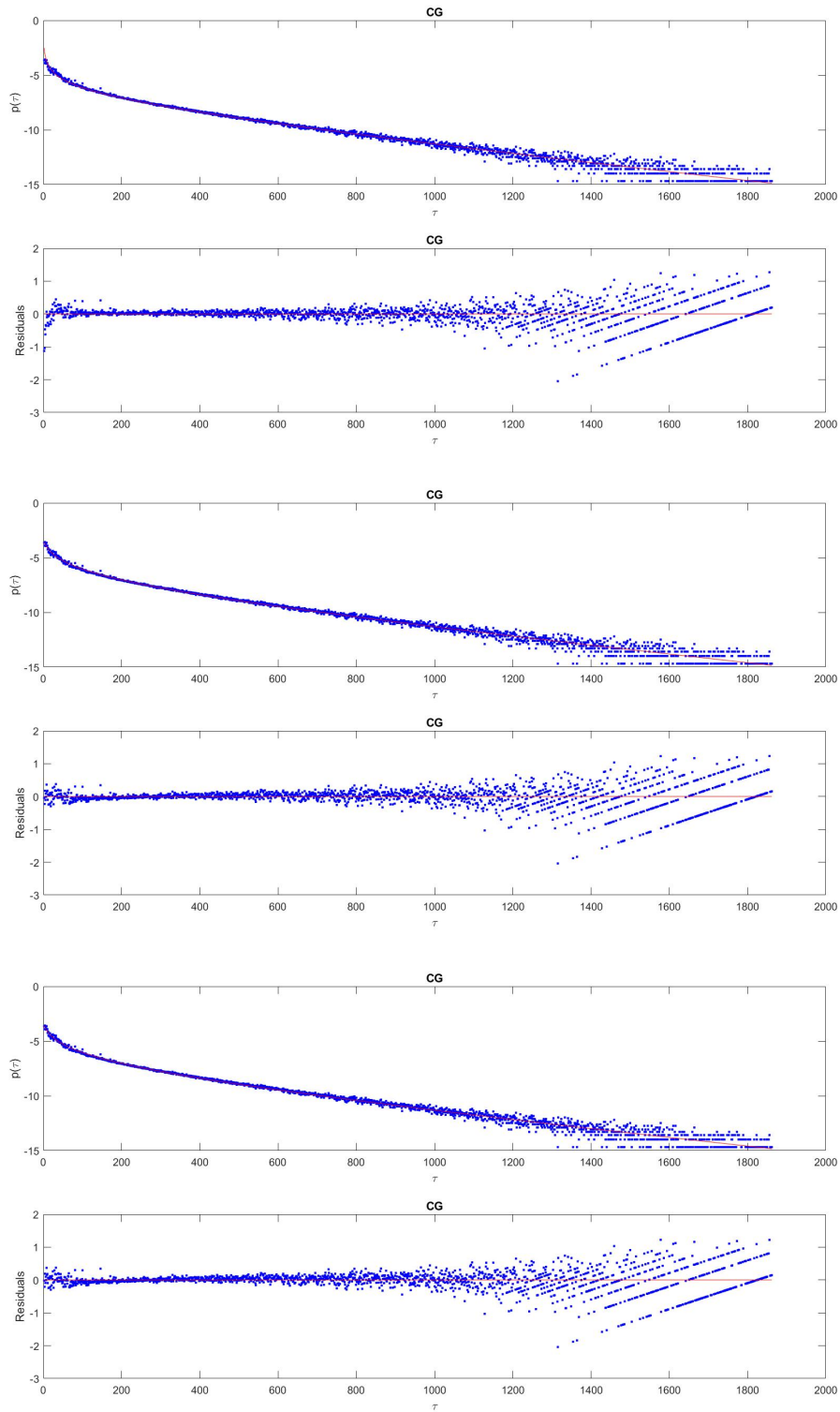


Figure 2.6: Log-linear plot of CG distance distribution within human chromosome 1, together with **gamma** (upper panel), **shifted gamma** (middle panel), and **shifted power-law with exponential tail** (lower panel) as fitting distributions. Fit results are accompanied by residual plots.

## 2.3 From CG to TA distance distribution

In order to compare CG with the remaining 15 dinucleotides within the human genome and measure how actually diverse their distance distributions are, they have been fitted to same function that provided the best results for CG - a shifted power-law with exponential tail - according to the principle that guided our analysis in section 2.2.3: we expect that dinucleotides with similar distance distributions would show similar fit results, reflecting similarities in their positioning process. Surprisingly, we found that TA was the only dinucleotide showing good fit results, as confirmed by Fig. 2.7 and by Figures in Appendix A. Furthermore, if we look at fit parameter values in Table 2.8, we can notice that  $b$ , which represents the characteristic distance between two consecutive TA, is equal 7000 bp, a value much larger than the highest distance considered for the fit ( $\sim 700$  bp). Therefore, we can conclude that the exponential decay in the tail is negligible and that the actual function describing TA distance distribution is a shifted power law, as confirmed by fit results shown in Fig. 2.7 and by the concordance between SPLE and SPL fit parameter values shown in Table 2.8.

$$p(\tau) = \log(c(\tau + d)^{-a}) \quad (2.10)$$

Distribution	a	b	c	d	$r^2$
SPL	$5.2 \pm 0.1$	–	$3 \cdot 10^7 \pm 2 \cdot 10^7$	$43 \pm 4$	0.989
SPLE	$5.0 \pm 0.5$	$7 \cdot 10^3 \pm 44 \cdot 10^3$	$1 \cdot 10^7 \pm 3 \cdot 10^7$	$39 \pm 8$	0.988

Table 2.8: *Fit parameter values obtained for TA distance distribution within human chromosome 1, corresponding to the two considered fitting functions: shifted power-law (SPL) and shifted power-law with exponential tail (SPLE).*

From a modeling point of view, this finding suggests that the same process could have given rise to both CG and TA distance distributions, with only one difference: the latter is characterized by a stronger long-term memory in TA positioning for all distance values along the sequence.

The presence of a connection between CG and TA has been already pointed out by Simmen [58] and Morozov et al. [59]. The former identified these two dinucleotides as the most underrepresented within the human genome, with an increase in CG relative abundance as GC content increases, corresponding to a decrease in TA relative abundance [58]. The latter identified these two dinucleotides as the most favorable to bend around the histone octamer, with TA showing the lowest elastic energy (i.e. the most prone to bending), followed by CG, thus allowing a tight wrapping of DNA sequence around the histone octamer [59].

### 2.3.1 TA random percolation

As we have seen in the previous sections, CG revealed peculiar properties in terms of distance distribution within the genome of higher-order organisms. These properties might be related to the effect that DNA methylation has on the sequence, since methylated cytosine undergo a spontaneous mutation, known as deamination, that transforms CG into TG [60]. In particular, we have seen that the exponential trend characterizing the tails of CG distance distribution can be associated to a positioning process without memory, regulating the displacement of consecutive CG separated by high distance values, and that it can be associated to a process of “memory loss”. Moreover, we have seen that, among the non-CG distance distributions, only TA is closely related to CG, as shown by the results of the fit to a shifted power-law with exponential tail. Therefore, we hypothesize that the striking difference observed by CG and non-CG in terms of distance distributions [4] can be due to the deamination process, that led to the loss of a considerable amount of CG. In this scenario, we are implicitly supposing that the original CG distance distribution followed a non-CG-like trend, and to test our hypothesis we selected TA as the best candidate for representing the original trend of CG distance distribution (given their relationship mentioned above) and we simulated the deamination process through a random percolation of TA, consisting in a random removal of  $N$  TA. In particular, we decided to remove a number  $N$  of TA equal to difference between the number of TA and the number of CG detected along the sequence, in order to have a final number of TA that approximately equals the number of CG.

The results show that, after the random percolation, TA distance distribution got closer to CG (see Fig. 2.8), as confirmed by fit parameter values in Table 2.9, with the value of the power-law exponent decreasing from  $\sim -5$  to  $\sim -3$ , as well as the value of the characteristic length  $b$  related to the exponential tail, which decreased from 7000 bp to 616 bp.

	<b>a</b>	<b>b</b>	<b>c</b>	<b>d</b>	$r^2$
<b>CG</b>	$1.02 \pm 0.08$	$298 \pm 8$	$0.4 \pm 0.2$	$12 \pm 6$	0.985
<b>TA</b>	$5.0 \pm 0.5$	$7 \cdot 10^3 \pm 44 \cdot 10^3$	$1 \cdot 10^7 \pm 3 \cdot 10^7$	$39 \pm 8$	0.988
<b>pTA</b>	$3.4 \pm 0.5$	$616 \pm 130$	$6 \cdot 10^5 \pm 2 \cdot 10^6$	$164 \pm 41$	0.981

Table 2.9: *Fit parameter values obtained for: CG, TA and percolated TA (pTA) distance distribution within human chromosome 1, using a shifted power-law with exponential tail (SPLE) as fitting function.*

Anyway, even if the simulated process brought TA distance distribution very close to CG, some significant differences still remain, as shown by the plot in Fig. 2.8. In particular, fit results of shown in Fig. 2.9, confirm that TA distance distribution, after the random percolation, is not well described by a shifted power-law with exponential tail, unlike CG. This might be due to the following reasons:



1. we do not know the real initial CG distribution before percolation, thus the final result of percolation could be affected by this unknown;
2. we applied the percolation by randomly selecting  $N$  TA from the distribution, without taking into account that there are some “TA excesses” deviating from the main trend (see next sections for further details);
3. we applied the percolation by randomly selecting  $N$  TA from the distribution, meaning that we are simulating a deamination process that acts with the same probability along the sequence, without taking into account that CpG islands are typically unmethylated [60].

### 2.3.2 TA in primates

Interestingly, if we consider the set of 9 mammal organisms analyzed in section 2.2.2, we can see that TA distance distributions of *Pan troglodytes* and *Macaca mulatta* show exactly the same trend identified for *Homo sapiens* (see Appendix B). In fact, it is well described by a shifted power-law, as confirmed by fit results represented in Fig. 2.10 and by fit parameter values shown in Table 2.10, with a power-law exponent  $a = -5.2$  and shift parameter  $d \sim 43$  bp. In particular, if we look at residual plots, we can clearly identify two deviations, characterizing all the three organisms: a sharp peak at 91 bp and gaussian-like peak centered around 140 bp, meaning that there is an excess of consecutive TA separated by a distance of 91 bp and by an interval of distance values around 140 bp. In order to better estimate the center and the shape of the latter peak, a Gaussian fit of residuals has been performed, providing a value of 142 bp for *Homo sapiens* and *Pan troglodytes*, and a value of 143 bp for *Macaca mulatta* (see Table 2.11). The goodness of fit is confirmed by the results shown in Fig. 2.11. Therefore, the next step will consist in deepening the characterization of these deviations with a focus on human genome.

Organism	$a$	$c$	$d$	$r^2$
Homo sapiens	$5.2 \pm 0.1$	$3 \cdot 10^7 \pm 2 \cdot 10^7$	$43 \pm 4$	0.989
Pan troglodytes	$5.2 \pm 0.1$	$3 \cdot 10^7 \pm 2 \cdot 10^7$	$42 \pm 4$	0.989
Macaca mulatta	$5.2 \pm 0.1$	$4 \cdot 10^7 \pm 3 \cdot 10^7$	$43 \pm 4$	0.989

Table 2.10: *Fit parameter values obtained by fitting TA distance distribution to a shifted power-law (SPL) within chromosome 1 of the three considered primates: Homo sapiens, Pan troglodytes and Macaca mulatta.*

<b>Organism</b>	$\mu$	$\sigma$
Homo sapiens	$142 \pm 3$	$20 \pm 3$
Pan troglodytes	$142 \pm 3$	$23 \pm 3$
Macaca mulatta	$143 \pm 3$	$22 \pm 3$

Table 2.11: *Gaussian fit parameter values for residual distribution of the three considered primates: Homo sapiens, Pan troglodytes and Macaca mulatta.*

### 2.3.3 Deviations from the expected trend within human genome

As we have seen in the previous section, there are two deviations from the expected trend, characterizing TA distance distribution: a sharp peak at 91 bp and a Gaussian peak centered around 142 bp, whose exceeding part with respect to the expected trend estimated by the fitting distribution, corresponds respectively to the 73% of the observed TA that are separated by a distance of 91 bp and to the 32% of the observed TA that are separated by a distance of 142 bp. Since it is known that TA has peculiar structural and mechanical properties, that make the sequence more flexible [61, 59], as first hypothesis, we tested whether these two excesses of TA distances might mark the positioning of nucleosomes, which are the fundamental packaging units of eukariotic genomes, constituted by a 147 bp DNA sequence wrapped around a histone octamer [62, 63], and characterized by a 10 bp periodic signal of AA/TT/TA that oscillates out of phase with a 10 bp periodic signal of GC [64, 65, 61]. These peculiar patterns of AA/TT/TA and GC signals make the sequence highly flexible, thus facilitating the bending around the histone octamer [61, 64]. Therefore, as a first check, we searched for a 10 bp periodicity within the sequences between consecutive TA separated by a distance of 91 bp (TA91), by plotting the occurrence probability (i.e. the fraction of a given dinucleotide at a specific position calculated using a 3-bp moving average [64]) of consecutive AA or TT along the sequences together with the occurrence probability of GC. The results, shown in Fig. 2.12, clearly show: (1) a symmetric signal with respect to the center of the considered sequences; (2) a periodicity close to 10 bp in AA/TT signal but not in GC signal; (3) a repetition of the same patterns in all chromosomes (see AppendixD). Furthermore, the expected occurrence probability of AA/TT for nucleosome sequences should show a sinusoidal shape in phase opposition with that of GC [65], a feature that is absent in our patterns.

These results suggest that the peak at 91 bp is likely to be associated with some kind repetitive sequences. Therefore, to further investigate this hypothesis, the sequences corresponding to TA91 were aligned to the sequences corresponding to SINE (Short Interspersed Nuclear Elements), which are mobile non-coding elements, characterized by short repetitive sequences ranging from 100 bp to 700 bp [66, 67]. The data were downloaded from the public database SINEBase [66] and the alignment was performed

through the Matlab function *localalign*, finding that the 40% of TA91 sequences have an identity percentage with SINEs greater than the 80%, thus corresponding to a good match [66]. In particular, we found that the 86% of TA91 sequences well matched with SINEs corresponded to Alu repeats, which are sequences extensively methylated and CG rich, often occurring within introns, and involved in chromosomal rearrangements [68]. Therefore, even though the match between TA91 sequences and SINE do not explain all the identified excess (73% of TA91), it seems non trivially related to genomic elements involved in epigenomic and structural processes.

A completely different trend was observed for sequences between consecutive TA separated by a distance of 142 bp (TA142), as confirmed by Fig. 2.12. In fact, even though the occurrence probability of AA/TT and GC still shows a symmetric signal with respect to the center of the considered sequences, an opposite behavior clearly emerges between the two. In particular, the former is characterized by high values at the borders of the sequences and by low values at the center, whereas the latter is characterized by high values at the center and low values at borders.

To test whether the observed patterns in Fig. 2.12 were actually different from those found for TA in line with the trend predicted by the shifted power-law distribution, we computed the occurrence probability of AA/TT and GC for sequences identified by consecutive TA separated by distance values not included in the two deviating peaks: 60 bp (TA60) and 100 bp (TA100). The results, shown in Fig. 2.13, reveal a completely different trend, with an average AA/TT occurrence probability of 0.11 and an average GC occurrence probability of 0.06 for both TA60 and TA100, confirming that the deviations from the trend predicted by the shifted power-law distribution for TA91 and TA142 distances show different patterns in the sequences within these intervals.

In order to have an estimate of the average periodicity associated with the peaks at 91 bp and 142 bp, we proceeded as follows:

- for each of the M sequences between consecutive TA separated by a distance X, we computed the average distance between consecutive AA/TT and between consecutive GC, thus obtaining a list of M values for AA/TT and GC.
- We computed the logarithm of these M values and fitted the corresponding histogram with a Gaussian distribution, thus obtaining an estimate of the mean values  $\mu_{AA/TT}$ ,  $\mu_{GC}$  and the standard deviations  $\sigma_{AA/TT}$ ,  $\sigma_{GC}$  of the average distance between consecutive AA/TT and consecutive GC. The mean values  $\mu_{AA/TT}$  and  $\mu_{GC}$  represent an estimate of the average periodicity associated with the M sequences between consecutive TA separated by a distance X. In particular, the lower the standard deviation, the stronger the periodicity.

The measures obtained through these steps are  $\mu_{AA/TT} \sim 12.8$  bp and  $\mu_{GC} \sim 14.1$  bp for TA91, and  $\mu_{AA/TT} \sim 9.7$  bp and  $\mu_{GC} \sim 12.3$  bp for TA142 (see Table 2.12), indicating

the sequences corresponding to TA142 as the closest to the 10 bp periodicity observed in nucleosome sequences.

This measure gives also the opportunity to easily compare several TA distances, thus allowing to understand if the values computed for TA91 and TA142 are actually a sign of something happening at those precise points of the distribution and not in all the others. Therefore, we considered all the sequences corresponding to TA distances ranging from 10 bp to 400 bp and computed the values of  $\mu$  and  $\sigma$ . As we can see from the results shown in Fig. 2.15, the average distance  $\mu$  between consecutive AA/TT increases as TA distance increases - as we would expect in a random null model - except for TA distances corresponding to the Gaussian peak centered around 142 bp, where  $\mu$  and  $\sigma$  decrease, with the former assuming values close to 10 bp. The same happens for GC, even if for TA distance values greater than  $\sim 180$  bp,  $\mu$  decreases as TA distance increases.

Interestingly, the same trend for  $\mu$  and  $\sigma$  has been observed for all the chromosomes within the human genome (see Appendix C), as well as both the deviations at 91 bp and around 142 bp, even if with some differences related to chromosome 11, 21, and Y, as shown by the percentages in Table 2.13 corresponding to the TA exceeding the expected trend estimated by the fitting distribution.

	$\mu_{AA/TT}$ [bp]	$\sigma_{AA/TT}$ [bp]	$\mu_{GC}$ [bp]	$\sigma_{GC}$ [bp]
<b>TA91</b>	$12.82 \pm 0.07$	$1.451 \pm 0.005$	$14.14 \pm 0.07$	$1.414 \pm 0.005$
<b>TA142</b>	$9.7 \pm 0.2$	$1.64 \pm 0.03$	$12.3 \pm 0.3$	$1.54 \pm 0.02$

Table 2.12: Mean values  $\mu_{AA/TT}$ ,  $\mu_{GC}$  and standard deviations  $\sigma_{AA/TT}$ ,  $\sigma_{GC}$  of the average distance between consecutive AA/TT and consecutive GC, obtained by the Gaussian fit shown in Fig. 2.14.

	<b>TA91</b>	<b>TA142</b>
<b>Chr1</b>	73.7%	32.4%
<b>Chr2</b>	71.6%	34.8%
<b>Chr3</b>	75.0%	37.7%
<b>Chr4</b>	74.4%	39.3%
<b>Chr5</b>	73.2%	30.8%
<b>Chr6</b>	74.0%	36.3%
<b>Chr7</b>	76.3%	32.8%
<b>Chr8</b>	72.2%	28.0%
<b>Chr9</b>	72.3%	28.5%
<b>Chr10</b>	73.4%	33.7%
<b>Chr11</b>	68.5%	20.1%
<b>Chr12</b>	77.4%	38.5%
<b>Chr13</b>	73.6%	34.6%
<b>Chr14</b>	73.8%	26.9%
<b>Chr15</b>	75.0%	37.2%
<b>Chr16</b>	74.8%	30.8%
<b>Chr17</b>	76.9%	33.5%
<b>Chr18</b>	71.9%	35.9%
<b>Chr19</b>	79.3%	31.3%
<b>Chr20</b>	70.2%	30.5%
<b>Chr21</b>	68.8%	20.4%
<b>Chr22</b>	72.1%	24.4%
<b>ChrX</b>	78.0%	30.7%
<b>ChrY</b>	67.1%	20.0%

Table 2.13: Percentages of observed TA91 and TA142 exceeding the expected trend estimated by the fitting distribution, for all chromosomes of human genome.

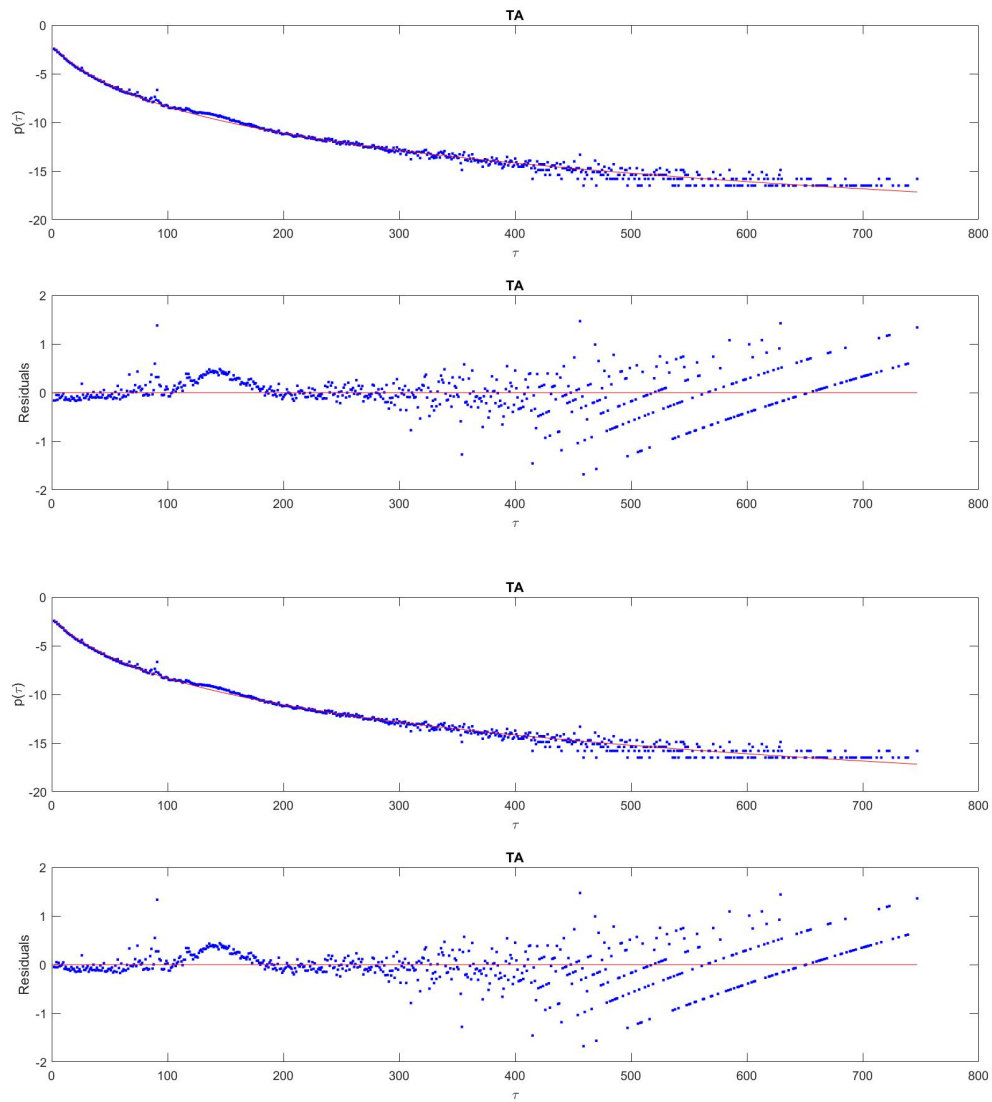


Figure 2.7: Log-linear plot of TA distance distribution within human chromosome 1, together with **shifted power-law with exponential tail** (upper panel) and **shifted power-law** (lower panel) as fitting distributions. Fit results are accompanied by residual plots.

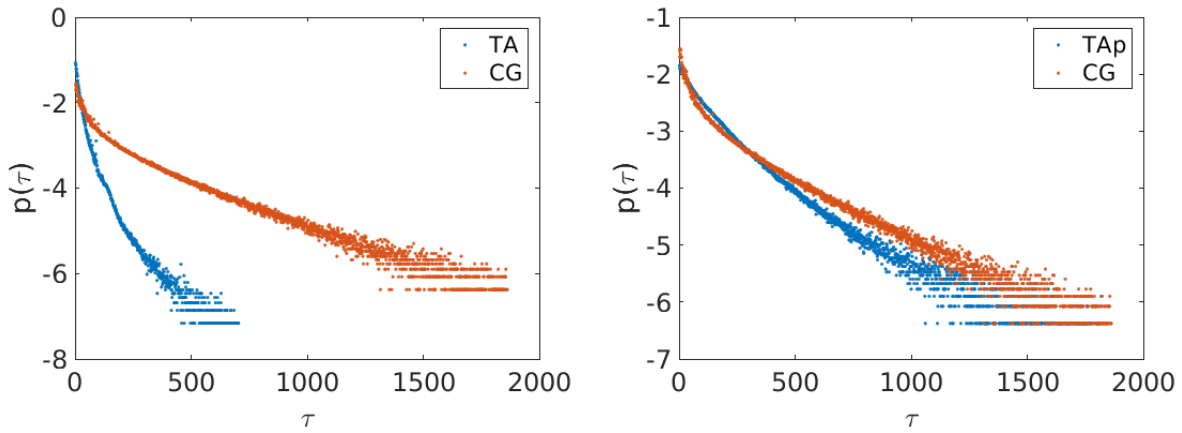


Figure 2.8: *TA distance distribution within human chromosome 1, before (left panel) and after (right panel) applying the random percolation. The distributions are represented together with CG's and using a shifted power-law with exponential tail as fitting function.*

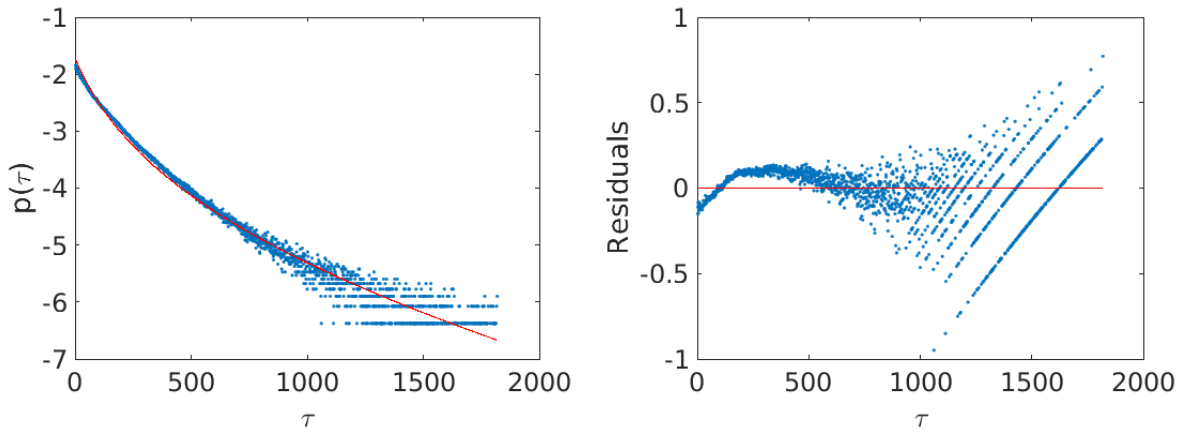


Figure 2.9: *Log-linear plot of TA distance distribution within human chromosome 1, after applying the random percolation, together with shifted power-law with exponential tail as fitting function. Fit results are accompanied by residual plots.*

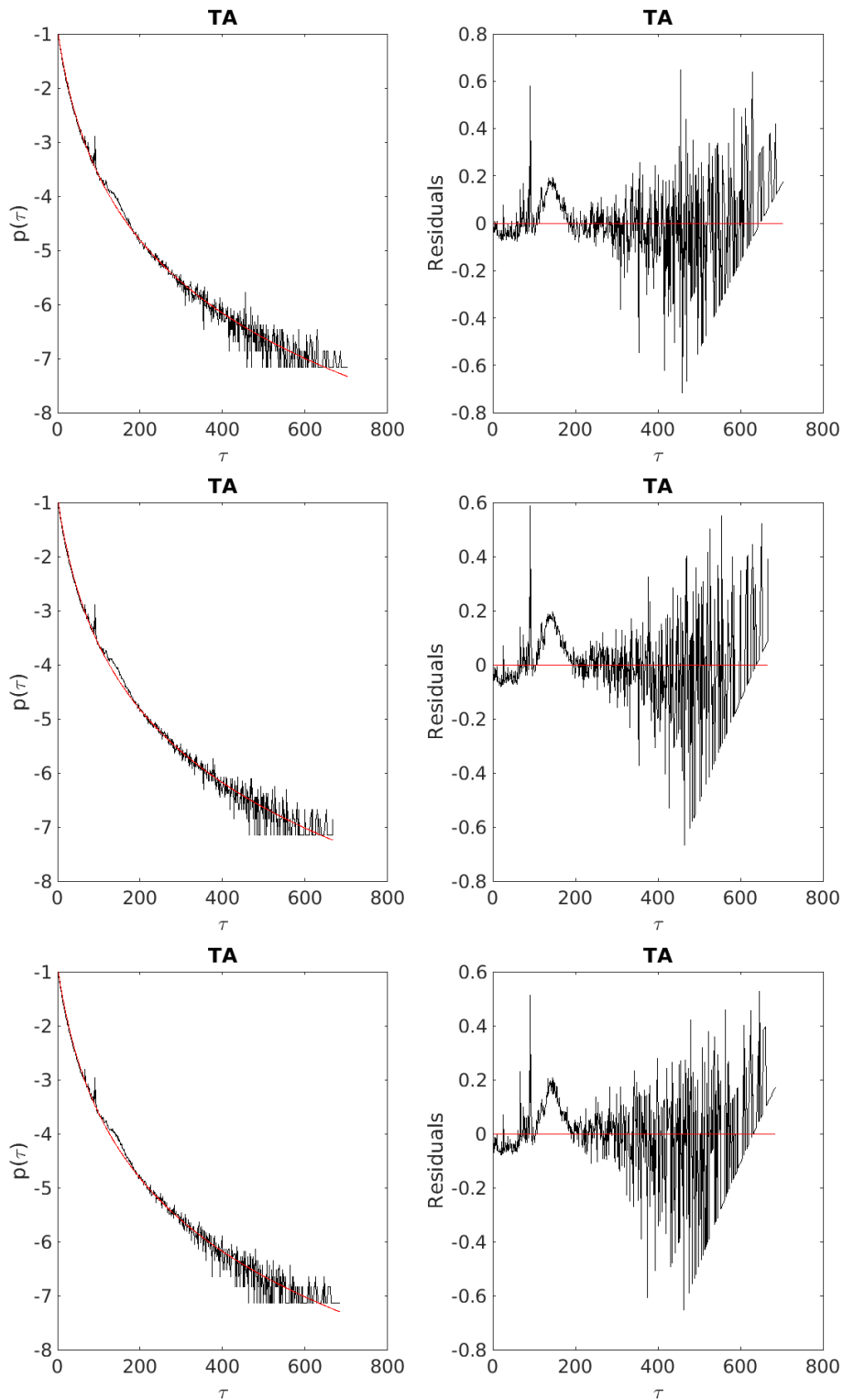


Figure 2.10: Log-linear plot of TA distance distribution within chromosome 1 of *Homo sapiens* (upper panel), *Pan troglodytes* (middle panel) and *Macaca mulatta* (lower panel) together with shifted power-law as fitting distribution. Fit results are accompanied by residual plots, which clearly show a sharp peak at 91 bp and a Gaussian peak centered around 142 bp.



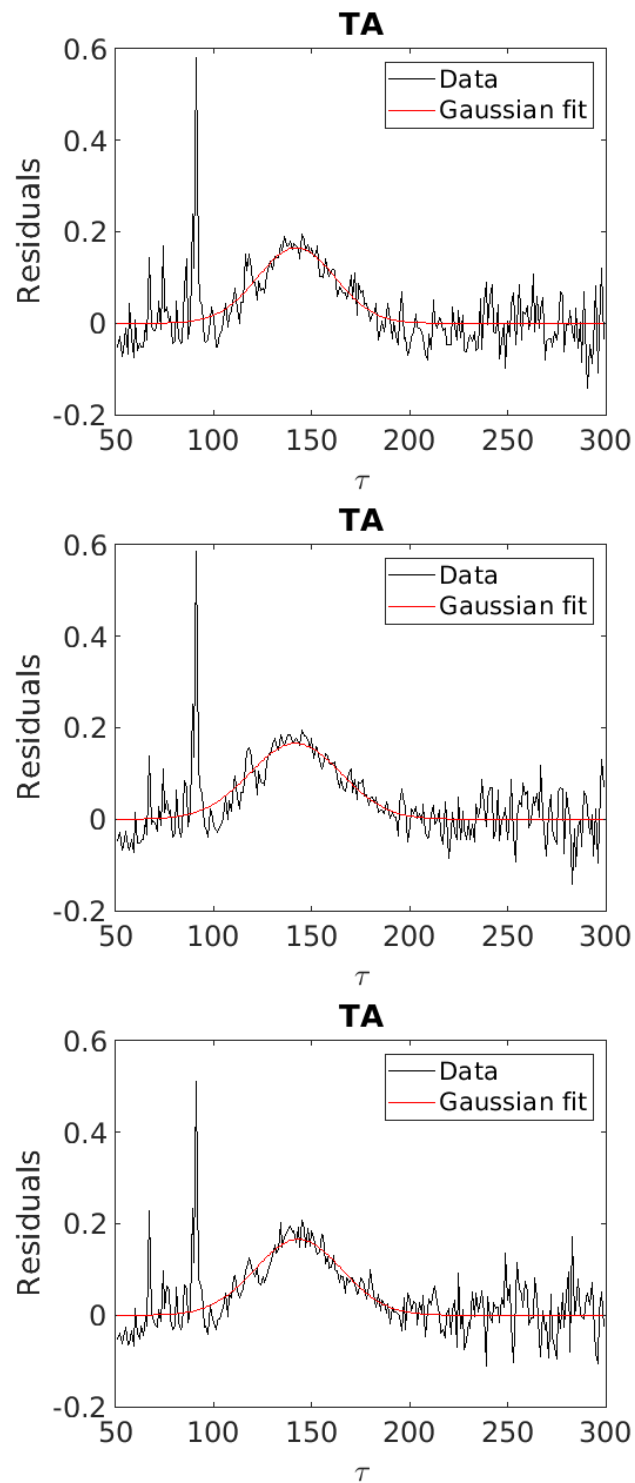


Figure 2.11: Gaussian fit of *Homo sapiens* (upper panel), *Pan troglodytes* (middle panel) and *Macaca mulatta* (lower panel) residuals shown in figure 2.10.

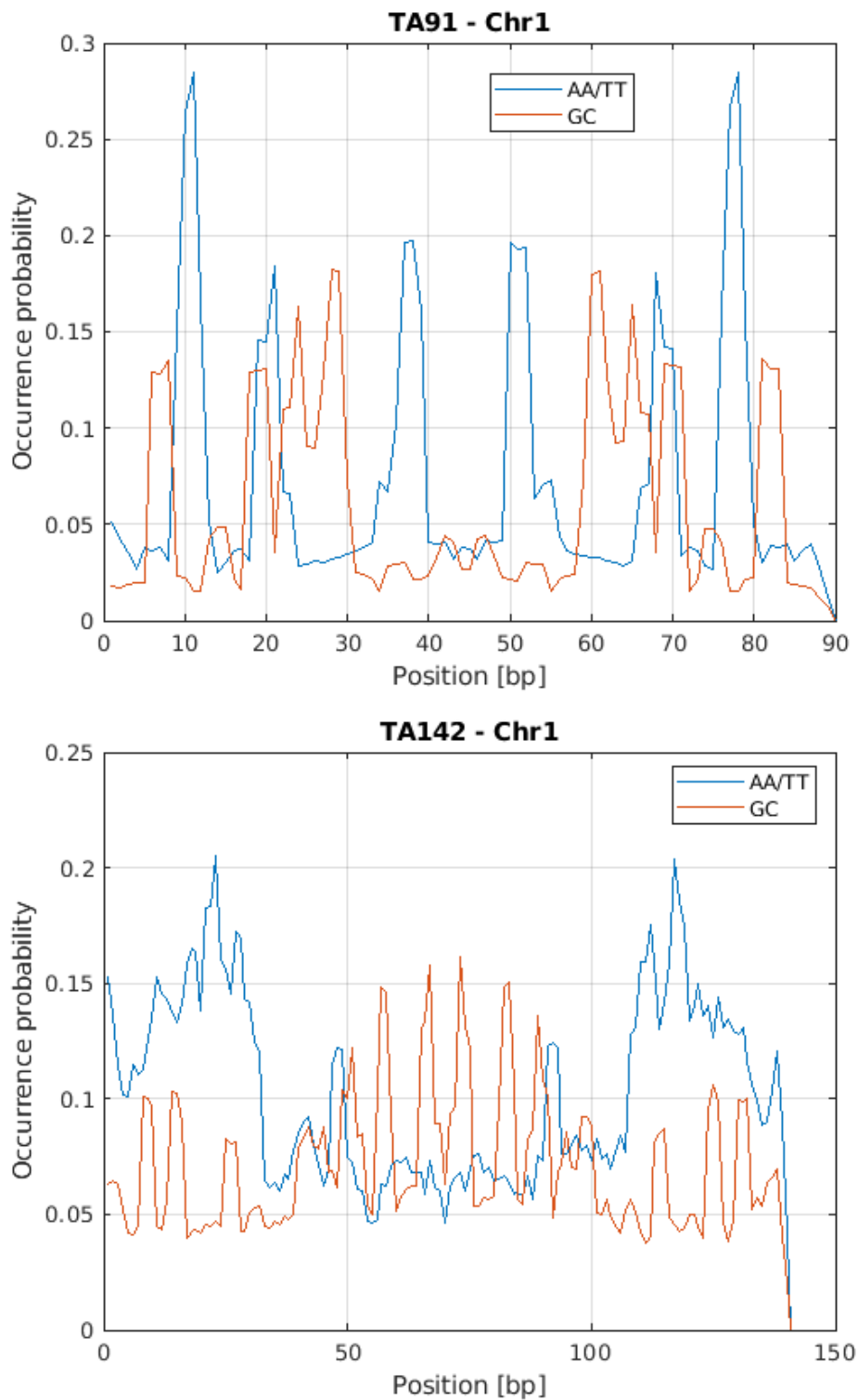


Figure 2.12: Occurrence probability as a function dinucleotide position within the sequences identified by consecutive TA separated by a distance of 91 bp (upper panel) and 142 bp (lower panel) within human chromosome 1.

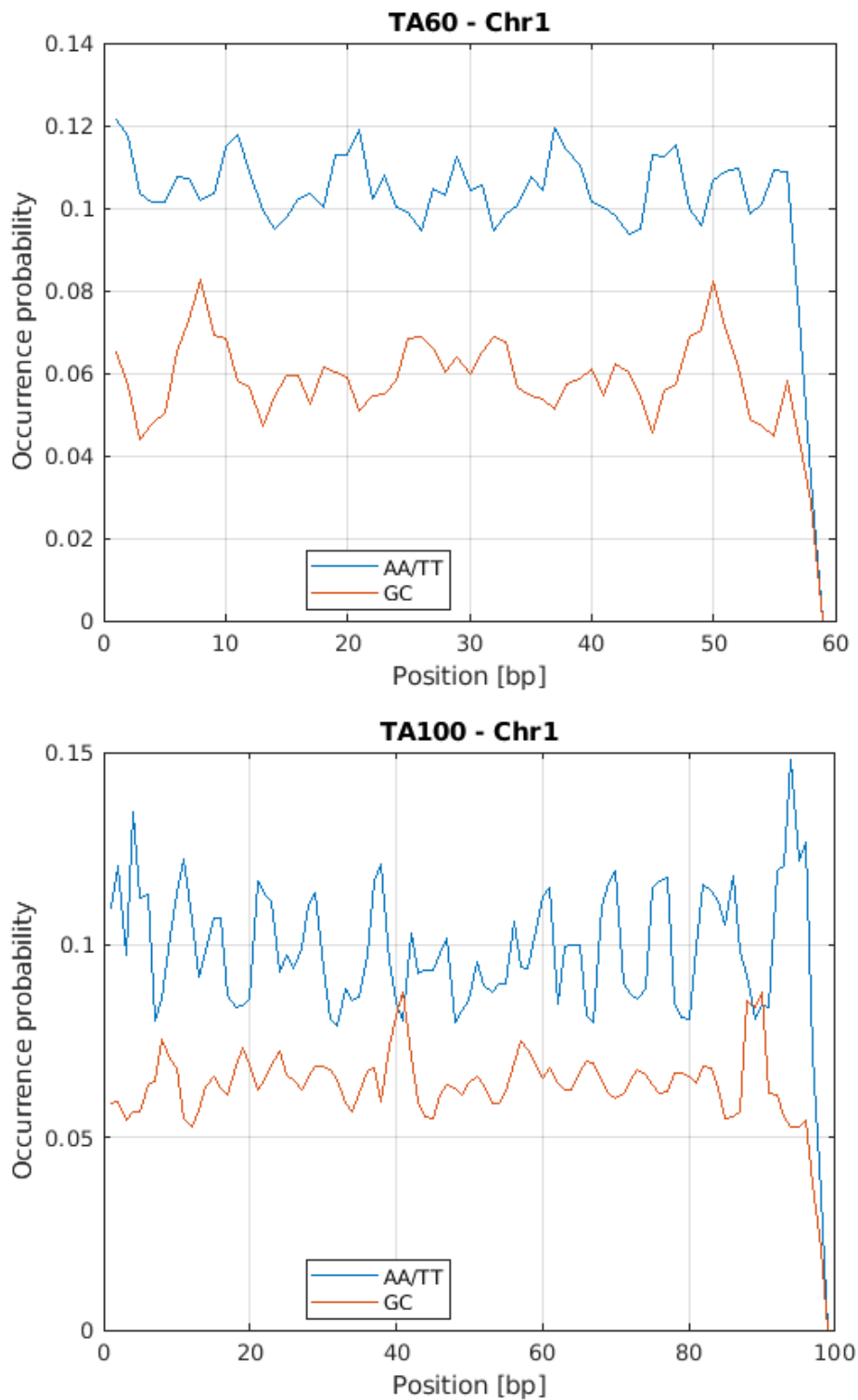


Figure 2.13: Occurrence probability as a function dinucleotide position within the sequences identified by consecutive TA separated by a distance of 60 bp (upper panel) and 100 bp (lower panel) within human chromosome 1.

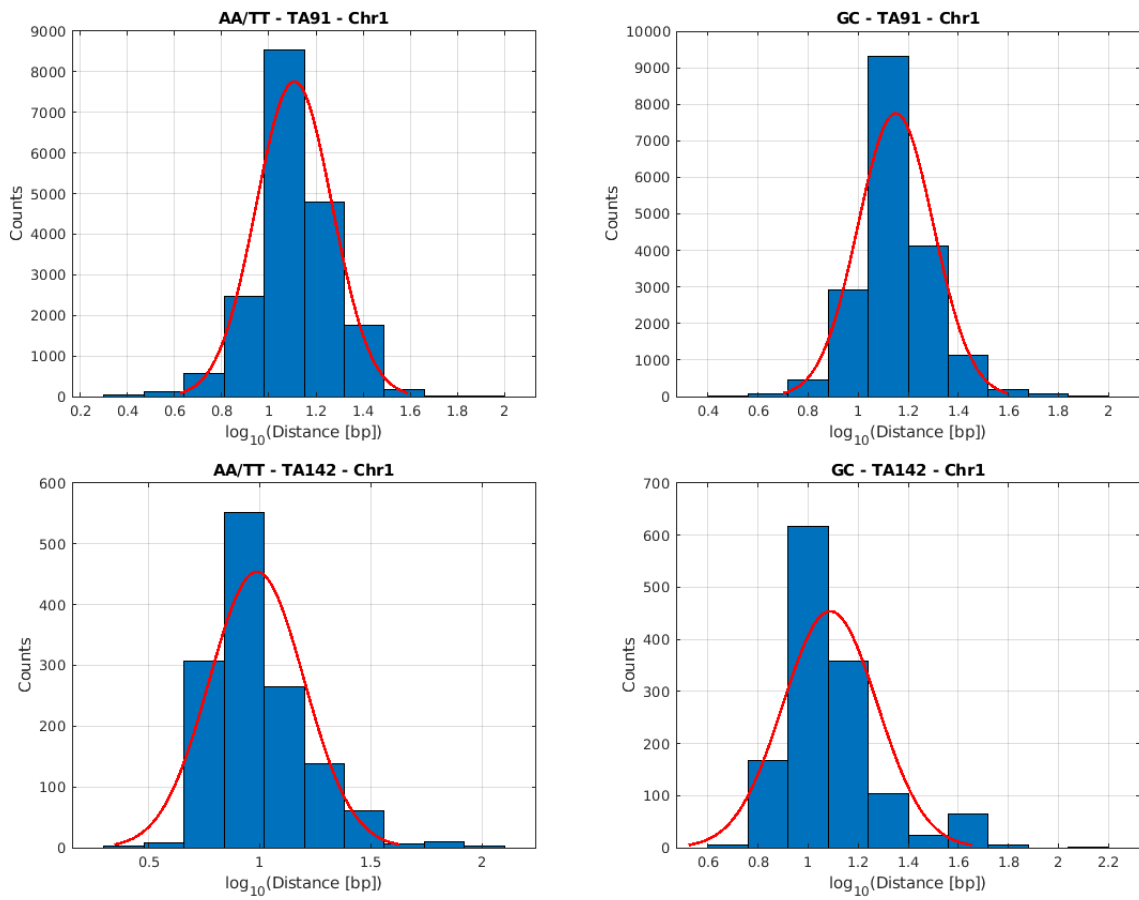


Figure 2.14: Histograms of the logarithm of the average distance between consecutive AA/TT and consecutive GC within the sequences corresponding to TA91 and TA142, together with Gaussian fit results.

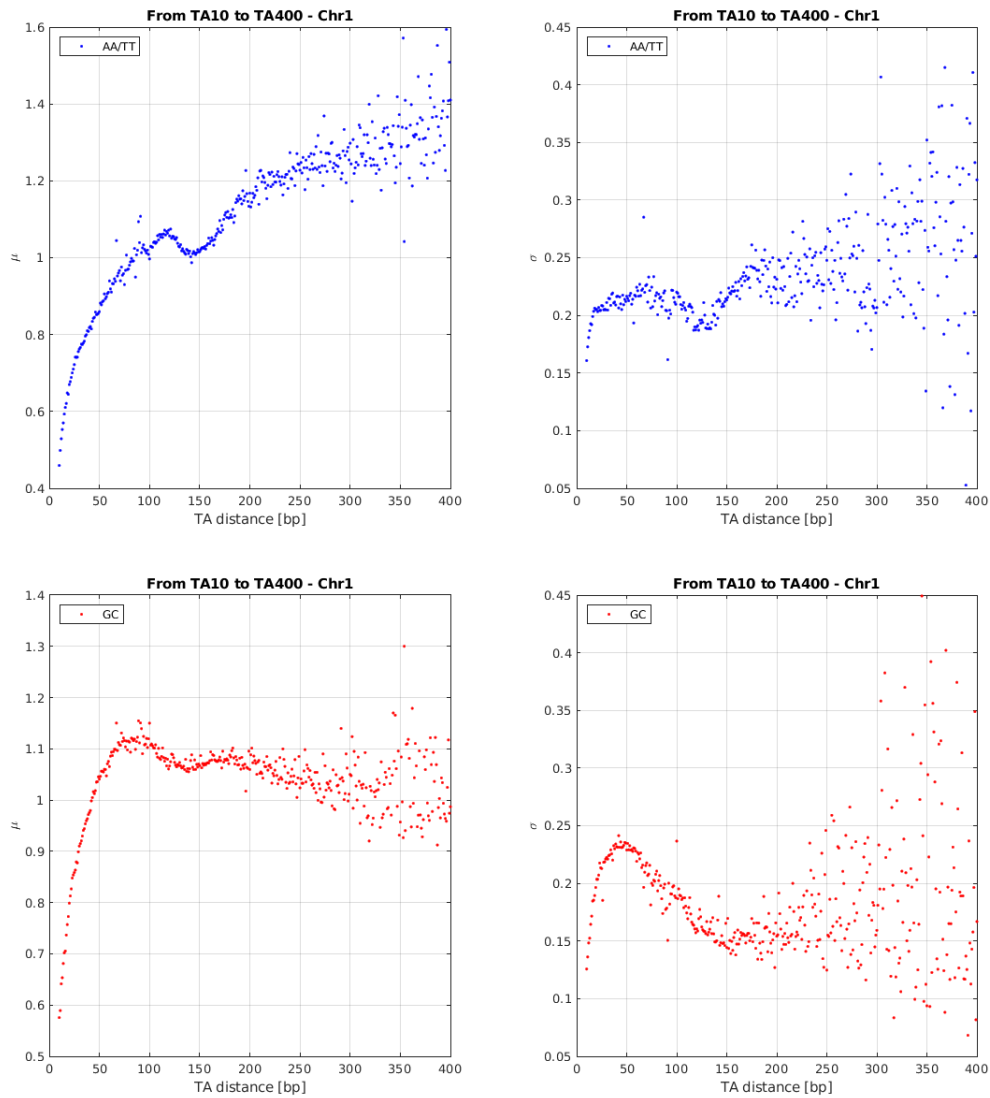


Figure 2.15: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT (upper panels) and consecutive GC (lower panels) along the sequences identified by TA pairs separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 1.

## 2.4 Conclusions

We considered several probability density functions to fit the CG distance distribution of a selected set of mammal organisms, and we initially observed that it is best described by a Gamma distribution. Applying this function on a wide set of organisms, taken from different taxonomic categories, we noticed that the scale parameter  $b$  of the Gamma distribution could be associated to the biological complexity of the organism category, increasing from bacteria to vertebrates. Moreover, we tested for possible factors affecting this parameter, like genome sequence length and CG density. While the first factor was not related to our observations, the second revealed stronger correlations; in particular, for a group of organisms, comprising those of minor biological complexity (bacteria, protozoa, fungi, invertebrates and plants), the relation between  $b$  and CG density could be explained by a minimal null model (i.e. a linear dependence on density), while for higher-order organisms (vertebrates) this null model did not explain the observations (depending on the square root of density). We argue that this difference could be related to the different role that CG methylation plays in these classes of organisms, affecting differently CG positioning and mechanisms that could modify this displacement.

Subsequently we focused on human genome, and we saw that gamma fit results of CG distance distribution show a systematic deviation at low distance values, therefore we introduced a shifted power-law with exponential tail as new fitting function, obtaining a better performance. Furthermore, we saw that the power-law trend characterizing low distance values, and the exponential trend characterizing high distance values, can be generated respectively by a positioning process with memory [47, 53, 54] and without memory [47, 48, 49, 50, 51, 52]. We hypothesized that the former process might be related to the mechanism of formation and conservation of CpG islands, while the latter might be due to an “erosion” of the original CG positioning, originated as a consequence of the deamination process occurring at methylated CGs outside CpG islands, which causes a spontaneous mutation  $CG \rightarrow TG$ .

In the end, we saw that CG and TA distance distributions can be well described by the same function, precisely a shifted power-law with an exponential tail, meaning that the same process could have given rise to both. In particular, we saw that the value of the characteristic distance  $b$  between two consecutive TA along the sequence estimated by the fitting procedure is much larger than the maximum distance actually considered for the fit, leading us to the conclusion that the exponential tail is negligible and that the actual fitting distribution is a pure shifted power-law. Furthermore, we saw that the same function well describes also TA distance distributions of *Pan troglodytes* and *Macaca mulatta*, providing almost the same values for the parameters  $a, c, d$  and showing exactly the same deviations from the trend estimated by the fitting function: a sharp peak at 91 bp and a Gaussian peak centered around 142 bp, that seems to unite the features of the genetic sequences of these primates.

Focusing again on human genome and investigating the properties of the sequences

between consecutive TA separated by a distance of 91 bp (TA91) and 142 bp (TA142), we identified specific patterns associated with the occurrence probability of AA/TT and CG. In particular, we saw that the 40% of TA91 sequences have a good match with SINE (Short Interspersed Nuclear Elements), and, in particular, with Alu sequences, which are mobile non-coding elements involved in epigenetic and structural processes [68]. Furthermore, we observed that the average distance between consecutive AA/TT increases as TA distance increases, as we would expect in a random null model, except for TA distances corresponding to the Gaussian peak centered around 142 bp, where it assumes values close to 10 bp, which corresponds to the typical distance between AA/TT/TA characterizing the 147bp-sequences wrapped around the histone octamer [61, 64, 65]. Therefore, these results suggest that the sequences corresponding to the sharp peak at 91 bp and the Gaussian peak centered around 142 bp have peculiar properties, concerning not only epigenetic and structural processes, but also DNA flexibility, that deserve further investigations.

## CHAPTER 3

---

### Unfolding the genome: from chromosome structure to sequence

---

In this chapter we will see an algorithm for building genome assembly starting from Hi-C contact maps, that is based on a network approach. Hi-C matrices, in fact, can be represented as undirected weighted networks where nodes correspond to chromosome bins and links to Hi-C contact values. In particular, we will see how the spectral properties of the Laplacian matrix will lead to identify the different chromosomes and reconstruct their sequences.

#### 3.1 Hi-C contact maps

Hi-C (High-throughput Chromosome conformation capture) is a technique that allows to identify chromatin interactions across the entire genome [14, 15], providing information about the spatial proximity between pairs of DNA fragments sampled from a *population of millions of cells* [16]. From an experimental point of view, the steps that lead to the production of this type of data are the following (see Fig. 3.1): the DNA is cross-linked with formaldehyde and subsequently cut by a restriction enzyme; the two ends of the obtained DNA are then ligated and marked with biotin, giving rise to a chimeric DNA molecule composed by two different fragments that were originally close in the nuclear space. This new sample is then purified and sheared, and the biotinylated junctions are identified by paired-end sequencing. Therefore, in the end, this experimental technique provides pairs of reads that allow to map onto a reference genome the two fragments composing the ligation products and see which parts of the DNA are interacting the most. In particular, the interacting parts can be defined by dividing the reference sequence into  $n$  regions of equal size - that we will call bins - and count the number  $m_{ij}$  of ligation



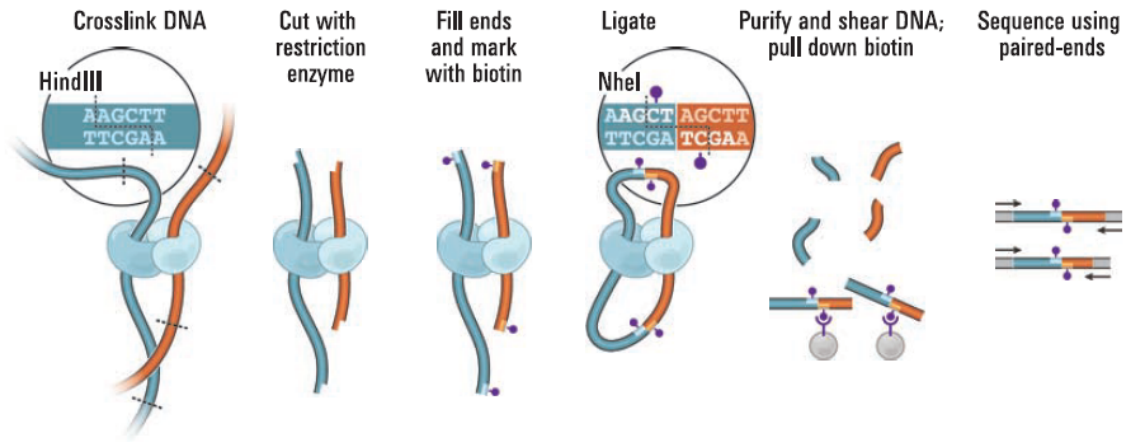


Figure 3.1: *Overview of Hi-C technique [14].*

products between bin  $i$  and bin  $j$ . If we consider all the possible interactions between pairs of bins, we can collect this information into a  $n \times n$  symmetric matrix where the value of a single entry is  $m_{ij}$ .

From the visualization of a Hi-C contact matrix (see Fig. 3.13), we can clearly see blocks along the diagonal representing the different chromosomes, each one showing a contact probability  $p$  that decreases as the genomic distance  $d$  (i.e. distance along the sequence) between a pair of bins increases, suggesting a polymer-like behavior in which the closest regions in the space are also the closest regions along the sequence [14]. In particular, the contact probability as a function of genomic distance, averaged across the genome, scales as  $1/d$  in the region between 500 kb and 7 Mb, if represented on log-log axes, thus showing a power-law trend that is compatible with a polymer conformation known as *fractal globule* [69, 70, 14, 71]. Fractal globules are interesting structures for modeling chromatin interactions since they are unknotted [70, 72] and this would facilitate the folding and the unfolding of chromatin during gene expression [14]. Furthermore, fractal globules are characterized by a structure organized into *territories*, meaning that contiguous regions of the DNA tend to cluster in close spatial proximity, giving rise to spatial sectors homogeneously populated in terms of chromosome composition [14, 71]. This is in line with the experimental observations that revealed a compartmentalized structure of the DNA within the nucleus [73, 74, 75, 71].

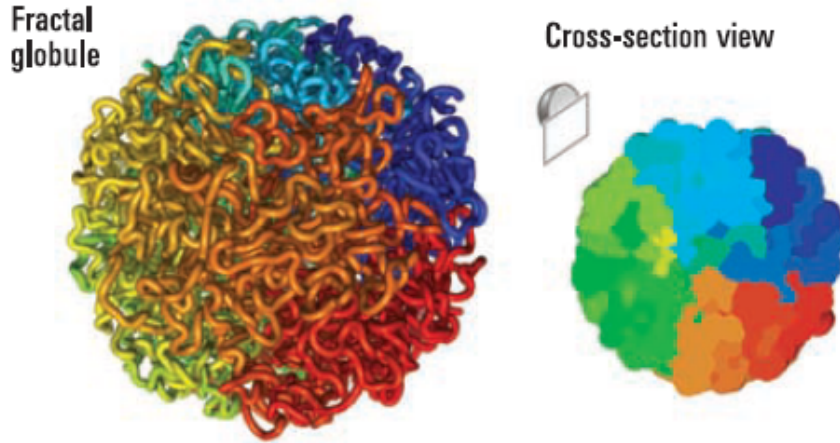


Figure 3.2: *Simulated fractal globule structure. Neighboring regions along the sequence (here represented with the same color) tend to be close also when folded in the 3D nuclear space, leading to a structure characterized by monochromatic blocks clearly visible both on the surface and in a cross section, and by the absence of knots [14].*

## 3.2 Hi-C data and genome assembly

By genome assembly we mean the process that leads to the reconstruction of the DNA sequence starting from its fragments, and it can be done in two ways: the fragments can be aligned to a reference genome or combined together by exploiting their mutual overlaps. Both the former and the latter (known as *de novo* assembly) procedures are usually applied to next-generation sequencing (NGS) data, which consist in DNA fragments that do not carry any information about their spatial proximity within the nucleus, thus not allowing to accurately reconstruct complex or polyploid genomes and to identify chromosomal rearrangements, which are particularly relevant for studying the majority of cancers [22, 23]. The advantage of using Hi-C data in the context genome assembly, consists exactly in having DNA fragments that carry information about their spatial proximity within the nucleus, thus allowing to identify chromosomal rearrangements [23].

In the last 7 years [17, 18], several tools have been developed for combining NGS and Hi-C data, exploiting, in particular, the latter for scaffolding contigs [19, 20, 21], which are small DNA subsequences obtained by merging DNA fragments (produced through NGS technique) that contain unique overlapping motifs [22]. In particular, the scaffolding process comprises four different steps: (1) karyotyping, that consists in assigning each contig to a chromosome; (2) contig ordering, that consists in determining the order of contigs within a chromosome; (3) contig positioning, that consists in calculating the

position of each contig within a chromosome; and (4) contig orientation, that consists in identifying the order of a contig with respect to all the others within a chromosome [22]. In the next sections, we will see two different implementations of the scaffolding procedure, based on network approaches, that I developed starting from my internship in Prof. Marc A. Marti-Renom's Lab, where I collaborated to the GENIGMA project.

### 3.3 My contribution in the Genigma project

GENIGMA is a citizen science project (<https://www.genigma.app>), part of ORION (Open Responsible research and Innovation to further Outstanding kNowledge) H2020 Initiative. The scientific goal of the project is to develop an app that people can install on their devices (smartphones or tablets) and use to identify possible chromosome translocations in human cancer cell-lines starting from Hi-C datasets. The players will receive a set of genome fragments that they have to order guided by a score, which is computed starting from the number of contacts observed in a Hi-C experiment: the higher score, the better the reconstruction of the DNA sequence. The novelty of the project consists in using as inputs Hi-C data sets, and in establishing a consensus over the structural features annotated during the game, thanks to the participation of the players.

My contribution consisted in building a bioinformatics pipeline that allows to prepare the input data and to analyze the outcome of the video game.

#### 3.3.1 The pipeline

My scientific activity has been focused on developing a bioinformatics pipeline that was divided into three main blocks (see Fig. 3.3):

1. Pre-game: assemble small pieces of genome (scaffolds) and group them into sets that will be the input of the game;
2. Game: mimic the players' activity by moving each piece of the genome to find the optimal order;
3. Post-game: merge together the scaffolds reordered by the players.

The pipeline was written by combining three different languages: Python, Bash and Awk, and its performance was tested by comparing the results obtained in two different scenarios: (1) using toy-model scaffolds produced by dividing the reference genome into bins of equal length and (2) using *de novo* scaffolds. The activity of the player was substituted with an automated search of the best combination, that, under certain assumptions, significantly reduces the computational cost of the procedure and provides quite accurate solutions.

## Pipeline

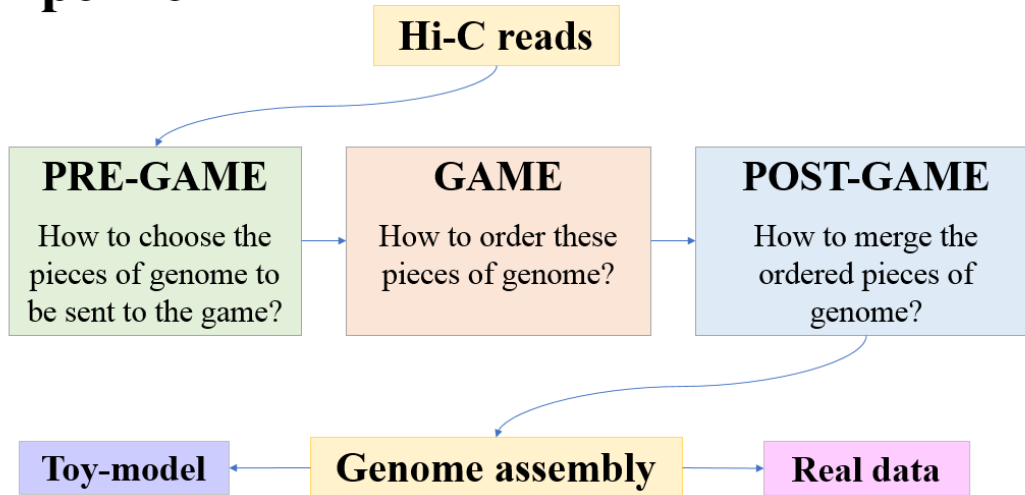


Figure 3.3: *Schematic representation of the pipeline.*

### Pre-game analyses

The pre-game part consisted in producing small pieces of genome, called scaffolds, and grouping them into appropriate sets that will constitute the input data for the video game. To do this, we focused on a small region of the genome so that we could develop and quickly test the prototype of the pipeline on a real case scenario. Specifically, the selected region was 1 Mbp long, going from 7Mbp to 8Mbp of human chromosome 19, and the Hi-C data [76] were downloaded from GEO database [77]. Firstly, we used Hi-C reads to produce a set of initial scaffolds using the MaSuRCA genome assembler [78]. Subsequently I mapped Hi-C reads onto the scaffolds using GEM mapper [79], and built an undirected weighted network where each node corresponded to a scaffold and each link weight to the number of contacts between scaffolds. Since the DNA molecule is a polymer, we expect that the higher the weight, the closer the scaffolds in the 3D conformation and along the sequence [20]. I characterized the network by computing: weight distribution, degree distribution and strength distribution. In particular, I focused on the relationship between strength (which measures the total number of contacts that a single scaffold has with all the others in the network) and scaffold length, finding a high correlation between the two (Pearson's correlation coefficient = 0.94, see Fig. 3.4), meaning that the total number of connections that a scaffold has depends on its length: the higher the length, the higher the number of contacts. This correlation is against our purpose, since the weight should be proportional to the proximity of the scaffolds

along the genomic sequence, without any bias related to the scaffold size. Therefore, I tested several weight normalizations to reduce this correlation, and applied the one that provided that lowest correlation value (i.e. Pearson's correlation coefficient = -0.15, see also Fig. 3.4):  $w_{ij} = c_{ij}/l_i l_j$ , where  $c_{ij}$ ,  $l_i$  and  $l_j$  correspond respectively to the number of contacts observed between scaffold  $i$  and scaffold  $j$ , to the length of scaffold  $i$  and to length of scaffold  $j$ .

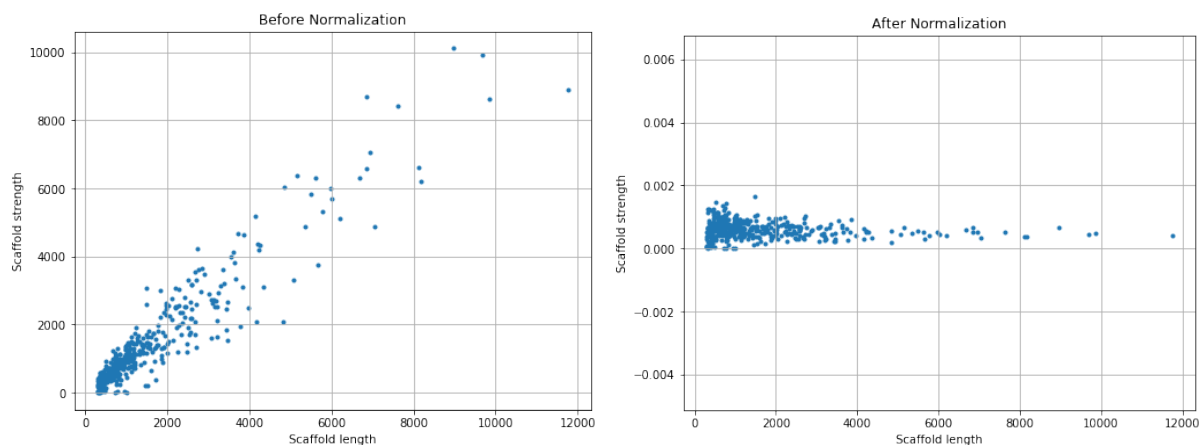


Figure 3.4: *Scatter plots of scaffold strength as a function of scaffold length (measured in bp), before (left panel) and after the normalization (right panel). The latter clearly shows that the correlation was removed.*

The next step consisted in extracting the subset of nodes to be sent to the game. To do that, I explored two possibilities: extracting fully and not fully connected subnetworks [80, 81]. First of all, I tried to understand if I could distinguish noise from signal, by identifying a specific weight threshold value under which I could remove links that were not relevant in terms of genomic contiguity. Therefore, I applied an iterative thresholding procedure by removing all links whose weight was lower than a threshold value  $t$ , ranging from the minimum weight value to the maximum. For each value  $t$ , I calculated the number of connected components and the size of the largest connected component, to check if I could identify a weight value beyond which the emerging structures were robust to link removal. If this value exists, when we plot the number of connected components as a function of the threshold, we should see an increasing trend that a certain point reaches a plateau and then increases again, until the number of connected components equals the number of nodes in the network. In the same way, if we plot the size of the largest connected component as a function of the threshold, we should see a decreasing trend that a certain point reaches a plateau and then decreases again, until the size of largest connected component equals the minimum possible value, that is 1 (i.e. the largest connected component is composed by only one node). As we can see from the results shown in Fig. 3.5, the expectations were not met. In particular, we notice that

the number of connected components quickly increases as the threshold value increases, and that the size of the largest connected component immediately decreases to values close to a few tens of nodes, meaning that by simply removing the links with the lowest weight, we are almost completely disconnecting the network. This is a clear sign that the level of noise is very high.

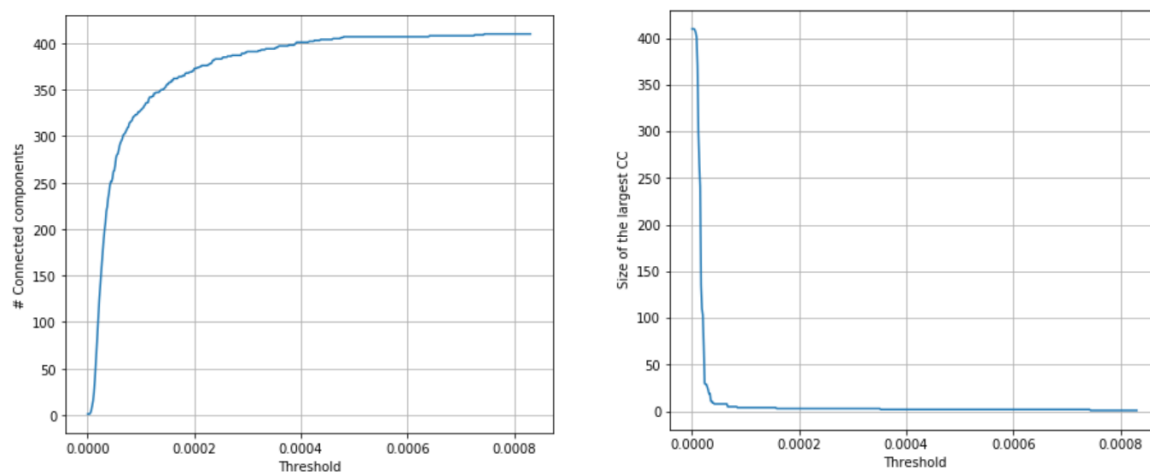


Figure 3.5: *Number of connected components (left panel) and size of the largest connected component (right panel) as a function of the threshold value. From the plot on the left we can see that the number of connected components quickly increases as the threshold value increases (left panel), and that the size of the largest connected component immediately decreases to values close to a few tens of nodes (right panel), meaning that by simply removing the links with the lowest weight, we are almost completely disconnecting the network. This is a clear sign that the level of noise is very high.*

Therefore, since the introduction of the normalization reduced the correlation between scaffold strength and scaffold length but did not help in distinguishing signal from noise, we opted for a different solution: we decided to divide scaffolds into bins of equal length (500 bp), that will represent the new nodes of the network, and to use the actual number of Hi-C contacts between pairs of bins as link weight (see Fig. 3.6).

Subsequently, we decided to extract the subnetworks by sorting the link weights in descending order and selecting the first link connecting bins belonging to different scaffolds, thus producing a submatrix composed by the bins of two scaffolds at a time, as shown in Fig. 3.7.

To resume, the steps followed by the pipeline corresponding to the pre-game block are:

1. Producing *de novo* scaffolds using the MaSuRCA genome assembler.
2. Divide each scaffold into 500 bp bins.

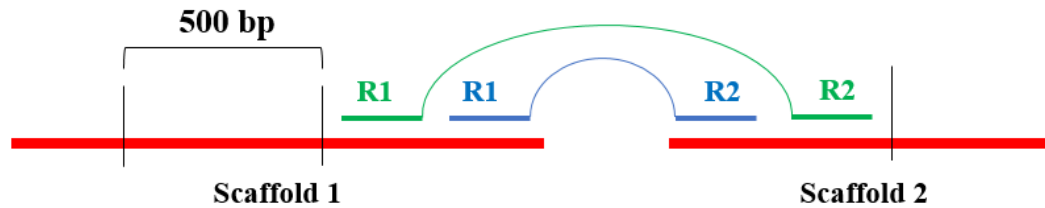


Figure 3.6: Representation of the procedure followed to identify bins (i.e. the nodes of the network) and Hi-C contact values between them (i.e. the weights assigned to the links connecting the nodes of the network). R1 and R2 are Hi-C reads, that corresponds to the DNA sequence fragments composing the ligation products obtained by Hi-C experiments.

3. Map Hi-C reads (i.e. R1 and R2) onto scaffolds.
4. Get a list of contacts between bin pairs, sorted in descending order.
5. Select the first link connecting bins that belong to different scaffolds.
6. Extract the matrix corresponding to the contacts between the bins composing the two scaffolds.

### Simulation of the game

The game consists in identifying the combination of nodes that corresponds to the best path within each subnetwork, which must satisfy the following requirements: (i) it must have the maximum length and (ii) it must visit each node only once. From a theoretical point of view, this corresponds to compute all the possible permutations of the  $N$  nodes composing the subnetwork (since we do not know which are the starting and ending nodes of the path), which scales as  $N!$ . In practice, this computation becomes very heavy as soon as  $N$  increases (e.g., for  $N = 10$  there are already 3,628,800 paths to evaluate), therefore we devised an algorithm, that, under certain assumptions, significantly reduces the computational cost of this procedure and provides a quite accurate solution through the following steps:

1. Select one of the  $N$  nodes composing the subnetwork.
2. Explore all the weights of the links connecting the selected node with its nearest neighbors.
3. Move to the node which is connected by the link with the highest weight.
4. Repeat step 2 without considering the links already selected.

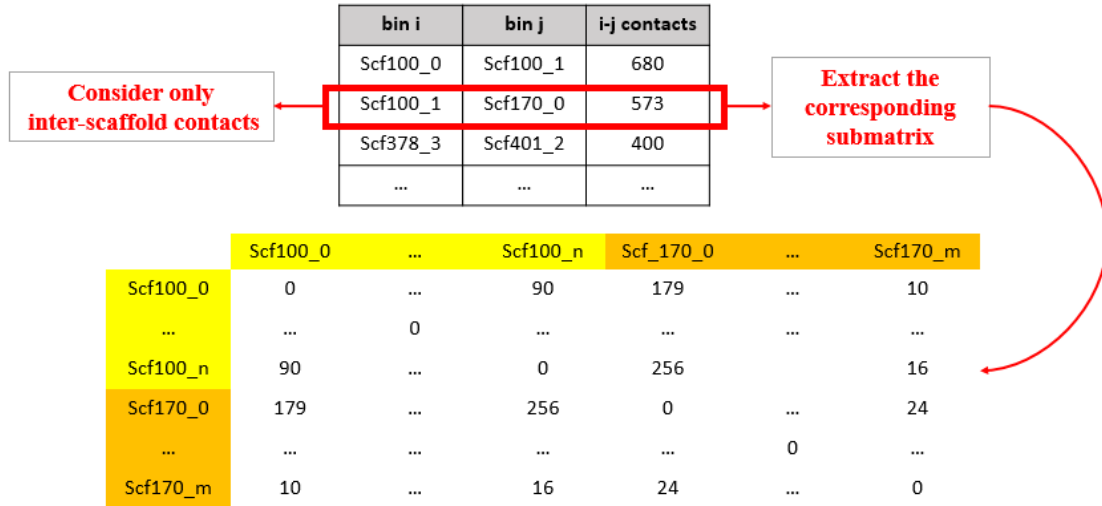


Figure 3.7: Representation of the procedure followed for extracting the subnetworks to be sent to the game. The steps are the following: sorting the link weights in descending order and selecting the first link connecting bins that belong to different scaffolds, thus producing a submatrix composed by the bins of two scaffolds at a time.


- When the N-th node has been reached, go back to step 1 until all the N nodes have been used as starting point.
- Among the N paths, select the one providing the longest path.

The idea is that the approximated solutions of this quick algorithm will be compared to the performance of the players that during the game will provide the actual best path. They, in fact, will search for the best path by moving each piece of the genome to find the combination that provides the highest score. After the best-path calculation, the reordered sequences were validated by mapping the pieces of the genome (nodes) onto the reference to check whether the positioning provided by the best path corresponded to the actual one (see figure Fig. 3.8 as an example).

### Post-game analyses

In the post-game analysis, we aim to merge scaffold pairs reordered by the game, by searching for overlaps between the termini of the two reordered scaffold sequences. To increase the probability of identifying a statistically significant overlap, we impose that its length must be greater than a given threshold value, set to 16 bp. In particular, we tested the procedure on both a toy-model and real data by applying the following steps:





Scaffold ID	Starting position
scf_1325	7407966
scf_1263	7717744
scf_1184	7108910
scf_1256	7167305
scf_1311	7331770
scf_1277	7187619
scf_1356	7353069
scf_1324	7360515
scf_1220	7385447
scf_1366	7085275

Figure 3.8: *Example of scaffold ordering provided by the fast algorithm developed to compute the longest path within a network. The second column indicates the positions of the first nucleotide of each scaffold within the reference genome. As we can see, there are some differences between the order identified by the algorithm and that provided by the reference genome.*

1. Search for an overlap between the termini of the two re-ordered scaffolds.
2. If an overlap was found, merge the two scaffolds.
3. Divide the new scaffold into 500 bp bins.
4. Map R1 and R2 reads onto it.
5. Get a new contact list.
6. Repeat pre-game, game and post-game steps until the highest inter-scaffold contact value in the list is lower than 10.

The toy-model was built by dividing the reference genome into scaffolds of equal length (1500 bp), with an overlap of 500 bp with the preceding and the following along the sequence, as shown in Fig. 3.9. The contacts between pairs of 500 bp bins are computed as explained in the pre-game section and shown in Fig. 3.6.

The results have been evaluated by comparing the assembly produced by our pipeline with the reference genome via D-GENIES dot-plots [82], which are graphs representing on the x-axis the sequence corresponding to the reference genome and on the y-axis the sequence corresponding to the obtained assembly; a dot is placed at a specific position

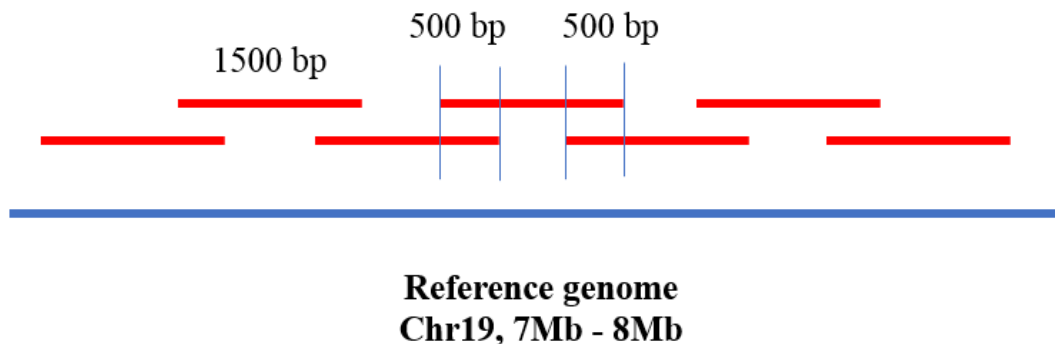


Figure 3.9: Representation of the procedure followed to build the scaffolds used as toy-model. The reference genome into scaffolds of equal length (1500 bp), with an overlap of 500 bp with the preceding and the following along the sequence.

when there is a match between the two: the closer the dots to the diagonal, the higher the similarity between the reference and the obtained assembly. Together with dot-plots, other statistical measures were computed on the initial scaffold file (initial state) produced by the pre-game block and on the final scaffold file produced after the game and post-game blocks (final state): (i) number of scaffolds, (ii) average scaffold length, (iii) largest scaffold length, (iv) sum of scaffold lengths, and (v) percentage of scaffold identity with the reference.

	Toy-model scaffolds		<i>de novo</i> scaffolds	
	Initial state	Final state	Initial state	Final state
$N$	1000	911	410	328
$\bar{l}$ (bp)	1499	1597	1627	2013
$l_{max}$ (bp)	1500	4500	11756	21935
$l_{tot}$ (bp)	1499500	1455000	667072	660443

Table 3.1: Statistical measures computed to evaluate the performance of the pipeline, on both the initial scaffold file (initial state) produced by the pre-game block of the pipeline and on the final scaffold file (final state) produced by the game and post-game blocks of the pipeline: number of scaffolds  $N$ , average scaffold length  $\bar{l}$ , largest scaffold length  $l_{max}$ , and sum of scaffold lengths  $l_{tot}$ .

The results obtained by these analyses highlighted that: (1) toy-model scaffolds provided an assembly having a better agreement with the reference genome, since its dot-plot shows a line close to the diagonal, and all the assembled scaffolds are represented in

dark green, meaning that the percentage of identity with the reference is greater than the 75%; whereas, some of the assembled *de novo* scaffolds are represented in orange and yellow, meaning that the percentage of identity with the reference is lower than the 50% and the 25%, respectively (see Fig. 3.10). (2) *De novo* scaffolds do not cover all the region of interest, since the sum of scaffold length is 667072 bp before applying the game and post-game blocks (see Table 3.1); (3) when dealing with both toy-model and *de novo* scaffolds, the overlaps are very few, thus retrieving a sequential order between one another is not always possible. To overcome this issue, I introduced a variant to the algorithm that aims to reorder the scaffolds through the following steps:

1. Convert the initial undirected network into a directed network, adding between each pair of scaffolds  $(i, j)$  a link going from scaffold  $i$  to scaffold  $j$  with weight  $w_{ij}$  and a link going from scaffold  $j$  to scaffold  $i$  with the same weight  $w_{ij}$ .
2. If there is an overlap between the last part of the sequence of scaffold  $i$  and the initial part of scaffold  $j$ , remove the link going from scaffold  $j$  to scaffold  $i$ .
3. If there are more than two consecutive overlapping scaffolds, remove the link going from the first to the last scaffold and the link going from the last to first scaffold.
4. Once all the links at point 3 and 4 have been removed, search for the best path on the directed network by imposing that we can move from node  $i$  to node  $j$  only through outgoing links.

Since also this variant did not significantly improve the assembly, I proposed an algorithm that performs the scaffolding process in a computationally efficient manner, by starting from a Hi-C contact map at higher resolution and exploiting the properties of the Laplacian matrix of a network.

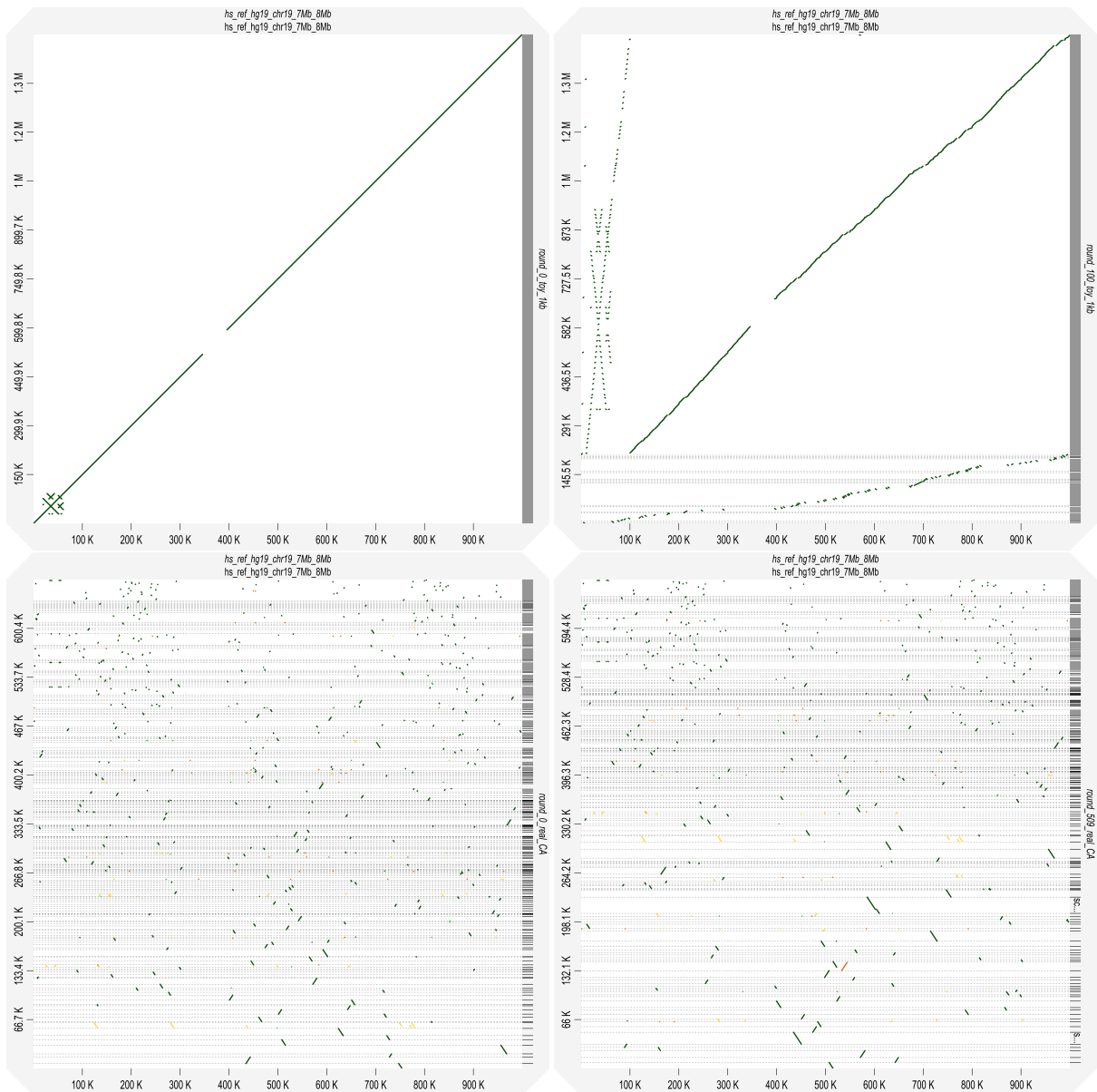


Figure 3.10: Dot-plots produced to evaluate the produced assembly (y-axis) by the comparison with the reference genome (x-axis). Top: dot-plot evaluating the toy model scaffolds before (left) and after (right) applying the pipeline. Bottom: dot-plot evaluating the de novo scaffolds before (left) and after (right) applying the pipeline. The colors of the points correspond to the different levels of identity between the reconstructed sequence and the reference: dark green for identity percentage greater than 75%, light green for identity percentage lower than 75%, orange for identity percentage lower than 50%, yellow for identity lower than 25%. These results show that the pipeline produces better results when applied to toy-model scaffolds.

### 3.4 The Laplacian matrix of a network

As we have seen in the previous sections, networks represent a powerful mathematical framework for studying *interactions* between *parts of a system*. They are constituted by two types of elements, nodes and links, representing respectively the parts of the analyzed system and their interactions. From an algebraic point of view, a network can always be represented by its *adjacency matrix*, which can be binary or weighted. In the former case, the matrix is symmetric and tells whether between each couple of nodes,  $i$  and  $j$ , there is a link or not. This can be translated into numbers by assigning to each entry  $A_{ij}$  of the matrix the following quantities:

$$A_{ij} = \begin{cases} 1 & \text{if there is a link between node } i \text{ and node } j \\ 0 & \text{if there is no link between node } i \text{ and node } j \end{cases} \quad (3.1)$$

In the latter case, the matrix is symmetric and not only tells whether between each couple of nodes,  $i$  and  $j$ , there is a link or not, but also how strong this connection is, namely which is the *weight* of the connection. This can be translated into numbers by assigning to each entry  $W_{ij}$  of the matrix the following quantities:

$$W_{ij} = \begin{cases} w_{ij} \geq 0 & \text{if there is a link between node } i \text{ and node } j \\ 0 & \text{if there is no link between node } i \text{ and node } j \end{cases} \quad (3.2)$$

In both cases, the defined matrices are symmetric, meaning that there is no preferential direction associated with the links. In other words, they represent *undirected* networks, and for both of them the Laplacian matrix  $L$  can be calculated. Specifically, for binary adjacency matrices  $A$ ,  $L$  is defined as:

$$L = D - A \quad (3.3)$$

where  $D$  represents the degree matrix, whose entries are defined as:

$$D_{ij} = \begin{cases} d_i & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

where  $d_i$  represents the degree of node  $i$ :

$$d_i = \sum_{j=1}^n A_{ij} \quad (3.5)$$

The same definition can also be extended to weighted networks as follows:

$$L = S - W \quad (3.6)$$

where  $W$  and  $S$  represent respectively the weighted adjacency matrix and the strength matrix of the network, whose entries are defined as:

$$S_{ij} = \begin{cases} s_i & \text{if } i=j \\ 0 & \text{otherwise} \end{cases} \quad (3.7)$$

where  $s_i$  represents the strength of node  $i$ :

$$s_i = \sum_{j=1}^n w_{ij} \quad (3.8)$$

As we have seen in section 3.1, Hi-C contact maps meet all the assumptions made for defining *undirected weighted networks*, since they are symmetric and their entries correspond to the number of ligation products between two bins, that are positive numbers. Moreover, the fact that the weight (i.e. contact value) is a decreasing function of the genomic distance  $d$ , allows to define a notion of proximity between nodes that distinguishes this network from a merely topological one, such as world wide web, where the connection between nodes and their proximity is not related to a “physical” distance.

### 3.4.1 Properties of the Laplacian matrix

Given an undirected network  $G$ , its Laplacian matrix  $L$  satisfies the following properties [83]:

1. Let  $G$  be a network corresponding to a lattice of  $n$  nodes in the Euclidean space  $\mathbb{R}^m$  and  $f$  be a function that assigns to each node of  $G$  a real number. Then  $f$  can be seen as a vector  $\in \mathbb{R}^n$  and the Laplacian  $L$  can be seen as an operator  $Lf = \nabla^2 f$  acting like the second derivative of  $f$  along the  $m$  axes of the lattice.
2. For every vector  $f \in \mathbb{R}^n$  we have

$$f^\top Lf = \frac{1}{2} \sum_{i,j=1}^n a_{ij} (f_i - f_j)^2 \quad (3.9)$$

where  $a_{ij}$  represents the entries of a general adjacency matrix, weighted or not. This means that the Laplacian matrix can always be seen a quadratic form corresponding to the squared difference of all the values on the nodes connected by a link.

3.  $L$  is symmetric and positive semi-definite.
4. The smallest eigenvalue of  $L$  is 0 and the corresponding eigenvector is the constant one vector  $\mathbb{1}$ .

5.  $L$  has  $n$  non-negative, real-valued eigenvalues  $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ .
6. Analyzing the spectrum of  $L$ , we can get the number of connected components in the network. In fact, the multiplicity  $m$  of the eigenvalue 0 equals the number of connected components  $A_1, \dots, A_m$  in the network, and the eigenspace of eigenvalue 0 is spanned by the indicator vectors  $\mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_m}$  of those components.
7. The first non-null eigenvalue is called Fiedler value and the corresponding eigenvector is called Fiedler vector. The Fiedler value represents the algebraic connectivity of the network, that is, the further from 0, the more connected is the network. Moreover, the multiplicity of the Fiedler value is always equal to 1.
8. The eigenvectors  $u_i$  of  $L$  of a connected network, form an orthonormal basis:  $u_i^\top u_j = \delta_{ij}$ .

### 3.4.2 Spectral clustering

Hi-C contact maps give us a picture of how all the chromosomes composing the genome are interacting in the nuclear space. Thus, the first step to do, in order to get their sequence (i.e. an assembly), is to identify the bins composing each chromosome. In terms of network, this means that we want to find a partition of the nodes such that the links between different groups have very low weights (i.e. few contacts) and the links within each group have very high weights (i.e. more contacts) [83]. This operation can easily be performed by an algorithm called spectral clustering, that allows to identify  $k$  clusters in a network through the following steps:

1. Compute the first  $k - 1$  eigenvectors corresponding to the smallest  $k - 1$  non-null eigenvalues of  $L$ .
2. Let  $U \in \mathbb{R}^{n \times (k-1)}$  be the matrix whose columns correspond to the first  $k - 1$  eigenvectors of  $L$ . Each node of the network is now represented by a point in a  $(k - 1)$ -dimensional space, whose coordinates are given by the rows of  $U$ . The  $k$  clusters can thus be identified by applying the k-means algorithm to the  $n$  points in  $\mathbb{R}^{k-1}$ .

Two delicate issues underlie these points:

1. The *number of eigenvectors* to be chosen in order to clearly distinguish the  $k$  clusters corresponding to the  $k$  different chromosomes. In order to address this issue, we will adopt a geometric approach and, in particular, we will consider the relationship between a network and a simplex  $S$ . In fact, it has been proved that every connected, undirected network of  $N$  nodes corresponds to a specific simplex in  $N - 1$  dimensions [84, 85] and that this correspondence can be studied through

the Laplacian matrix [86, 85]. In the specific case of a similarity matrix, such as a Hi-C contact map, where nodes cluster according to the  $k$  blocks formed along the diagonal, we hypothesized that this property can be extended as follows: every network, whose adjacency matrix is characterized by  $k$  blocks along the diagonal, corresponds to a specific simplex  $S$  in  $(k - 1)$  dimensions. This led us to conclude that in order to distinguish  $k$  clusters, the nodes need to be represented in a  $k - 1$  space, and therefore, the optimal number of eigenvectors for our purposes is  $k - 1$ .

2. *Consistency*, since spectral clustering can fail to converge [87]. In order to avoid this problem, we have to be sure that the  $k - 1$  eigenvalues of  $L$  corresponding to the  $k - 1$  eigenvectors used for the spectral clustering are all much smaller than the minimum degree (or strength) in the network [83].

### 3.4.3 Laplacian embedding

From a geometric point of view, computing a genome assembly starting from its 3D configuration (i.e. its Hi-C contact map), corresponds to map the 3D-contact-network of a genome onto a line (i.e. its sequence), so that highly connected nodes in the space stay as close as possible also along the sequence. From an algebraic point of view, this operation corresponds to computing the following minimization [88]:

$$\operatorname{argmin}_f f^\top L f \quad \text{with } f^\top f = 1 \text{ and } f^\top \mathbf{1} = 0 \quad (3.10)$$

whose solution is given by the eigenvector associated with the smallest non-zero eigenvalue of the eigenvalue problem  $Lf = \lambda f$ , namely the Fiedler vector. Therefore, once we have identified each chromosome through spectral clustering, we will use the Fiedler vector as a guide to get the sequence.

## 3.5 Chromosome identification through spectral clustering

As we have shown so far, a Hi-C contact map can be seen as the adjacency matrix of an undirected weighted network, where nodes represent bins and the weight of the links represents their spatial proximity: the higher the weight  $m_{ij}$ , the closer they are in the nuclear space and the closer they are along the sequence. Therefore, in order to get an assembly, the proposed algorithm will use weights as a guide to identify the actual signal associated with intra-chromosomal contacts and extract the clusters corresponding to the different chromosomes. In particular, this task will be performed through the following steps:

1. Compute the  $\log_{10}$  of the contact values.



2. Remove isolated nodes.
3. Add a value  $c_0$  to all non-zero contact values so that the weights of the links satisfy the condition  $w_{ij} > 0$ .
4. Identify, within an interval of selected values  $T_w$ , an optimal threshold value  $T$  that allows to: (a) preserve only those links whose weight correspond to the *actual chromatin-interaction signal* (i.e. to remove links corresponding to noise), without partitioning the network into disconnected components. This was achieved by accepting as a product of the thresholding procedure, only networks composed by one connected component or by one connected component and a set of isolated nodes. (b) Provide the best possible clustering, by maximizing the *silhouette value*. In fact, after applying each threshold  $T_w$ , spectral clustering is performed and the obtained grouping is evaluated through silhouette value, which is defined as  $S_{tot} = \sum_i S_i$ , with

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)} \quad i = 1, \dots, n \quad (3.11)$$

where  $i$  is the index associated to each point to be clustered,  $a_i$  is the average distance from point  $i$  to the other points in the same cluster, and  $b_i$  is the average distance from point  $i$  to the points in a different cluster, minimized over clusters.  $S_i$  ranges between -1 and 1, with high values (close to 1) indicating that point  $i$  is well matched to the points in its own cluster, and poorly matched to the points in the other clusters [89]. Therefore, in an ideal scenario, where each point is perfectly clustered, the silhouette value  $S_{tot}$  is equal to number of points to be grouped.

5. Remove all links whose weight is lower than the optimal threshold  $T$ .
6. Apply spectral clustering algorithm.

### 3.5.1 Results on a toy-model

The procedure was tested on a simple toy-model Hi-C contact map, composed by four diagonal blocks of different sizes: 20, 15, 10, and 8 bins respectively (the number of synthetic chromosomes is 4 because this is the largest number of clusters that we can actually visualize in a plot). The contact values within each block decay as  $c/d$ , where  $d$  represents the genomic distance and  $c$  is an arbitrary constant value, in this case set to 20. Outside the blocks, a constant value equal to 0.1 has been assigned to all the links and, in the end, a gaussian noise with mean value equal to 5 and variance equal to 1 has been added to all the links, getting the contact map represented in Fig. 3.11, whose contact values are distributed as in Fig. 3.14, clearly showing a peak centered around 0.7, which corresponds to  $10^{0.7} = 5$ , that is exactly the average value set for generating the gaussian noise. Therefore, we can conclude that the main peak on the

left, centered around 0.7, describes the noise within the contact map and that the actual signal, corresponding to simulated chromatin interactions, can be found on the right tail of the distribution. In order to identify the optimal threshold value that allows to separate clusters corresponding to different chromosomes (i.e. to distinguish signal from noise), all the set of possible values was explored, identifying the optimal  $T$  at 0.8, where the silhouette value reaches its absolute maximum (see Fig. 3.14).

After removing all links whose weight was lower than 0.8, spectral clustering algorithm was applied, obtaining the result shown in Fig. 3.11, where the four colors correspond to the four clusters identified by k-means algorithm as the four chromosomes. Both cluster representation and confusion matrix (see Tab. 3.2) confirm that all chromosomes were correctly detected.

		PREDICTED			
		chr1	chr2	chr3	chr4
TRUE	chr1	20	0	0	0
	chr2	0	15	0	0
	chr3	0	0	10	0
	chr4	0	0	0	8

Table 3.2: *Confusion matrix evaluating spectral clustering performance in identifying the different chromosomes starting from a toy-model Hi-C contact map.*

### 3.5.2 Results on a real Hi-C contact map

Publicly available Hi-C data of GM12878 cell line [76] were downloaded from GEO database [77] and processed by Marti-Renom’s lab via TADbit [90], obtaining an ICE normalized Hi-C contact map at 1Mb resolution. As a first step, chromosome Y was excluded from the analysis, since the chosen cell line is a female one. Subsequently, the computation of the  $\log_{10}$  of contact values was performed, followed by the removal of the isolated nodes that led to the loss of 117 bins (see Fig. 3.13 for the representation of the obtained contact map). Since the computation of the  $\log_{10}$  of contact values produced a negative weight for some of the links, a value of 3 was added to all non-null contact values, in order to get only positive weights  $w_{ij} > 0$  (as the assumptions of spectral clustering require), whose distribution is shown in Fig. 3.14. As we can see, the shape of the histogram is similar to the one obtained for the toy-model (see Fig. 3.12), with a less pronounced peak on the right tail of the distribution. This comparison suggests that the main peak, here centered around 2.1, could represent the noise associated with ICE normalized Hi-C data and that the actual chromatin-interaction signal is represented by the lower peak, observed in the right tail of the distribution. Therefore, the search of the optimal threshold value, was performed within an interval between 2.6 and 3.8, obtaining

as result  $T = 3.1$  (see Fig. 3.14). After the removal of the links with weight lower than 3.1, the spectral clustering algorithm was applied imposing the search of 23 clusters to k-means, obtaining the result shown in Table 3.3. As we can see, most of the clusters are homogeneous in their composition, with 4 exceptions: cluster 1, 2, 4 and 5, which are composed by bins belonging to different chromosomes. In particular, cluster 1 is composed by 229 bins of chromosome 1 and 1 bin of chromosome 10; cluster 2 is composed by 242 bins of chromosomes 2, 3 bins of chromosome 21 and 1 bin of chromosome 1; cluster 4 is composed by 190 bins of chromosome 4 and 1 bin of chromosome 1; cluster 5 is composed by 179 bins of chromosome 5 and 1 bin of chromosome 19. This heterogeneous composition can represent the presence of potential *translocations*, meaning that there has been an exchange between parts of two non-homologous chromosomes. In particular, the identified rearrangements consist in:

- 1Mb of chromosome 10 translocated on chromosome 1;
- 3Mb of chromosome 21 translocated on chromosome 2;
- 1Mb of chromosome 22 translocated on chromosome 2;
- 1Mb of chromosome 1 translocated on chromosome 4;
- 1Mb of chromosome 19 translocated on chromosome 5.

In order to verify whether the algorithm succeeded or not in this identification, the contact map of the chromosome pairs involved in the potential translocations were represented (see Figures 3.15, 3.16, 3.17, 3.18, and 3.19). The composition of cluster 1 is telling us that 1 bin of chromosome 10 is translocated onto chromosome 1, and this finding is confirmed by the pattern observed in the contact map shown in Fig. 3.15: the translocated bin corresponds to the 41<sup>st</sup> of chromosome 10, which shows weak contacts with the bins of chromosome 10 and strong contacts with almost all the bins of chromosome 1. A similar pattern was also observed for the potential translocations detected between chromosome 1 and chromosome 4 (see Fig. 3.18), between chromosome 22 and chromosome 2 (see Fig. 3.17), and between chromosome 19 and chromosome 5 (see Fig. 3.19), confirming that the algorithm actually did a good job. The potential translocation between chromosome 21 and chromosome 2, instead, deserves a separate discussion; in fact, the representation of the contact map between this pair of chromosomes shows a brighter block of contacts between the first 3 bins of chromosome 21 and a limited number of bins of chromosome 2. Neither corresponding intra-chromosomal weak contacts are observed, nor strong contacts with all (or at least the majority of) the bins of chromosome 2. This might be due to two reasons: (1) the translocation does not involve the 100%, or at least the majority, of the cell population; (2) the observed signal is associated to another chromosomal aberration, such as an insertion.

Cluster	Bins	Chr	Cluster	Bins	Chr
1	229	chr1	13	97	chr13
	1	chr10	14	89	chr14
2	242	chr2	15	83	chr15
	3	chr21	16	81	chr16
	1	chr22	17	80	chr17
3	196	chr3	18	77	chr18
4	190	chr4	19	57	chr19
	1	chr1	20	61	chr20
5	179	chr5	21	35	chr21
	1	chr19	22	35	chr22
6	167	chr6	23	151	chrX
7	158	chr7			
8	145	chr8			
9	125	chr9			
10	133	chr10			
11	133	chr11			
12	132	chr12			

Table 3.3: Clusters identified as chromosomes through spectral clustering algorithm and their composition. Cluster 1, 2, 4 and 5 (highlighted in red) are composed by two types of chromosomes, revealing the presence of potential translocations.

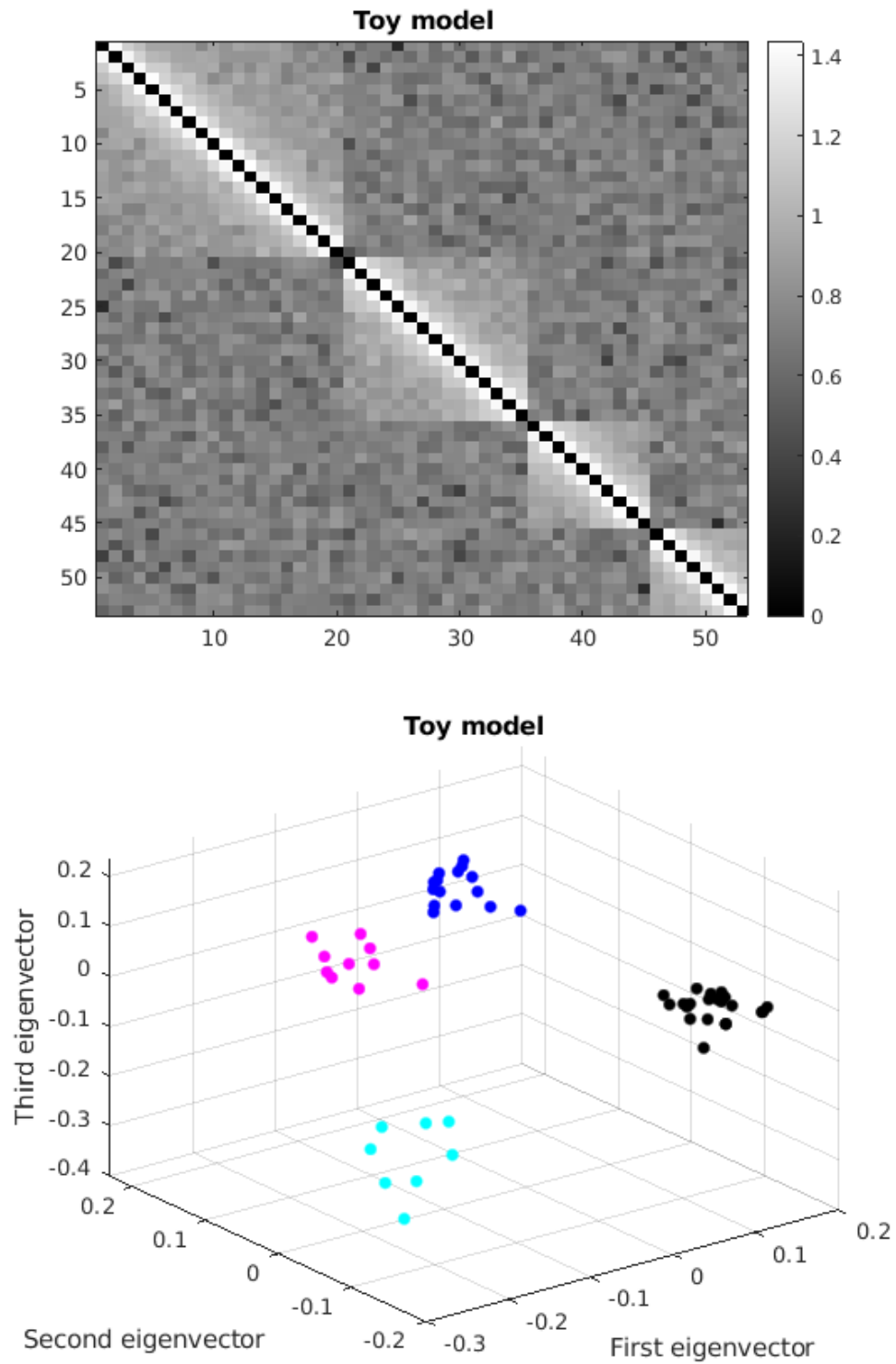


Figure 3.11: *Toy-model Hi-C contact map and its spectral clustering results, where the four colors correspond to the four clusters identified by k-means algorithm as the four chromosomes.*

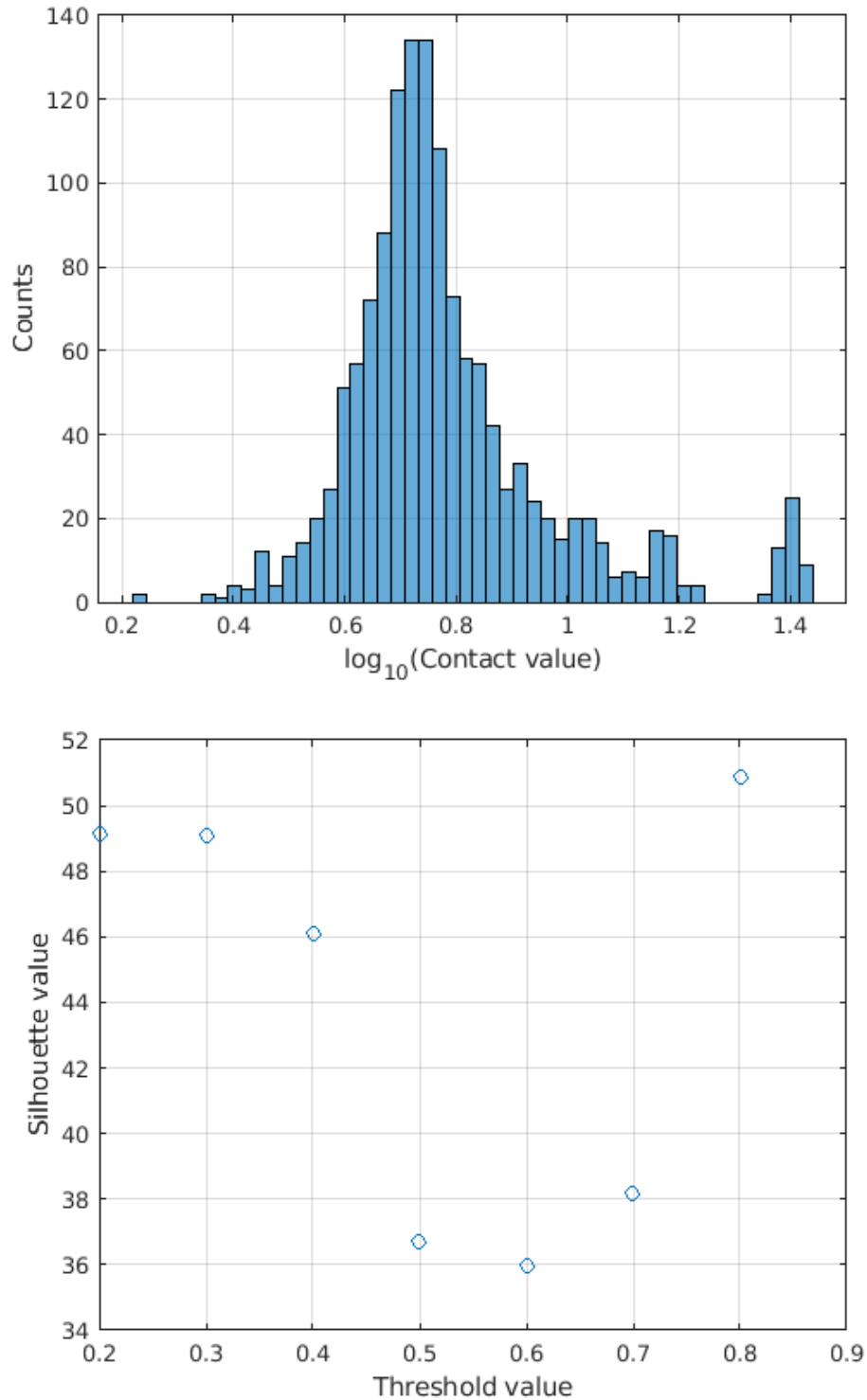


Figure 3.12: Histogram of the logarithm of contact values for the toy-model Hi-C contact map, together with silhouette value as function of the threshold. The threshold value chosen as optimal is 0.8, since it corresponds to the maximum silhouette value.

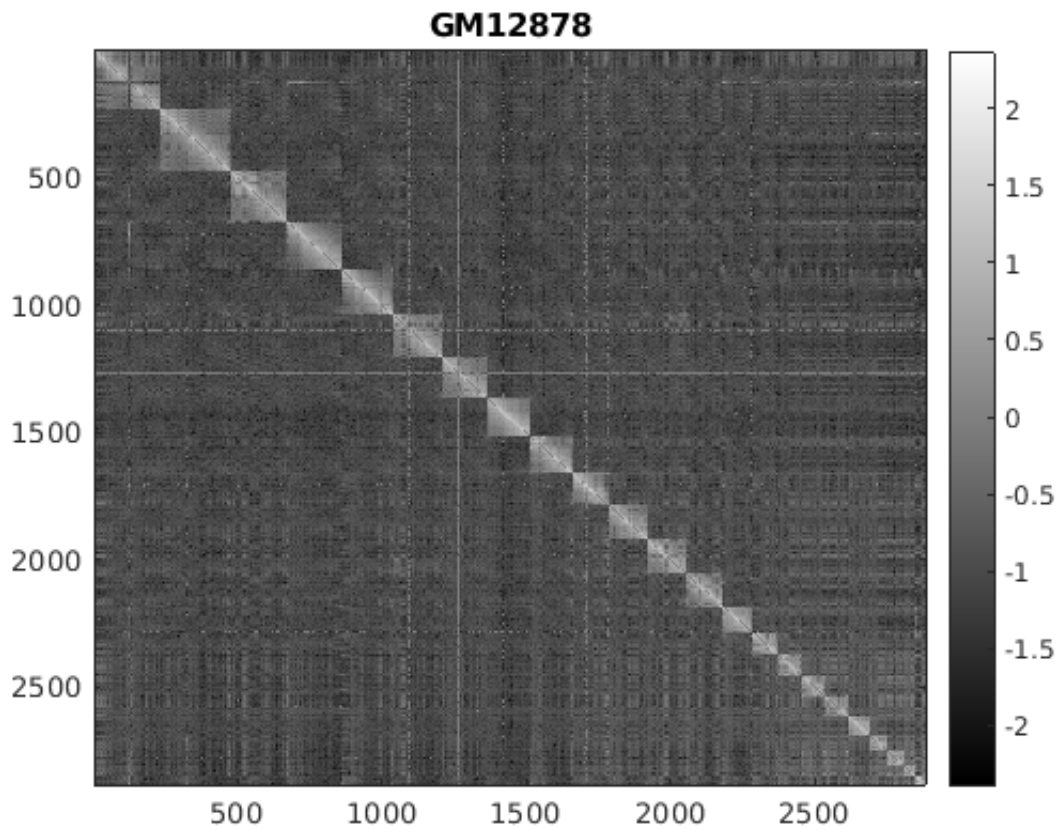


Figure 3.13: Representation of Hi-C contact map of GM12878 cell after considering the  $\log_{10}$  of contact values and after removing isolated nodes. The brighter blocks along the diagonal correspond to the different chromosomes, here sorted from the longest to the shortest: chr1, chr2, chr3, chr4, chr5, chr6, chr7, chrX, chr8, chr9, chr10, chr11, chr12, chr13, chr14, chr15, chr16, chr17, chr18, chr20, chr19, chr22, chr21. Chromosome Y was excluded from the analysis, since the considered cell line is a female one.

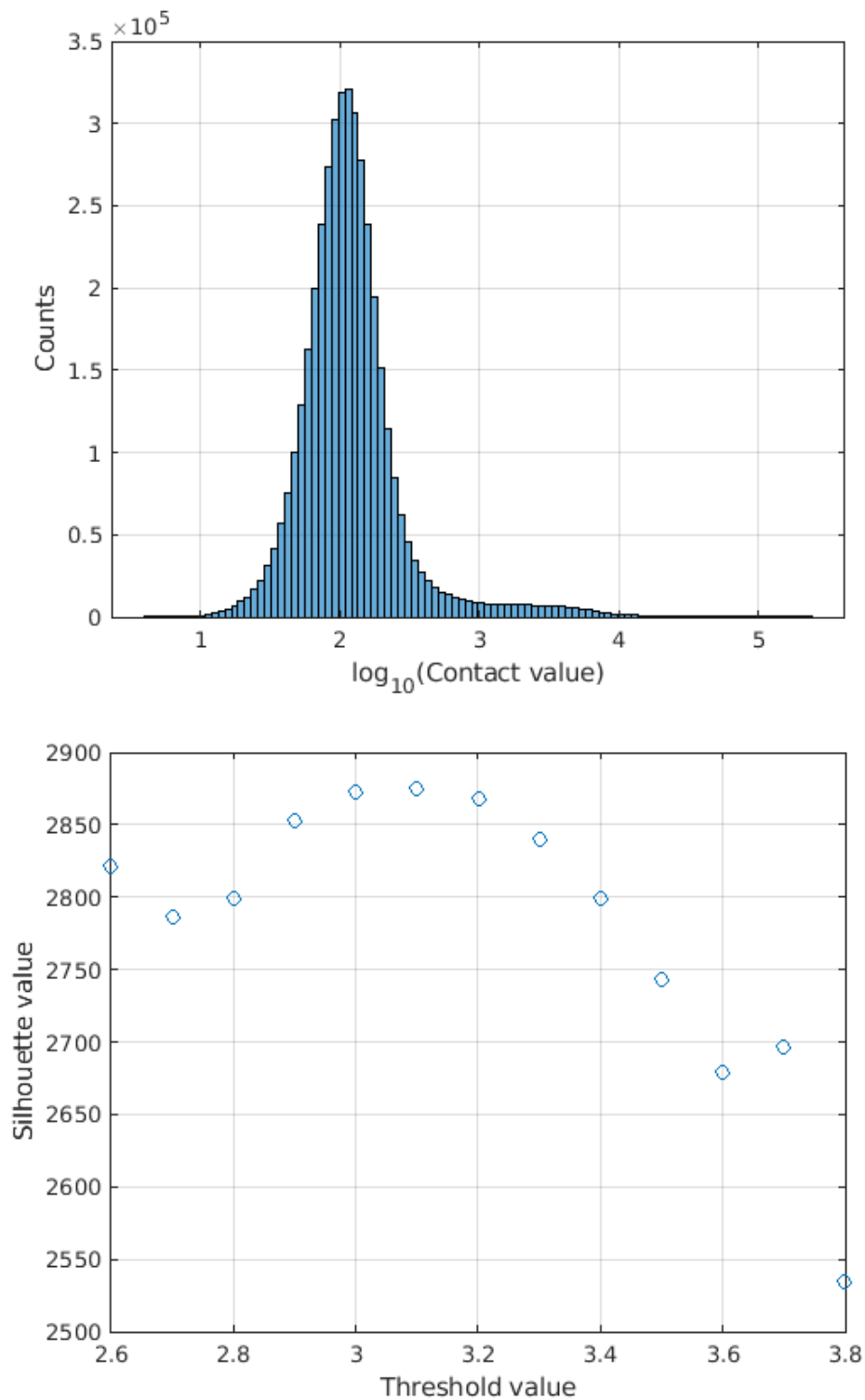


Figure 3.14: Histogram of the logarithm of contact values for the Hi-C contact map of GM12878 cell line, together with silhouette value as function of the threshold. The threshold value chosen as optimal is 3.1, since it corresponds to the maximum silhouette value.



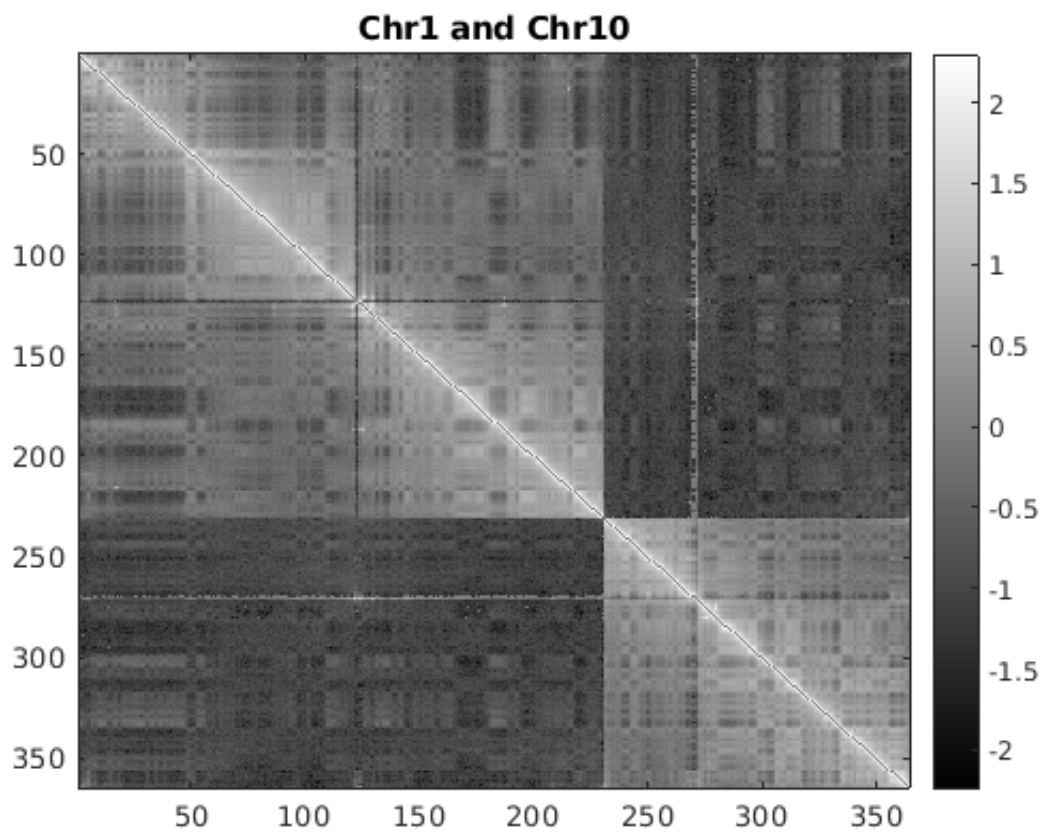


Figure 3.15: *Hi-C* contact map of chromosome 1 (first block on the left) and chromosome 10 (second block on the right). The algorithm identifies a translocation of the 41<sup>st</sup> bin of chromosome 10 onto chromosome 1. This finding is confirmed by the pattern observed in the contact map.

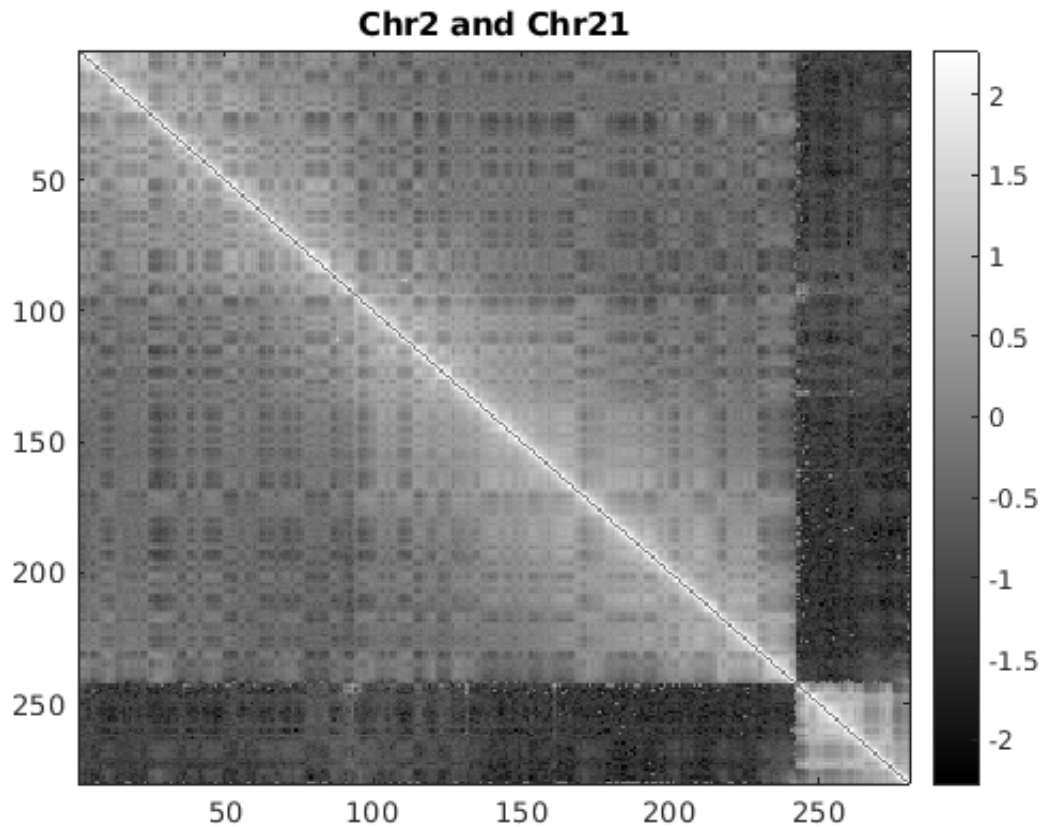


Figure 3.16: *Hi-C contact map of chromosome 2 (first block on the left) and chromosome 21 (second block on the right). The algorithm identifies a translocation of the 1<sup>st</sup>, the 2<sup>nd</sup> and the 3<sup>rd</sup> bin of chromosome 21 onto chromosome 2. The contact map shows a brighter block of contacts between the first 3 bins of chromosome 21 and a limited number of bins of chromosome 2, indicating that the translocation would not be present in the 100%, or at least the majority, of the cell population; or that the observed signal could be associated to an insertion rather than a translocation.*

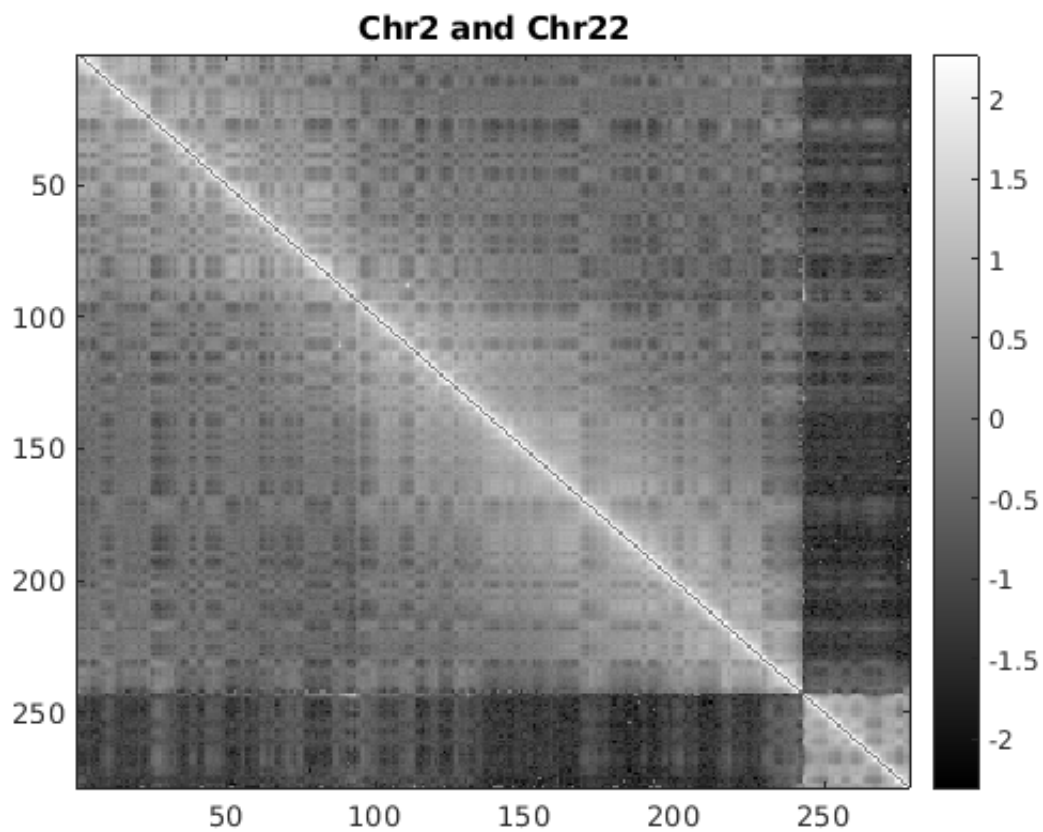


Figure 3.17: *Hi-C* contact map of chromosome 2 (first block on the left) and chromosome 22 (second block on the right). The algorithm identifies a translocation of the 1<sup>st</sup> bin of chromosome 22 onto chromosome 2. This finding is confirmed by the pattern observed in the contact map.

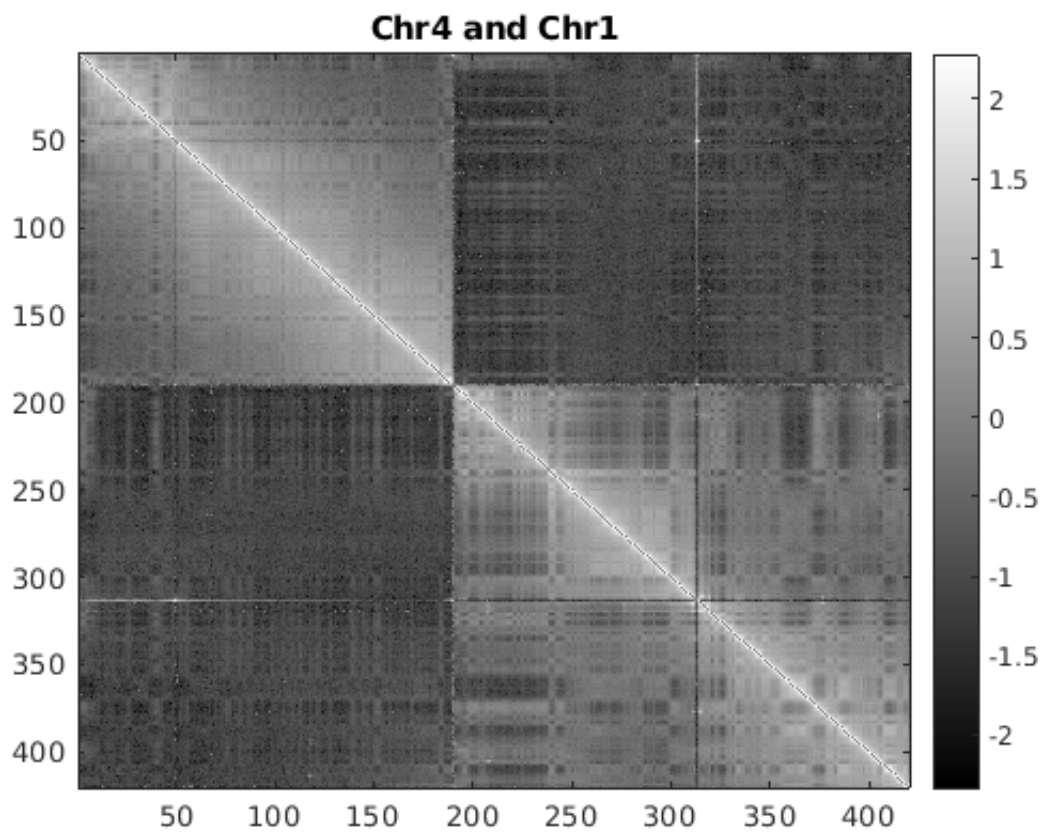


Figure 3.18: *Hi-C contact map of chromosome 4 (first block on the left) and chromosome 1 (second block on the right). The algorithm identifies a translocation of the 123<sup>rd</sup> bin of chromosome 1 onto chromosome 4. This finding is confirmed by the pattern observed in the contact map.*

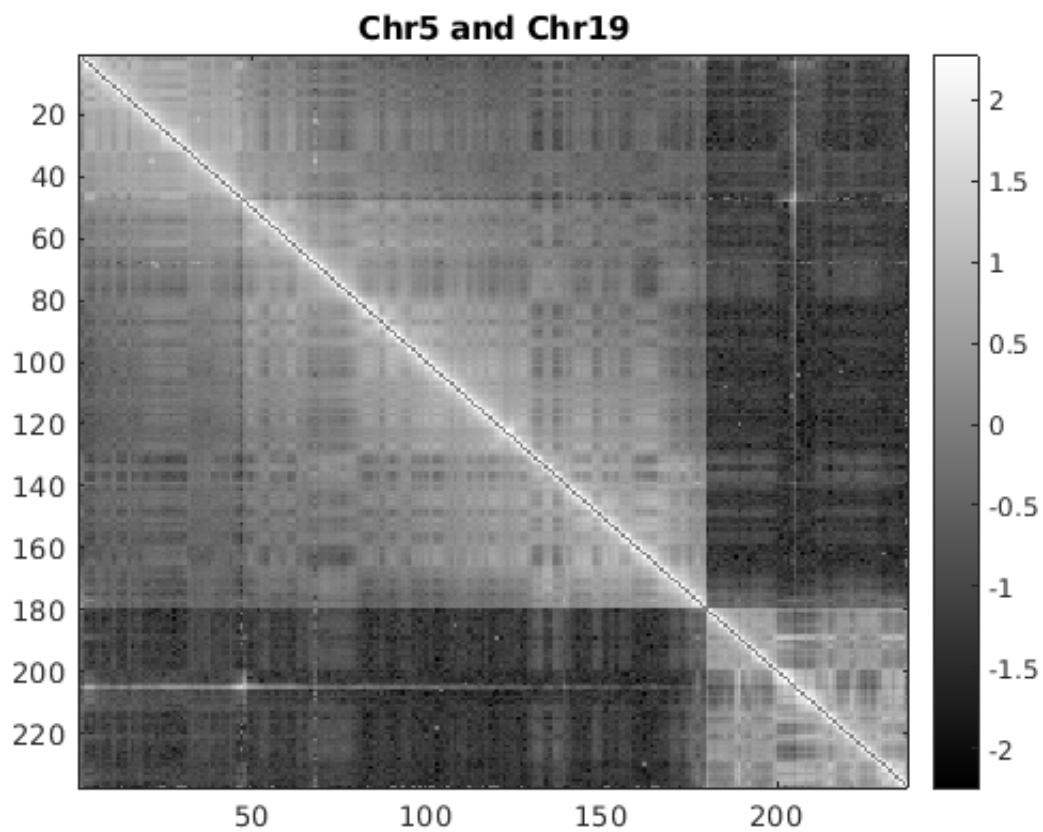


Figure 3.19: *Hi-C* contact map of chromosome 5 (first block on the left) and chromosome 19 (second block on the right). The algorithm identifies a translocation of the 26<sup>th</sup> bin of chromosome 19 onto chromosome 5. This finding is confirmed by the pattern observed in the contact map.

## 3.6 Sequence reconstruction through Fiedler vector

Once the clusters corresponding to the different chromosomes have been identified, we can go further and get the order to be assigned to each bin (i.e. point) within a cluster. This order will allow to reconstruct the sequence for each chromosome. To do this, we will work on single clusters, independently, according to the following steps:

1. Identify an optimal threshold  $(t_i)_{i=1,\dots,k} \geq T$  for each of the  $k$  cluster, in order to preserve only the links with the highest possible weights, without partitioning the network into disconnected components. This was achieved by accepting as a product of the thresholding procedure, only networks composed by one connected component and a set of isolated nodes.
2. Apply the thresholds  $t_i$  to the Hi-C contact maps corresponding to the different clusters, by removing all the links with weight lower than  $t_i$ .
3. Compute the Laplacian matrix and the corresponding Fiedler vector.
4. Order the bins so that the Fiedler vector's components are sorted in ascending order.

### 3.6.1 Results on a toy-model

The optimal thresholds  $t_i$  identified for each of the four chromosomes are all the same, with a value equal to 1.3. As we can see from Fig. 3.20, in an ideal case, where the intra-chromosomal contacts follow a perfect  $1/d$  decay, the Fiedler vector's components show a monotonic trend when represented as a function of bin index along the reference sequence. Therefore, it will constitute the landmark trend to get a good assembly when dealing with more complex case studies, such as real Hi-C contact maps, where inversions, deletions, insertions, duplications or translocations may occur. Of course, in this case, sorting Fiedler vector's components lead to a perfect match between the original Hi-C map and the assembled one, as confirmed by Fig. 3.21.

### 3.6.2 Results on a real Hi-C contact map

The optimal thresholds  $t_i$  identified for each of the 23 clusters range between a minimum value of 3.9 for chromosome 2 and a maximum value of 4.8 for chromosome 21 (see Table 3.4), producing only 1 isolated node belonging to chromosome 6 that will be classified as unplaced in terms of sequence. As we can see from Fig. 3.22, the Fiedler vector's components show, in general, a monotonic trend when represented as a function of bin index along the reference sequence, with some exceptions concerning chromosome 1, 2, 4, 5, 6 and 19. Furthermore, the points corresponding to the potentially translocated bins

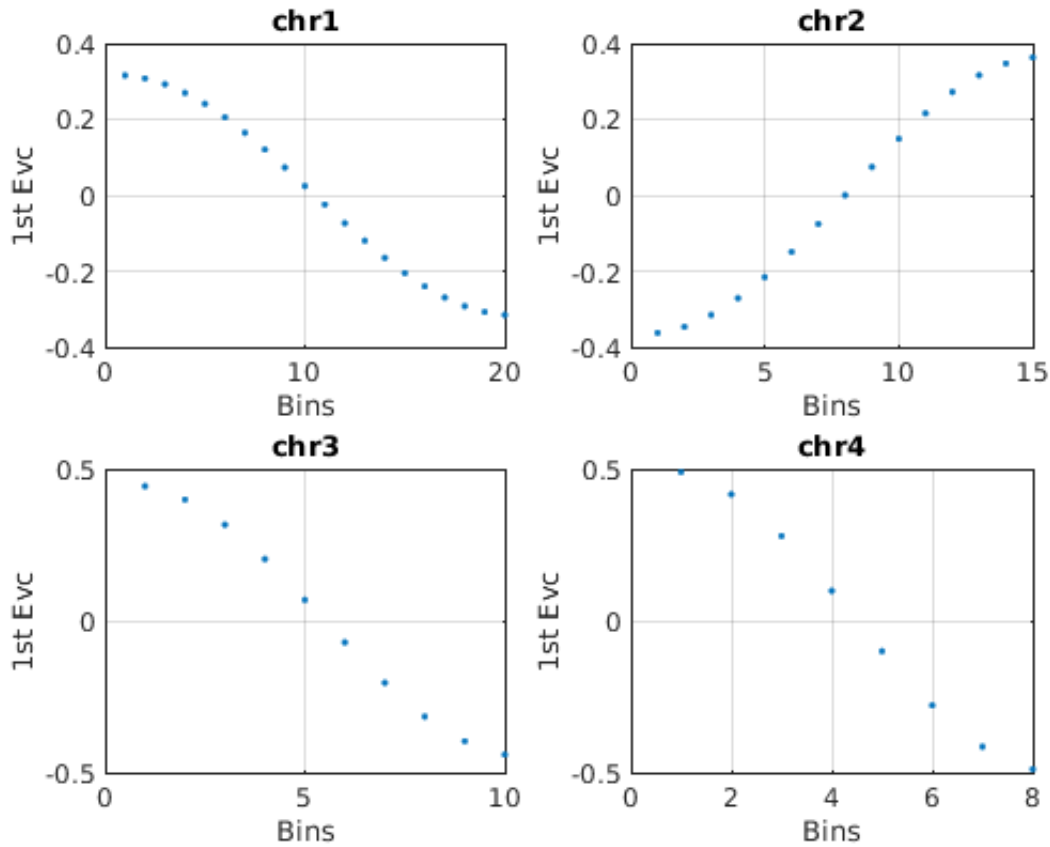


Figure 3.20: *Fiedler vector's components (indicated as 1st Evc) for each of the identified clusters of the toy-model Hi-C map as a function of bin index along the reference sequence.*

are clearly visible: the bin of chromosome 10 associated to chromosome 1 is represented by the last point on the right in the 'chr1' plot; the 3 bins of chromosome 21 and the bin of chromosome 22, associated to chromosome 2 are represented by the last points on the right in the 'chr2' plot; the bin of chromosome 1 associated to chromosome 4 is represented by the first point on the left in the 'chr4' plot; and the bin of chromosome 19 associated to chromosome 5 is represented by the last point on the right in the 'chr5' plot. An interesting and unexpected trend was observed for Fiedler vector's components of the cluster corresponding to chromosome 4, where a sharp disruption of the monotonicity divides the two arms. This peculiar trend led to an incorrect sequence reconstruction (see Fig. 3.23), as shown by the Pearson's correlation coefficient in Table 3.5 and by the reordered Hi-C contact map represented in Fig. 3.24. From a visual inspection of the contact map (see Fig. 3.25), it seems as if the last bin of chr4 was more much connected to the first one rather than its nearest neighbor.

Cluster	$t$	Cluster	$t$
Chr1	4.2	Chr13	4.7
Chr2	3.9	Chr14	4.5
Chr3	4.0	Chr15	4.5
Chr4	4.1	Chr16	4.4
Chr5	4.1	Chr17	4.3
Chr6	3.9	Chr18	4.2
Chr7	4.5	Chr19	4.3
Chr8	4.5	Chr20	4.3
Chr9	4.6	Chr21	4.8
Chr10	4.2	Chr22	4.6
Chr11	4.7	ChrX	4.7
Chr12	4.5		

Table 3.4: *Thresholds obtained for each of the 23 clusters identified for Hi-C contact map of GM12878 cell line. Cluster names were chosen according to the name of the chromosome that is mainly present within the cluster.*

Chr	$\rho$	Chr	$\rho$
1	1.00	13	1.00
2	0.95	14	-0.99
3	-0.98	15	1.00
4	-0.24	16	1.00
5	0.98	17	1.00
6	1.00	18	-0.97
7	-0.99	19	-0.89
8	-0.99	20	1.00
9	-0.97	21	1.00
10	1.00	22	-1.00
11	-1.00	X	1.00
12	1.00		

Table 3.5: *Pearson's correlation coefficient between Fiedler vectors's components before and after sequence reconstruction.*



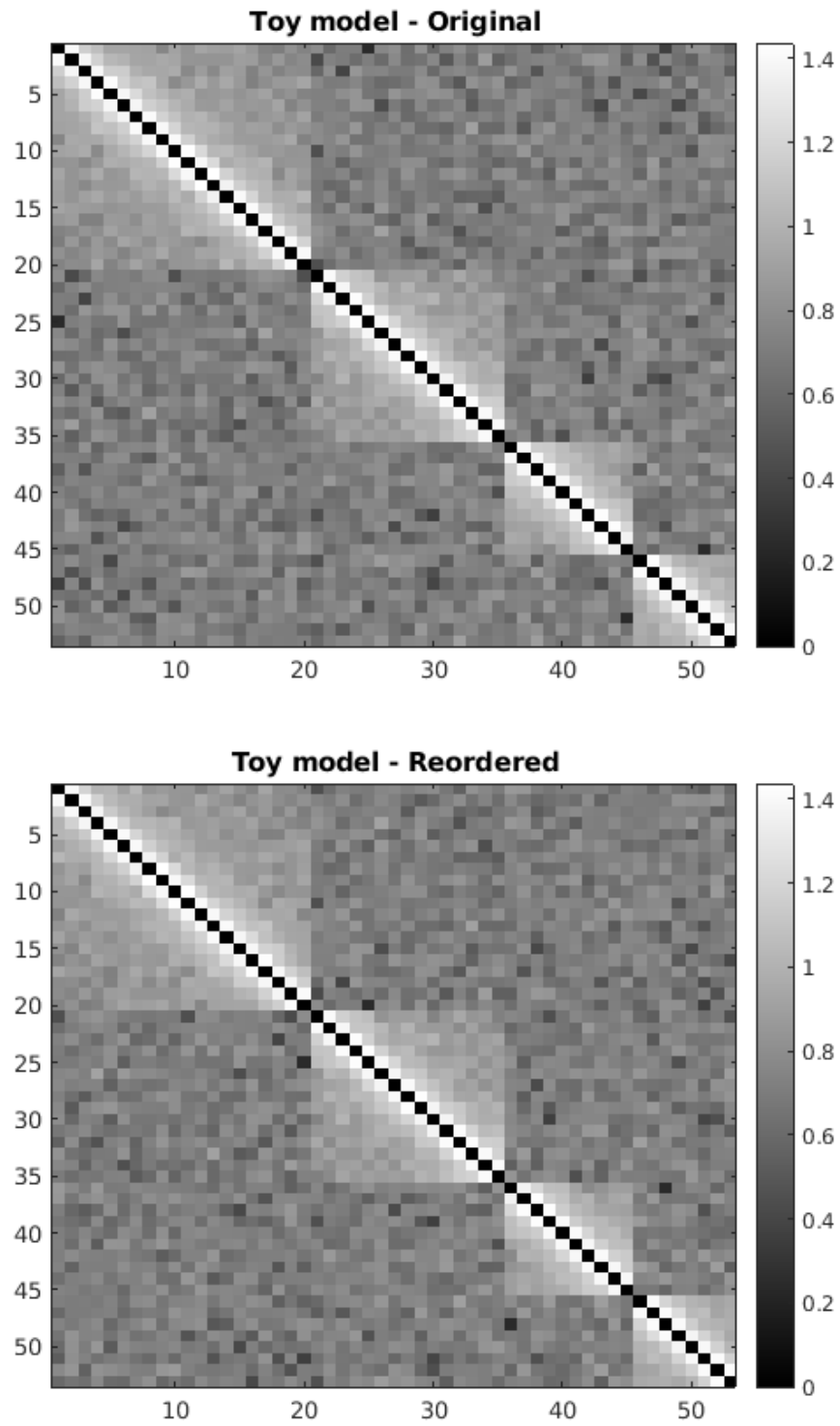


Figure 3.21: Comparison between the original toy-model Hi-C contact map and the one obtained after sequence reconstruction.

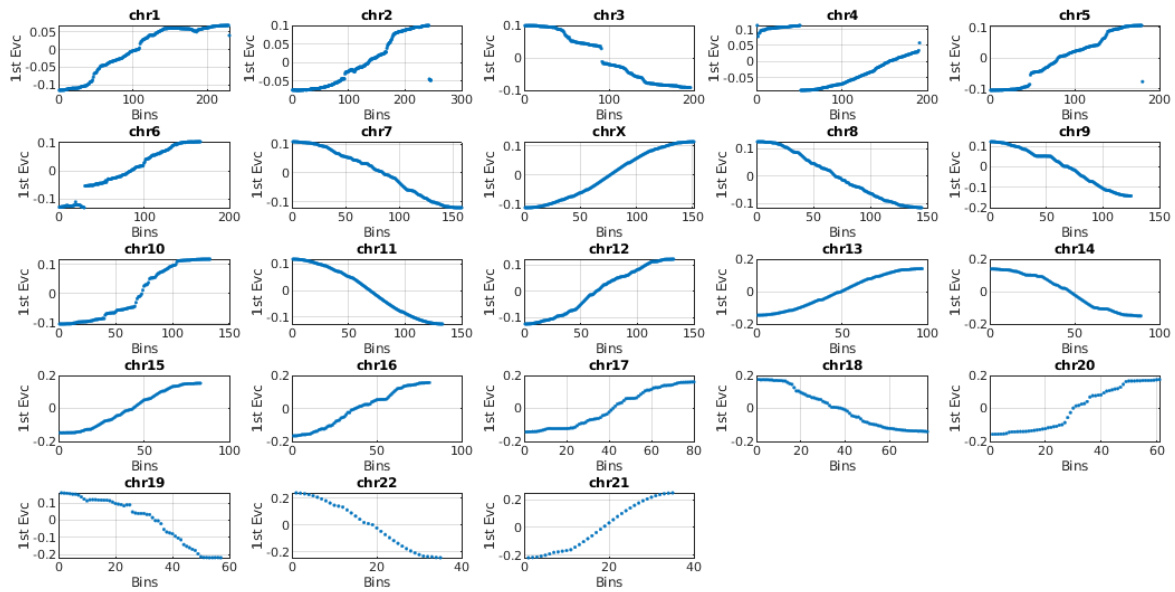


Figure 3.22: *Fiedler vector's components of all the 23 clusters identified for Hi-C contact map of GM12878 cell line as a function of bin index along the reference sequence.*

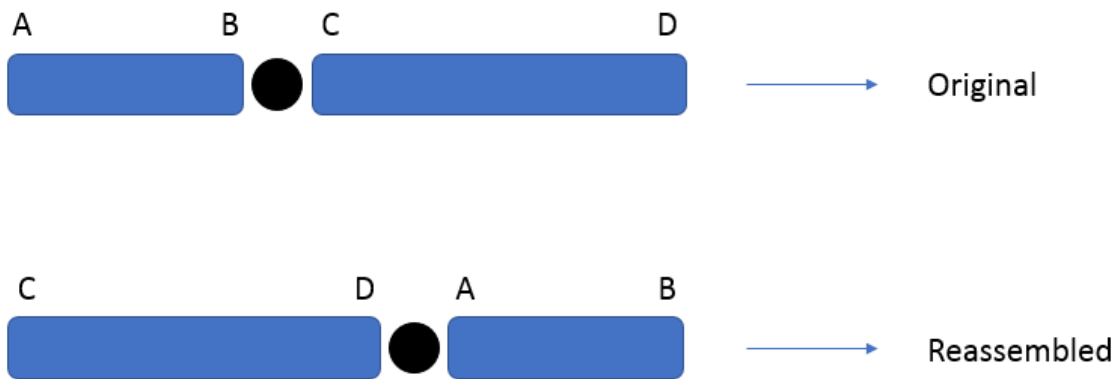


Figure 3.23: *Schematic representation of the original structure of chromosome 4 and the one obtained after sorting Fiedler vector's components.*

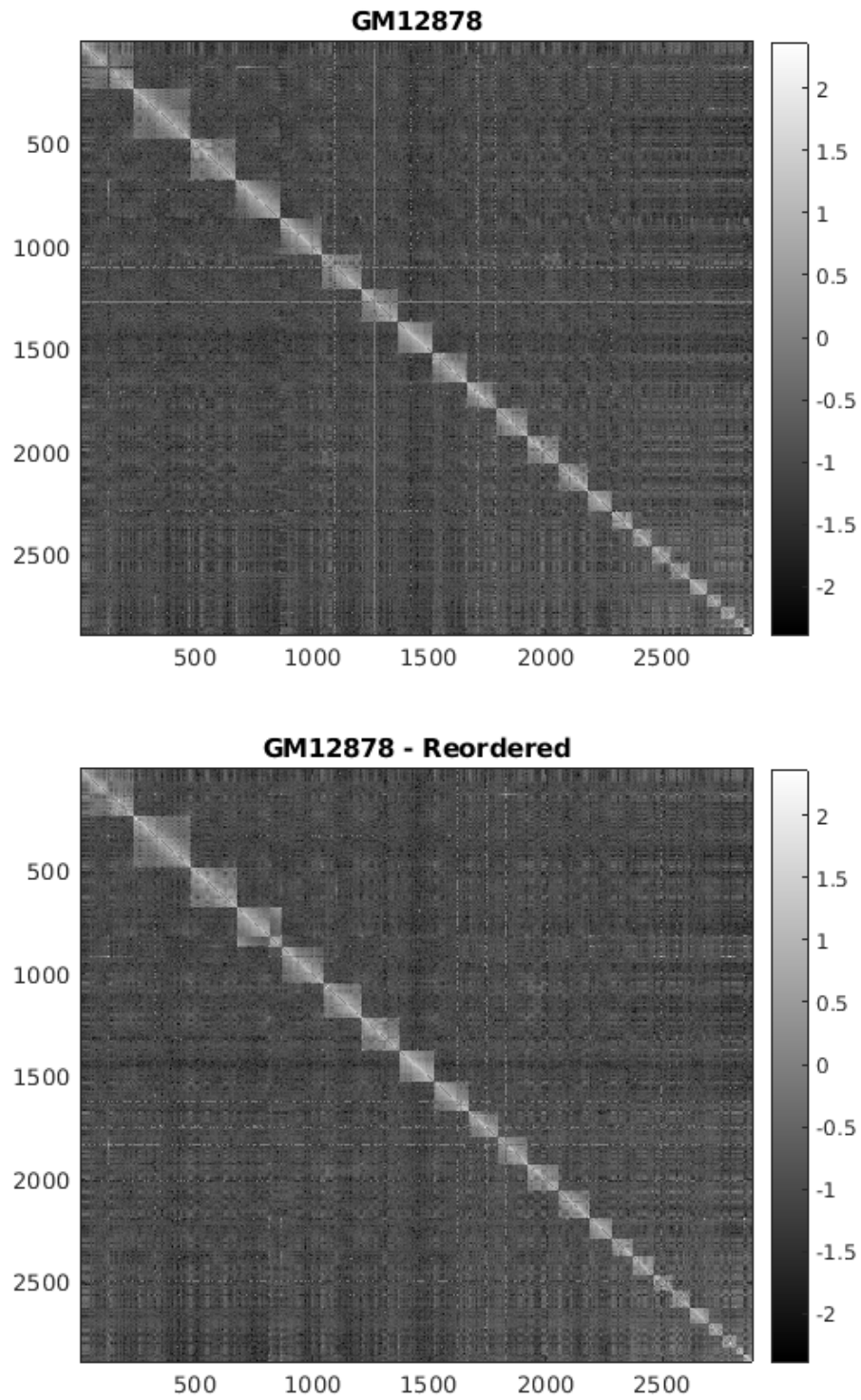


Figure 3.24: Comparison between the original GM12878 cell line Hi-C contact map and the one obtained after sequence reconstruction.

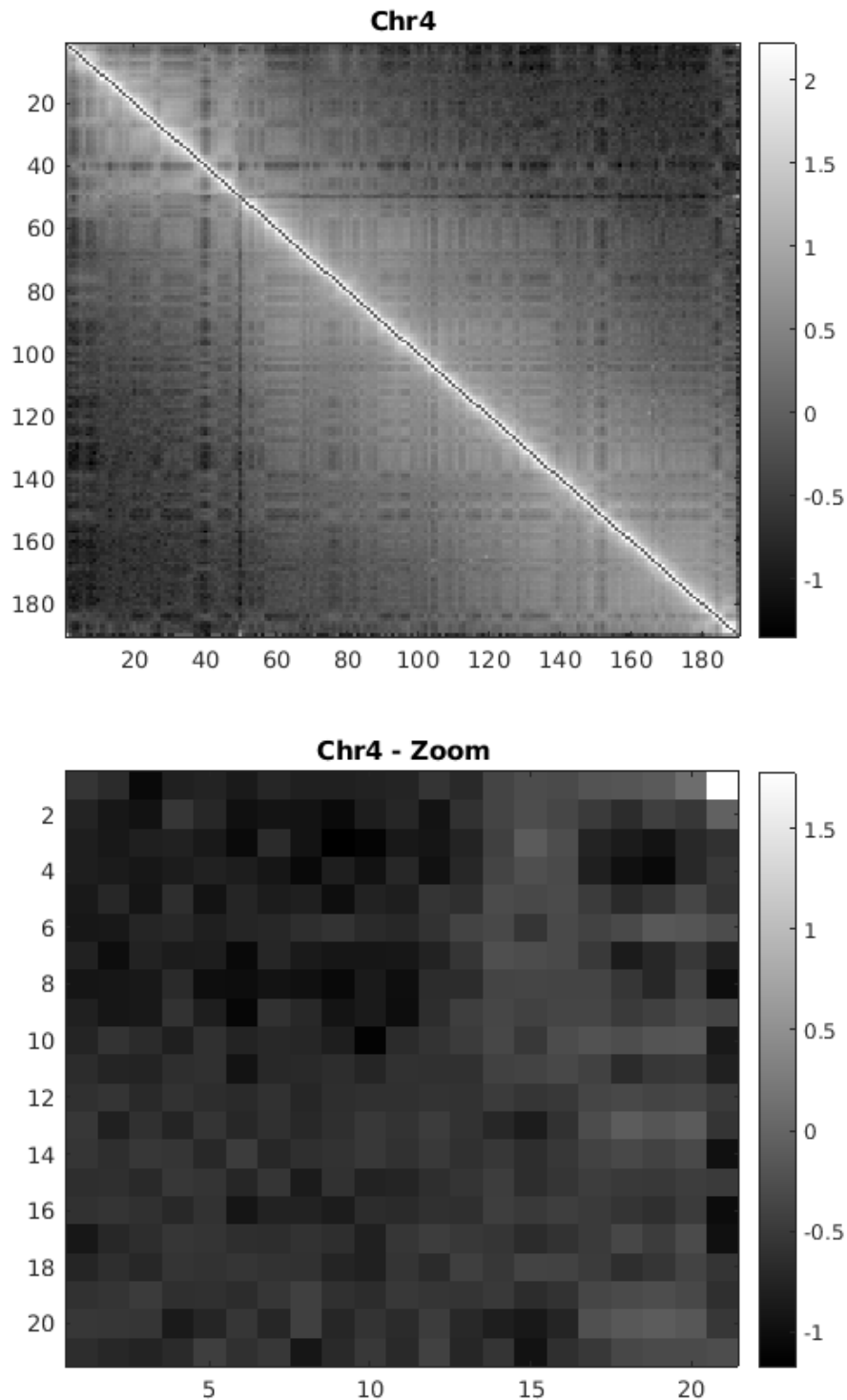


Figure 3.25: *Hi-C* contact map of chromosome 4. A bright point on the top-right corner of the matrix indicates that the last bin is much more connected to the first rather than its nearest neighbors along the sequence, as confirmed by the zoom. This is why the corresponding Fiedler vector's components led the algorithm to produce the structure shown in Fig. 3.23.

## 3.7 Conclusions

A Hi-C contact map can always be represented as an undirected weighted network, given its symmetric structure and its positive contact values. In this way, all the properties of network theory can be exploited to understand the relationships between nodes (i.e. bins). In particular, the proposed algorithm showed that the properties of the Laplacian matrix of a network allow to map the DNA spatial proximity into a DNA sequence proximity, thus obtaining a genome assembly. The two main steps composing the proposed algorithm are: (1) spectral clustering, that allows to identify the different chromosomes and their potential rearrangements; (2) the reordering of the Fiedler vector's components in ascending order, that allows to get an assembly where the closest bins along the sequence are also the closest in the space.

From a computational point of view, the algorithm takes  $\sim 54$  seconds to get the assembly, starting from a whole genome Hi-C contact map at 1Mb resolution, thus being quite fast if compared to the latest tools for reference-guided scaffolding, such as RaGOO [91], which requires  $\sim 12$  min for scaffolding a human draft assembly. The main limitation consists in the resolution at which the algorithm works, which is determined by the resolution of a Hi-C contact map, whose maximum value is currently around 1kb [76].

## CHAPTER 4

---

### From sequence to host-bacteria interactions

---

In this chapter we will see a novel method for source clustering and source attribution, based on a network approach, that exploits the high discriminatory power of whole-genome sequencing data. In fact, the problem of attributing the correct source (e. g. food origin) to bacteria found in human infections is an enduring challenge, allowing fast identification of possible cures or the start of an outbreak. We will see how the single genetic profiles of the infecting agent can be exploited to improve the source attribution task. The major results shown in this chapter are published in [27].

#### 4.1 The source attribution problem

The operation of attributing a human case of foodborne disease to the putative source of infection is called source attribution [24, 25]. Many methods have been developed to estimate the relative contribution of different food sources to human foodborne diseases worldwide, including microbial subtyping, comparative exposure assessment, epidemiological analysis of sporadic cases, analysis of data from outbreak investigations, and expert elicitation [24, 26]. Each of these approaches has strengths and limitations, and the usefulness of each depends on the public health questions being addressed [26]. Recently, phenotyping and molecular data are increasingly replaced by genomic data with high discriminatory power, such as that required to distinguish strains of a monomorphic serovar like the *S. typhimurium* monophasic variant [92, 93, 94]. In particular, in the following study, three different types of whole-genome sequencing (WGS) data were used: single-nucleotide polymorphisms (SNP), core-genome multi-locus sequence typing (cgMLST) and whole-genome multi-locus sequence typing (wgMLST), detecting respect-

ively all the variants of a single letter along the sequence, the variants within a subset of common genes and the variants within all genes. These measures were then used to calculate pairwise distances between the different genomes, isolated from human and animal samples, as shown in Fig. 4.1, thus obtaining three different distance matrices, namely SNP distance matrix, cgMLST distance matrix and wgMLST distance matrix.

<b>Genome A</b>	ATT <b>C</b> AGTA		<b>A</b>	<b>B</b>	<b>C</b>
<b>Genome B</b>	AT <b>G</b> CAG <b>T</b> C	<b>A</b>	0	2	3
<b>Genome C</b>	AT <b>G</b> CA <b>A</b> T <b>C</b>	<b>B</b>	2	0	1
		<b>C</b>	3	1	0

Figure 4.1: *Toy-model example showing the process that lead to the computation of pairwise distance between genomes.*

Usually, source attribution studies are conducted by using frequency-matching models like the Dutch and Danish models based on phenotyping data (serotyping, phage-typing, and antimicrobial resistance profiling) [24, 25]. In this chapter, a different approach to source attribution based on the theory of weighed networks will be shown and evaluated. There are many examples of network modeling applications in different fields, such as computer and information sciences, social sciences, and biology [95]. In biomedical fields, networks are powerful tools to perform characterization, classification, and ranking of interacting elements in a complex biological system [96]. Specifically, for source attribution, pairwise distance matrices can be interpreted as fully connected networks where nodes correspond to bacterial isolates and links to a function of genetic distances (i.e., number of different nucleotides along DNA sequences): the weaker the link, the higher the genetic distance within two isolates. The aim is to extract disconnected components corresponding to different animal sources and subsequently see how the human isolates bind to the different clusters. The probability of a human isolate to be associated with a specific animal source is computed as a function of the number of links the human isolate has with other isolates of specific animal sources. The network approach is useful also in investigating which structural features of a data set play a fundamental role in determining the internal coherence of clusters. Apart from animal sources, in fact, the country of origin of imported food samples and the year of collection might impact the clustering formation.

### 4.1.1 The case of *Salmonella enterica* serovar Typhimurium and its monophasic variant

*Salmonella enterica* subspecies *enterica* serovar Typhimurium and its monophasic variant (STm) are among the top three serovars in confirmed human cases of salmonellosis each year in Europe [97]. Although in 2017, in comparison to the previous year, the percentage of confirmed human cases associated with the monophasic variant was similar, specific concern arose in recent years due to the emergence of outbreaks worldwide since 2006 [98, 99, 100, 101, 102, 103, 104]. Therefore, attributing cases of Salmonellosis to specific sources is crucial to identify and prioritize targeted interventions in the food chain, as well as to evaluate the effectiveness of each intervention.

In general, *Salmonella typhimurium* can infect humans from different sources; however, in 2017 the most reported matrices were broiler, pig, turkey, and layers in decreasing order for *S. typhimurium*, and pig and broilers for its monophasic variant accounting for 49.7 and 35.3%, respectively [97].

## 4.2 A Network-based method to identify host-bacteria interactions

The symmetric structure and the positive numbers constituting the entries of these distance matrices, make them representable as *undirected weighted networks*, where nodes correspond to isolates, and links correspond to a function of pairwise distance  $d_{ij}$ , calculated as the number of different nucleotides or number of different alleles between two isolate DNA sequences  $i$  and  $j$ . In order to find an association between distances and animal sources, the following assumption was made: genomes coming from the same source should show smaller distance values. Therefore, a fully connected weighted network  $W$  (in which the weight  $w_{ij} = 1/d_{ij}$  was assigned to each link between samples  $i$  and  $j$ ) was built. Subsequently, a threshold was applied to the constructed weighted matrices in order to remove weaker links (associated with larger genetic distances). In the resulting binarized network, nodes were linked by an edge only if their weight was greater than a given threshold value, and source clusters were identified as the disconnected components (i.e., groups of nodes with links within each other but not with the nodes of other components) obtained by the thresholding procedure. The threshold value was chosen in order to maximize internal coherence of clusters and minimize the number of isolated nodes. Precisely, the best threshold value  $t$  was found through a 70/30 cross-validation procedure applied on animal source data, aiming to maximize the following score function on distance matrices:

$$score = \left(1 - \frac{N_{ISO}}{N_{TOT}}\right) CSC \quad (4.1)$$



where  $N_{TOT}$  represents the total number of nodes in the network,  $N_{ISO}$  represents the number of isolated nodes (i.e., not forming any link with other nodes), and  $CSC$  represents the coherent source clustering, the parameter that estimates algorithm clustering performance, computed as follows:

$$CSC = \frac{\sum TP_i}{\sum T_i} 100 \quad (4.2)$$

where  $TP_i$  represents the number of true positives inside the  $i$ -th cluster (i.e., the isolates from the same source found in the same cluster) and  $T_i$  the total number of nodes inside the  $i$ -th cluster. Specifically, the 70/30 cross-validation procedure is structured as follows: 70% of the animal origin samples were randomly selected in order to build a network (training set) on which the best threshold value  $t$  was computed by maximizing the score function and then applied to the network constructed with the remaining 30% samples (test set), in which the clustering performance was evaluated. This procedure was repeated 100 times, with different random 70/30 data set subdivisions, and the most frequent value was selected as the global best threshold  $t$  for source clustering. It was then applied to the distance matrix obtained from the whole data set, including both human and animal *Typhimurium* genomes, so that human samples could be attributed to a putative source according to the following rule:

$$\max_j (l_j/L) \quad j = 1, \dots, N \quad (4.3)$$

where  $n$  represents the number of different animal origin sources,  $l_j$  represents the number of links between a single human isolate  $h$  and all nodes belonging to the  $j$ -th animal source and  $L$  represents the number of all neighbors of animal origin of the human sample. The ratio  $l_j/L$  can be considered as the best estimate of the probability that a human isolate  $h$  is attributed to the  $j$ -th animal origin source, given the available dataset. Graphical representations of networks (Figures 4.3, 4.4, 4.5, 4.6) were generated using MathWorks Matlab *plot* function with a force-directed graph layout [105].

### 4.2.1 Data set

The method was applied to a data set comprising 141 human and 210 food and animal isolates of pig, broiler, layer, duck, and cattle collected in Denmark from 2013 to 2014 (see Table 4.1) [106]. Another important stratification of data is provided by serotype, since, as explained in section 4.1.1, the percentage of confirmed cases associated with the monophasic variant is increasing over the years, showing peculiar characteristics in terms of genomic variability [107]. The number of isolates belonging to *S. Typhimurium* serotype and its monophasic variant are respectively 108 and 102 for animal origin samples, and 73 and 68 for human origin samples (see Table 4.2). Another important distinction is determined by the geographical origin of the analyzed samples. In the considered data

set, all human isolates were from Denmark, while food isolates were from Denmark as well as imported to Denmark from four different countries: Germany, Ireland, United Kingdom, and others (see Table 4.3).

	<b>2013</b>	<b>2014</b>	<b>Total</b>
Broilers	13	21	34
Pigs	104	55	159
Ducks	0	11	11
Cattle	1	1	2
Layers	3	1	4
Human	29	112	141
<b>Total</b>	150	201	351

Table 4.1: *Data set composition according to primary source and sampling year.*

	<b>Monophasic</b>	<b>Typhimurium</b>	<b>Total</b>
Broilers	16	18	34
Pigs	84	75	159
Ducks	1	10	11
Cattle	1	1	2
Layers	0	4	4
Human	68	73	141
<b>Total</b>	170	181	351

Table 4.2: *Data set composition according to primary source and serotype.*

## 4.2.2 Source clustering results

The best threshold values, obtained by the cross-validation procedure, were 412, 24.7, and 32.79 for SNP, cgMLST, and wgMLST matrices respectively, since they maximized the score function and corresponded to the most probable values obtained from 100 cross-validation runs (Figure 4.2). In particular, the structure of the disjoint connected components shown in Figures 4.3, 4.4, 4.5, 4.6 could be achieved by considering only pairwise genomic distances lower than these threshold values. The method reaches 90% of coherent source clustering for animal source on SNP and wgMLST matrices and 89% on the cgMLST matrix, showing that animal source type is the main factor driving cluster formation, followed by country of origin, serotype, and sampling year (Table 4.4). Although the overall algorithm performance is good, broilers and cattle represent the most difficult sources to detect: 18 out of 34 among the former as well as 1 out of 2 of the latter are classified as pigs (see 4.5 for SNP and wgMLST distance matrices and Table 4.6

	Denmark	Germany	Others	Ireland	UK	Total
Broilers	34	0	0	0	0	34
Pigs	125	32	0	1	1	159
Ducks	0	0	11	0	0	11
Cattle	1	0	1	0	0	2
Layers	4	0	0	0	0	4
Human	141	0	0	0	0	141
<b>Total</b>	305	32	12	1	1	351

Table 4.3: *Data set composition according to primary source and country of sample origin.*

for cgMLST distance matrix), being included in the same network component. Figures 4.2 and 4.3, on the left-hand sides, show that most of the confusion between broilers and pigs arises from cluster 1, which is mainly composed of isolates of monophasic variant: their peculiar low variability at the genomic level could be the reason for encountered difficulties in distinguishing the two different sources [107]. In terms of cluster structure, most of the subnetworks are composed of the same type of animal source except for cluster 1, where pig and broiler isolates are mixed together with one of the cattle samples (Figure 4.3, left panel). Regarding country of origin of imported food samples, regionality affects cluster formation since most of import isolates tend to group apart from those from Denmark, as confirmed by cluster 2, mainly composed of pig isolates from Germany and by clusters 3, 4, and 5, mainly composed of import ducks and cattle isolates (Figure 4.3, right panel). Another relevant parameter for cluster formation is serotype, as confirmed by subnetwork composition, since a clear separation between genomes of Typhimurium and its monophasic variant was observed. In particular, genomes of *S. typhimurium* monophasic variant cluster all together in subnetwork 1, whereas *S. typhimurium* showed a more heterogeneous behavior, especially for pig isolates, which appeared stratified in more than one group (Figure 4.4, left panel). Finally, sampling year had no impact on cluster formation, as confirmed by the high variability in terms of cluster composition (Figure 4.4, right panel).

### 4.2.3 Source attribution results

Adding human isolate genomes to the network (Fig. 4.6), a percentage between 93.6 and 97.2% clustered with the existing animal network components, and only a percentage between 2.8% (from cgMLST distance matrix) and 6.4% (from SNPs distance matrix) appeared as not linked to any animal Typhimurium genome (Table 4.7). The majority of attributable human genomes were associated with pigs with probabilities ranging from 83.9 (SNP matrix) to 84.5% (cgMLST and wgMLST matrices), followed by broilers,

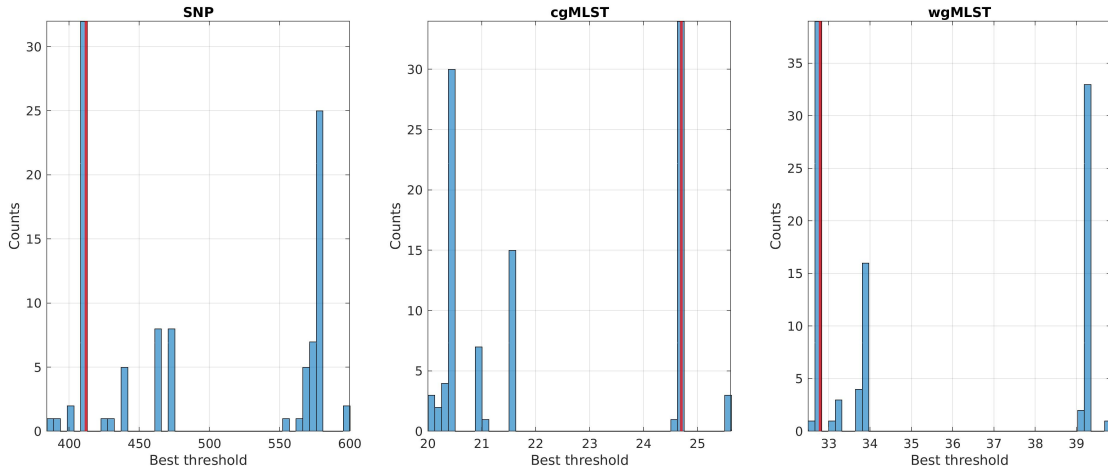


Figure 4.2: *Best threshold values obtained by the 70/30 cross-validation procedure on the training sets of SNP, cgMLST, and wgMLST distance matrices. The red line corresponds to the global best threshold value used for source clustering (Figures 4.3, 4.4, 4.5) and source attribution (Figure 4.6)*

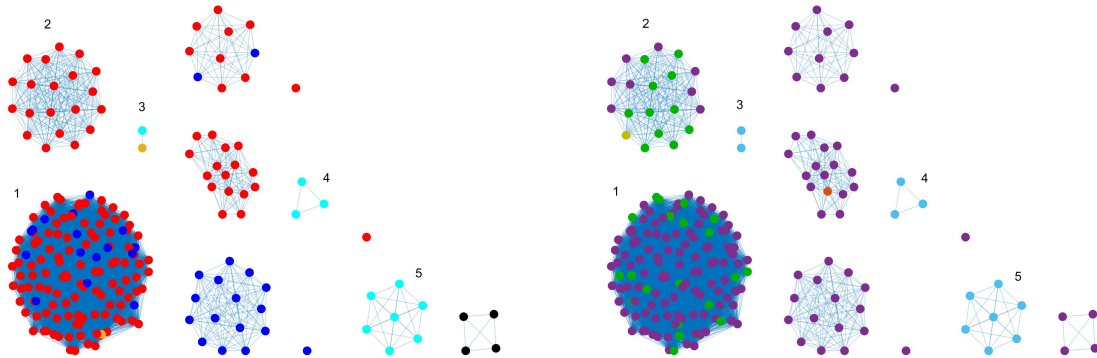


Figure 4.3: *Clustering results (force-directed graph drawing algorithm) obtained by SNP distance matrix, where different node colors represent different **animal sources** (left) and different **countries of origin** (right). Legend for left-hand figure: pigs, red; broilers, blue; cattle, yellow; ducks, cyan; layers, black. Legend for right-hand figure: purple, Denmark; green, Germany; light blue, others; orange, United Kingdom; yellow, Ireland.*

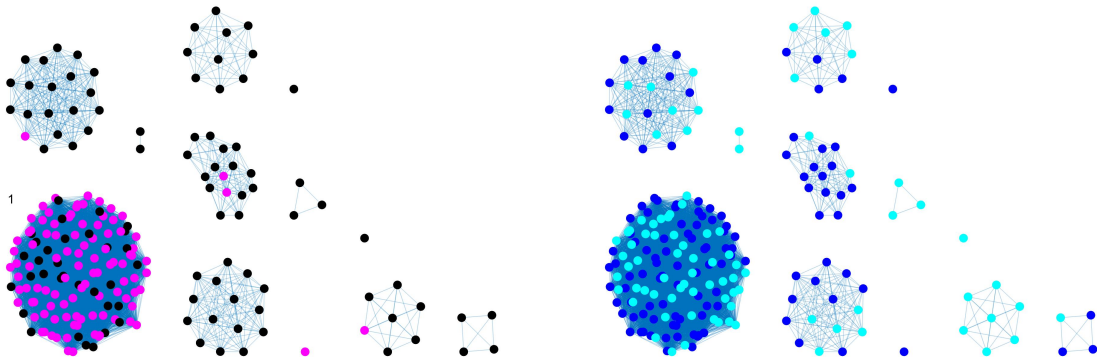


Figure 4.4: *Clustering results (force-directed graph drawing algorithm) obtained by SNP distance matrix, where different node colors represent different **serotypes** (left) and different **sampling years** (right). Legend for left-hand figure: pink, monophasic; black, Typhimurium. Legend for right-hand figure: 2013, blue; 2014, cyan*

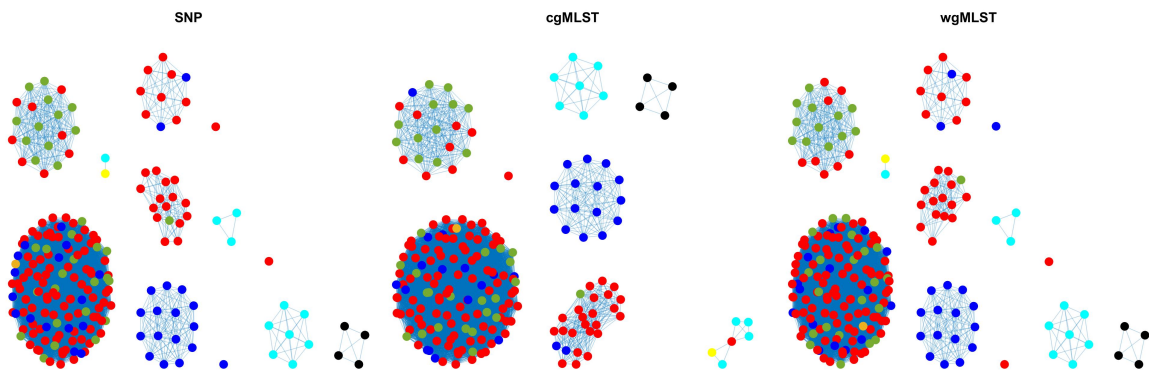


Figure 4.5: **Source clustering** results (force-directed graph drawing algorithm) obtained by SNP, cgMLST, and wgMLST distance matrices. Legend: pigs from Denmark, red; import pigs, green; broilers, blue; cattle from Denmark, darker yellow; import cattle, lighter yellow; ducks, cyan; layers, black.

	SNP	cgMLST	wgMLST
Animal source	90%	89%	90%
Country of origin	85%	84%	85%
Serotype	82%	82%	82%
Sampling year	65%	63%	65%

Table 4.4: *Coherent source clustering (CSC) for SNP, cgMLST, and wgMLST distance matrices, computed on animal origin isolates, according to the following parameters: animal source, serotype, country of origin, and sampling year.*

		PREDICTED				
		Broilers	Cattle	Ducks	Layers	Pigs
TRUE	Broilers	16	0	0	0	18
	Cattle	0	0	1	0	1
	Ducks	0	0	10	0	0
	Layers	0	0	0	4	0
	Pigs	0	0	0	0	159

Table 4.5: *Confusion matrix obtained from source clustering results on **SNP** and **wgMLST** distance matrices.*

ducks, cattle, and layers in descending order (Table 4.8). We remark that even if the data set presents a large abundance of pig and broiler samples, we also found human isolates with 100% links toward less abundant animal sources, such as layers and ducks, reflecting the fact that our analysis does not seem heavily affected by such source representation imbalance.

Moreover, if we further stratify pigs by distinguishing between import and non-import (Table 4.9), we can notice that most of the human isolates are associated with non-import pigs, with probabilities ranging from 66.4 (SNP matrix) to 66.7% (cgMLST and wgMLST matrices). The same stratification can be applied also to cattle, but the probability that a human isolate is associated with an import and non-import cattle is almost the same due to the very low number of genomes. Furthermore, links between human samples and animal sources showed high specificity in all the three networks (SNP, cgMLST, and wgMLST) as confirmed in Figures 4.7, 4.8, making the putative originating animal source clearly attributable.

Pigs were by the far most frequent animal source to which human genomes of *S. typhimurium* and its monophasic variant were attributed in this study. This result is not surprising. Excluding traveling, other authors have been describing pigs as the main source of human Salmonella infections in Denmark as well as in Southern Europe for two decades with estimated percentages ranging from 15% (Denmark) to 44% (Italy) [108, 26]. The higher values in the present study are linked to the dataset which exclus-

		PREDICTED				
		Broilers	Cattle	Ducks	Layers	Pigs
TRUE	Broilers	15	0	0	0	19
	Cattle	0	0	1	0	1
	Ducks	0	0	10	0	0
	Layers	0	0	0	4	0
	Pigs	0	0	1	0	158

Table 4.6: *Confusion matrix obtained from source clustering results on cgMLST distance matrix.*

ively includes serovar *S. typhimurium* and its monophasic variant, historically associated with pig reservoir. In particular, in the period 2013–2014 *S. typhimurium* was the most frequently detected serovar in pigs and pig meat in Europe [109, 110].

In principle, the discriminatory power of the subtyping method is of crucial importance in source attribution studies. In fact, the high discriminatory power of the genomic subtyping used to produce the analyzed distance matrices, might lead to unjustified differentiation with the identification of too many clusters and a higher number of not attributable human isolates [111, 26]. Although highly discriminatory, the three genomic subtyping data sets used in the present study as input data (SNP calling, cgMLST, and wgMLST) showed that this was not the case, revealing a good discriminatory power that led to maximize cluster coherence and minimize the number of human isolated nodes corresponding to not attributable human genomes. Besides discriminatory power, the output of the network analysis revealed that, although wgMLST generally offers higher resolution than cgMLST, source attribution results did not differ significantly, demonstrating the robustness of the approach.

One of the major concerns on all source attribution approaches is the estimate percentages of human infections of unknown sources or not attributable human infections. With the network approach, less than 7% of human genomes were not attributed to any animal source. This value is lower than those previously reported for other microbial subtyping methods for source attribution [26]. However, along with the model approach, the dataset itself might strongly influence this estimate especially in case the dataset does not fully represent the real temporal and spatial distribution of human and animal subtypes/isolates.

Human isolates	SNP	cgMLST	wgMLST
Attributed	93.6%	97.2%	95.0%
Not attributed	6.4%	2.8%	5.0%

Table 4.7: *Percentage of attributed and not attributed human isolates.*

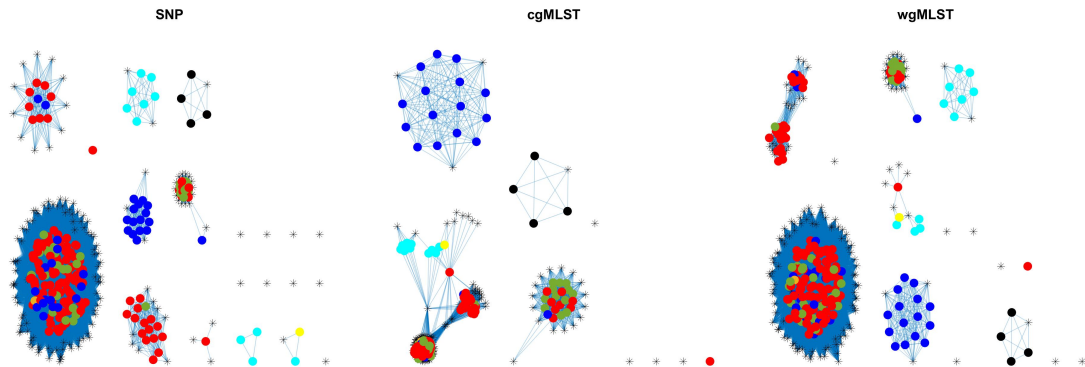


Figure 4.6: **Source attribution** results (force-directed graph drawing algorithm) obtained by SNP, cgMLST, and wgMLST distance matrices. Legend: pigs from Denmark, red; import pigs, green; broilers, blue; cattle from Denmark, darker yellow; import cattle, lighter yellow; ducks, cyan; layers, black; humans, asterisk.

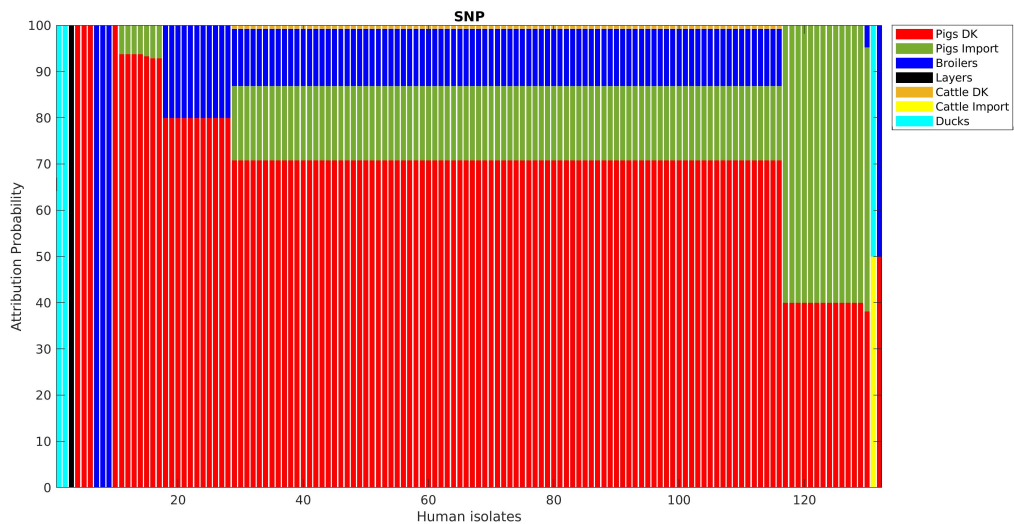


Figure 4.7: *Human isolate probability to originate from each source as determined by source attribution analysis via the network-based approach on the SNP pairwise distance matrix.*



<b>Source</b>	<b>SNP</b>	<b>cgMLST</b>	<b>wgMLST</b>
Broilers	12.5 (10.0–15.1)	11.8 (9.4–14.2)	11.7 (9.2–14.1)
Cattle	0.9 (0.1–1.6)	1.2 (0.3–2.0)	0.9 (0.3–1.4)
Ducks	1.9 (0–4.1)	1.8 (0–3.7)	2.2 (0–4.6)
Layers	0.7 (0–2.2)	0.7 (0–2.2)	0.7 (0–2.2)
Pigs	83.9 (80.3–87.5)	84.5 (81.3–87.8)	84.5 (80.1–88.0)

Table 4.8: Mean probability (expressed in percentage) of a human isolate to be attributed to a source, together with 95% confidence intervals, calculated for each of the considered pairwise distance matrices (SNP, cgMLST, and wgMLST).

<b>Source</b>	<b>SNP</b>	<b>cgMLST</b>	<b>wgMLST</b>
Broilers	12.5 (10.0–15.1)	11.8 (9.4–14.1)	11.7 (9.2–14.1)
Cattle	0.5 (0.4–0.6)	0.6 (0.5–0.6)	0.5 (0.4–0.6)
Cattle import	0.4 (0–1.1)	0.6 (0–1.5)	0.4 (0–0.9)
Ducks	1.9 (0–4.1)	1.7 (0–3.7)	2.2 (0–4.6)
Layers	0.7 (0–2.2)	0.7 (0–2.2)	0.7 (0–2.2)
Pigs	66.4 (62.9–69.9)	66.7 (63.3–70.1)	66.7 (63.2–70.2)
Pig import	17.5 (14.7–20.2)	17.9 (15.3–20.4)	17.7 (15.0–20.5)

Table 4.9: Mean probability (expressed in percentage) of a human isolate to be attributed to a source (taking into account the stratification of pigs and cattle into import and non-import), together with 95% confidence intervals, calculated for each of the considered pairwise distance matrices (SNP, cgMLST, and wgMLST).

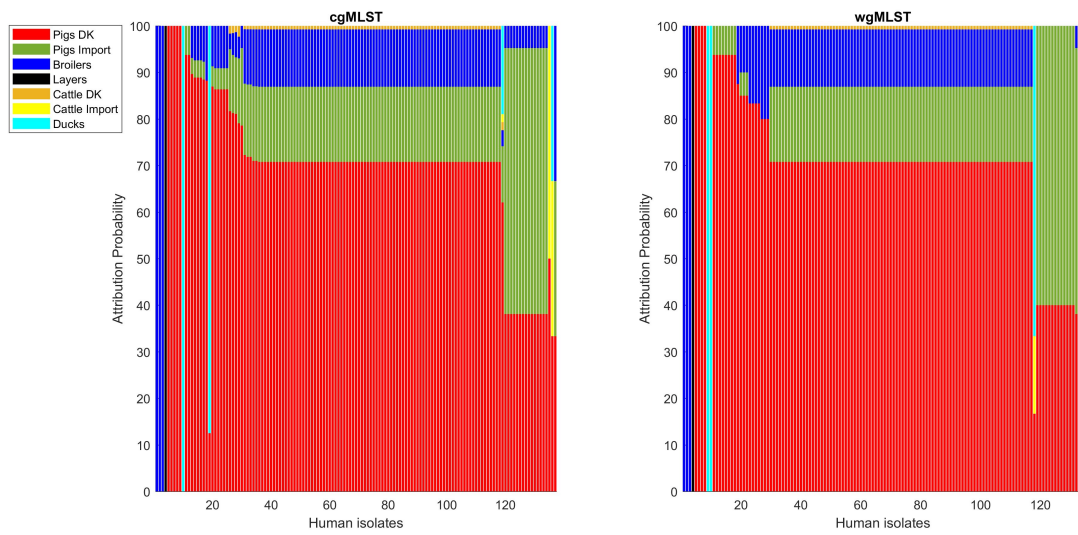


Figure 4.8: *Human isolate probability to originate from each source on the cgMLST pairwise distance matrix (left) and on the wgMLST pairwise distance matrix (right).*

### 4.3 Validation of the method

In order to validate the method, SNP, cgMLST and wgMLST distance matrices were calculated on an independent data set, with the same predominant animal sources characterizing the Danish data set: pigs, broilers and layers. The thresholds identified on the previous data set were then used to evaluate the algorithm’s performance in terms of source clustering and source attribution. In particular, in order to assess the strength of the results obtained on the training set, it is fundamental that source clustering procedure provides highly homogeneous clusters (i.e. high coherence source clustering values), meaning that these threshold values actually allow to distinguish the considered animal sources, and that the percentage of attributed isolates is comparable.

#### 4.3.1 Data set

The validation data set is composed by isolates coming from Germany [106], comprising 161 human and 121 food and animal samples of pig, broiler, layer, bird, and game, collected in 2014, 2015, and 2016. Originally the data set included also a significant number of cattle that were excluded from the validation, since the algorithm was trained on a data set that was cattle poor (only 2 isolates out of 210). Moreover, the data set can be further stratified according to serotype into 57 and 72 isolates belonging respectively to animal and human origin samples of *S. Typhimurium*, and into 64 and 89 isolates belonging respectively to animal and human origin samples of its monophasic variant. In this case, no regionality effect will impact source clustering, since the geographical origin of all the isolates was the same.

	2014	2015	2016	Total
Birds	0	0	1	1
Broilers	5	4	1	10
Game	1	0	0	1
Layers	5	20	12	37
Pigs	25	26	21	72
Human	43	49	69	161
<b>Total</b>	79	99	104	282

Table 4.10: *Data set composition according to primary source and sampling year.*

#### 4.3.2 Results

After applying the same threshold values obtained by training the algorithm on the Danish data set, we got a good clustering performance, even though not as high as in

	<b>Monophasic</b>	<b>Typhimurium</b>	<b>Total</b>
Birds	0	1	1
Broilers	6	4	10
Game	1	0	1
Layers	10	27	37
Pigs	47	25	72
Human	89	72	161
<b>Total</b>	153	129	282

Table 4.11: *Data set composition according to primary source and serotype.*

the previous case (see Table 4.12 and Table 4.4), reaching the 82% of internal cluster coherence. The high concordance between the two data sets is also reflected in the distribution distance values among animal origin samples, as shown by Fig. 4.12 and 4.13. In fact, both distributions are characterized by a major peak at low distances (centered around 50 for SNP matrix and around 11 and 13 for cgMLST and wgMLST respectively) and a set of lower peaks at higher distances (centered around 700 for SNP matrix and around 35 and 50 for cgMLST and wgMLST respectively), with the threshold values always positioned right before the latter.

As shown by the results in Table 4.12 and by the confusion matrix in Table 4.13, it is confirmed that there is no difference among the three types of distance matrices, and that sampling year does not affect cluster formation (see Fig. 4.11). The only difference that emerges from the comparison between Table 4.12 and Table 4.4 is related to the coherence source clustering values computed according to serotype, which are slightly higher than those computed according to animal source for the validation data set. This might be due to the fact that serotype has a strong impact on cluster formation and that the training part of the algorithm is fundamental to identify the best solution for source clustering and subsequently for source attribution purposes.

As happened for Danish data set, also in this case there is a clear separation between *S. Typhimurium* isolates and those corresponding to its monophasic variant, which mainly populate two clusters: a bigger one on the bottom-left corner of the panels in Fig. 4.10 and a smaller one on the top-left corner of the panels in Fig. 4.10.

Unlike the results achieved regarding broilers in the training data set, in this case, none of them form a cluster where they represent the predominant source (see Table 4.13 and Fig. 4.9). Furthermore, also game and bird isolates were not correctly grouped in terms of clusters, but this is most likely due to their poor statistical representation, since there is only one isolate per class.

In the end, the percentage of attributed human isolates is comparable to that obtained for the training set, showing no great differences among the three matrices (see Table 4.14).

	SNP	cgMLST	wgMLST
Animal source	82%	82%	82%
Serotype	85%	83%	83%
Sampling year	49%	49%	49%

Table 4.12: *Coherent source clustering (CSC) for SNP, cgMLST, and wgMLST distance matrices, computed on animal origin isolates, according to the following parameters: animal source, serotype, and sampling year.*

		PREDICTED				
		Birds	Broilers	Game	Layers	Pigs
TRUE	Birds	0	0	0	1	0
	Broilers	0	0	0	3	7
	Game	0	0	0	0	1
	Layers	0	0	0	32	5
	Pigs	0	0	0	5	67

Table 4.13: *Confusion matrix obtained from source clustering results on the validation SNP, cgMLST, and wgMLST distance matrices.*

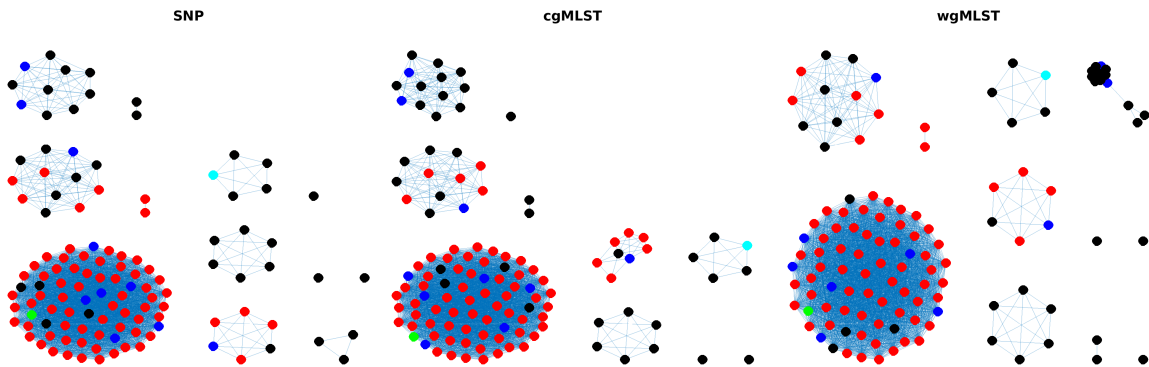


Figure 4.9: **Source clustering** results (force-directed graph drawing algorithm) obtained by applying the threshold values calculated for the Danish data set on SNP, cgMLST and wgMSLT distance matrices chosen as validation data set. Nodes are colored according to the different **animal sources**: pigs, red; broilers, blue; layers, black; birds, cyan; game, green.

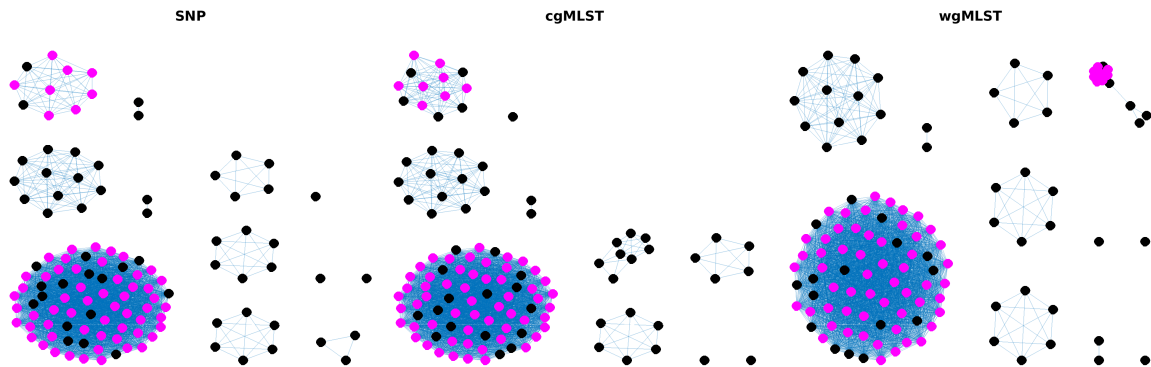


Figure 4.10: *Source clustering* results (force-directed graph drawing algorithm) obtained by applying the threshold values calculated for the Danish data set on SNP, cgMLST and wgMSLT distance matrices chosen as validation data set. Nodes are colored according to *serotypes*: black for *Typhiumrium* isolates and pink for Monophasic ones.

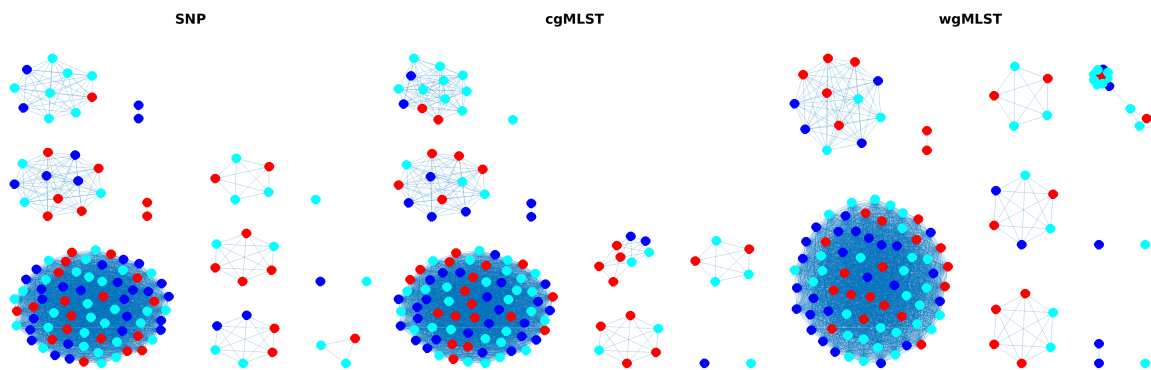


Figure 4.11: *Source clustering* results (force-directed graph drawing algorithm) obtained by applying the threshold values calculated for the Danish data set on SNP, cgMLST and wgMSLT distance matrices chosen as validation data set. Nodes are colored according to *sampling years*: 2014, blue; 2015, cyan ; 2016, red.

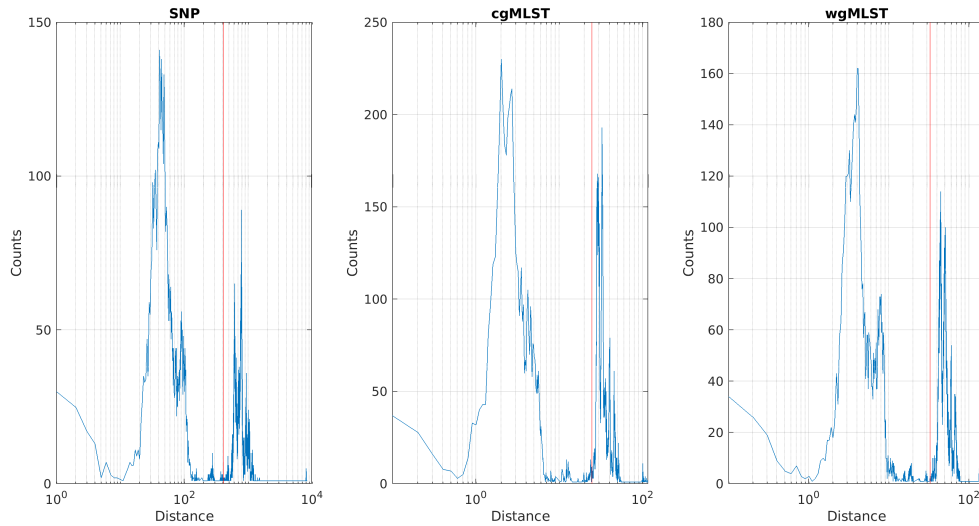


Figure 4.12: Comparison between distance value distributions of SNP, cgMLST and wgMLST validation matrices, comprising only animal source isolates. The red lines indicate the thresholds applied for source clustering and obtained by training the algorithm on the Danish data set.

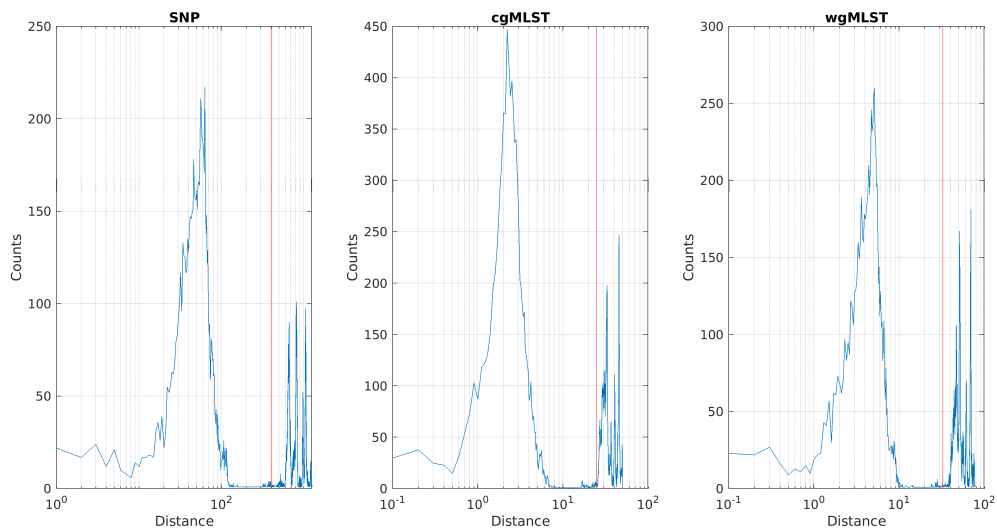


Figure 4.13: Comparison between distance value distributions of Danish SNP, cgMLST and wgMLST matrices, comprising only animal source isolates. The red lines indicate the thresholds applied for source clustering.

<b>Human isolates</b>	<b>SNP</b>	<b>cgMLST</b>	<b>wgMLST</b>
Attributed	96.3%	97.5%	97.5%
Not attributed	3.7%	2.5%	2.5%

Table 4.14: *Percentage of attributed and not attributed human isolates on the validation data set.*



## 4.4 Conclusions

This approach allows to extract threshold values that lead to obtain disconnected subnetworks characterized by the highest homogeneity in terms of animal source composition and, more in general, to evaluate the driving force of each parameter (animal source, serotype, country of origin, and sampling year) in cluster formation, by comparing the obtained clusters with the parameter labels distributed on the nodes. In the training data set (i.e. Danish data set), animal source was the major driving force of clustering formation, followed by the country of origin of imported food samples and serotype. The year of isolation did not impact significantly, although it must be underlined that the time period was only 2 years. A higher impact of the year of isolation would have been probably observed in the case of longer time frames (i.e., 10 or more years). Results on the impact of the country of origin are in line with the high relatedness of *S. typhimurium* subtypes in isolates of the same geographic area, recently highlighted also in a study describing a geographical segregated genomic clade of the *S. typhimurium* monophasic variant in Italy [94]. Interestingly, this approach allows also to distinguish between the two serotypes, as shown by the results obtained on both training and validation data sets, highlighting this label as an important parameter to take into account in order to understand the process of cluster formation, since the low variability at genomic level characterizing the monophasic variant might represent a source of confusion for animal source distinction. Lastly, the number of not attributed human isolates is lower than those previously reported for other microbial subtyping methods for source attribution [26], being less than 7% for the Danish data set and less than the 4% for the validation data set. Thus this simple approach, based on the whole network structure of bacterial similarities, at difference with the typical phylogenetic trees that discard a lot of proximity relations, revealed very powerful for the characterization of bacterial strains and their relation to human hosts.

## CHAPTER 5

---

### Conclusions

---

The results shown in the three chapters of this thesis show that the DNA sequence is constantly shaped by the interactions with its environment at multiple levels, revealing footprints of DNA methylation, of its 3D organization, and in the case of bacteria such as *Salmonella enterica* serovar Typhimurium, of the interaction with their host organisms. In the first chapter, we considered several probability density functions to fit CG distance distribution of a selected set of mammal organisms, and we observed that it is best described by a Gamma distribution. Applying this function on a wide set of organisms, taken from different taxonomic categories, we noticed that the scale parameter  $b$  of the Gamma distribution could be associated to the biological complexity of the organism category, increasing from bacteria to vertebrates. Moreover, we tested for possible factors affecting this parameter, like genome sequence length and CG density. While the first was not related to our observations, the second revealed stronger correlations; in particular, for a group of organisms, comprising those of minor biological complexity (bacteria, protozoa, fungi, invertebrates and plants), the relation between  $b$  and CG density could be explained by a minimal null model, while for higher order organisms (vertebrates) this null model did not explain the observations. We argue that this difference could be related to the different role that CG methylation plays in these classes of organisms. Subsequently we focused on human genome and we saw that gamma fit results of CG distance distribution show a systematic deviation at low distance values, therefore we introduced a shifted power-law with exponential tail as new fitting function, obtaining a better agreement by residue analysis. Furthermore, we saw that the power-law trend characterizing low distance values, and the exponential trend characterizing high distance values, can be generated respectively by a positioning process with [47, 53, 54] and without memory [47, 48, 49, 50, 51, 52], with the former process able to introduce the

long-range correlations characterized by the power-law tail in the distributions studied. We hypothesized that the former process might be related to the mechanism of formation and conservation of CpG islands, while the latter might be due to an “erosion” of the original CG positioning (corresponding to a *random percolation process* in physical terms) originated as a consequence of the deamination process occurring at methylated CGs outside CpG islands, which causes a spontaneous mutation  $CG \rightarrow TG$ .

In the end, we saw that CG and TA distance distributions can be well described by the same function, precisely a shifted power-law with an exponential tail, meaning that the same process could have given rise to both. In particular, we saw that the value of the characteristic distance  $b$  between two consecutive TA along the sequence, estimated by the fitting procedure, is much larger than the maximum distance actually considered for the fit, leading us to the conclusion that the exponential tail is negligible and that the actual fitting distribution is a shifted power-law. Furthermore, we saw that the same function well describes also TA distance distributions of *Pan troglodytes* and *Macaca mulatta* (all the primates found in our genome database), providing almost the same values for the parameters  $a, c, d$  and showing exactly the same deviations from the trend estimated by the fitting function: a sharp peak at 91 bp and a Gaussian peak centered around 142 bp.

Focusing again on human genome and investigating the properties of the sequences between consecutive TA separated by a distance of 91 bp (TA91) and 142 bp (TA142), we identified specific patterns associated with the occurrence probability of AA/TT and CG. In particular, we saw that the 40% of TA91 sequences have a good match with SINE (Short Interspersed Nuclear Elements), and, in particular, with Alu sequences, which are mobile non-coding elements involved in epigenetic and structural processes [68]. Furthermore, we observed that the average distance between consecutive AA/TT increases as TA distance increases, as we would expect in a random null model, except for TA distances corresponding to the Gaussian peak centered around 142 bp, where it assumes values close to 10 bp, which corresponds to the typical distance between AA/TT/TA characterizing the 147bp-sequences wrapped around the histone octamer [61, 64, 65]. Therefore, these results suggest that the sequences corresponding to the sharp peak at 91 bp and the Gaussian peak centered around 142 bp have peculiar properties, concerning not only epigenetic and structural processes, but also DNA flexibility, that deserve further investigations.

In the second chapter, we saw that a Hi-C contact map can always be represented as an undirected weighted network, given its symmetric structure and its positive contact values. In this way, all the properties of network theory can be exploited to understand the relationships between nodes (i.e. bins in which DNA sequence has been divided within the Hi-C experiment considered). In particular, the proposed algorithm showed that the properties of the Laplacian matrix of a network allow to map the DNA spatial proximity into a DNA sequence proximity, thus obtaining a genome assembly. The two main steps composing the proposed algorithm are: (1) spectral clustering, that allows to

identify the different chromosomes and their potential rearrangements; (2) the reordering of the Fiedler vector's components in ascending order, that allows to get an assembly where the closest bins along the sequence are also the closest in the space.

From a computational point of view, the algorithm takes  $\sim 54$  seconds to get the assembly, starting from a whole genome Hi-C contact map at 1Mb resolution, thus being quite fast if compared to the latest tools for reference-guided scaffolding, such as RaGOO [91], which requires  $\sim 12$  min for scaffolding a human draft assembly. The main limitation consists in the resolution at which the algorithm works, which is determined by the resolution of a Hi-C contact map, whose maximum value is currently around 1kb [76].

In the third chapter, we designed a novel method for source clustering and source attribution of *Salmonella* bacterial strains: by identifying the optimal threshold for the genome distance matrix, we transformed the initial genome distance matrix into a set of disconnected subnetworks characterized by the highest homogeneity in terms of animal source composition. This approach allowed us to evaluate the role of several factors (animal source, serotype, country of origin, and sampling year) in cluster formation, by comparing the obtained clusters with the parameter labels distributed on the nodes. In particular, we saw that, in the training data set (i.e. a data set of Danish strains), animal source was the major driving force of cluster formation, followed by the country of origin of imported food samples and serotype. The year of isolation did not impact significantly, although it must be underlined that the time period was only 2 years. Results on the impact of the country of origin are in line with the high relatedness of *S. typhimurium* subtypes in isolates of the same geographic area, recently highlighted also in a study describing a geographical segregated genomic clade of the *S. typhimurium* monophasic variant in Italy [94]. Interestingly, this approach allows also to distinguish between the two serotypes, as shown by the results obtained on both training and validation data sets, highlighting this label as an important parameter to take into account in order to understand the process of cluster formation, since the low variability at genomic level characterizing the monophasic variant might represent a source of confusion for animal source distinction. Lastly, the number of not attributed human isolates is lower than those previously reported for other microbial subtyping methods for source attribution [26], being less than 7% for the Danish data set and less than the 4% for the validation data set. Thus this simple approach, based on the whole network structure of bacterial similarities at difference with the typical phylogenetic trees that discard a lot of proximity relations, revealed very powerful for the characterization of bacterial strains and their relation to human hosts, that might find further applications when applied to larger data sets and to different bacterial species, in order to fully grasp the complexity of the bacterial ecosystem.

## CHAPTER 6

---

### Appendix A

---

### **6.1 Comparison among the 16 dinucleotide distance distributions within human genome**

Through the comparison among the fit results of a shifted power-law with exponential tail (SPLE) to the 16 dinucleotide distance distributions within human chromosome 1 represented below, we can clearly see a relationship between CG and TA, since they are the only ones providing good fit results, as confirmed by residual plots.

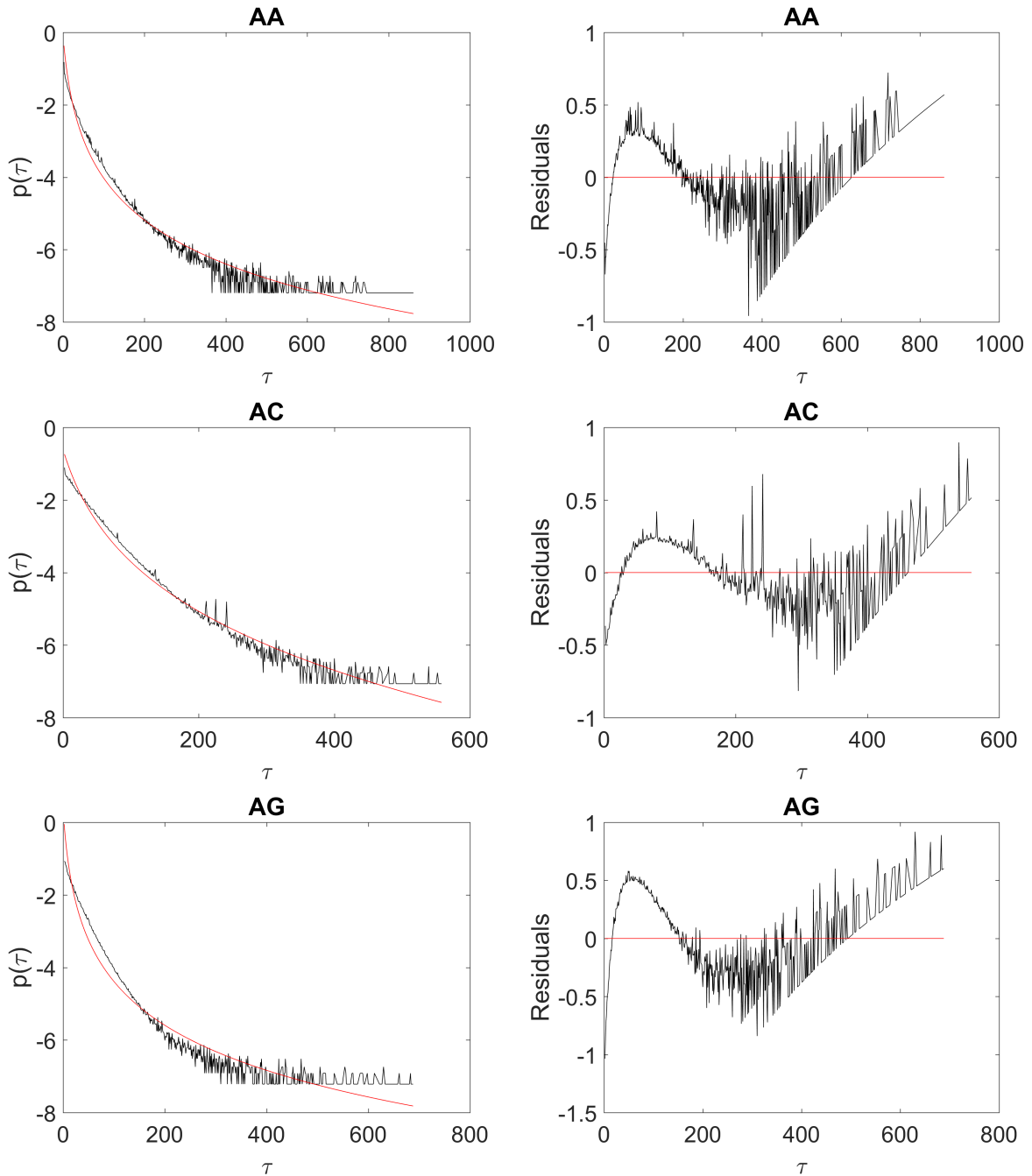


Figure 6.1: *Log-linear plot of AA, AC and AG distance distributions within human chromosome 1, together with shifted power-law with exponential tail as fitting distribution (left panel). Fit results are accompanied by residual plot (right panel).*

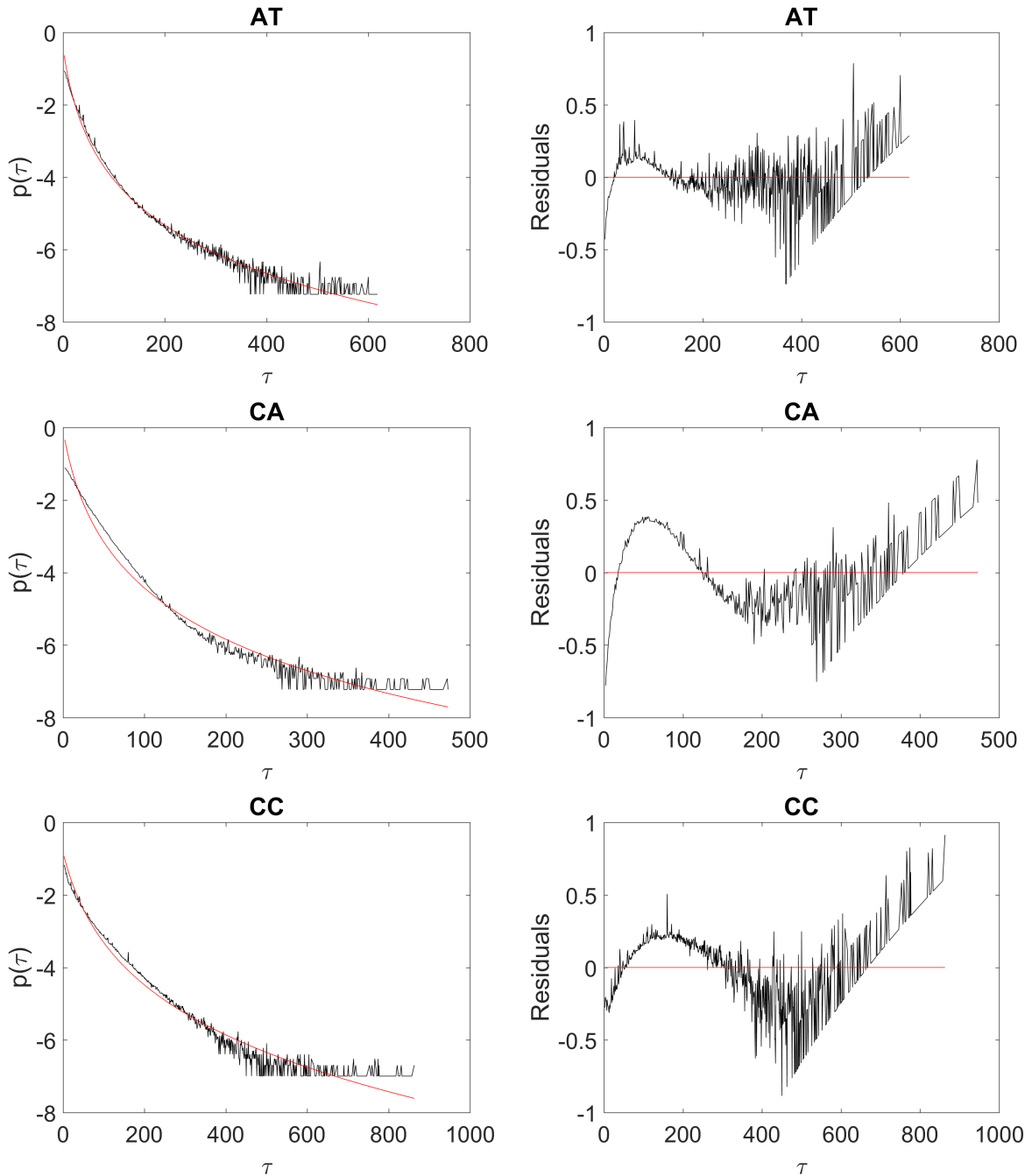


Figure 6.2: *Log-linear plot of AT, CA and CC distance distributions within human chromosome 1, together with shifted power-law with exponential tail as fitting distribution (left panel). Fit results are accompanied by residual plot (right panel).*

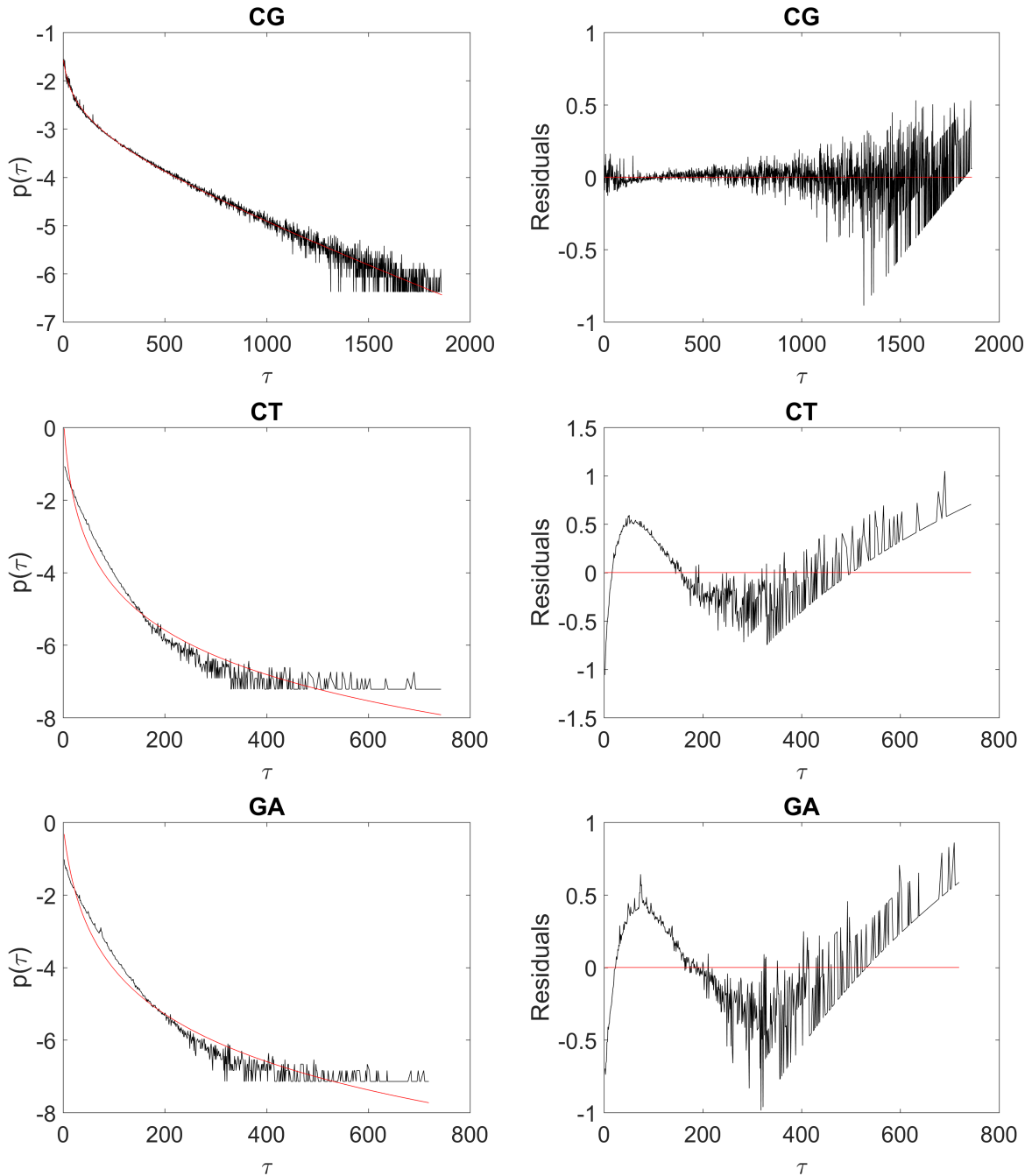


Figure 6.3: *Log-linear plot of CG, CT and GA distance distributions within human chromosome 1, together with shifted power-law with exponential tail as fitting distribution (left panel). Fit results are accompanied by residual plot (right panel).*



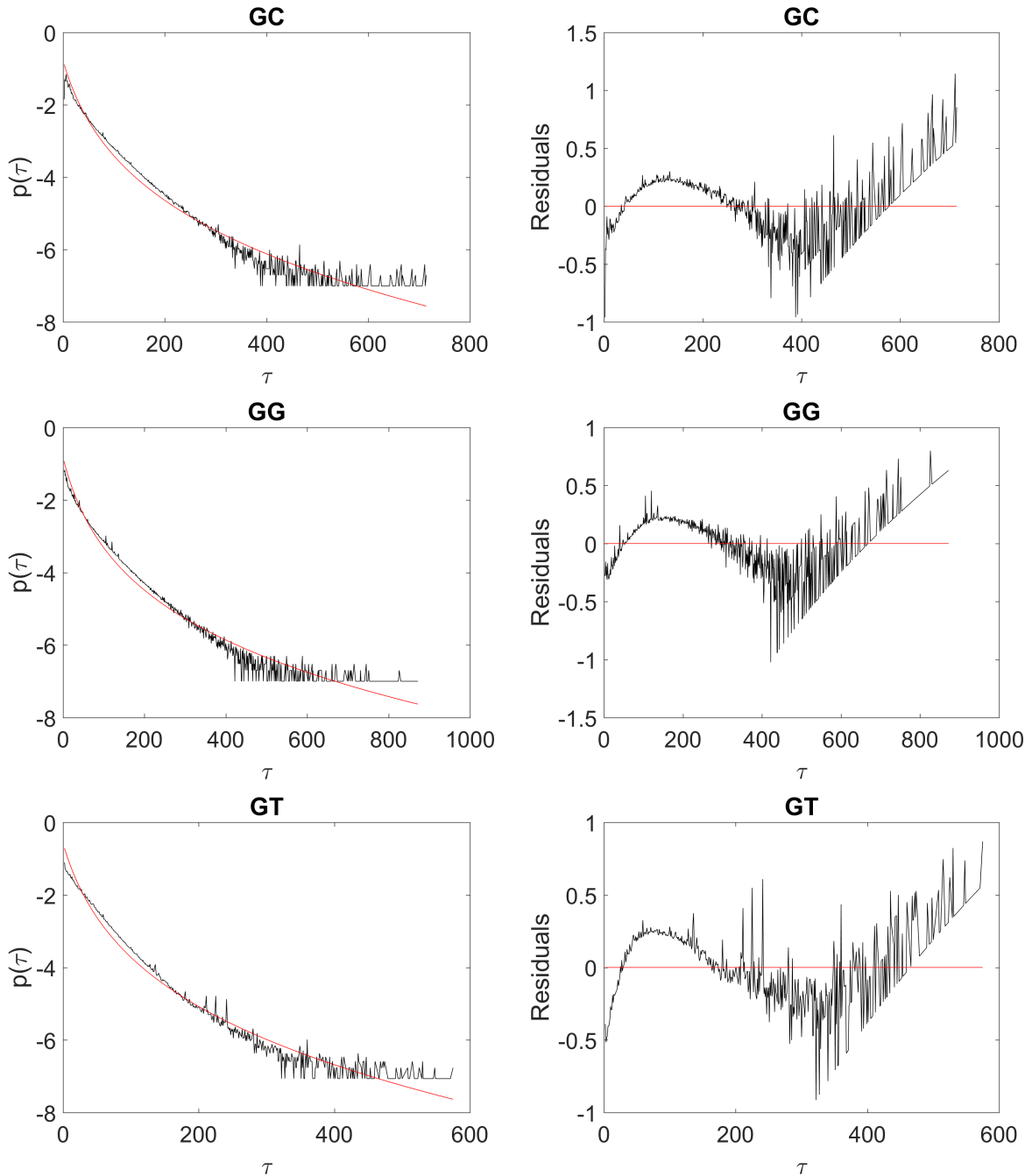


Figure 6.4: *Log-linear plot of GC, GG and GT distance distributions within human chromosome 1, together with shifted power-law with exponential tail as fitting distribution (left panel). Fit results are accompanied by residual plot (right panel).*

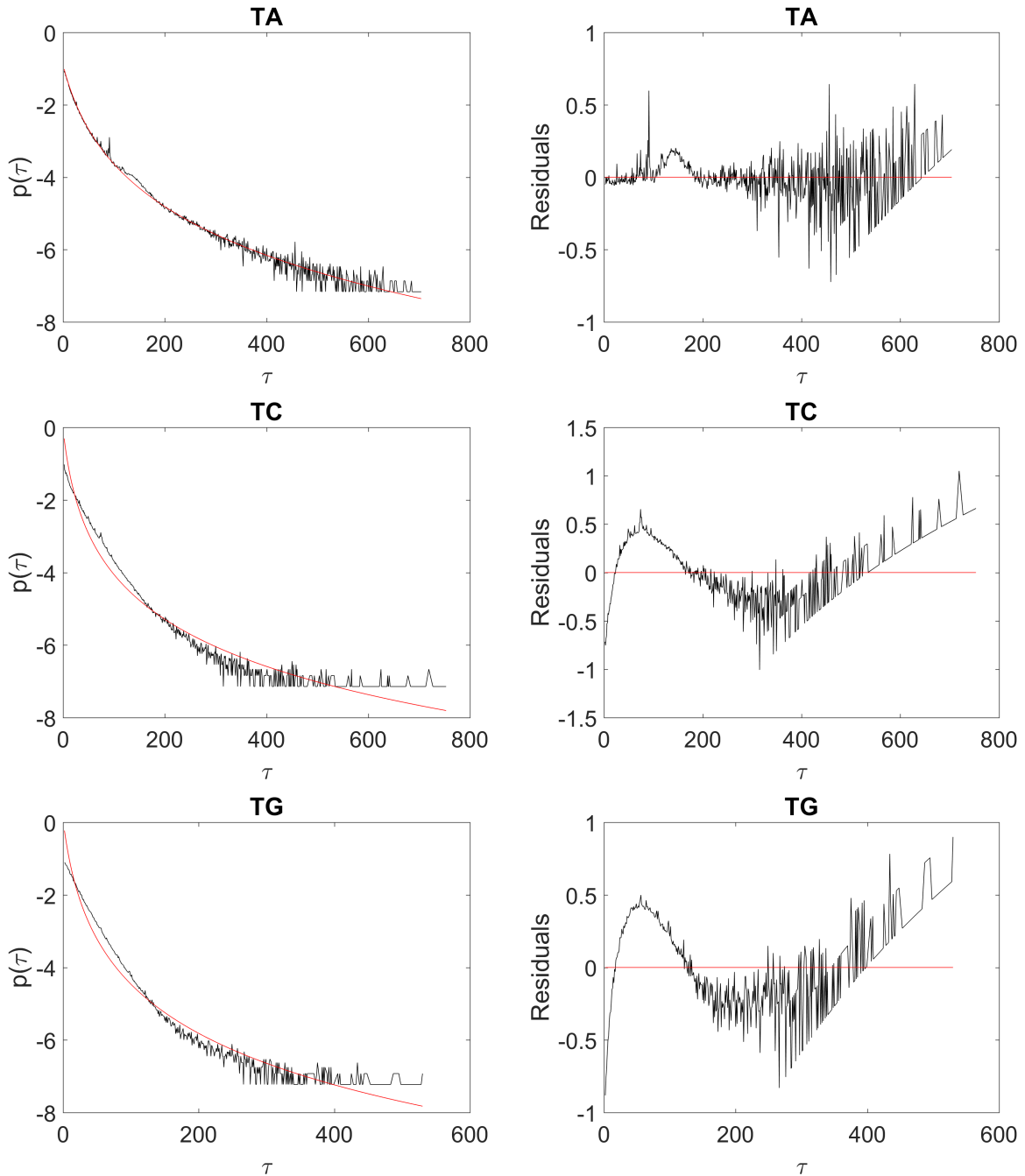


Figure 6.5: *Log-linear plot of TA, TC and TG distance distributions within human chromosome 1, together with shifted power-law with exponential tail as fitting distribution (left panel). Fit results are accompanied by residual plot (right panel).*

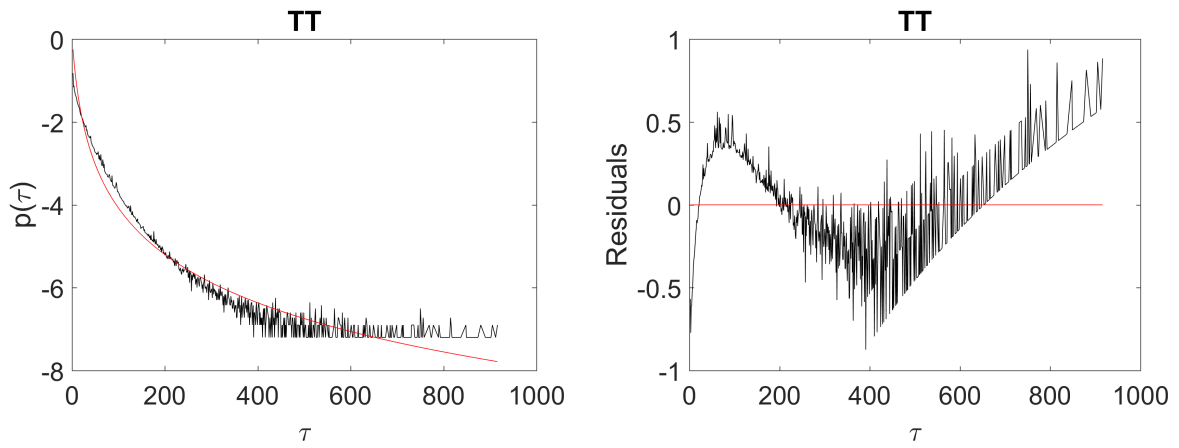


Figure 6.6: *Log-linear plot of TT distance distribution within human chromosome 1, together with shifted power-law with exponential tail as fitting distribution (left panel). Fit results are accompanied by residual plot (right panel).*

## CHAPTER 7

---

**Appendix B**

---

**7.1 TA distance distribution within mammal genomes**

We represented TA distance distributions for the 9 mammal considered in this study (*Bos taurus*, *Canis familiaris*, *Equus caballus*, *Homo sapiens*, *Macaca mulatta*, *Mus musculus*, *Ornithorhynchus anatinus*, *Pan troglodytes* and *Rattus norvegicus*), together with shifted power-law as fitting function. All fit results are accompanied by residual plots.

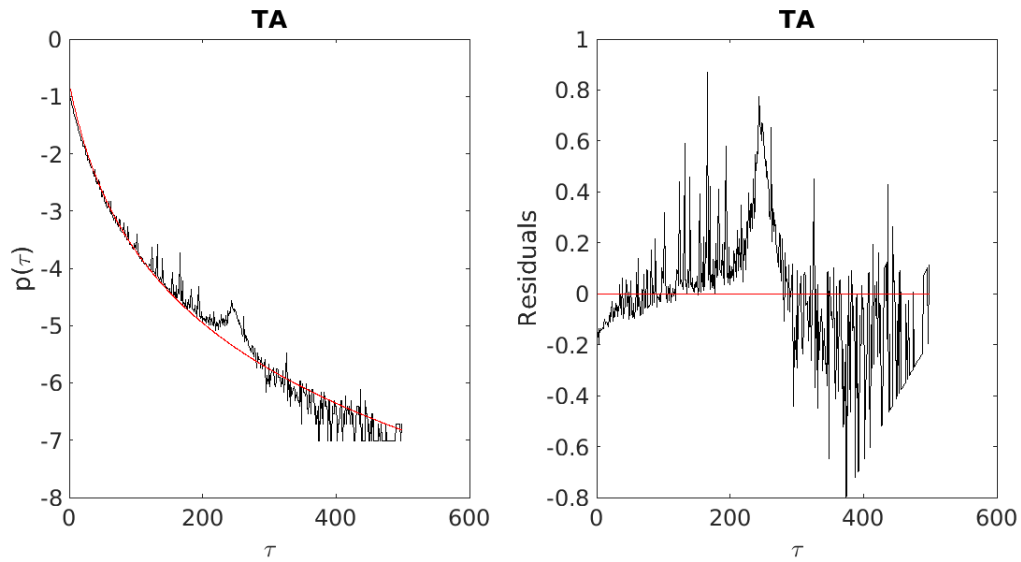


Figure 7.1: Log-linear plot of TA distance distribution within chromosome 1 of *Bos taurus*, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).

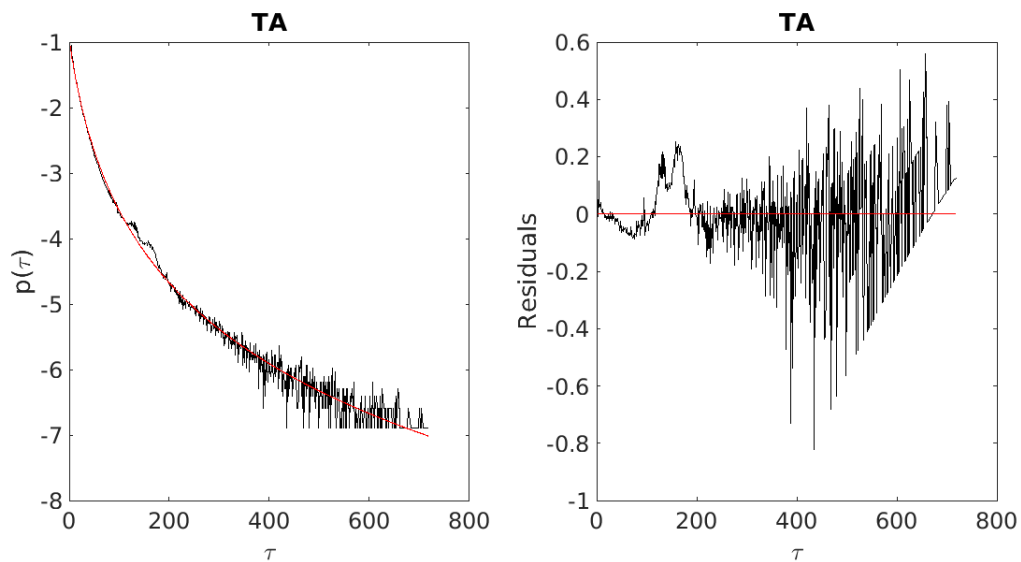


Figure 7.2: Log-linear plot of TA distance distribution within chromosome 1 of *Canis familiaris*, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).

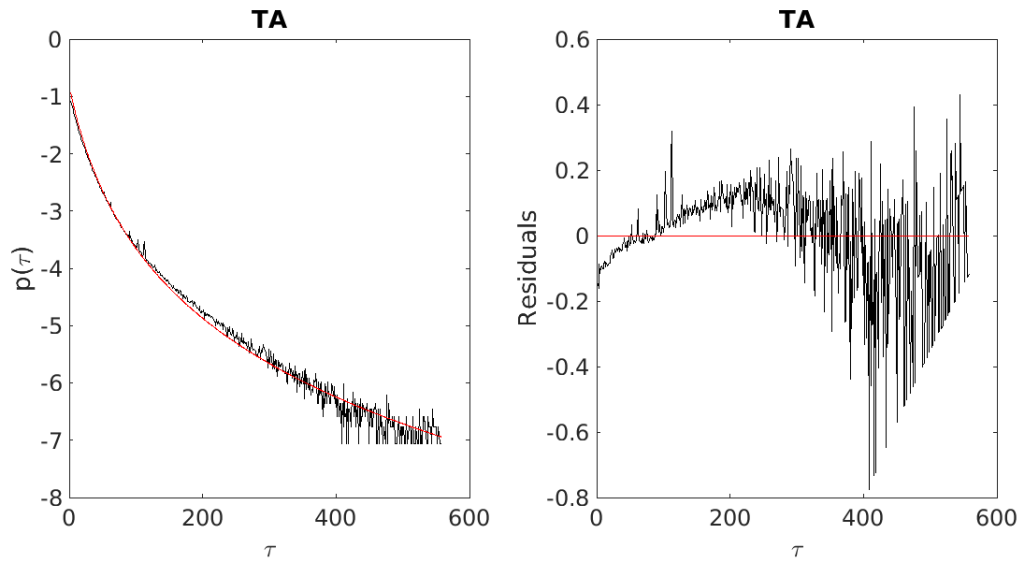


Figure 7.3: Log-linear plot of TA distance distribution within chromosome 1 of *Equus Caballus*, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).

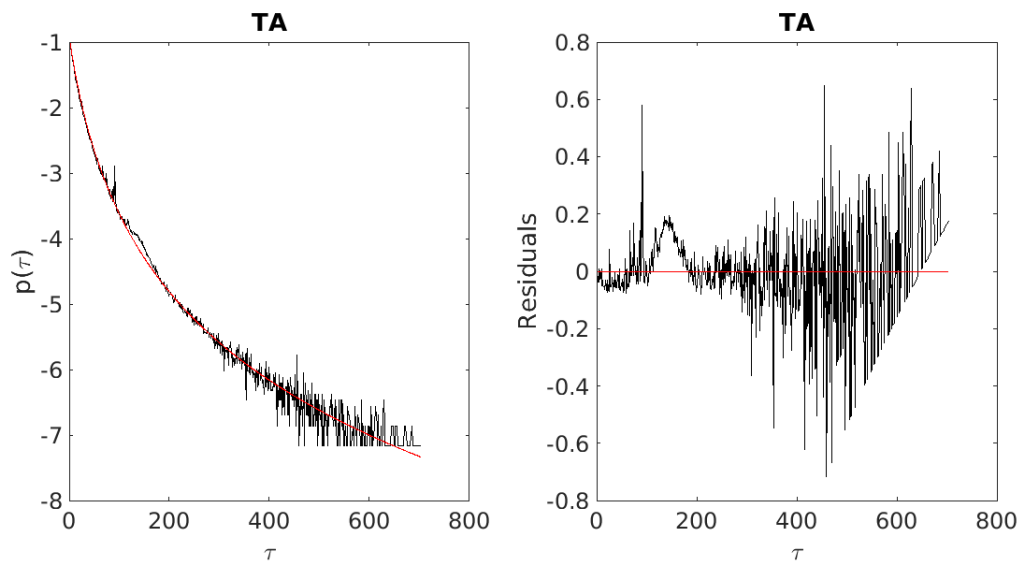


Figure 7.4: Log-linear plot of TA distance distribution within chromosome 1 of *Homo sapiens*, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).

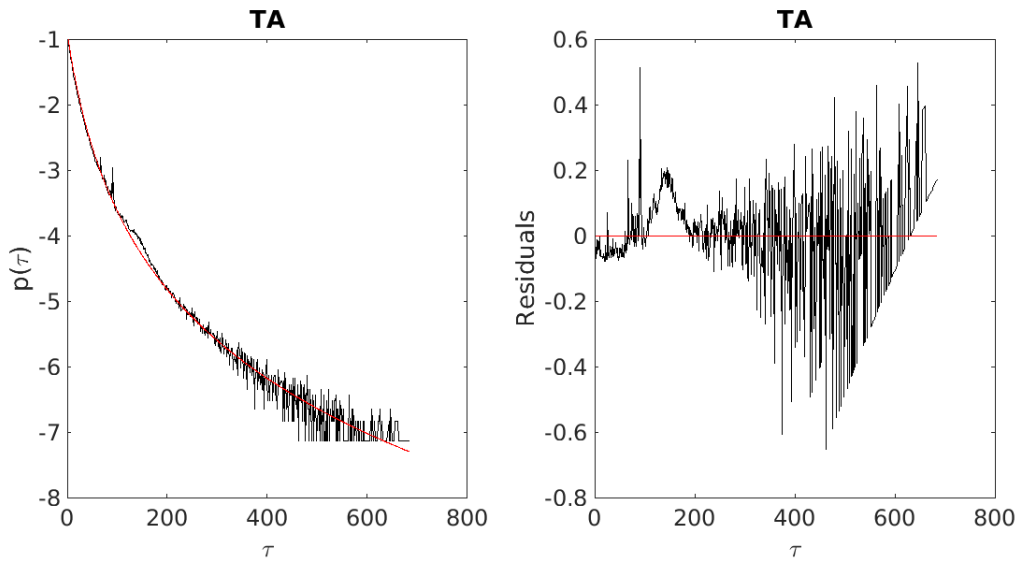


Figure 7.5: *Log-linear plot of TA distance distribution within chromosome 1 of **Macaca mulatta**, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).*

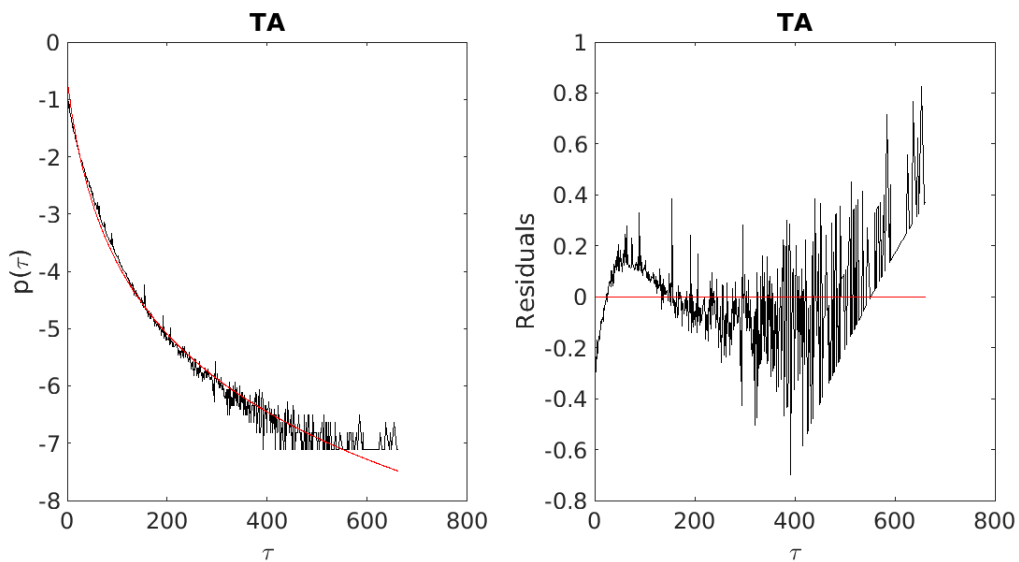


Figure 7.6: *Log-linear plot of TA distance distribution within chromosome 1 of **Mus musculus**, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).*

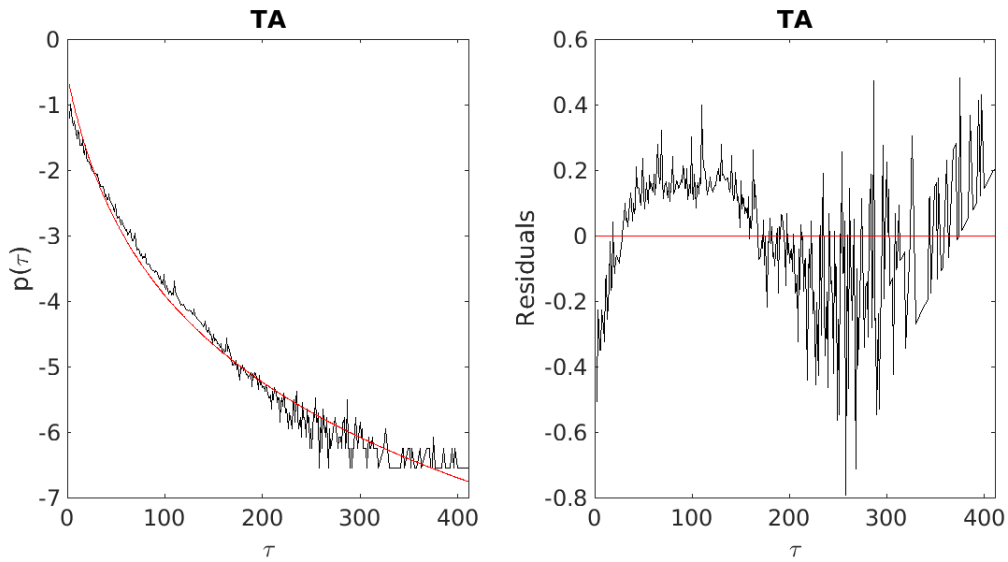


Figure 7.7: Log-linear plot of TA distance distribution within chromosome 1 of *Ornithorhynchus anatinus*, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).

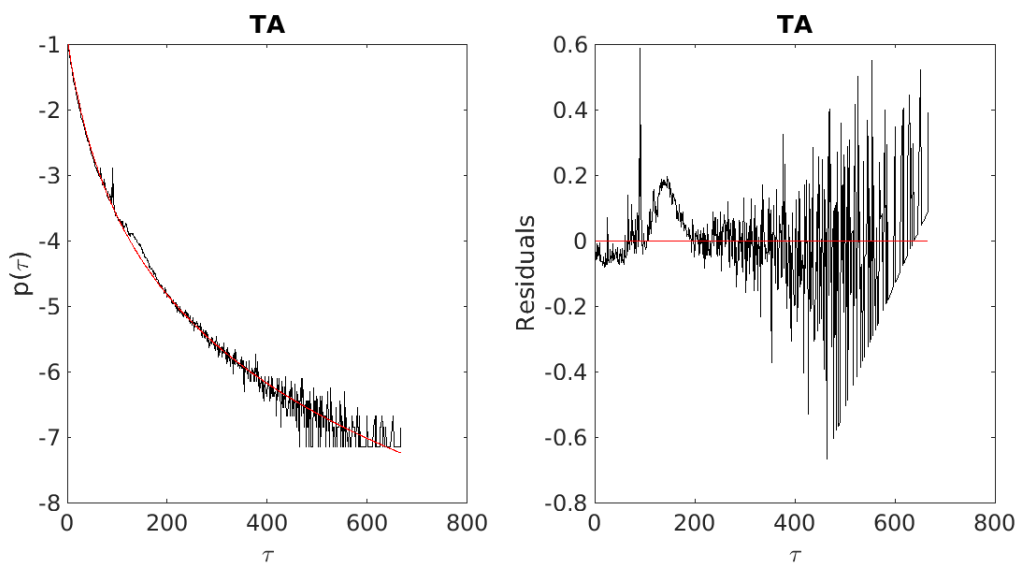


Figure 7.8: Log-linear plot of TA distance distribution within chromosome 1 of *Pan troglodytes*, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).



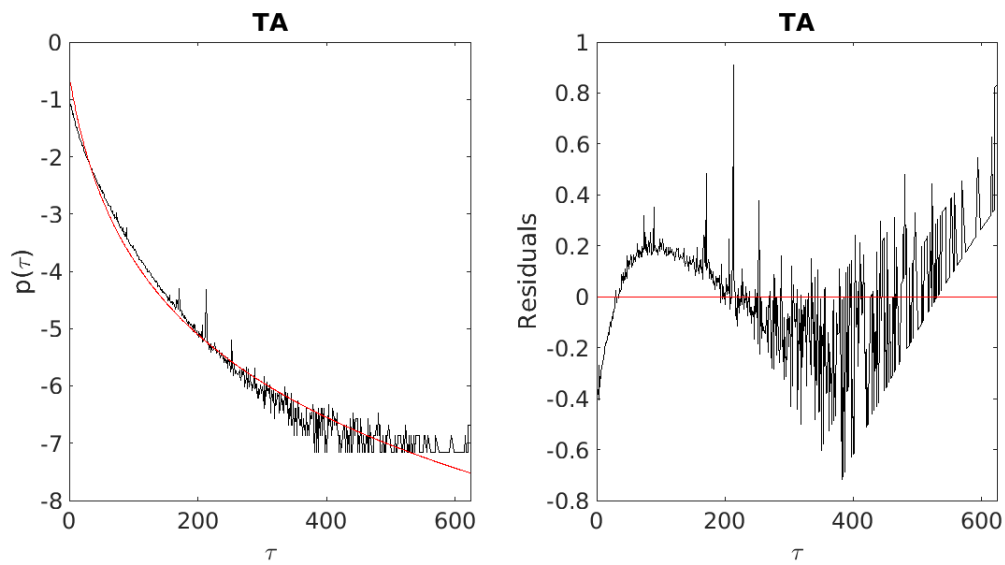


Figure 7.9: *Log-linear plot of TA distance distribution within chromosome 1 of **Rattus norvegicus**, together with shifted power-law as fitting distribution (upper panel). Fit results are accompanied by residual plot (lower panel).*

## CHAPTER 8

---

### Appendix C

---

We considered all the sequences between consecutive TA separated by distances ranging from 10 bp to 400 bp and computed the mean distance  $\mu$  and the standard deviation  $\sigma$  between consecutive AA/TT and consecutive GC along the sequence (as explained in section 2.3.3), across all chromosomes of the human genome.

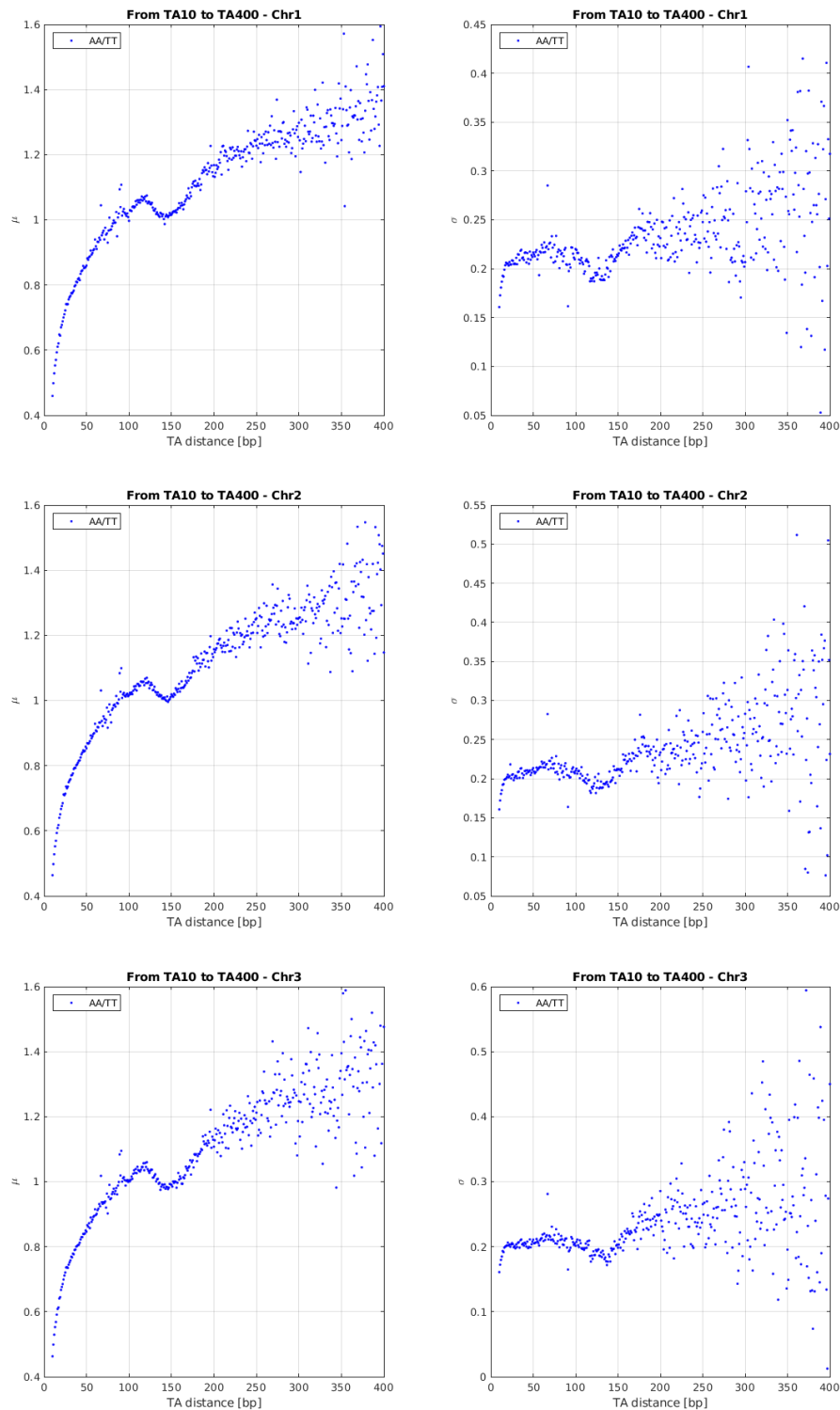


Figure 8.1: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 1, 2 and 3.

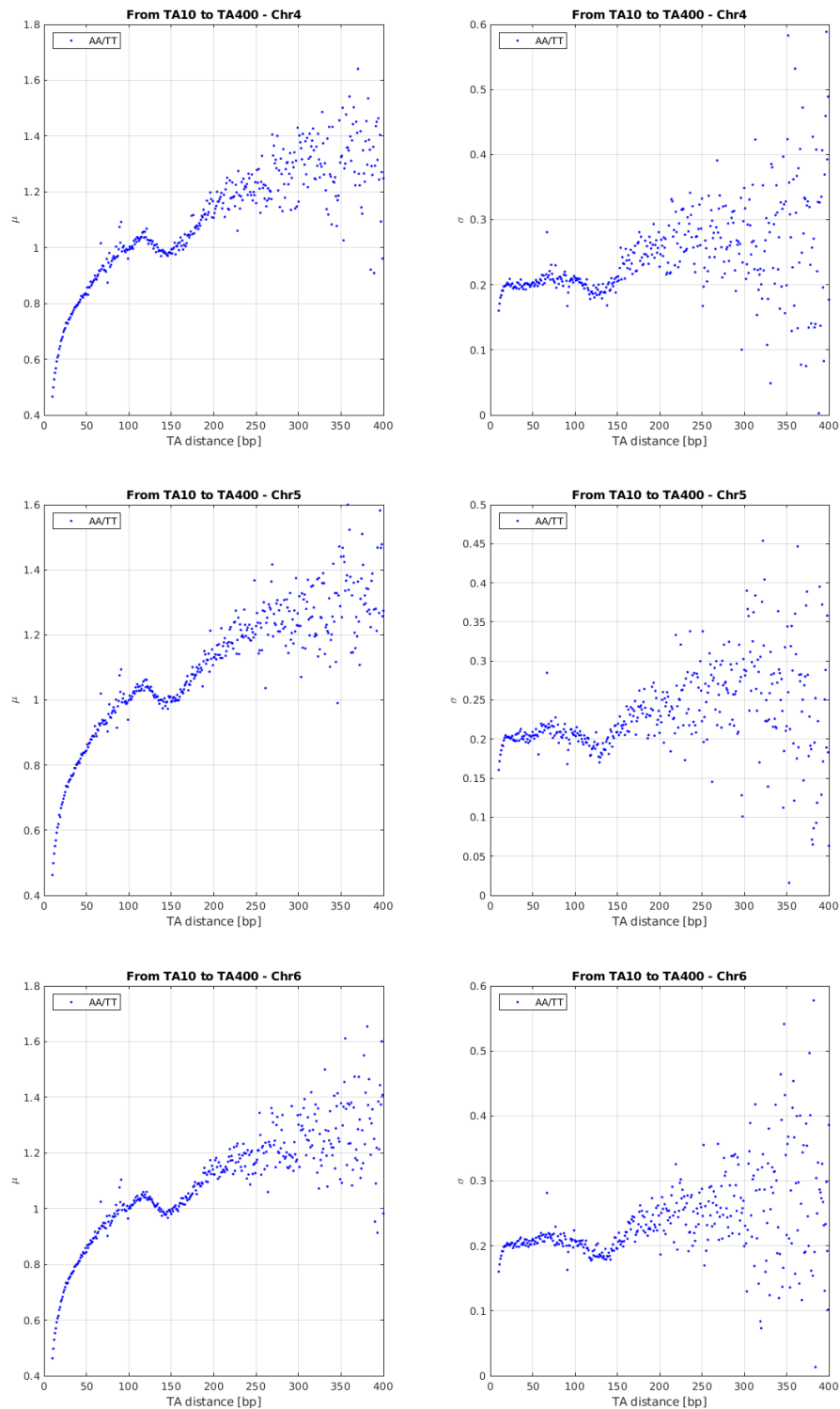


Figure 8.2: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 4, 5 and 6.

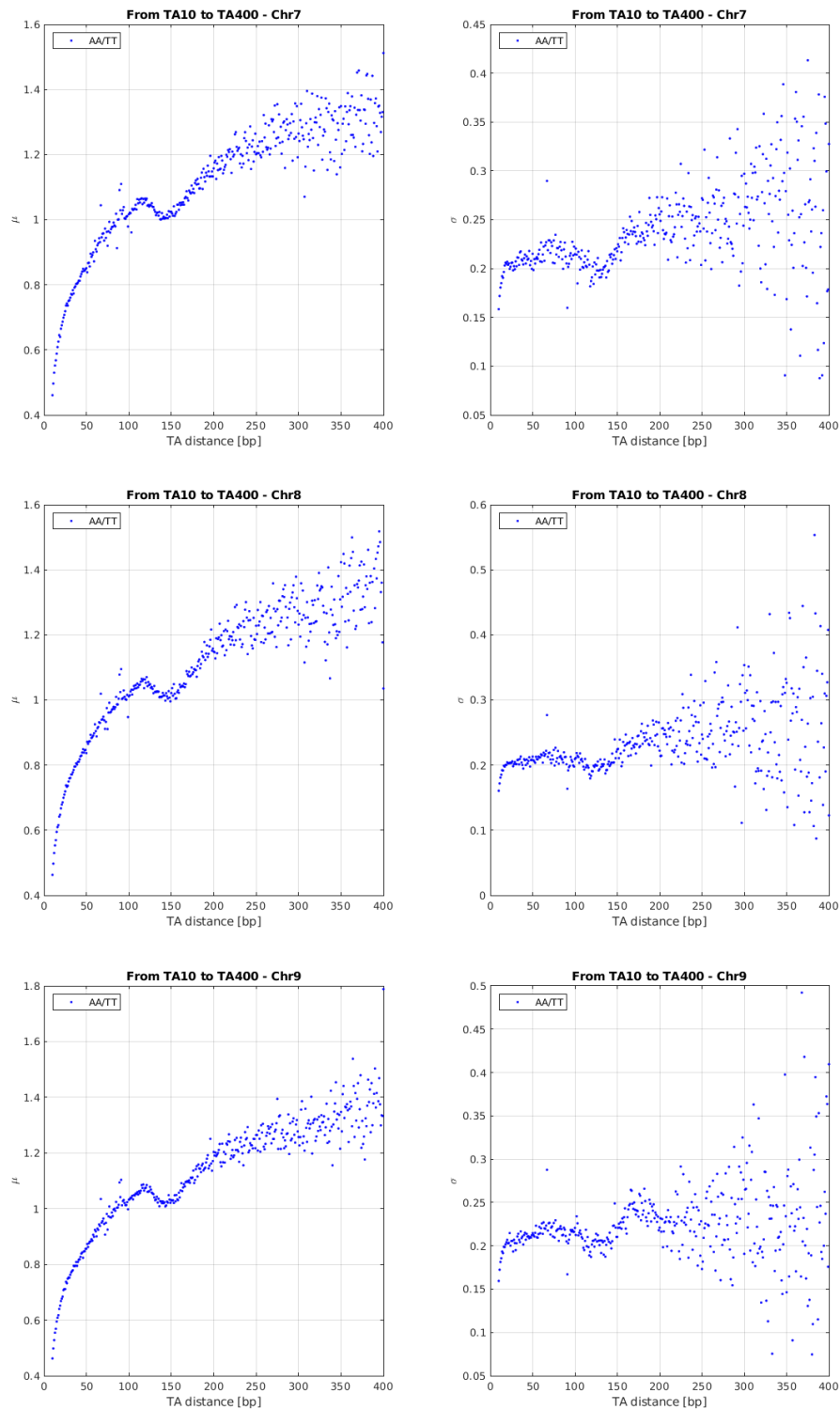


Figure 8.3: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 7, 8 and 9.

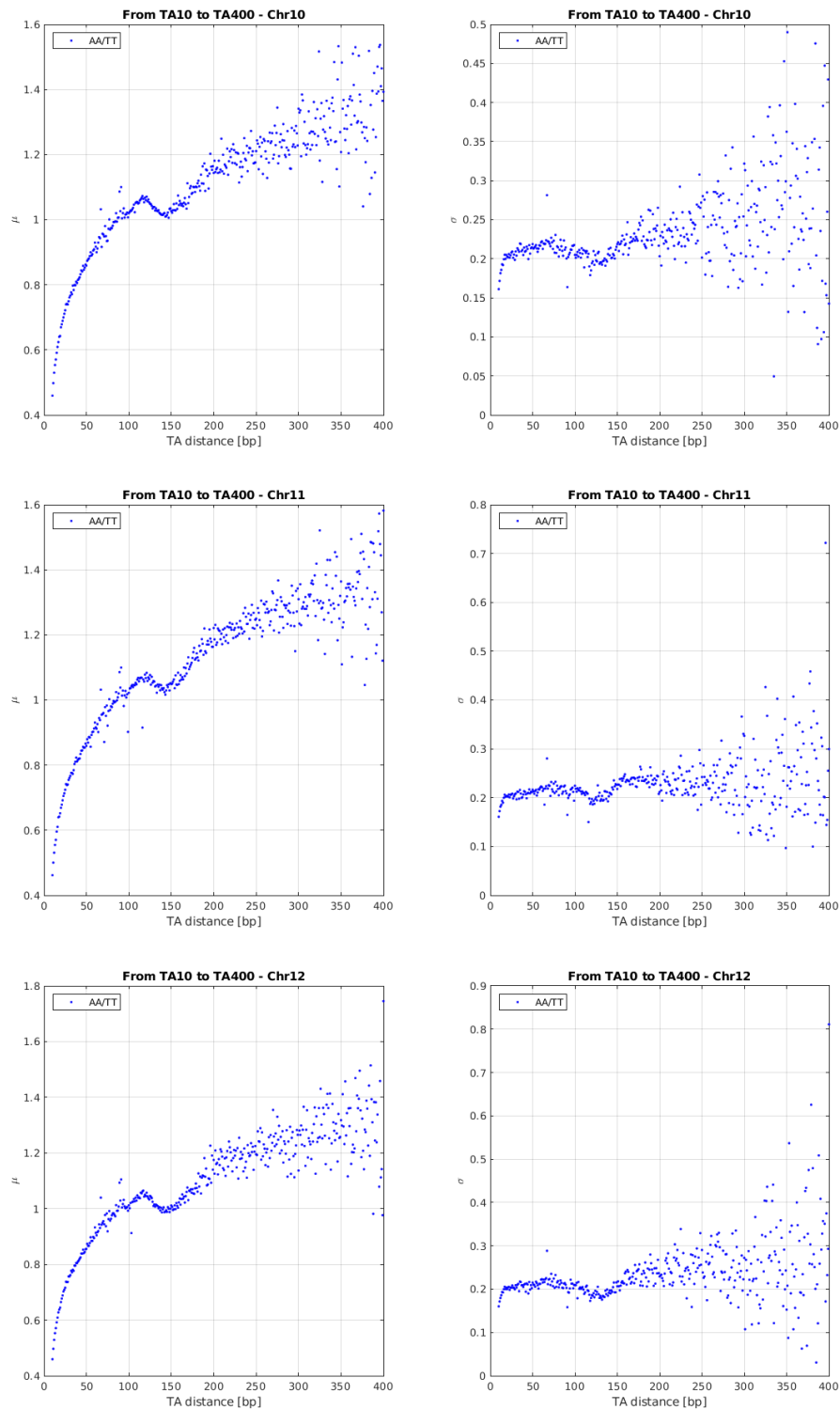


Figure 8.4: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 10, 11 and 12.

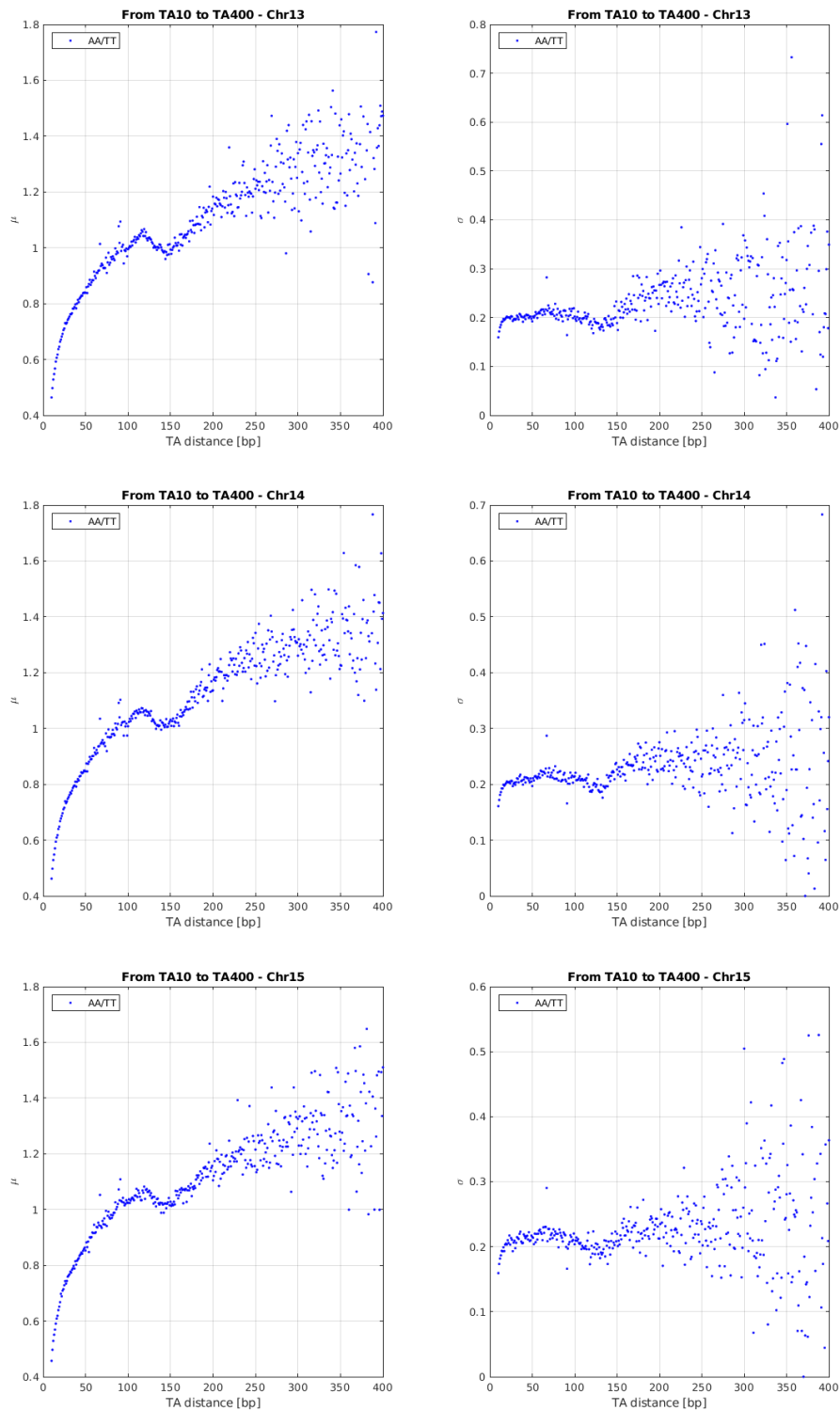


Figure 8.5: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 13, 14 and 15.

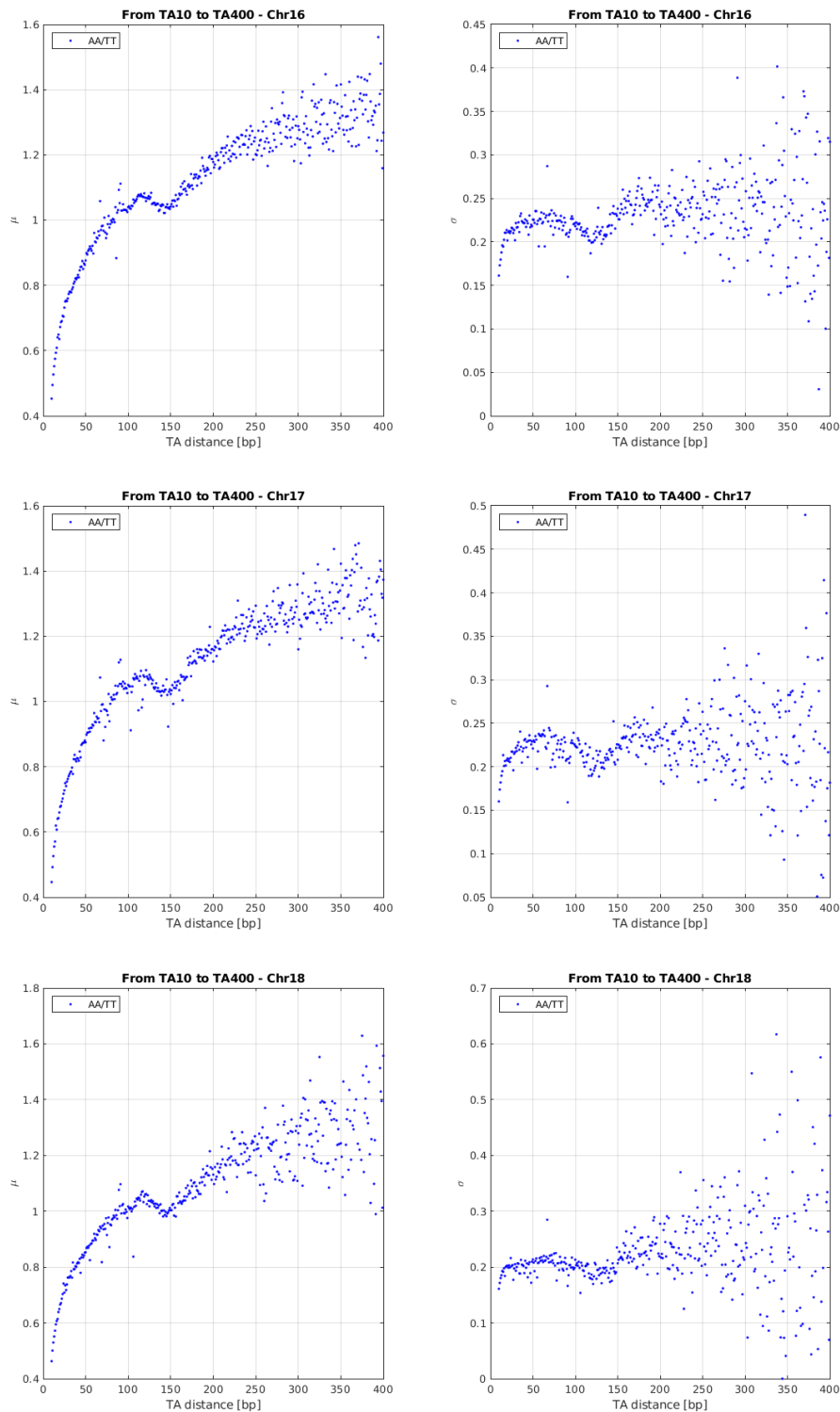


Figure 8.6: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 16, 17 and 18.



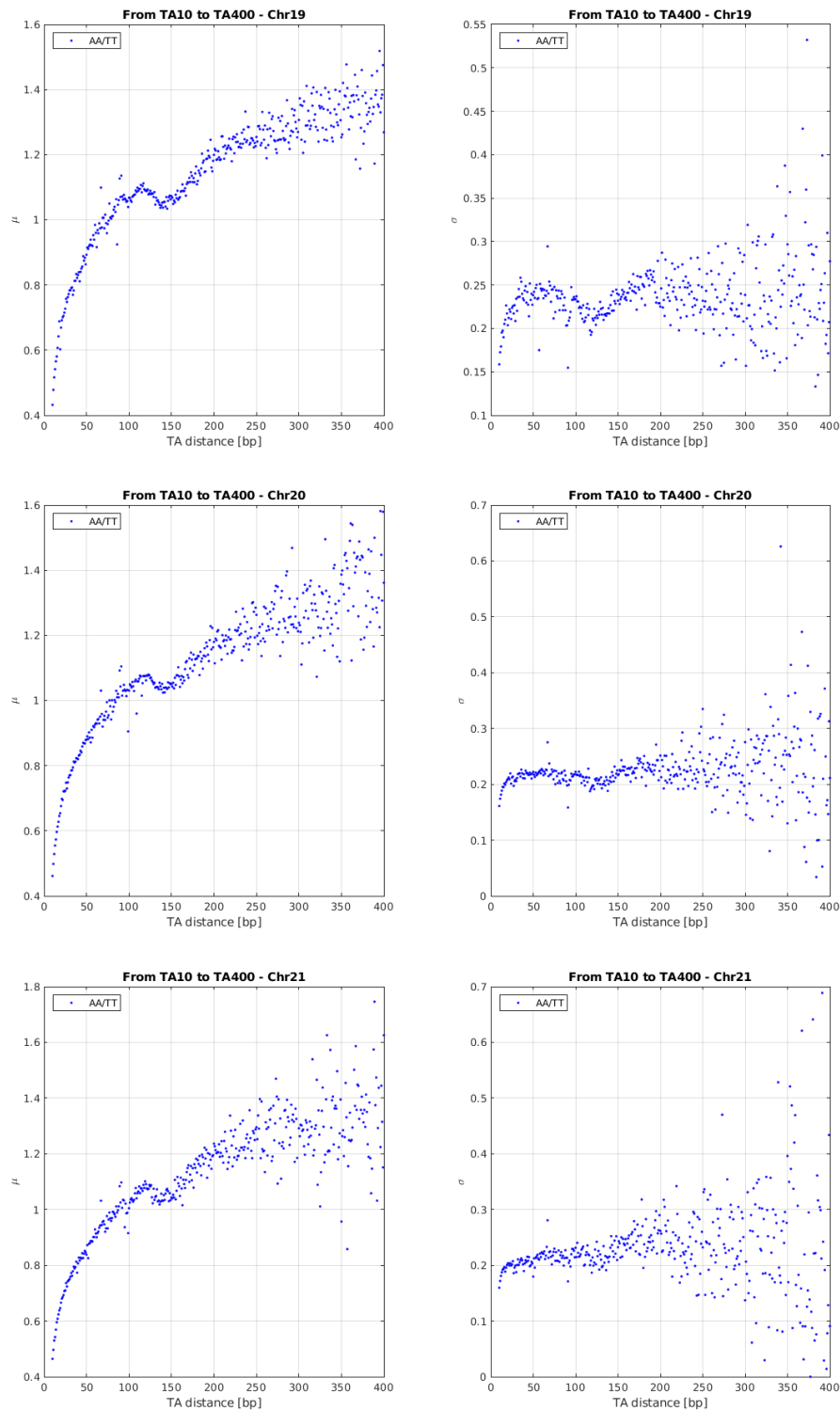


Figure 8.7: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 19, 20 and 21.

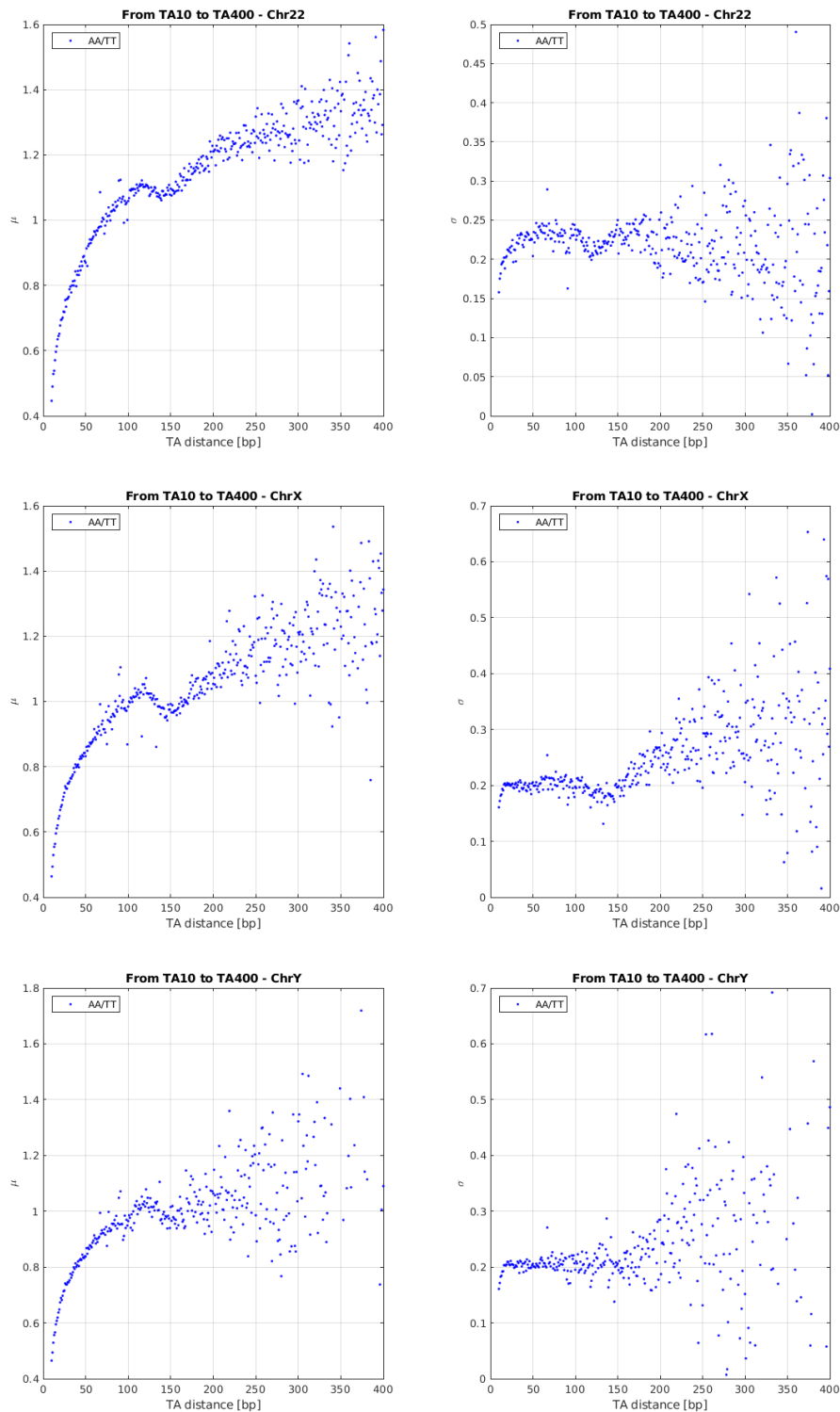


Figure 8.8: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive AA/TT along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 19, 20 and 21.

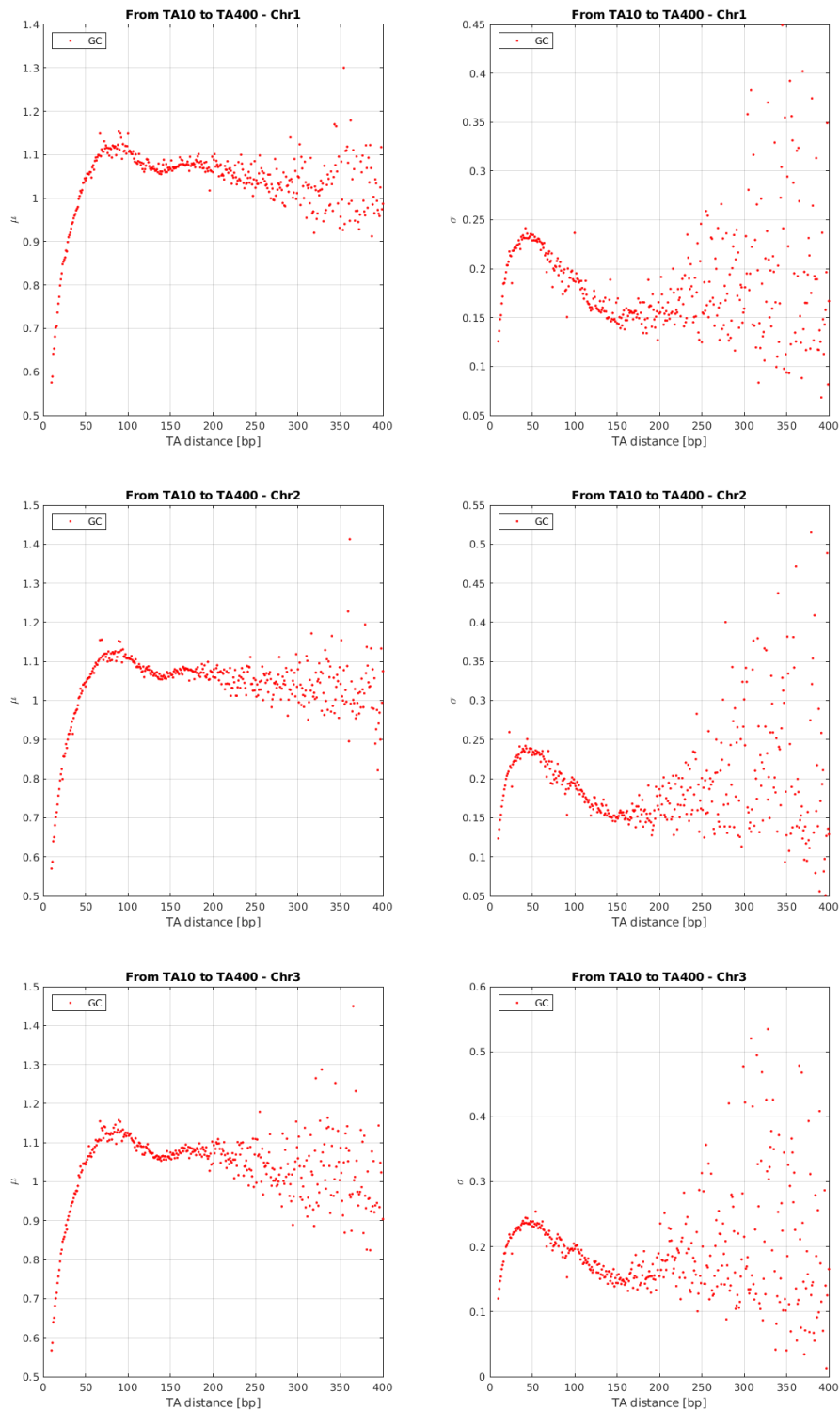


Figure 8.9: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 1, 2 and 3.

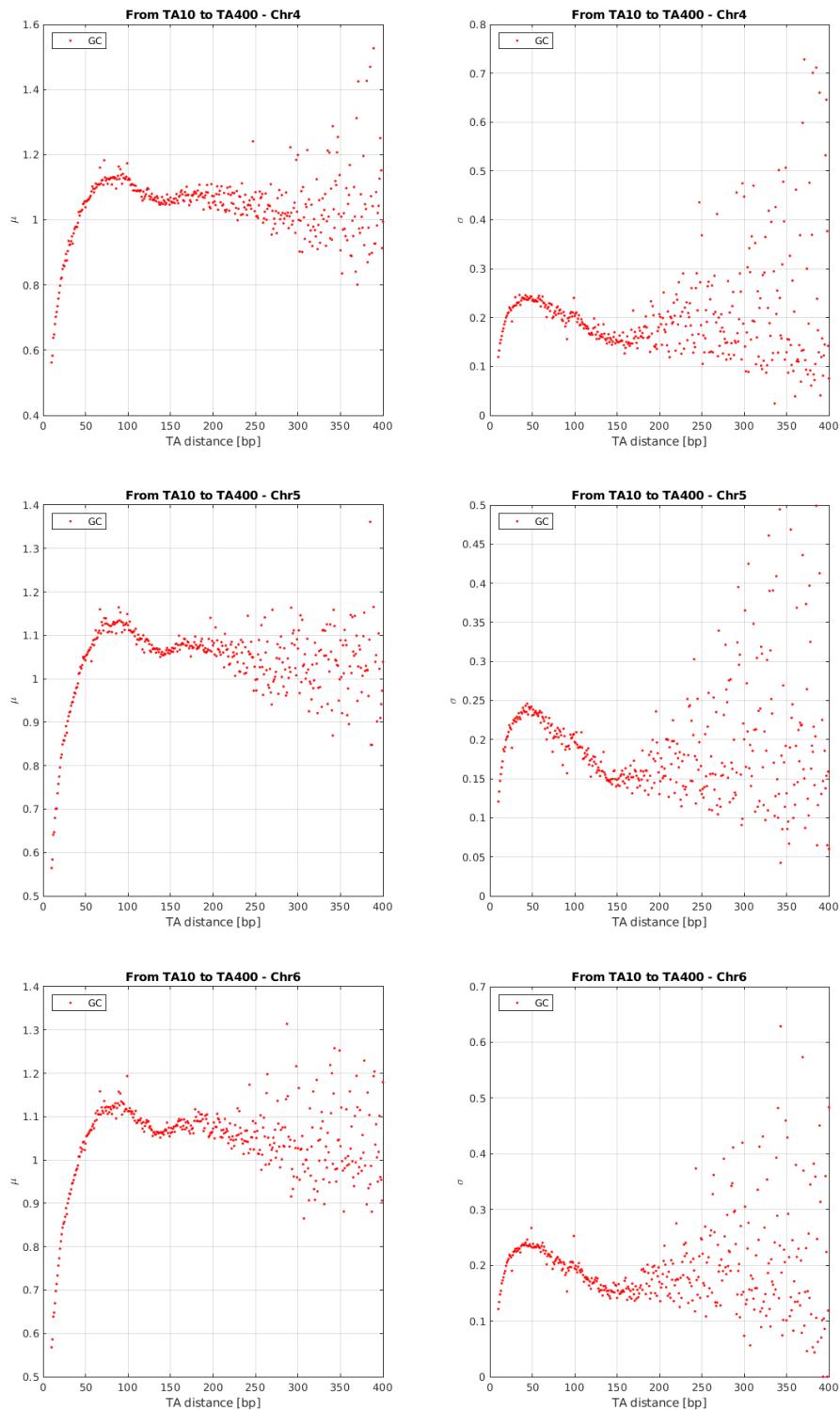


Figure 8.10: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 4, 5 and 6.

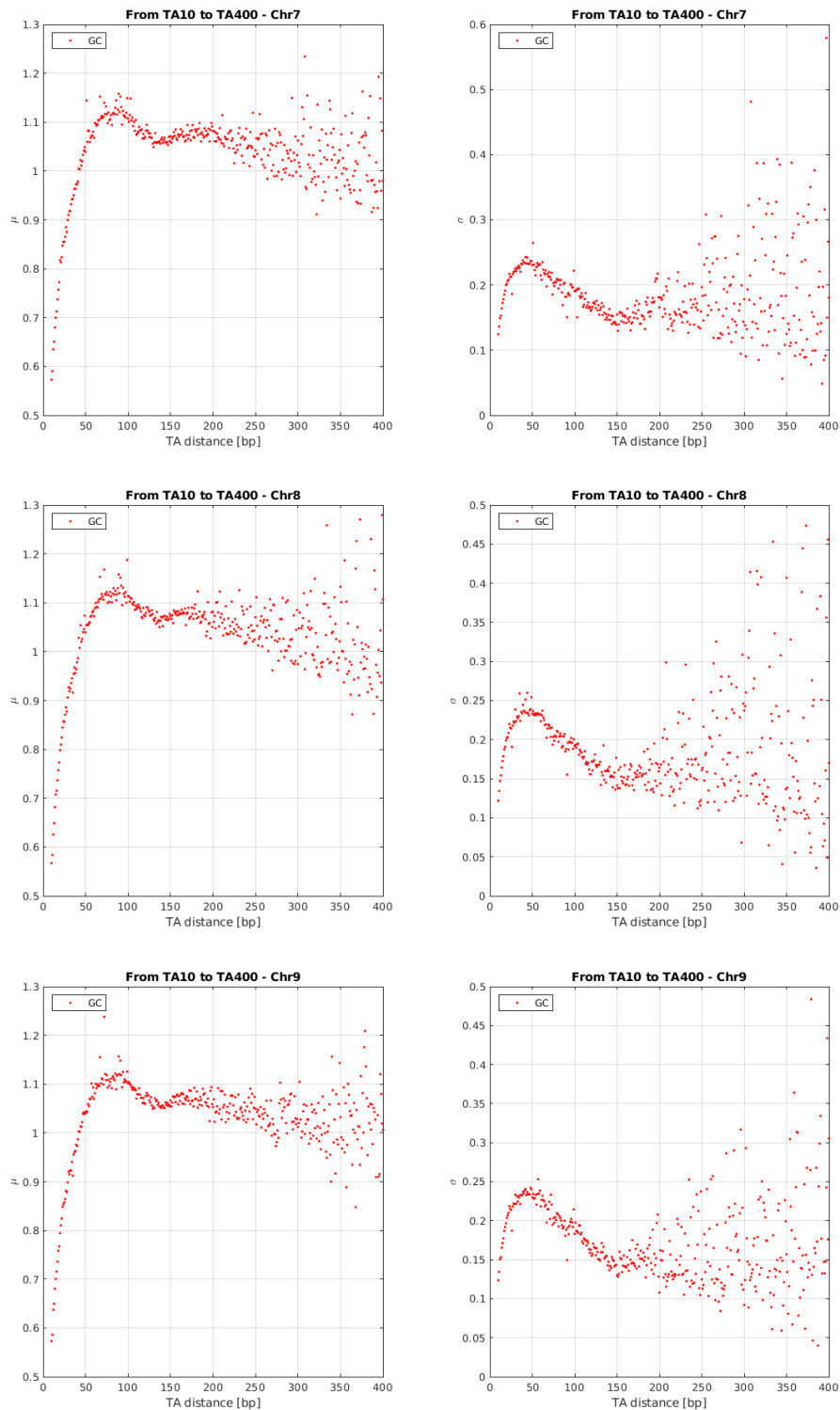


Figure 8.11: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 7, 8 and 9.

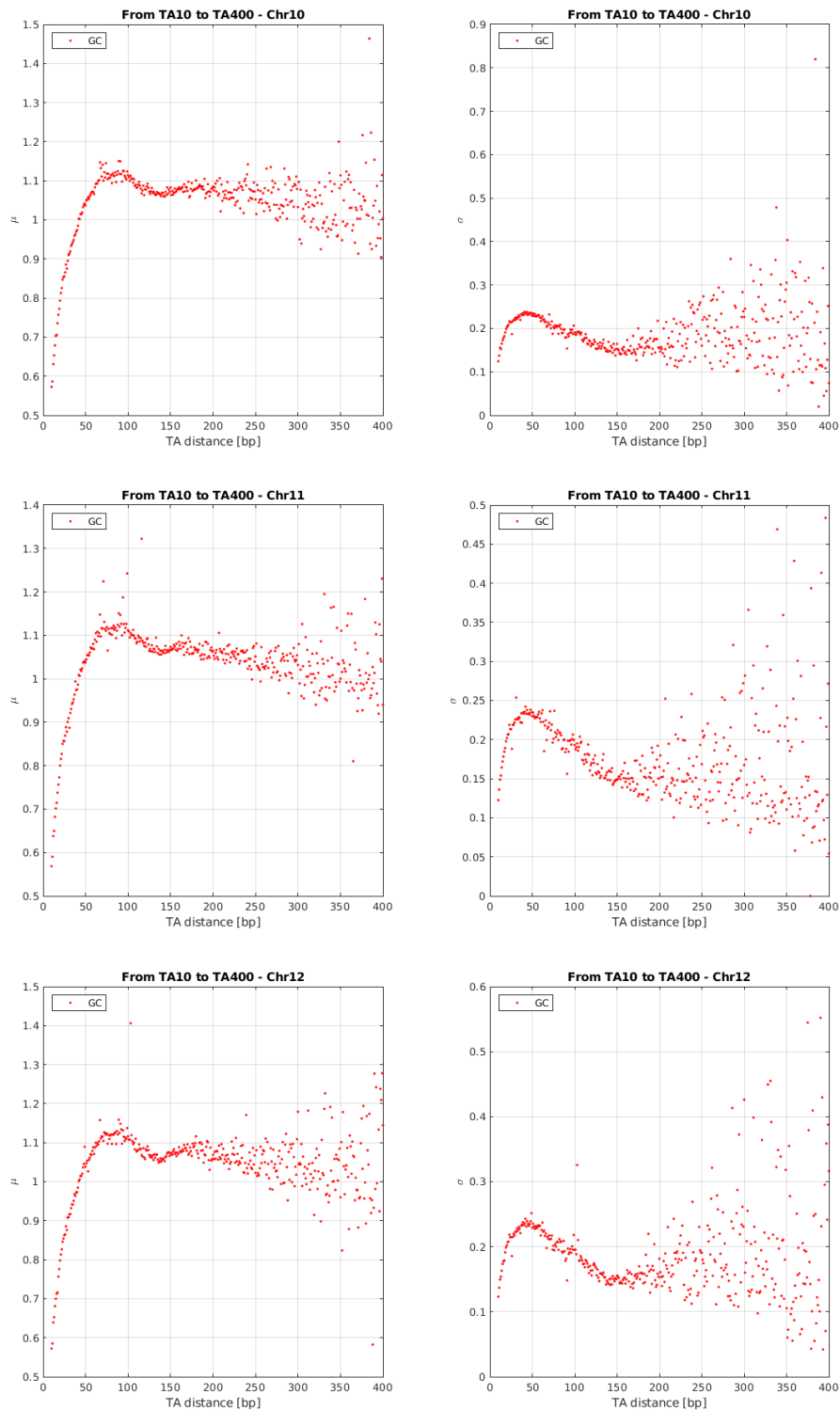


Figure 8.12: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 10, 11 and 12.

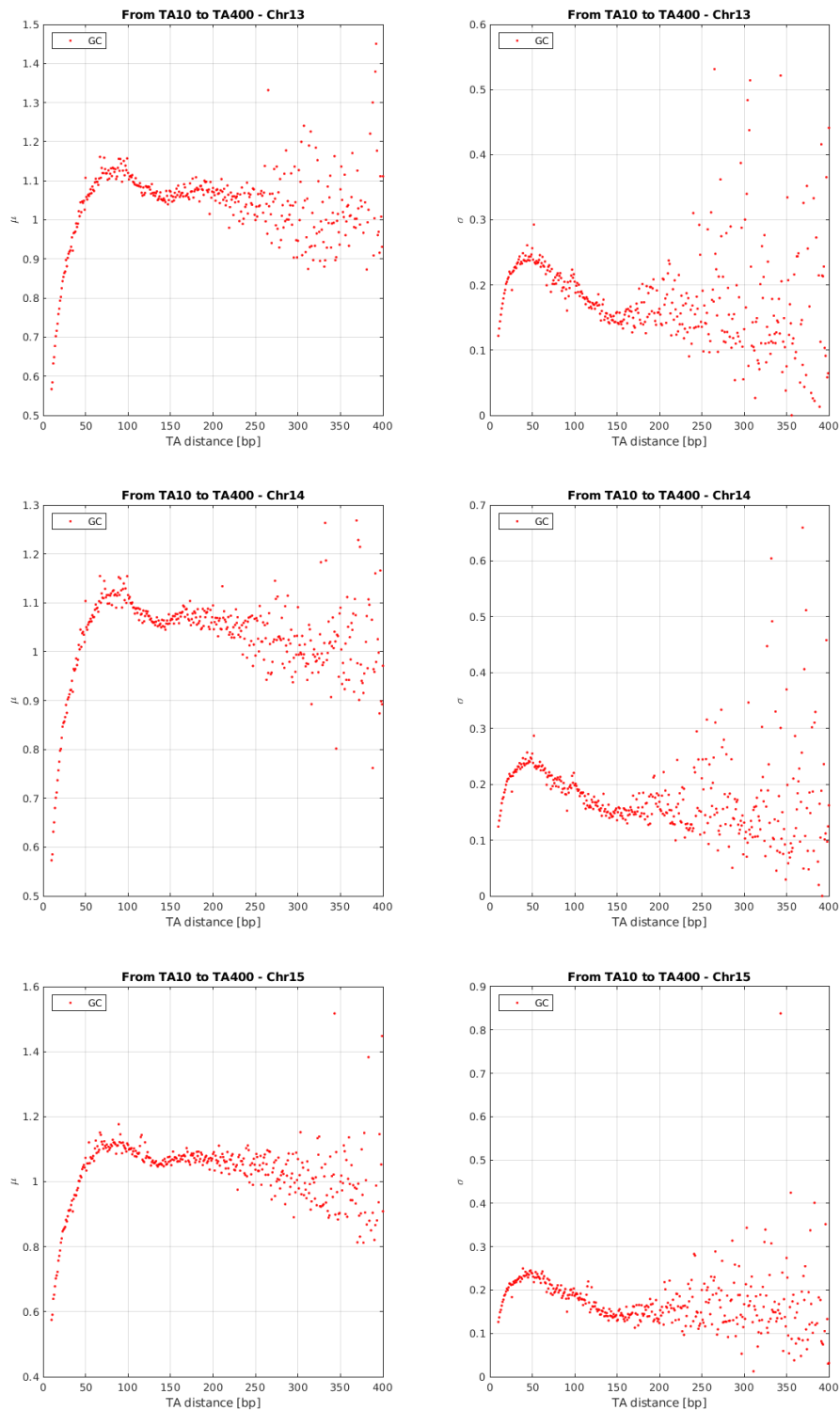


Figure 8.13: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 13, 14 and 15.

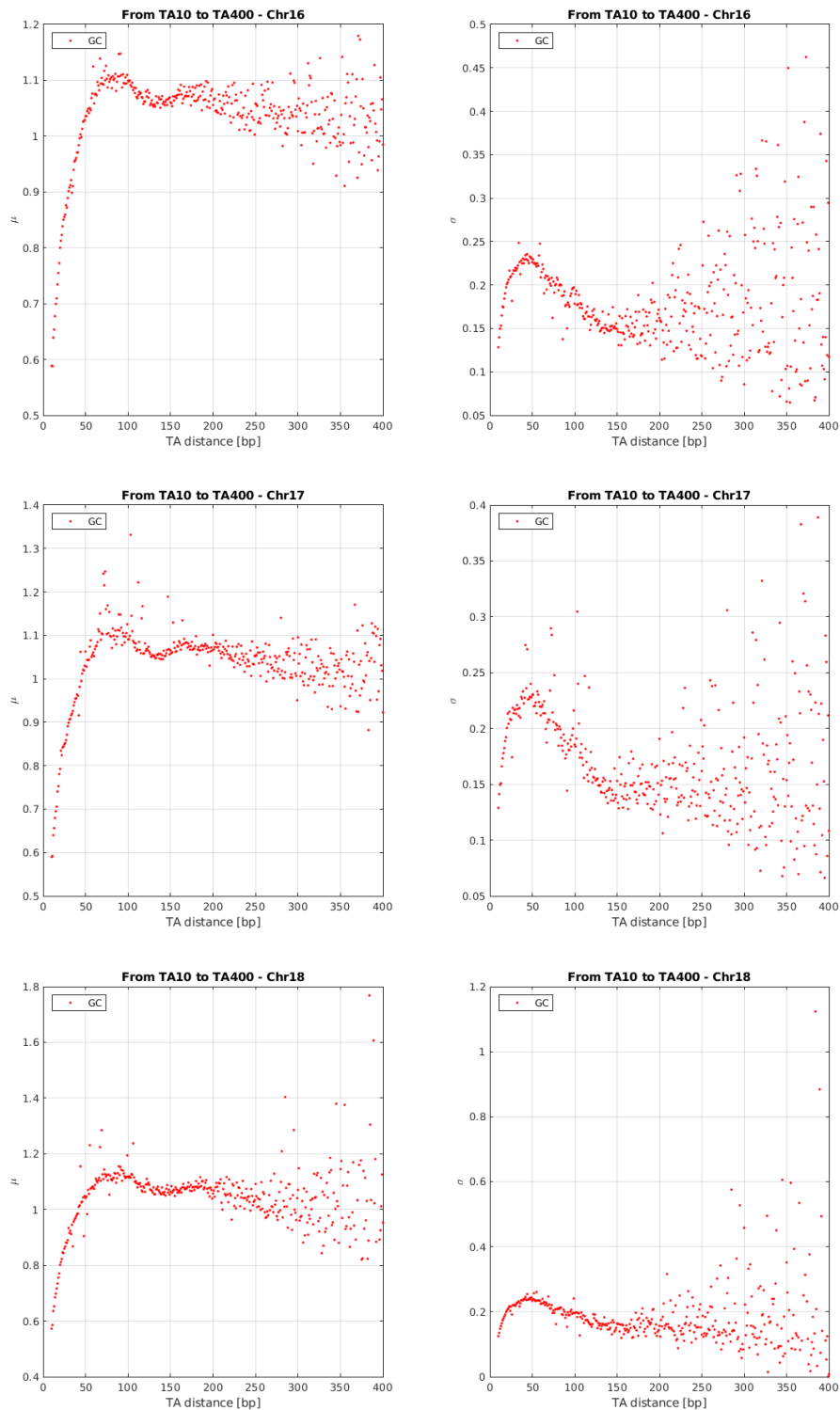


Figure 8.14: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 16, 17 and 18.



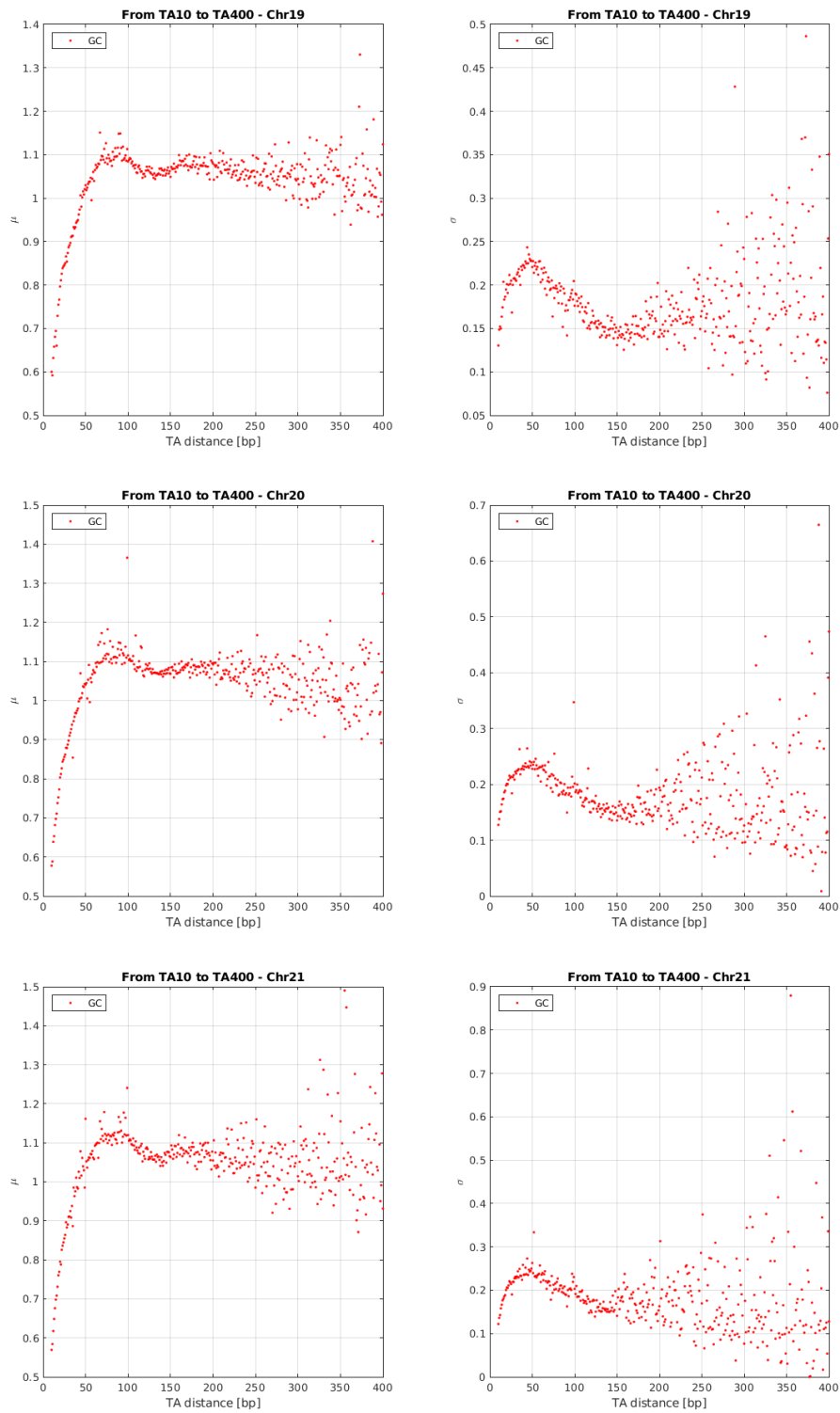


Figure 8.15: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 19, 20 and 21.

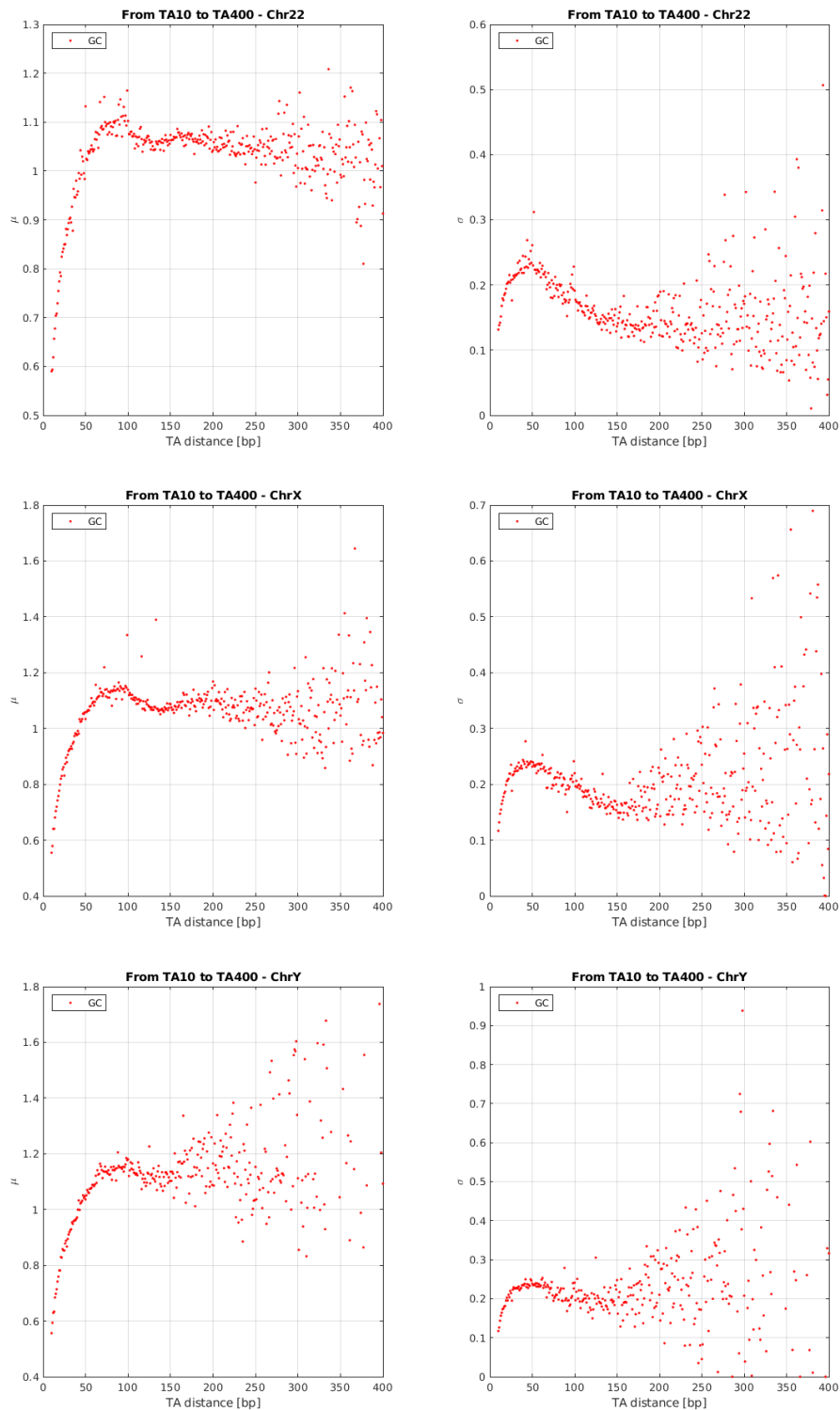


Figure 8.16: Gaussian fit parameters  $\mu$  and  $\sigma$  representing respectively the mean value and the standard deviation of the logarithm of the distance between consecutive GC along the sequences identified by consecutive TA separated by a minimum distance of 10 bp up to a maximum distance of 400 bp, within human chromosome 22, X and Y.

## CHAPTER 9

---

### Appendix D

---

### **9.1 Deviation at 91 bp from TA distance distribution**

The representation of the occurrence probability of consecutive AA/TT and consecutive GC along the sequences between two consecutive TA separated by a distance of 91 bp, clearly shows the same patterns for all chromosomes of human genome.

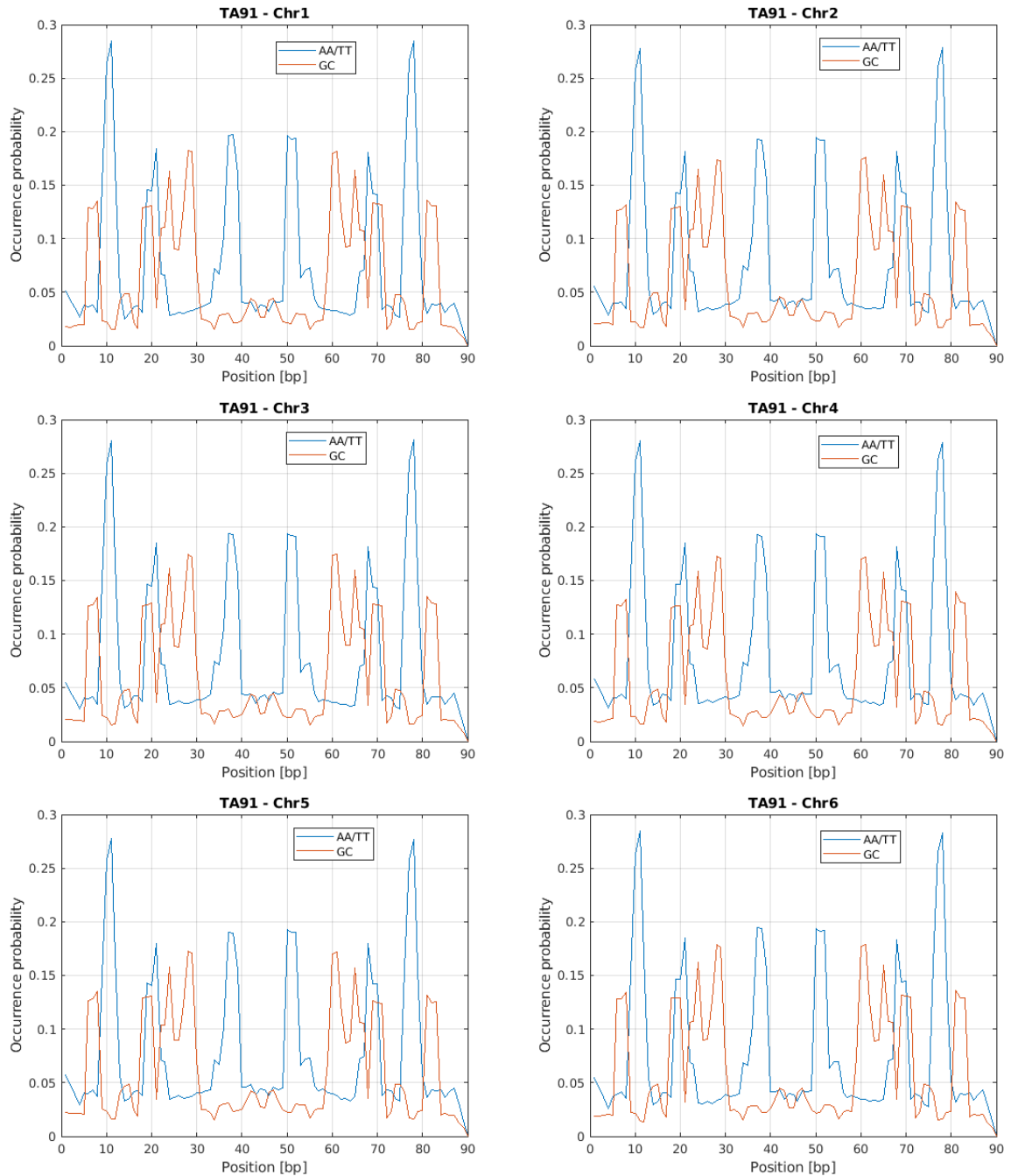


Figure 9.1: Occurrence probability as a function dinucleotide position within the sequences identified by consecutive TA separated by a distance of 91 bp within human chromosomes 1, 2, 3, 4, 5 and 6.

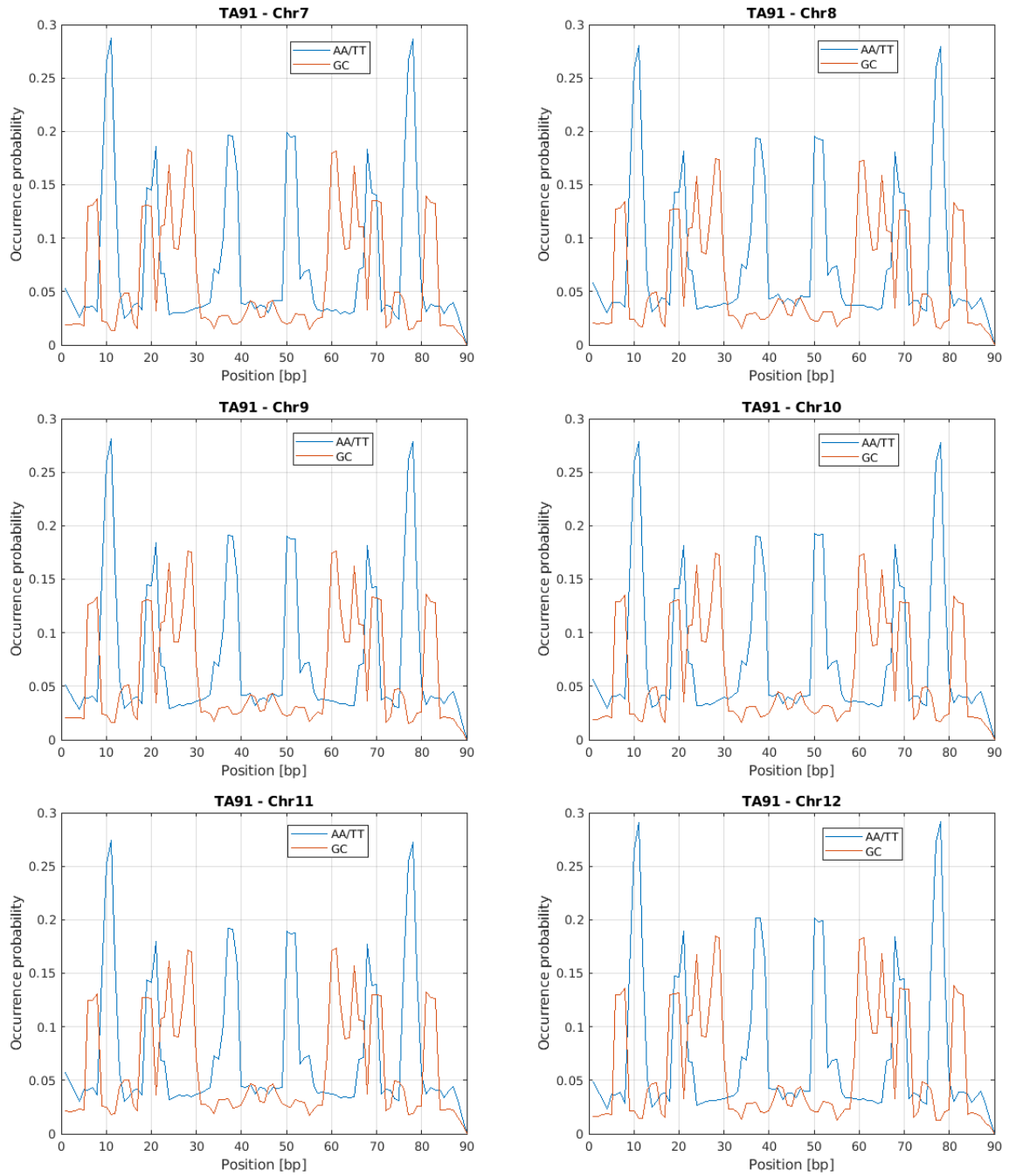


Figure 9.2: Occurrence probability as a function dinucleotide position within the sequences identified by consecutive TA separated by a distance of 91 bp within human chromosomes 7, 8, 9, 10, 11 and 12.

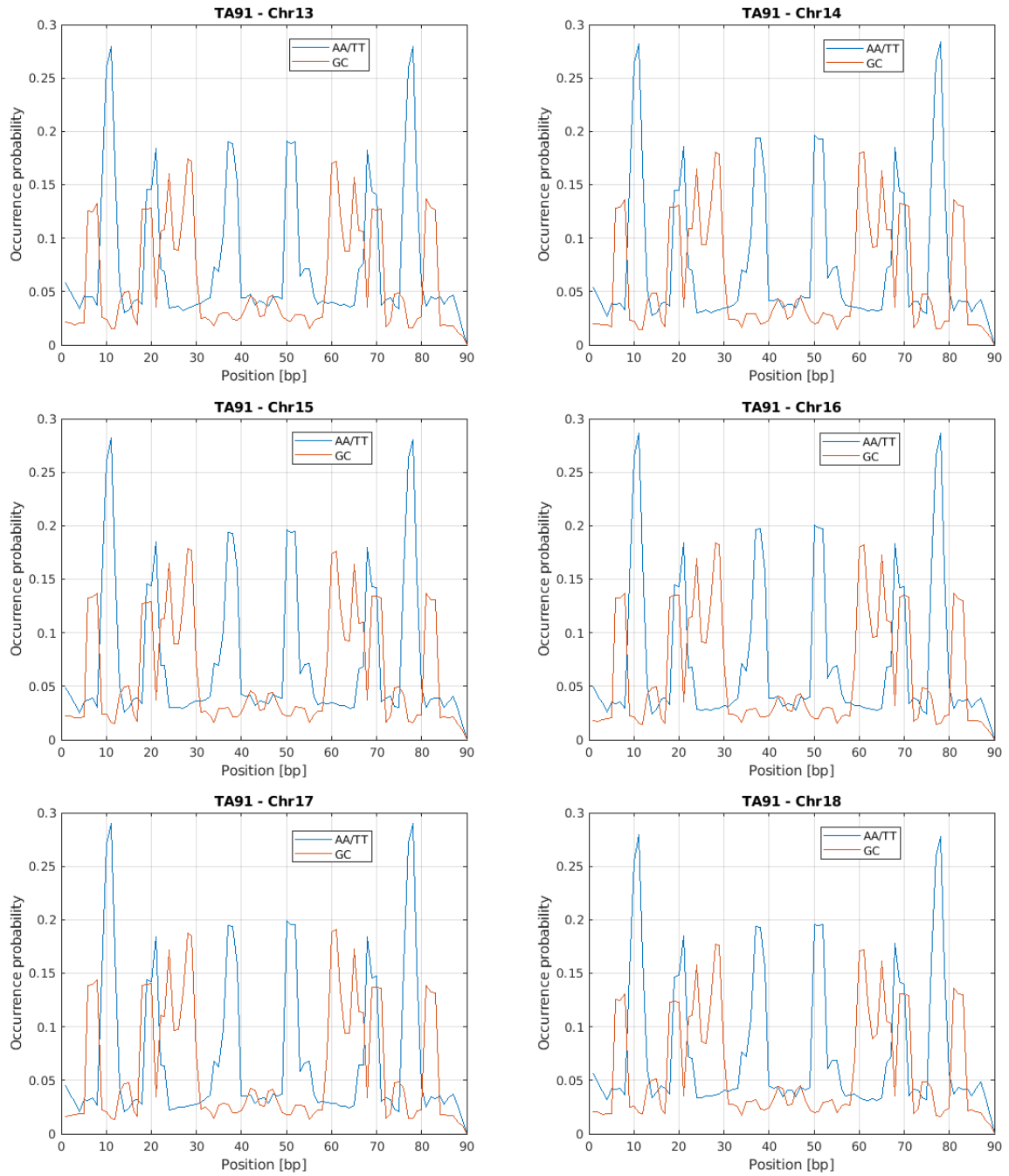


Figure 9.3: Occurrence probability as a function dinucleotide position within the sequences identified by consecutive TA separated by a distance of 91 bp within human chromosomes 13, 14, 15, 16, 17 and 18.

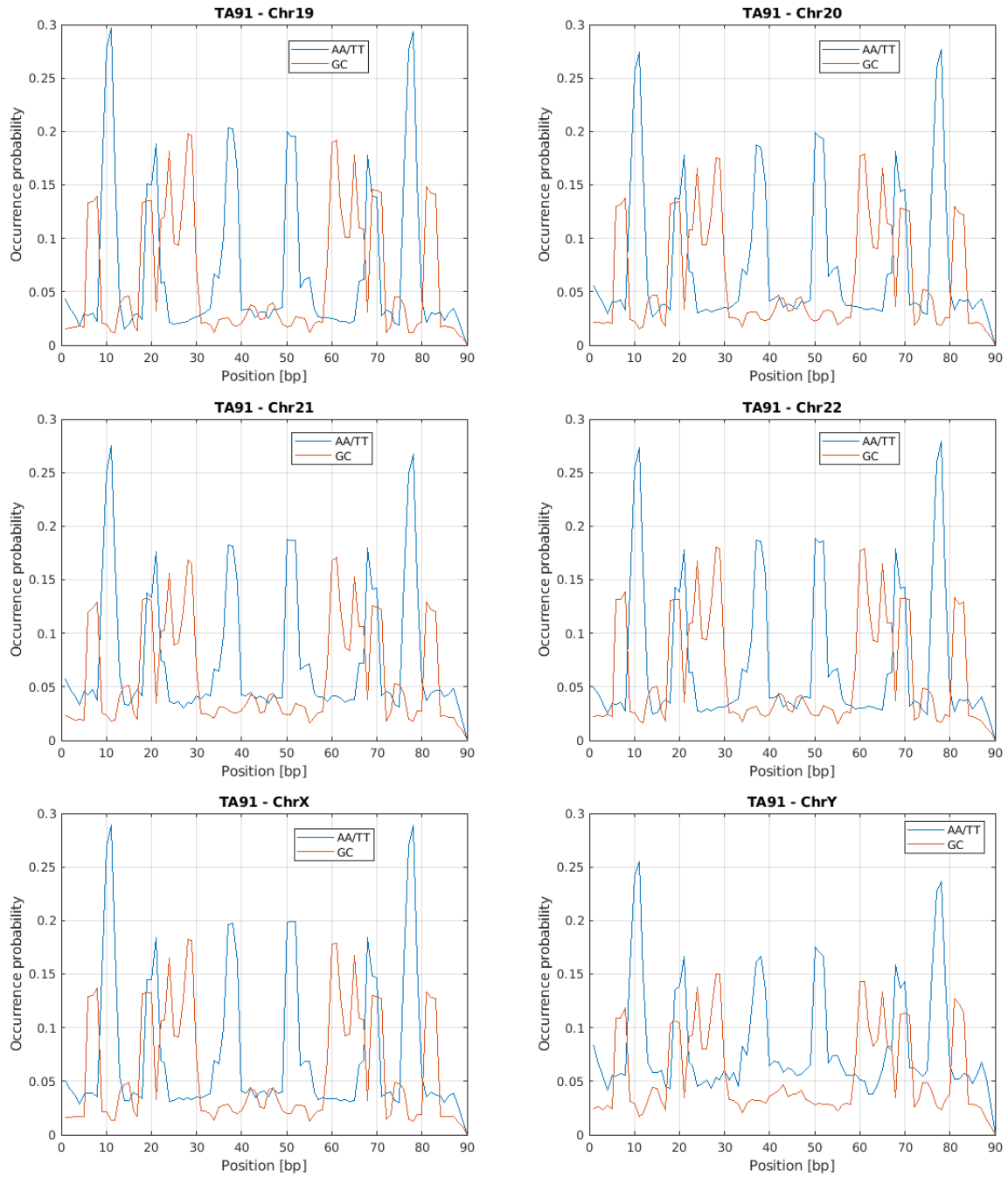


Figure 9.4: Occurrence probability as a function dinucleotide position within the sequences identified by consecutive TA separated by a distance of 91 bp within human chromosomes 19, 20, 21, 22, X and Y.

---

## Bibliography

---

- [1] C. A. Bastos et al. *Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions*. Journal of Integrative Bioinformatics, 2011.
- [2] K. M. Frahm and D. L. Shepelyansky. *Poincaré recurrences of DNA sequences*. Phys. Rev. E, 2012.
- [3] V. Afreixo et al. *Identification of DNA CpG islands using inter-dinucleotide distances*. Commun. Comput. Inform. Sci., 2015.
- [4] G. Paci et al. *Characterization of DNA methylation as a function of biological complexity via dinucleotide inter-distances*. Philosophical transactions A, 2015.
- [5] H. Moghaddasi, K. Khalifeh and A. H. Darooneh. *Distinguishing Functional DNA Words; A Method for Measuring Clustering Levels*. Scientific reports, 2017.
- [6] A. Zemach et al. *Genome-wide evolutionary analysis of eukaryotic DNA methylation*. Science, 2010.
- [7] S. Bagga. *Introduction to DNA methylation*. BioFiles, 2012.
- [8] I. Hernando-Herraez et al. *DNA methylation: insights into human evolution*. PLOS genetics, 2015.
- [9] A. A. Pai and Y. Gilad. *Comparative studies of gene regulatory mechanisms*. Curr. Opin. Genet. Dev., 2014.
- [10] T. M. Devlin. *Biochimica con aspetti clinico-farmaceutici*. EdiSES, 2013.
- [11] R. Cortini, M. Barbi and B. R. Care. *The physics of epigenetics*. Reviews of Modern Physics, 2016.



- [12] X. Zhong. *Comparative epigenomics: a powerful tool to understand the evolution of DNA methylation*. New Phytologist, 2016.
- [13] Merlotti A. et al. ‘Statistical modelling of CG interdistance across multiple organisms’. In: *BMC Bioinformatics* 19 (2018), pp. 355–362.
- [14] Erez Lieberman-Aiden et al. ‘Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome’. In: *Science* 326.5950 (2009), pp. 289–293.
- [15] Jon-Matthew Belton et al. ‘Hi-C: A comprehensive technique to capture the conformation of genomes’. In: *Methods* 58.3 (2012), pp. 268–276.
- [16] T. Nagano, Y. Lubling and T. et al. Stevens. ‘Single-cell Hi-C reveals cell-to-cell variability in chromosome structure’. In: *Nature* 502 (2013), pp. 59–64.
- [17] N. Kaplan and J. Dekker. ‘High-throughput genome scaffolding from in vivo DNA interaction frequency’. In: *Nat Biotechnol* 31 (2013), pp. 1143–1147.
- [18] J. Burton, A. Adey and R. et al. Patwardhan. ‘Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions’. In: *Nat Biotechnol* 31 (2013), pp. 1119–1125.
- [19] Ghurye J., Pop M. and Koren S. et al. ‘Scaffolding of long read assemblies using long range contact information’. In: *BMC Genomics* 18 (2017).
- [20] Jay Ghurye et al. ‘Integrating Hi-C links with assembly graphs for chromosome-scale assembly’. In: *PLOS Computational Biology* 15 (2019), pp. 1–19.
- [21] Baudry L. et al. ‘instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder’. In: *Genome Biol.* 21 (2020).
- [22] Sivan Oddes, Aviv Zelig and Noam Kaplan. ‘Three invariant Hi-C interaction patterns: Applications to genome assembly’. In: *Methods* 142 (2018), pp. 89–99.
- [23] L. Harewood, K. Kishore and M.D. et al. Eldridge. ‘Hi-C as a tool for precise detection and characterisation of chromosomal rearrangements and copy number variation in human tumours’. In: *Genome Biol* 18 (2017), p. 125.
- [24] S. M. Pires et al. ‘Attributing the human disease burden of foodborne infections to specific sources’. In: *Foodb. Pathog. Dis.* 6 (2009), pp. 417–424.
- [25] L. Mughini-Gras, E. Franz and W. Van Pelt. ‘New paradigms for Salmonella source attribution based on microbial subtyping’. In: *Food Microbiol.* 71 (2018), pp. 60–67.
- [26] S. R. Pires et al. ‘Source attribution of human salmonellosis: an overview of methods and estimates’. In: *Foodb. Pathog. Dis.* 11 (2014), pp. 667–676.

- [27] Alessandra Merlotti et al. ‘Network Approach to Source Attribution of Salmonella enterica Serovar Typhimurium and Its Monophasic Variant’. In: *Frontiers in Microbiology* 11 (2020), p. 1205.
- [28] GenBank. <http://www.ncbi.nlm.nih.gov/genbank/>. Accessed 01 February 2017.
- [29] U. Frisch and D. Sornette. *Extreme deviations and applications*. J. Phys. I France, 1997.
- [30] J. Laherrere and D. Sornette. *Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales*. Eur. Phys. J. B2, 1998.
- [31] D. Sornette. *Critical phenomena in natural sciences*. Springer, 2000.
- [32] M. E. J. Newman. *Power laws, Pareto distributions and Zipf's law*. Contemporary Physics, 2006.
- [33] R. J. Aldler, R. E. Feldman and M. S. Taqqu. *A practical guide to heavy tails: statistical techniques and applications*. Birkhauser, 1998.
- [34] C. K. Peng et al. *Long-range correlations in nucleotide sequences*. Nature, 1992.
- [35] C. K. Peng and S. V. Buldyrev. *Finite-size effects on long-range correlations: implications for analyzing DNA sequences*. Physical Review E, 1993.
- [36] L. Rossi and G. Turchetti. *Poincaré recurrences and multifractal properties of genomic sequences*. Physica A, 2004.
- [37] S. Milojević. *Power-law distributions in information science - Making the case for logarithmic binning*. JASIST, 2010.
- [38] R. Durbin et al. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, 2002.
- [39] P. A. Jones and D. Takai. *The role of DNA methylation in mammalian epigenetics*. Science, 2001.
- [40] M. J. Blow et al. *The epigenomic landscape of prokaryotes*. PLOS Genetics, 2016.
- [41] J. A. Head. *Patterns of DNA methylation in animals: an ecotoxicological perspective*. Integrative and comparative biology, 2014.
- [42] E. Sacrano et al. *The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos*. Proc. Natl. Acad. Sci. USA, 1967.
- [43] B.F. Vanyushin. *DNA methylation in plants*. Curr Top Microbiol Immunol., 2006.
- [44] A. Jeltsch. *Phylogeny of Methylomes*. Science, 2010.
- [45] ENCODE. <https://www.encodeproject.org/>. Accessed 16 March 2018.
- [46] R. Hubley et al. *The Dfam database of repetitive DNA families*. Nucleic Acids Research, 2016.

- [47] Marković D. and Gros C. ‘Power laws and self-organized criticality in theory and nature’. In: *Physics Reports* 536 (2014), pp. 41–74.
- [48] M.S. Wheatland, P.A. Sturrock and J.M. McTiernan. ‘The waiting-time distribution of solar flare hard X-ray bursts’. In: *Astrophys. J. Davidsen, J., Goltz, C.,* 509 (1998), p. 448.
- [49] G. Boffetta et al. ‘Power laws in solar flares: Self-organized criticality or turbulence?’ In: *Phys. Rev. Lett.* 83 (1999), pp. 4662–4665.
- [50] M.P. Freeman, N.W. Watkins and D.J. Riley. ‘Power law distributions of burst duration and interburst interval in the solar wind: Turbulence or dissipative self-organized criticality?’ In: *Phys. Rev. E* 62 (2000), pp. 8794–8797.
- [51] X. Yang, S. Du and J. Ma. ‘Do Earthquakes Exhibit Self-Organized Criticality?’ In: *Phys. Rev. Lett.* 92 (2004), p. 228501.
- [52] J. Davidsen and C. Goltz. ‘Are seismic waiting time distributions universal?’ In: *Geophys. Res. Lett.* 31 (2004), pp. L21612 1–4.
- [53] S. Lennartz et al. ‘Long-term memory in earthquakes and the distribution of interoccurrence times’. In: *Europhys. Lett.* 81 (2008), p. 69001.
- [54] M. Paczuski, S. Boettcher and M. Baiesi. ‘Interoccurrence Times in the Bak-Tang-Wiesenfeld Sandpile Model: A Comparison with the Observed Statistics of Solar Flares’. In: *Phys. Rev. Lett.* 95 (2005), p. 181102.
- [55] Maté Ongenaert. ‘9 - Epigenetic Databases and Computational Methodologies in the Analysis of Epigenetic Datasets’. In: *Epigenetics and Cancer, Part B*. Ed. by Zdenko Herceg and Toshikazu Ushijima. Vol. 71. Advances in Genetics. Academic Press, 2010, pp. 259–295.
- [56] Netta Mendelson Cohen, Ephraim Kenigsberg and Amos Tanay. ‘Primate CpG Islands Are Maintained by Heterogeneous Evolutionary Regimes Involving Minimal Selection’. In: *Cell* 145 (2011), pp. 773–786.
- [57] Long H.K. et al. ‘Protection of CpG islands from DNA methylation is DNA-encoded and evolutionarily conserved’. In: *Nucleic Acids Res.* 44 (2016), pp. 6693–706.
- [58] Martin W. Simmen. ‘Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals’. In: *Genomics* 92 (2008), pp. 33–40.
- [59] Alexandre V. Morozov et al. ‘Using DNA mechanics to predict in vitro nucleosome positions and formation energies’. In: *Nucleic Acids Research* 37 (2009), pp. 4707–4722.
- [60] Vinson C. and Chatterjee R. ‘CG methylation’. In: *Epigenomics* 4 (2012), pp. 655–663.

- [61] Widom J. ‘Role of DNA sequence in nucleosome stability and dynamics’. In: *Q Rev Biophys* 34 (2001), pp. 269–324.
- [62] C. Bustamante, Z. Bryant and S. Smith. ‘Ten years of tension: single-molecule DNA mechanics’. In: *Nature* 421 (2003), pp. 423–427.
- [63] Behrouz Eslami-Mossallam, Helmut Schiessel and John van Noort. ‘Nucleosome dynamics: Sequence matters’. In: *Advances in Colloid and Interface Science* 232 (2016), pp. 101–113.
- [64] E. Segal, Y. Fondufe-Mittendorf and L. et al. Chen. ‘A genomic code for nucleosome positioning’. In: *Nature* 442 (2006), pp. 772–778.
- [65] Behrouz Eslami-Mossallam et al. ‘Multiplexing Genetic and Nucleosome Positioning Codes: A Computational Approach’. In: *PLOS ONE* 11 (2016), pp. 1–14.
- [66] Vassetzky N.S. and Kramerov D.A. ‘SINEBase: a database and tool for SINE analysis’. In: *Nucleic Acids Res* 41 (2013), pp. D83–D89.
- [67] Kanhayuwa L. and Coutts R.H. ‘Short Interspersed Nuclear Element (SINE) Sequences in the Genome of the Human Pathogenic Fungus *Aspergillus fumigatus* Af293’. In: *PLoS One* 11 (2016), e0163215.
- [68] Mighell A.J., Markham A.F. and Robinson P.A. ‘Alu sequences’. In: *FEBS Letters* 417 (), pp. 1–5.
- [69] A. Yu. Grosberg, S.K. Nechaev and E.I. Shakhnovich. ‘The role of topological constraints in the kinetics of collapse of macromolecules’. In: *Journal de Physique* 49 (1998), pp. 2095–2100.
- [70] A Grosberg et al. ‘Crumpled Globule Model of the Three-Dimensional Structure of DNA’. In: *Europhysics Letters (EPL)* 23.5 (Aug. 1993), pp. 373–378.
- [71] Rosa A., Becker N.B. and Everaers R. ‘Looping probabilities in model interphase chromosomes’. In: *Biophys J.* 98 (2010), pp. 2410–2419.
- [72] Vasilyev O.A. and Nechaev S.K. ‘Topological Correlations in Trivial Knots: New Arguments in Favor of the Representation of a Crumpled Polymer Globule’. In: *Theoretical and Mathematical Physics* 134 (2003), pp. 142–159.
- [73] T. Cremer and C. Cremer. ‘Chromosome territories, nuclear architecture and gene regulation in mammalian cells’. In: *Nat. Rev. Genet.* 2 (2001), pp. 292–301.
- [74] A. Bolzer et al. ‘Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes’. In: *Plos Biol.* 3 (2005), e157.
- [75] Cremer T. and Cremer M. ‘Chromosome territories’. In: *Cold Spring Harb Perspect Biol.* 3 (2010), a003889.
- [76] Rao S.S. et al. ‘A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping’. In: *Cell* 159 (2014), pp. 1665–1680.

- [77] Edgar R., Domrachev M. and Lash A.E. ‘Gene Expression Omnibus: NCBI gene expression and hybridization array data repository’. In: *Nucleic Acids Research* 30.1 (2002), pp. 207–210.
- [78] Zimin A.V. et al. ‘The MaSuRCA genome assembler’. In: *Bioinformatics* 29 (2013), pp. 2669–77.
- [79] Marco-Sola S., Sammeth M. and Guigó R. et al. ‘The GEM mapper: fast, accurate and versatile alignment by filtration’. In: *Nat Methods* 9 (2012), pp. 1185–1188.
- [80] Santo Fortunato and Darko Hric. ‘Community detection in networks: A user guide’. In: *Physics Reports* 659 (2016), pp. 1–44.
- [81] Fortunato S. ‘Community detection in graphs’. In: *Physics Reports* 486 (2010), pp. 75–174.
- [82] Cabanettes F. and Klopp C. ‘D-GENIES: dot plot large genomes in an interactive, efficient and simple way.’ In: *PeerJ* 6 (2018), e4958.
- [83] U. von Luxburg. ‘A tutorial on spectral clustering’. In: *Statistics and Computing* 17 (2007), pp. 395–416.
- [84] Fiedler M. ‘Aggregation in graphs’. In: *Combinatorica* (1976), pp. 315–330.
- [85] Karel Devriendt and Piet Van Mieghem. ‘The simplex geometry of graphs’. In: *Journal of Complex Networks* 7.4 (2019), pp. 469–490.
- [86] Fiedler M. ‘A geometric approach to the Laplacian matrix of a graph’. In: *Combinatorial and Graph-Theoretical Problems in Linear Algebra* (1993), pp. 73–98.
- [87] Ulrike von Luxburg, Mikhail Belkin and Olivier Bousquet. ‘Consistency of spectral clustering’. In: *Annals of Statistics* 36.2 (2008), pp. 555–586.
- [88] Golub G. H. and Van Loan G. F. *Matrix Computations*. The Johns Hopkins University Press, 1996.
- [89] Kaufman L. and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ: John Wiley Sons, Inc., 1990.
- [90] François Serra et al. ‘Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors’. In: *PLOS Computational Biology* 13.7 (2017), pp. 1–17.
- [91] Alonge M. et al. ‘RaGOO: fast and accurate reference-guided scaffolding of draft genomes’. In: *Genome Biology* 20 (2019).
- [92] P. Gymoese et al. ‘Investigation of outbreaks of *Salmonella enterica* Serovar typhimurium and its monophasic variants using whole-genome sequencing, Denmark’. In: *Emerg. Infect. Dis.* 23 (2017), pp. 1631–1639.

- [93] M. Morganti et al. ‘Rise and fall of outbreak-specific clone inside endemic pulso-type of Salmonella 4,[5],12:i:-; insights from high-resolution molecular surveillance in Emilia-Romagna, Italy, 2012 to 2015’. In: *Euro. Surveill.* 23 (2019), p. 375.
- [94] F. Palma, G. Manfreda and M. et al. Silva. ‘Genome-wide identification of geographical segregated genetic markers in Salmonella enterica serovar Typhimurium variant 4’. In: *Sci Rep* 8 (2018).
- [95] M. E. J. Newman. *Networks An Introduction*. Oxford University Press, 2010.
- [96] M. Bersanelli et al. ‘Methods for the integration of multi-omics data: mathematical aspects’. In: *BMC Bioinformatics* 17 (2016), p. 15.
- [97] EFSA and ECDC. ‘The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2017’. In: *EFSA J.* 16 (2018), p. 5500.
- [98] J. Mossong et al. ‘Outbreaks of monophasic Salmonella enterica serovar 4,[5],12:i:- in Luxembourg’. In: *Euro. Surveill.* 12 (2006), p. 719.
- [99] A. Bone et al. ‘Nationwide outbreak of Salmonella enterica serotype 4,12:i:- infections in France, linked to dried pork sausage’. In: *Euro. Surveill.* 15 (2010), p. 19592.
- [100] M. E. Raguenaud et al. ‘Epidemiological and microbiological investigation of a large outbreak of monophasic Salmonella typhimurium 4,5,12:i:- in schools associated with imported beef in Poitiers, France, October 2010’. In: *Euro. Surveill.* 17 (2012), p. 20289.
- [101] L. Barco et al. ‘Molecular characterization of Salmonella enterica serovar 4,[5],12:i:- DT193 ASSuT strains from two outbreaks in Italy’. In: *Foodb. Pathog. Dis.* 11 (2014), pp. 138–144.
- [102] F. Cito et al. ‘Outbreak of unusual Salmonella enterica serovar Typhimurium monophasic variant 1,4 [5],12:i:-, Italy, June 2013 to September 2014’. In: *Euro. Surveill.* 21 (2016), p. 194.
- [103] M. de Frutos et al. ‘Monophasic Salmonella typhimurium outbreak due to the consumption of roast pork meat’. In: *Rev. Esp. Quimioter.* 31 (2018), pp. 156–159.
- [104] D. Dewey-Mattia et al. ‘Surveillance for foodborne disease outbreaks — United States, 2009-2015’. In: *MMWR Surveill. Summ.* 67 (2018), pp. 1–11.
- [105] T. Fruchterman and E. Reingold. ‘Graph drawing by force-directed placement’. In: *Softw. Pract. Exper.* 21 (1991), pp. 1129–1164.
- [106] N. Munck, P. Leekitcharoenphon and E. et al. Litrup. ‘Four European Salmonella Typhimurium datasets collected to develop WGS-based source attribution methods’. In: *Sci Data* 7 (2020), p. 75.

- [107] Eleonora Mastrorilli et al. ‘A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic Salmonella Serovar 4,[5],12:i:-’. In: *Frontiers in Microbiology* 9 (2018), p. 715.
- [108] L. Mughini-Gras et al. ‘Attribution of human Salmonella infections to animal and food sources in Italy (2002–2010): adaptations of the Dutch and modified Hald source attribution models’. In: *Epidemiol. Infec.* 142 (2014), pp. 1070–1082.
- [109] EFSA and ECDC. ‘The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2013’. In: *EFSA J.* 13 (2015), p. 3991.
- [110] EFSA and ECDC. ‘The European Union summary report on trends and sources of zoonoses, zoonotic agents and food-borne outbreaks in 2014’. In: *EFSA J.* 13 (2015), p. 4329.
- [111] L. Barco et al. ‘Salmonella source attribution based on microbial subtyping’. In: *Int. J. Food Microbiol.* 163 (2013), pp. 193–203.