Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Corso di Laurea Magistrale in Fisica

Statistical methods for the analysis of DNA sequences: application to dinucleotide distribution in the human genome

Relatore:
Prof.
Daniel Remondini

Correlatore:
Dott.
Giampaolo Cristadoro

Presentata da:
Giulia Paci

Sessione II
Anno Accademico 2013/14

Abstract

Questa tesi si inserisce nell'ambito delle analisi statistiche e dei metodi stocastici applicati all'analisi delle sequenze di DNA. Nello specifico il nostro lavoro è incentrato sullo studio del dinucleotide CG (CpG) all'interno del genoma umano, che si trova raggruppato in zone specifiche denominate CpG islands. Queste sono legate alla metilazione del DNA, un processo che riveste un ruolo fondamentale nella regolazione genica. La prima parte dello studio è dedicata a una caratterizzazione globale del contenuto e della distribuzione dei 16 diversi dinucleotidi all'interno del genoma umano: in particolare viene studiata la distribuzione delle distanze tra occorrenze successive dello stesso dinucleotide lungo la sequenza. I risultati vengono confrontati con diversi modelli nulli: sequenze random generate con catene di Markov di ordine zero (basate sulle frequenze relative dei nucleotidi) e uno (basate sulle probabilità di transizione tra diversi nucleotidi) e la distribuzione geometrica per le distanze. Da questa analisi le proprietà caratteristiche del dinucleotide CpG emergono chiaramente, sia dal confronto con gli altri dinucleotidi che con i modelli random. A seguito di questa prima parte abbiamo scelto di concentrare le successive analisi in zone di interesse biologico, studiando l'abbondanza e la distribuzione di CpG al loro interno (CpG islands, promotori e Lamina Associated Domains). Nei primi due casi si osserva un forte arricchimento nel contenuto di CpG, e la distribuzione delle distanze è spostata verso valori inferiori, indicando che questo dinucleotide è clusterizzato. All'interno delle LADs si trovano mediamente meno CpG e questi presentano distanze maggiori. Infine abbiamo adottato una rappresentazione a random walk del DNA, costruita in base al posizionamento dei dinucleotidi: il walk ottenuto presenta caratteristiche drasticamente diverse all'interno e all'esterno di zone annotate come CpG island. Riteniamo pertanto che metodi basati su questo approccio potrebbero essere sfruttati per migliorare l'individuazione di queste aree di interesse nel genoma umano e di altri organismi.

II

Contents

Introduction

There's real poetry in the real world.
Science is the poetry of reality.
-Richard Dawkins

Methylation of DNA is one of the most important epigenetic processes, sometimes referred to as the "fifth base" for its key role in functions such as gene regulation. CpG islands are important genomic regions, which are believed to be protected from methylation but can also show aberrant methylation patterns in different diseases, including cancer; it is therefore important to obtain "maps" of these regions in order to develop appropriate assays to monitor DNA methylation where this could be linked to diseases.
However, since the first formal definition of a CpG island, given by Gardner, Gardiner and Frommer in 1987 [18], many alternative definitions and algorithms for CGI annotation have been proposed (see for example [42] and [24]), and the matter is still much debated.

As physicists approaching a biological problem new to us, we decided to take a step back and first of all aim at characterising in detail the positioning of CpG dinucleotides in the whole genome, comparing them with the different dinucleotides and with appropriate random models. To do this, we employed a method based on the computation of distances between successive occurrences of dinucleotides and we considered a zeroth and first order Markov chain models for random reference sequences.

The first part of our work is focused on a global analysis of dinucleotides in the whole human genome: we compare the relative frequencies and the distance distributions of all 16 dinucleotides with each other and with reference sequences generated with both a Markov 0 and Markov 1 models, as well as a null model of distance distributions given by the geometric distribution. We then focus on genomic regions of interest, specifically CpG islands (we compare two different annotations), promoters and LADs (Lamina Associated Domains, regions of DNA which bind to the nuclear membrane and control the three dimensional structure of the DNA sequence) to gather insight on the mutual relationship between these areas and their CpG content and distributions. In the final part of this thesis we employ a random walk based representation of the DNA sequence, which is constructed based on dinucleotide positioning, and we discuss the possibility of using simple methods based on DNA walks for the search of candidate CpG islands.

In the first chapter we introduce the biological background and motivation of this work, with a focus on DNA methylation and the concept of CpG islands. The second chapter is devoted to the main methods employed in our study, namely Markov chain models, inter-dinucleotide distance distribution analysis and the random walk representation of DNA. In the third chapter we illustrate the results of the characterisation of dinucleotide abundances and distribution in the whole genome. The fourth chapter is focused on the analysis inside genomic regions of interest and on the results of our DNA walk representation applied to CpG islands.

CHAPTER 1

CpG islands and DNA methylation

In this chapter we briefly review the biological background of this work. In the first two sections we give an overview of the human genome and of the most important sequencing techniques which allow the analysis of the DNA bases sequence. We then give a brief description of the methylation process in DNA, which is an important epigenetic modification and the main reason for the interest in CpG dinucleotides, the subject of this study. In the last section we describe the main definitions of CpG islands and we give a survey of methods and algorithms for CpG islands annotation present in literature.

1.1 The human genome

In all living organisms hereditary information is stored, transmitted and expressed with the help of nucleic acids DNA (deoxyribonucleic acid) and RNA (ribonucleic acid). DNA is the genetic material that organisms inherit from their parent: it provides directions for its own replication, directs RNA synthesis and, through RNA, controls protein synthesis.

DNA and RNA composition and structure

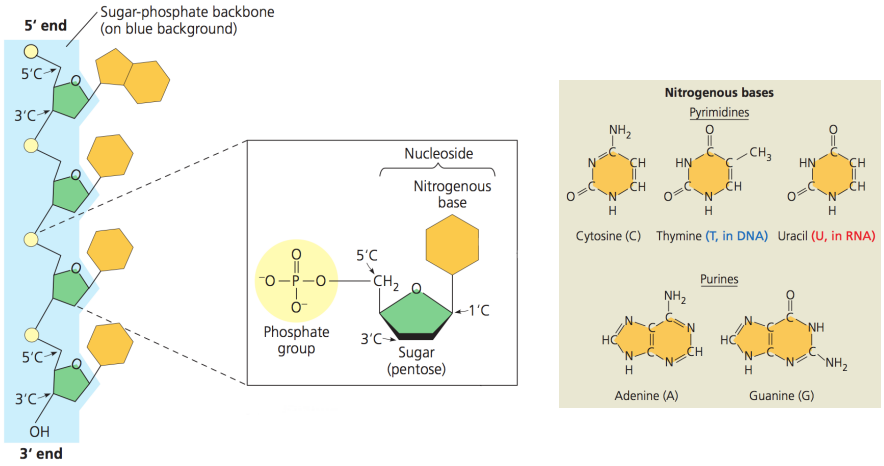
Nucleic acids are polymers, constituted of monomers called nucleotides. Each nucleotide is composed of a phosphate group, a sugar (deoxyribose in DNA, ribose in DNA) and a nitrogenous bases. There are two families of

nitrogenous bases: pyrimidines and purines. The first ones include cytosine, thymine and uracil (which is present in RNA instead of T) and are characterized by one six-membered ring of carbon and nitrogen atoms. Purines, which are larger molecules formed by a six-membered ring fused to a five-membered ring, include adenine and guanine (see Figure 1.1b). In the polymer structure, adjacent nucleotides are joined by a phosphodiester bond: a phosphate group links the sugars of two nucleotides. This bonding results in a backbone with a repeating pattern of sugar-phosphate units characterized by an intrinsic directionality: one end has a phosphate attached to the 5' carbon, whereas the other has a hydroxyl group on a 3' carbon (see Figure 1.1a).

RNA molecules usually exist as single polynucleotide chains, on the other hand DNA molecules have two polynucleotides, or strands, that spiral around an imaginary axis, forming a double helix structure (see Figure 1.2). The two strands are antiparallel, with sugar-phosphate backbones on the outside of the double helix running in the 5'-3' direction opposite from each other. The nitrogenous bases are paired in the interior of the helix, and hydrogen bonds between them hold the two strands together (see Figure1.2). Only certain bases in the double helix are compatible with each other: adenine always pairs with thymine by two hydrogen bonds and cytosine with guanine by three hydrogen bonds. The two strands are therefore complementary: if we were to read the sequence of bases along one strand of the double helix, we would know the sequence of bases along the other strand. This unique feature of DNA allows the creation of two identical copies of each DNA molecule in a cell that is preparing to divide, making daughter cells genetically identical to the parent. Base pairing also occurs in RNA among bases in two different RNA molecules or on the same molecule: for example it is responsible for the three-dimensional functional structure of transfer RNA.

1.2 DNA sequencing: history and techniques

The goal of DNA sequencing is the determination of the precise order of nucleotides within a DNA molecule: this knowledge is essential for the advancement of biological and medical research, and can help the understanding of many diseases, for example with the identification of oncogenes and mutations linked to different forms of cancer. The main efforts towards the sequencing of the Human Genome started in 1990 with the launch of the Human Genome Project through funding from the the National Institutes of Health (NIH) and



(a) (b)

Figure 1.1: (a) Structure of a polynucleotide and nucleotide components. (b) Nitrogenous bases. Figures from [43].

the US Department of Energy, whose labs joined with international partners in a quest to sequence all 3 billion letters, or base pairs, in the human genome in just 15 years. A first draft of the human genome was released in 2000, and in 2003 the human DNA sequence was deemed complete[7]. This version of the human genome actually consists of 99 percent of the gene-containing sequence, with the missing parts essentially contained in less than 400 defined gaps. Since then, the Genome Reference Consortium (GRC) has been working to improve the sequence by closing gaps, fixing errors and representing complex variation. Sequencing technologies played a vital role in the Human Genome Project, and the project itself stimulated the development of new, faster and cheaper methods. We will now briefly review the most important sequencing techniques, from Sanger sequencing, which was widely exploited by the HGP, to the so called next-generation methods.

Early DNA sequencing technologies

Early efforts at DNA sequencing were extremely labor intensive and time consuming (e.g the Gilbert & Maxam technique), and a huge improvement occurred around mid 1970 with the methods developed by Sanger (who later received the Nobel Prize in Chemistry in 1980) and his colleagues. Sanger sequencing is also called chain-termination method and it was the most widely used sequencing technique for approximately 25 years.

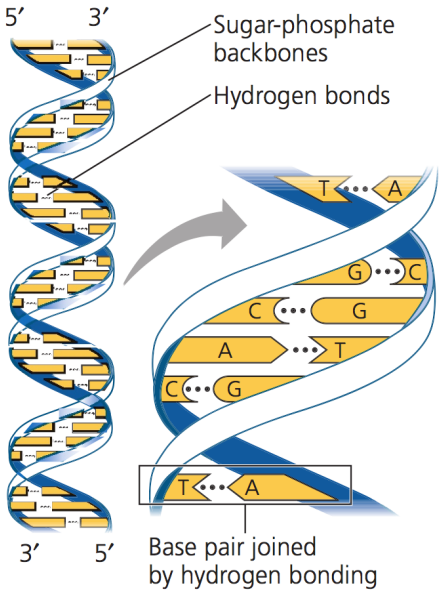


Figure 1.2: Structure of the DNA double helix. Figure from [43].

This method is characterized by the use of dideoxynucleotides triphosphates (ddNTP) which lack the 3' hydroxyl (OH) group needed to form the phosphodiester bond between nucleotides along the strand of DNA: due to this unique feature, incorporation of a dideoxynucleotide in the growing strand inhibits further strand extension. In the standard Sanger process, four parallel sequencing reactions are used for a single sample: each reaction involves a single-stranded template, a specific primer, the four standard deoxynucleotides and DNA polymerase. The polymerase adds bases to a DNA strand that is complementary to the single-stranded sample template. One of the four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is then added to each reaction at a lower concentration than the standard deoxynucleotides. Because the dideoxynucleotides lack the 3' OH group, whenever they are incorporated by DNA polymerase, the growing DNA terminates. Four different ddNTPs are used such that the chain doesn't always terminate at the same nucleotide (i.e., A, G, C, or T). This produces a variety of strand lengths for analysis. Then, by putting the resulting samples through four columns on a gel (according to which dideoxynucleotide was added), researchers can see the fragments line up by size and know which base is at the end of each fragment. If the four terminators are labelled with fluorescent dyes, each of which emit light at different wavelengths, sequencing can be performed in a single reaction (dye-terminator sequencing). Due to its greater expediency and speed, this method is now the mainstay in automated

sequencing.

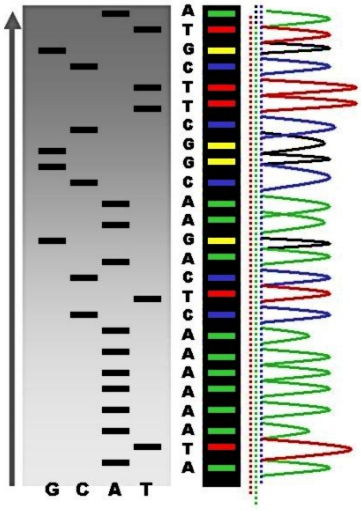


Figure 1.3: Results of traditional (on the left) and dye-terminator (on the right) Sanger sequencing. Figure from [6].

Next-generation sequencing technologies

Since around 2005 there has been a gradual shift away from automated Sanger sequencing towards completely new methods. In fact, despite many technical improvements, the limitations of automated Sanger sequencing showed a need for new and improved technologies for sequencing large numbers of human genomes. Newer methods are commonly referred to as next-generation sequencing (NGS), and include different techniques for template preparation, sequencing and imaging, and data analysis. The most widespread methods include 454-Pyrosequencing, Illumina/Solexa and SOLiD: these are all different implementation of cyclic-array sequencing, which is the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and image-based data collection.

The major advance offered by NGS is the ability to produce an enormous volume of data cheaply: this allows for example performing large-scale comparative and evolutionary studies by sequencing the whole genome of related organism, and resequencing of human genomes to improve our understanding of the impact of genetic differences on health and disease.

Pyrosequencing-454 The first next-generation DNA sequencer on the market was the GS20 machine, released in 2005 by the 454 Life Sciences Company (now owned by Roche Diagnostics). Pyrosequencing relies on the luminometric detection of pyrophosphate that is released during primer-directed DNA polymerase catalyzed nucleotide incorporation.

The workflow of 454 sequencing can be summarized as follows (see also Figures 1.4 and 1.5):

- Library preparation: the double stranded DNA is broken into short segments, which are joined with an adaptor at either end. The fragments are then separated into single stranded DNA and joined with micro-sized beads.

- Emulsion formation: the DNA-bead complexes are mixed with emulsion oil, so that the water forms droplets around the beads, called an emulsion.

- Emulsion PCR: each droplet contains only one DNA molecule, which is amplified producing million copies of each DNA fragment on the surface of each bead.

- Beads loading: the droplets are broken and the beads are loaded into a picoliter plate, designed such that one well only fits one bead ($\sim 28\mu m$ in diameter). Smaller beads are also added, bearing immobilized enzymes also required for pyrosequencing (ATP sulfurylase and luciferase).

- Pyrosequencing: the sequencing reagents are delivered across the wells of the plate. These include ATP sulfurylase, luciferase, apyrase, the substrates adenosine 5 phosphosulfate (APS) and luciferin and the four deoxynucleoside triphosphates (dNTPs). The latter are added sequentially in a fixed order during a sequencing run. During the nucleotide flow, millions of copies of DNA bound to each bead are sequenced in parallel. When a nucleotide complementary to the template strand is added into a well, the polymerase extends the existing DNA strand by adding nucleotide(s). Via ATP sulfurylase and luciferase, incorporation events immediately drive the generation of a burst of light, which is detected by the CCD camera as corresponding to the array coordinates of specific wells. Across multiple cycles (e.g., A-G-C-T-A-G-C-T...), the pattern of detected incorporation events reveals the sequence of templates represented by individual beads. The signal strength is proportional to the number of nucleotides: homopolymer stretches, incorporated in a single

nucleotide flow, generate a greater signal than single nucleotides. However, the signal strength for homopolymer stretches is linear only up to eight consecutive nucleotides after which the signal falls-off rapidly.

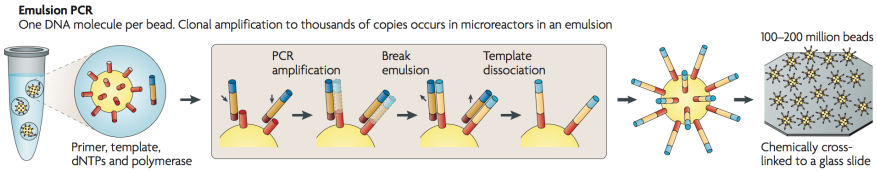


Figure 1.4: Emulsion PCR: bead-DNA complexes are incapsulated into single aqueous droplets, PCR amplification is performed within these droplets to create beads containing several thousand copies of the same template sequence. The beads can than be chemically attached to a glass slide or deposited into PicoTiterPlate wells. Figure from [28].

Illumina-Solexa This sequencing technique is based on the inventions of S. Balasubramanian and D. Klenerman of Cambridge University, who subsequently founded Solexa, a company later acquired by Illumina. In this method, DNA fragments are amplified by bridge PCR which produces local clusters and sequencing occurs by addition of fluorescently labeled reversible terminate bases, which compete for binding sites on the template DNA to be sequenced, and are detected by laser excitation.

The workflow of Illumina can be summarized as follows (see also Figures 1.6 and 1.7):

- Library preparation: the DNA samples are sheared into a random library of 100-300 base-pair long fragments. After fragmentation the ends of the obtained DNA-fragments are repaired and an A-overhang is added at the 3'-end of each strand. Then, adaptors which are necessary for amplification and sequencing are ligated to both ends of the DNA-fragments. These fragments are then selected according to their size and purified.

- Cluster generation: single DNA-fragments are attached to the flow cell by hybridization to oligos on its surface that are complementary to the ligated adaptors. The DNA molecules are then amplified by a so-called bridge amplification which results in a hundred of millions of unique clusters. Finally, the reverse strands are cleaved and washed away and the sequencing primer is hybridized to the DNA-templates.

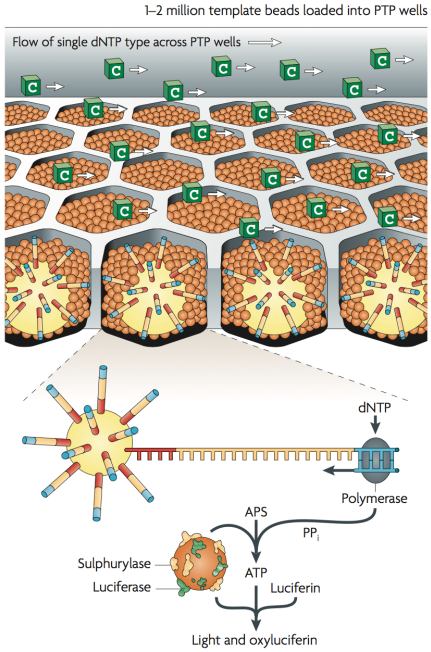


Figure 1.5: Pyrosequencing: DNA-amplified beads are loaded into individual PicoTiterPlate wells, and additional beads, coupled with sulphurylase and luciferase, are added. In this example, a single type of 2-deoxyribonucleoside triphosphate (dNTP) is shown flowing across the wells. The fibre-optic slide is mounted in a flow chamber, enabling the delivery of sequencing reagents to the bead-packed wells. The underneath of the fibre-optic slide is directly attached to a high-resolution CCD camera, which allows detection of the light generated from each well undergoing the pyrosequencing reaction. Figure from [28].

- Sequencing: the DNA templates are copied base by base using the four nucleotides (ACGT) which are fluorescently-labeled and reversibly terminated. After each synthesis step, the clusters are excited by a laser which causes fluorescence of the last incorporated base. After that, the fluorescence label and the blocking group are removed allowing the addition of the next base. The flourescence signal after each incorporation step is captured by a built-in camera, producing images of the flow cell.

SOLiD The SOLiD technology, developed by Life Technologies, has been commercially available since 2006 and is based on sequencing by ligation, which is driven by a DNA ligase rather than a polymerase. In this method, clonal sequencing features are generated by emulsion PCR, with amplicons captured

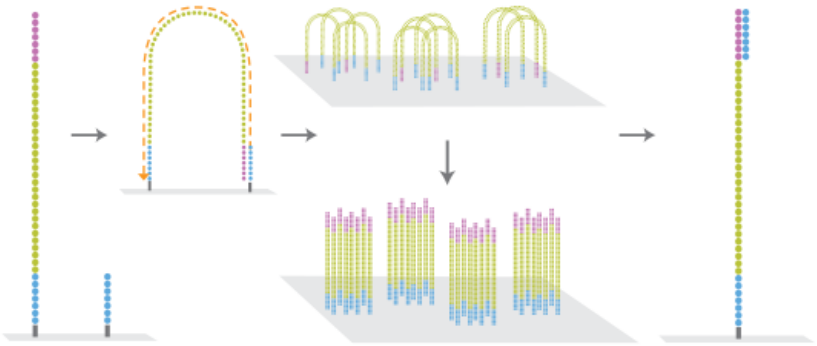


Figure 1.6: Attachment of DNA fragments to the flow cell, bridge amplification with formation of clusters, cleavage of reverse strands and hybridization of the sequencing to the DNA-templates. Figure from [4].

to the surface of $1\mu m$ paramagnetic beads. After breaking the emulsion, the beads bearing amplification products are selectively recovered and immobilized to a solid planar substrate in order to generate a dense, disordered array. Each sequencing cycle needs the following: a bead, a degenerate primer (which can bind all the four bases), a ligase and four dNTP 8-mer probe. The latter are eight bases in length, with a free hydroxyl group at the 3' end, a fluorescent dye at the 5' end and with a cleavage site between the fifth and sixth nucleotide. The first two bases (starting at the 3' end) are complementary to the nucleotides being sequenced, while bases 3 through 5 are degenerate and able to pair with any nucleotides on the template sequence. After ligation, images are acquired in four channels, collecting data for the same base positions across all template-bearing beads. Finally, the fluorescent label is removed by cleavage of the 8-mer between positions 5 and 6. Several cycles as the one described will iteratively interrogate an evenly spaced, discontiguous set of bases.

1.3 Genome assembly

For all the different sequencing techniques described in the previous section DNA is sequenced in small pieces, called "reads" which then need to be aligned and merged in order to reconstruct the original sequence. This problem is often compared to the one of reconstructing the text in a book just by looking at the shredded pieces obtained by fragmenting many copies of the same book. In the case of a genome assembly there are further issue to consider, such as the presence of repeats which are especially difficult to reconstruct, and the possible

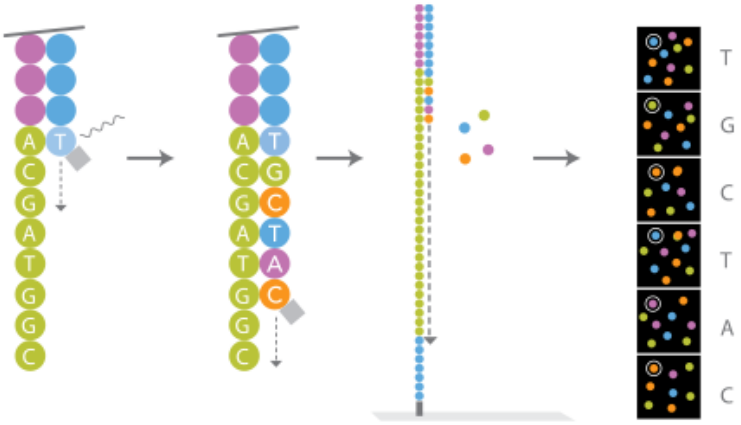


Figure 1.7: The first base is extended, read and deblocked; the above step is repeated on the whole strand; the fluorescent signals are read. Figure from [4].

occurrence of errors and gaps in the sequencing.
Two different types of assembly tasks can be distinguished:

- *De novo*: the goal is to reconstruct a whole sequence (which may even be novel) by assembling the reads.
- Mapping: the reads are assembled against a reference backbone sequence, building a sequence that is similar but not necessarily identical to the reference one.

The first task is evidently much more complex and computationally intensive than the second one, therefore we will focus on *de novo* assembly algorithms. In the main step of a *de novo* assembly task, an *assembler* program reconstructs sequences, based on sequence similarity between reads.

We will now describe the most important assembly methods for the reconstruction of genomic sequences.

Shortest common superstring - greedy method

One of the first approaches to genome assembly formulates the problem as one of finding the shortest common superstring of a set of sequences: namely, given a collection of strings S find $SCS(S)$ which contains all strings in S as substrings. Without the requirement of obtaining the shortest superstring, the problem is trivial and the solution would be to simply concatenate the strings.

We can picture this task with an overlap graph such as the one shown in Figure 1.8, where each node represents a read and two nodes are connected by a directed edge if there is an overlap between the suffix of the source (first node) and the prefix of the sink (second node) by at least l characters. It is also possible to construct a weighted graph, in which each edge weight corresponds to the length of the overlap.

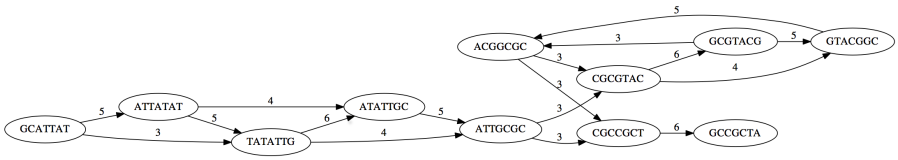


Figure 1.8: Example of an overlap graph with $l = 3$: the edge labels represent the overlap lengths. Figure from [5].

In this framework the SCS problem corresponds to finding a path that visits every node once, minimizing the total cost (which can be pictured as minus the edge weight - or length of the overlap) along the path: this is the Travelling Salesman problem, which is known to be NP-hard. If we simplify the problem and do not consider edge weights, only looking for a path that visits all the nodes exactly once, we have an Hamiltonian Path problem, which is still NP-complete. If we give up on finding the shortest possible superstring, suboptimal solutions can be found with a greedy algorithm: all possible overlaps between the strings are computed and a score is assigned to each potential overlap. The algorithm then merges strings iteratively by combining those strings whose overlap has the highest score, and the procedure continues until no more strings can be merged. This method is easy to implement but it ignores long-range relationships between reads, which could be useful in detecting and resolving repeats and its computational cost limits its applicability only to short genomes (such as the ones of bacteria). In order to overcome these limitations new algorithms have been developed, which are more tractable and avoid collapsing repeats. Two of the most widely employed approaches are the overlap-layout-consensus method and the De Bruijn graph method, both of which exploit different techniques developed in the field of graph theory. Unresolvable repeats are treated by leaving them out: therefore with these methods we do not obtain a whole assembly but fragments called “contigs” which must be further assembled by a *scaffolding* program.

Overlap-layout-consensus (OLC) method

In this method, the following graph representation of the assembly problem is adopted: a node corresponds to a read, an edge denotes an overlap between two reads. In this framework, each contig is represented as a path through the graph that contains each node at most once. The main steps of the OLC algorithm are:

- Overlap: Construction of the overlap graph by computing all the possible alignments between reads. Different approaches, more or less efficient, are possible (dynamic programming, suffix trees);
- Layout: "clean up" of the graph by removing transitive edges and resolving ambiguities (i.e due to sequencing errors);
- Consensus: generation of a *consensus* sequence for each contig by constructing the multiple alignment of the reads that is consistent with the chosen path (for example by majority vote).

The "clean up" phase of the algorithm includes removing transitively-inferrible edges that skip one or more nodes such as the one pictured in Figure 1.9.

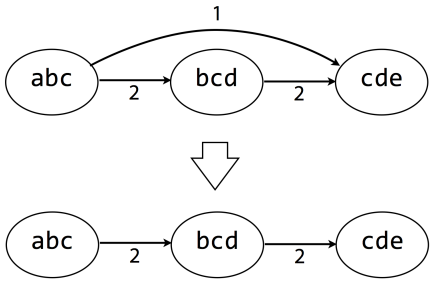


Figure 1.9: Example of transitively-inferrible edges removal. Figure from [5].

The OLC assembly method is recommended when there is a limited number of reads but significant overlap, on the other hand it is computationally intensive for short reads. An example of assembler based on this OLC paradigm is the Celera Assembler, developed at Celera Genomics for the first *Drosophila* whole genome shotgun sequence[31] and the first diploid sequence of an individual human[27]. The development of this assembler later continued as an open source project at the J.Craig Venter Institute (available at [2]).

De Bruijn graph method

This method is inspired by Sequencing by Hybridization (SBH) technique, in which an unknown fragment of the single stranded DNA labelled either fluorescently or radioactively is hybridized to a DNA chip that holds the oligonucleotide library to be used. A reduction of SBH to an easy-to-solve Eulerian Path Problem in the de Bruijn graph was proposed in 1989 by Pevzner[35] and an algorithm for DNA fragment assembly (EULER) based on a similar method was later described by Pevzner and others (see [34]). The first step of this algorithm consists in the construction of a de Brujin graph, as follows:

- Pick a substring length k;
- take each k mer and split into left and right $k-1$ mers;
- add $k-1$ mers as nodes to de Bruijn graph (if not already there), and add an edge from the left $k-1$ mer to right $k-1$ mer

The result of these steps is a directed multigraph such as the one depicted in Figure 1.10 (the multiple edges are represented there as numbers) in which a given sequence of length k–1 can appear only once as a node. The assembly problem is now translated into finding a path that uses all the edges: according to Euler's theorem this Eulerian path exists if the graph is balanced (the indegrees are equal to the outdegrees for all nodes). If the sequencing is perfect the de Brujin graph must be balanced, as the node for $k-1$mer from the left end is semi-balanced with one more outgoing edge than incoming and the node for $k-1$mer at the right end is semi-balanced with one more incoming than outgoing (all the other nodes are balanced). The Eulerian walk can be found in a time proportional to the number of edges, thus this method is is computationally far more efficient compared with the OLC. The problem is more complex when analysed from a less idealised perspective: repeats yield different possible walks and sequencing errors can make the graph non-Eulerian, therefore additional refinements and error correction steps are required. An improved formulation is the de Bruijn Superwalk Problem (DBSP), in which we seek a walk over the De Bruijn graph, and the walk contains each read as a subwalk; but this problem has been proven to be NP-hard[25].

Scaffolding

We have seen that boh the OLC and the de brujin graph methods give rise to stretches of unambiguously assembled sequence called "contigs", which must

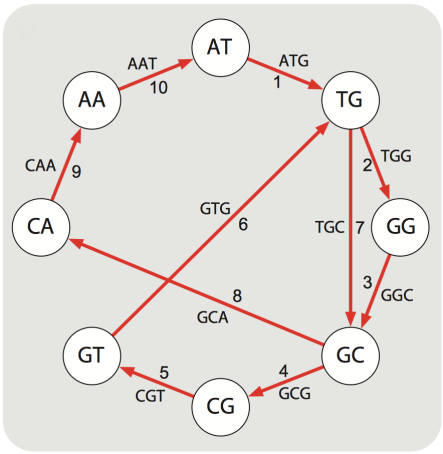


Figure 1.10: Example of a de Brujin graph in which nodes represent k-mer prefixes and suffixes and edges represent k-mers having a particular prefix and suffix. For example, the k-mer edge ATG has prefix AT and suffix TG. Figure from [13].

then be further assembled in a continuous sequence. The main task of a scaffolder is to orient and order contains with respect to each other. In paired-end sequencing we have a pair of reads taken from either end of a longer fragment: these *mates* might overlap in the middle of the fragment as shown in Figure 1.11. These "spanning pairs" can be exploited to get information about the relative orientation and ordering of contigs. Another strategy is to use the sequence of a closely related organism as another source of scaffolding information.

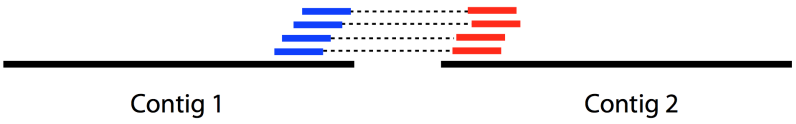


Figure 1.11: Example of spanning pairs between two contigs. Figure from [5].

Assembly quality assessment

Assessing the quality of assembled genomes is a very difficult task, as it needs to take into account different aspects and because the concept of "assembly quality" itself largely depends on the specific cases. Tools to evaluate and compare quantitatively the completeness and exactness of assembled genomes

are nevertheless essential and a lot of effort in this direction has been made in the last years thanks to collaborative projects such as the *Assemblathon* competition [1]. Some of the most widely employed quality metrics are:

- Number of contigs: in most cases, a low number is preferred.
- The total sum of bases in all contigs: ideally, this number should be close to the expected size of the target sequences (i.e. the size of the target genome for whole genome sequencing).
- Number of gaps: in most cases, a low number is preferred.
- N50 statistics: defined as the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter also contains at least half of the sum of the lengths of all contigs.

As an example, a comparison of N50 statistics can be found in Table 1.1 for the human genome release hg19 (employed in this work) and hg38 (recently released)[1].

N50 for all chromosomes	hg19 release	hg38 release
Placed scaffolds	46,395,641	70,114,165
Unplaced scaffolds	172,149	176,845
All scaffolds	46,395,641	67,794,873

Table 1.1: Quality measures for two different human genome releases. Data taken from the Human Reference Consortium [3].

Comparison of different assembly and scaffolding algorithms can also be made by "benchmarking" them against predefined tasks and evaluating their capability to correctly resolve repeats and deal with sequencing errors.

1.4 DNA methylation

DNA methylation is one of the most important epigenetic modifications, which are defined as "mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence"[38]. In mammals, DNA methylation is essential for normal development and evidence of

[1] Unplaced scaffolds are scaffolds for which the chromosome they belong to is currently unknown.

alterations in methylation profiles in cancer and aging have brought a lot of attention on the study of this epigenetic modification, which is sometimes referred to as a "fifth base" in the DNA sequence because of its importance.

The process of DNA Methylation

Methylation consists in the addition of a methyl group (containing one carbon atom bonded to three hydrogen atoms -CH3): in the DNA of vertebrates it typically occurs at CpG sites (cytosine-phosphate-guanine sites, that is, where a cytosine is directly followed by a guanine in the DNA sequence). This process results in the conversion of the cytosine to 5-methylcytosine (see Figure 1.12). Methylated C residues spontaneously deaminate to form T residues over time; hence CpG dinucleotides steadily deaminate to TpG dinucleotides, which is evidenced by the under-representation of CpG dinucleotides in the human genome[12]. The remaining CpG sites are spread out across the genome where they are heavily methylated with the exception of CpG islands (stretches of DNA that have a higher CpG density than the rest of the genome, and will be described in detail in the next section).

DNA methylation is catalyzed by a family of DNA methyltransferases (Dnmts), which share a similar structure with a large N-terminal regulatory domain and a C-terminal catalytic domain, but have unique functions and expression patterns. Dnmt3a and Dnmt3b can establish a new methylation pattern to unmodified DNA and are thus known as "*de novo*" Dnmt; on the other hand Dnmt1 functions during DNA replication to copy the DNA methylation pattern from the parental DNA strand onto the newly synthesized daughter strand (see Figure 1.13). Additionally, Dnmt1 also has the ability to repair DNA methylation[30]. For these reasons, Dnmt1 is called the "maintenance" Dnmt because it maintains the original pattern of DNA methylation in a cell lineage: without it, the replication machinery would produce daughter strands that are unmethylated and, over time, this would lead to passive demethylation.

DNA methylation and gene regulation

DNA methylation is a major epigenetic factor influencing gene activities, and it can repress transcription both directly and indirectly. In the first case, it can physically impede the binding of transcription factors to the gene, whereas in the second methylated DNA may be bound by methyl-CpG-binding domain proteins which recruit additional proteins that modify histones forming compact, inactive heterochromatin. In addition to this, DNA methylation also cooperates

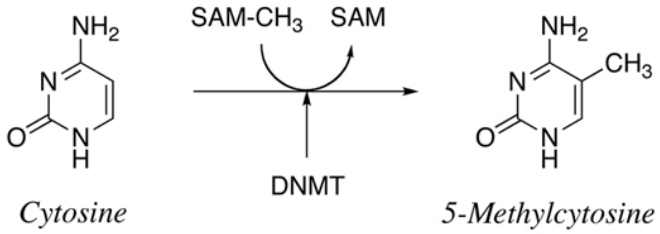


Figure 1.12: Conversion of cytosine in 5-methylcytosine.

with histone modifications to impose a repressive state on a gene region.

Due of its vital role in gene regulation, a correctly established DNA methylation is essential for the normal development and functioning of organisms, and an increasing number of diseases have been found to be associated with aberrant methylation patterns: among these, cancer is one of the most studies one. The first evidence of a link between DNA methylation and cancer was demonstrated in 1983, when it was shown that the genomes of cancer cells are hypomethylated relative to their normal counterparts[16]. Tumor cells, in fact, are characterized by a loss of methylation in the repetitive regions of the genome, which causes genomic instability. In addition to this genome-wide demethylation, gene-specific hypermethylation events are also observed in cancer: these typically occur at CpG islands, most of which are not methylated in normal somatic cells, and result in a silenced transcription (see Figure 1.14). For example, genes involved in cell-cycle regulation, tumour cell invasion, DNA repair, chromatin remodelling, cell signalling, transcription and apoptosis are known to become aberrantly hypermethylated and thus silenced in nearly every tumour type (see [37], supplementary information S2 and S3).

1.5 CpG islands

In the previous section, we have seen the relevance of CpG sites as methylation targets. We have also introduced the concept of CpG islands as CpG-rich stretches of DNA which are usually unmethylated and can become methylated in diseases such as cancer. We will now discuss the biological importance of CpG islands in more detail and review possible definitions of CpG islands that have been proposed in literature, as well as the corresponding problematics.

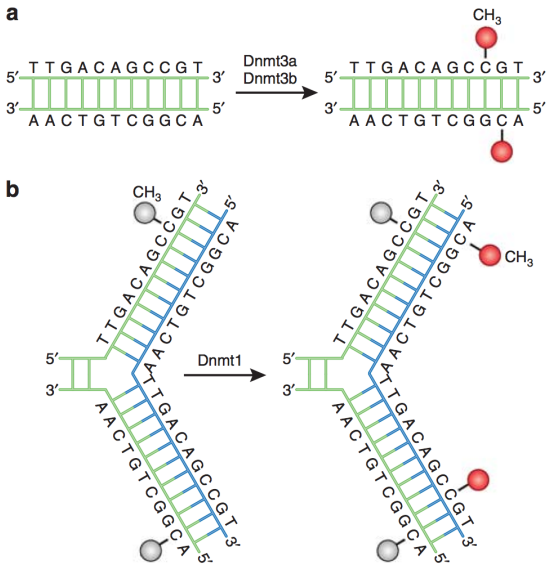


Figure 1.13: Dnmt3a and Dnmt3b are the *de novo* Dnmts and transfer methyl groups (red) onto naked DNA. (b) Dnmt1 is the maintenance Dnmt: when DNA undergoes semiconservative replication, the parental DNA stand retains the original DNA methylation pattern (gray). Dnmt1 associates at the replication foci and precisely replicates the original DNA methylation pattern by adding methyl groups (red) onto the newly formed daughter strand (blue). Figure from [29].

Biological relevance of CpG islands

Interest in CGIs first grew in the 1980s when it was demonstrated that, in vertebrates, they are enriched in regions of the genome involved in gene transcription referred to as "promoters"[10]. Saxonov and others[40] found in 2005 that promoters could be classified in two classes according to their CpG content: 72% of promoters belong to the class with high CpG content, and 28% are in the class whose CpG content is characteristic of the overall genome (low CpG content). In addition to this, CGIs have been shown to colocalize with the promoters of all constitutively expressed genes and approximately 40% of those displaying a tissue restricted expression profile[26].

As far as the methylation status of CpG islands is concerned, we have already discusses altered DNA methylation of CGIs (which are usually unmethylated) in development and cancer (see ref[17]). A study of CGI methylation on chromosome 21 by Yamada and others[46] found that although most CGIs (103 out of 149) escape methylation, a sizable fraction (31 out of 149) are fully methylated

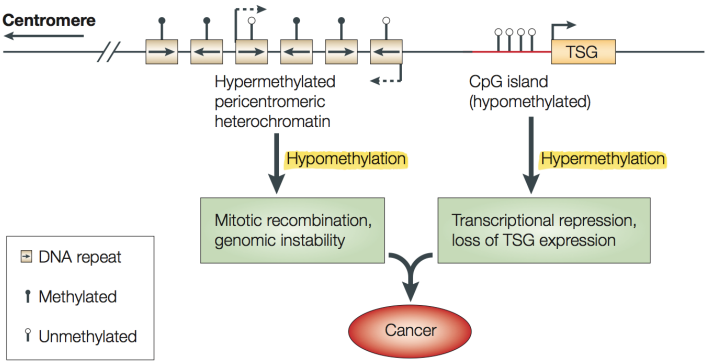


Figure 1.14: The diagram shows a representative region of DNA in a normal cell. The region shown contains repeat-rich, hypermethylated heterochromatin and an actively transcribed tumour suppressor gene (TSG) associated with a hypomethylated CpG island (indicated in red). In tumour cells, repeat-rich heterochromatin becomes hypomethylated, contributing to genomic instability. *De novo* methylation of CpG islands also occurs in cancer cells, resulting in the transcriptional silencing of growth-regulatory genes. Figure from [37].

even in normal peripheral blood cells. This result may however have been influenced by the low specificity of the CGI definition employed (as will be discussed later).

From the above considerations it is evident that knowledge of CpG islands locations plays an important role both in promoter prediction and for the identification of candidate regions for aberrant DNA methylation.

Main definitions and problematics

The first formal definition of a CpG islands was given by Gardner-Gardiner and Frommer in 1987[18], as a region of at least 200 bp with the proportion of Gs or Cs, referred to as "GC content", greater than 50%, and observed to expected CpG ratio (O/E) greater than 0.6. This ratio is computed by dividing the proportion of CpG dinucleotides in the region by what is expected by chance, when bases are assumed to be independent outcomes of a multinomial distribution:

$$O/E = \frac{\#CpG/N}{\#C/N * \#G/N}$$

where N is the number of base pairs in the segment under consideration. Based on this definition, multiple algorithms for CpG islands identification have been developed, sometimes employing different parameter values. For example, Takai and Jones[42] considered slightly different cutoffs: a minimum length of 500 bp, a minimum GC content of 55% and a minimum O/E of 0.65. With these parameters they were able to exclude most repetitive *Alu* elements from the CpG island list, which otherwise required to be filtered out in the pre-processing of data, as many of them tend to be mistaken for CpG islands due to their base composition. In general, definition-based methods such these lack specificity and robustness, as an arbitrary variation of parameters can give a substantially different list of CpG islands.

The "CG cluster" algorithm described by Glass and others[20] extracts overlapping sequence fragments containing a fixed number of CGs and having variable length. The histogram of these lengths shows a bimodal distribution and it is possible to select a cutoff and find regions associated with the first mode (CG clusters, or CpG islands). The distance-based "CpGcluster" algorithm by Hackenberg and others looks for clusters of CG dinucleotides according to their distance compared to a threshold distance (the median of the distance distribution is often employed as a reference). A *p-value* is then associated to each cluster according to a randomization test on the DNA sequence or by means of a theoretical probability function.

A different method was proposed by Bock and others[11]: it combines an initial, sequence-based mapping of CpG islands with subsequent prediction of CpG island strengths, calculated as a combination of epigenome prediction and which express their tendency to exhibit an unmethylated, open, and transcriptionally competent chromatin structure. Lastly, Irizarry and others[24] developed an algorithm for CpG annotation based on Hidden Markov Models: their work was motivated by recent high-throughput measurement of epigenetic events such as differentially methylated regions (DMRs), which showed that many are DMRs not associated with CGIs but that are nevertheless in the shores of CpG-enriched sequences[23]. Therefore they developed an HMM-based approach for a more flexible definition of CpG islands which would include these newly discovered regions, based on the assumption that the property that defines a CGI is not the CpG density *per se* but the CpG density conditioned on GC content. In this framework, the CGI and baseline regions are the hidden states and the CpG counts are the observations that depend on these states, and CpG counts are

modeled in small intervals. Thus, the genome is divided into non overlapping segments of length s and the posterior probability of being a CGI state is estimated for each segment, after fitting of the model. The authors found that the CGI list, created with their method, covered 94% of the DMRs reported in [23], in comparison with the 65% covered by the GenomeBrowser CGI.

The two tracks providing mapping of CpG islands currently available on the Genome Browser are based respectively on the method of Bock and on the HMM-based algorithm of Irizarry. The main characteristics of these two different CpG islands annotations are listed in Table 1.2.

Measure	CGI Genome Browser	CGI HMM-based
Total CGI n°	28.691	65.535
Total CGI length	21.842.742	39.958.086
Min length	201	18
Max length	45.712	44.214
Mean length	761,3	609,7
Median length	559	408
Mean O/E ratio	0,862	0,747

Table 1.2: Comparison of two different CpG islands definitions.

CHAPTER 2

Sequence analysis methods

In this chapter we illustrate the main methods employed in this thesis. In the first section we describe different types of random reference models used in this work, both for genomic sequences and distance distributions. In the second section we introduce the details of our distance-based analysis, used to compare inter-dinucleotide distance distributions across the whole human genome. The final section is devoted to the description of a random walk representation of DNA sequences.

2.1 Random reference models

When working with statistical analyses on genomic sequences it is essential to compare the results obtained with an appropriate null model: in this way, in fact, it is possible to reject the hypothesis that the features observed could also happen by chance, and thus confirm that they are biologically significant. In our case this translates into the generation of "synthetic" DNA sequences, in which the nucleotides are picked in a random fashion following different rules. We will see that it is also possible to model the distance distributions in the random case directly, though this is derived from a sequence null model as well.

DNA sequence random models

A suitable model for the generation of random DNA sequences is necessary in any study involving the human genome: in fact, any finding has to be compared and validated with what would be observed if the sequence were to be randomly generated. Naturally, many possible definitions of a "random" sequence and different models (from the naive to the more refined ones) exist: it is therefore necessary to carefully consider the best method for each study, taking into account the tradeoff between complexity and plausibility. One of the most widely used random model for DNA sequences is based on Markov chains.

A sequence $\{X_n\}$ of discrete random variables is called a Markov chain if it satisfies the Markov property: for all $n \geq 1$ and $(x_1, x_2, ...x_n)$:

$$P(X_{n+1} = x_{n+1}|X_1 = x_1, ..., X_n = x_n) = P(X_{n+1} = x_{n+1}|X_n = x_n)$$

That is, the conditional probability at step $n+1$, given the value X_n at step n, is uniquely determined and is not affected by any knowledge of the values at earlier times. A Markov chain is defined by:

- $S = \{$ possible $x_n, \forall n\}$: the state space;
- X_0: the initial state;
- $P(X_{n+1} = j|X_n = i), i, j \in S$: the transition probabilities.

If for all n and $i, j \in S$, $P(X_{n+1} = j|X_n = i) = p_{ij}$ independently of n, the chain is said to be *homogeneous*, and we have a transition probability matrix:

$$\begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{bmatrix}$$

where $m = |S|$.

The transition probabilities determine entirely the behavior of the chain: their knowledge, together with the initial state, is sufficient to generate the whole sequence. This concept can be generalised to a $m - th$ order Markov chain, where the state at step/time n depends on the states at the m preceding steps and not on any of the other states.

Markov chain models have been widely employed in the context of genomic sequences analysis as they can capture short-range correlations among bases, however it should be noted that these are still very simple models that cannot reproduce many complexities of DNA sequences, such as long-range correlations. A pictorial representation of a DNA (first-order) Markov chain is depicted in Figure 2.1: the state space consists of the four nucleotides, and the arrows represent the transition probabilities from one state to the other (i.e. from A to A itself and to the other three nucleotides, and so on).

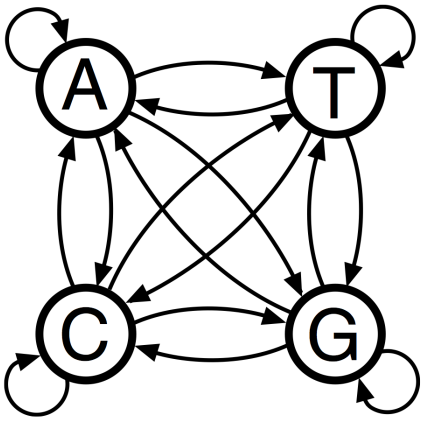


Figure 2.1: Example of a DNA Markov chain. Figure from [14]

The two types of random sequence models employed in this work are:

- **Zeroth order Markov chain or Bernoulli model:** this is the simplest possible model, in which the sequence is constructed by picking one of the four letters A, C, G, T with a fixed probability, independent from the preceding values. The nucleotide probabilities are determined according to the relative frequencies of the bases in the DNA sequence.

- **First order Markov chain model:** in this model the synthetic DNA sequence is generated taking into account the first order transition probabilities (namely the probability to pick a nucleotide X when we picked Y in the previous step). The entries of the transition matrix can be derived from the relative frequencies of nucleotides and dinucleotides in the sequence.

The transition matrix for a first order Markov chain DNA model is:

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} p(A|A) & p(C|A) & p(G|A) & p(T|A) \\ p(A|C) & p(C|C) & p(G|C) & p(T|C) \\ p(A|G) & p(C|G) & p(G|G) & p(T|G) \\ p(A|T) & p(C|T) & p(G|T) & p(T|T) \end{bmatrix} \end{array}$$

The transition probabilities are estimated from the biological sequence as follows: the probability $p(C|A)$ of "picking" a C when we picked an A in the previous step is approximated by the ratio of the observed frequencies of CA dinucleotides and A nucleotides in the DNA sequence:

$$p(C|A) = \frac{\#CA/\#dinucleotides}{\#A/\#nucleotides}$$

As an example, the transition probabilities estimated for chromosome 1 are listed below:

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} 0.3265 & 0.1726 & 0.2449 & 0.2560 \\ 0.3484 & 0.2610 & 0.0486 & 0.3420 \\ 0.2864 & 0.2116 & 0.2608 & 0.2411 \\ 0.2179 & 0.2053 & 0.2499 & 0.3269 \end{bmatrix} \end{array}$$

Higher-order models would give rise to sequences that are more "biologically" correct: for example in a second order model we would have transition probabilities such as $p(A|AT)$, and this "memory" could allow the reproduction of three-letter words in DNA, which correspond to codons (pieces of sequence that specify a single amino acid). However, for the present interest a first order model is sufficient, as we wish to compare our results with a reference sequence which is characterised by a similar dinucleotide bias, but with randomly distributed dinucleotides across the sequence.

Random model for the distance distribution

It is also possible to model the distance distribution in the random case directly: in fact, if CpGs were distributed totally at random along the chromosome sequence, the distances between neighboring CpG dinucleotides should follow the geometric distribution (depicted in Figure 2.2):

$$f_x(k) = p_x(1 - p_x)^{k-1}$$

Where p_x is the probability of x in the sequence, estimated from its frequency in the biological sequence (i.e. p_{CG}). The mean and variance of the distribution are equal to: $E[k] = \frac{1}{q}$, $var(k) = \frac{1-q}{q^2}$.

We can easily understand why this is the case, if we reckon that the geometric distribution is used to model the probability that the kth trial (out of k trials) is the first success, given that the success probability in a single trial is p. In this work we employ the geometric distribution as a random reference for the inter-dinucleotide distance distributions, and it proves to be very convenient as we can directly compare them without having to generate a whole synthetic sequence (which can be extremely computationally intensive).

Note that this model is not independent from the previously introduced random sequence models: the geometric distribution for inter-dinucleotide distances, in fact, can be derived from a random sequence in which we pick dinucleotides with a fixed probability, estimated from dinucleotide frequencies in the real DNA sequence. This corresponds to a zeroth order Markov model in which, instead of picking nucleotides, the sequence is constructed by picking dinucleotides.

2.2 Inter-dinucleotide distance analysis

In this section we describe a distance-based approach for the characterisation of dinucleotides distribution inside the human genome. Note that this method could also be formulated in the context of dynamical systems in terms of "return times" or "Poincaré recurrences" of symbolic trajectories.

The outline of the method is the following: the sequence of interest is read and all the inter-dinucleotide distances are computed for the 16 different dinucleotides as shown in Figure 3.1. We do not divide the sequence rigidly in

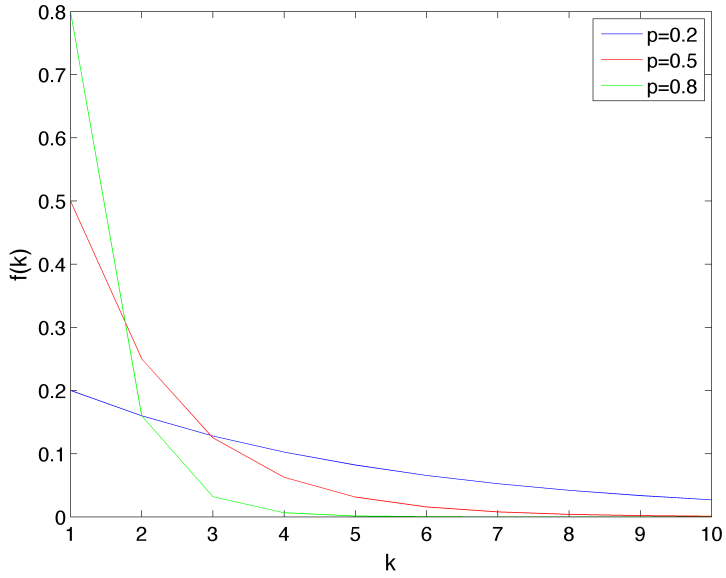


Figure 2.2: Example of geometric distribution for different parameters value.

dinucleotides as done in other works (see for example [8]) but, starting from the first nucleotide in the sequence, we look for every occurrence of the dinucleotide of interest and store the corresponding distances, which are computed in nucleotide units. These two methods are compared in Figure 3.2: notice how dividing the sequence rigidly into dinucleotides introduces the problem of having two different frames of reference, which as shown detect different CpG couples. This could pose a problem when dealing with CpG islands, because one reading frame may miss on CpG islands that are found in the second one: this is our main motivation for using a different method.

In addition to this we decided to employ a non-overlapping approach: namely, AAAA is evaluated as two AA dinucleotides with a distance of two (however, notice that this difference is irrelevant when focusing on CpGs), and all distances are subtracted by one at the end of the process in order to obtain a minimum distance of 1. In this way, all dinucleotides have a minimum distance of one (using an overlapping approach would give rise to a minimum distance of zero for "double" dinucleotides only).

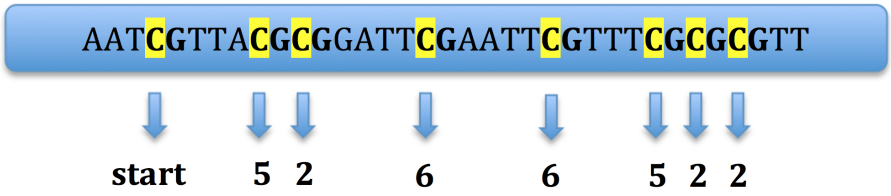


Figure 2.3: Example of inter-dinucleotide distance computation for CG dinucleotides: the start point is indicated and C nucleotides in a CpG couple are highlighted for clarity. The distances are counted in single nucleotides units, and will later be subtracted by one to obtain a minimum distance of 1 for adjacent couples.

Figure 2.4: Comparison of two methods for dinucleotide analysis in the sequence: in the first case it is divided rigidly in dinucleotides, obtaining two reading frames which are not equivalent (pictures a and b). In the second case distances are computed in single nucleotides, so there are no issues with different reading frames (picture c).

In order to quantitatively compare the distributions obtained for different dinucleotides, we employ the Kullback-Leibler divergence (also known as relative entropy): this is a measure of the difference between two probability distributions $f(x)$ and $g(x)$. It represents the information lost when $g(x)$ is used to approximate $f(x)$: more precisely it measures the number of additional bits required when encoding a random variable with a distribution $f(x)$ using the alternative distribution $g(x)$. For two probability distributions $f(x)$ and $g(x)$ and a random variable X, the KL divergence is defined as:

$$D(f||g) = \sum_{x \in X} f(x) log \frac{f(x)}{g(x)}$$

The KL divergence has the following properties:

- $D(f||g) > 0$;
- $D(f||g) \neq D(g||f)$: this measure is therefore asymmetric;
- $D(f||g) = 0$ iff $f(x) = g(x)$ for all $x \in X$.

Considerations on the treatment of N nucleotides

When working with the Human Genome assembly, an observation which could at first come as a surprise to non-biologists is the presence of a fifth letter (besides the usual A, C, G and T) in the sequence: N. This letter stands for "aNy" base - an unknown nucleotide which could not be correctly classified: Ns are often found clustered together in areas corresponding to repetitive elements, the centromere, or other non coding regions. In most works these parts are discarded without any further consideration, but here it was found appropriate to perform an initial analysis in order to evaluate the distribution and abundance of these interspersed sequences and, most importantly for our application, the effect of their presence on the evaluation of inter-nucleotide distances. To this end, three different approaches for computing inter-dinucleotides distances in presence of Ns are compared. While looping on the DNA sequence looking for specific occurrences in order to calculate their distances, if a N is encountered three different actions are possible:

- the counting continues as if a typical nucleotide was encountered,
- the distance counter is not updated: this is effectively the same as removing the Ns from the sequence,
- the distance is calculated in the standard way but then it is discarded from the distribution.

The approaches described above are more easily compared using an example: take for instance the following DNA sequence sample, for which we want to compute the distances between CpG occurrences.

$$ATCGAANNNCGTTATCG$$

Using the first method we would get the distance values (taken from the C to the next C) $d_1 = (7, 6)$; with the second method (removing the Ns) the distances are $d_2 = (4, 6)$ and the third method would only give one distance $d_3 = (6)$.

The method described above was applied to the CpG dinucleotide in the different chromosomes: the most interesting finding in this initial analysis is the fact that only a very little number of distances are discarded in method 3 due to the presence of Ns, and this number is also not directly proportional to N content in the chromosome (see Table 2.1). This unique feature is probably due to the fact that Ns are usually present in blocks, so they affect only a few distance counts, and in some chromosomes with high N content these are distributed mostly at the start and end points of the chromosome: a perfect example is chromosome 14, which has a high N content but thanks to their distribution in the sequence there is no effect on the dinucleotide distances computed. Due to all the above consideration it is therefore found here that the effect of discarding the Ns is actually negligible, and this will be the starting point for the following analyses.

Chromosome	N content (%)	Δ num dist
1	9.6	37
2	2.1	20
3	1.6	5
4	1.8	10
5	1.8	5
6	2.2	9
7	2.4	15
8	2.4	7
9	14.9	38
10	3.1	22
11	2.9	7
12	2.5	11
13	17.0	4
14	17.8	0
15	20.3	9
16	12.7	4
17	4.2	7
18	4.4	9
19	5.6	4
20	5.6	5
21	27.1	14
22	32.0	9
X	2.69	21
Y	56.79	16

Table 2.1: N content for all chromosomes and corresponding difference in the number of CG distances computed with methods 2 and 3.

2.3 DNA walk

We review here different methods based on the representation of DNA sequences as random walks, and introduce the specific conversion of the sequence into a walk employed in this thesis.

One of first works to introduce a mapping of the nucleotide sequence into a walk, termed "DNA walk", is the one by Peng, Buildyrev and others on long range correlations in genomic sequences [32]. They define a DNA walk based on a conventional one dimensional random walk, in which a walker movers either up ($u(i) = +1$) or down ($u(i) = -1$) one unit length (u) for each step i of the walk. The net displacement (y) of the walker after l steps is the sum of the unit steps $u(i)$ for each step i:

$$y(l) = \sum_{i=1}^{l} u(i)$$

In the case of the DNA walk, the walker steps up when a pyrimidine (C or T nucleotide) occurs along the DNA chain and down when a purine is encountered (A or G nucleotide). Two examples of the walks obtained with this method are shown in Figure 2.5. An analysis of the root mean square fluctuation about the average of the displacement, which is related to the auto correlation of the sequence, emphasised the presence of long-range correlation in nucleotide sequences, especially intron-containing genes. Furthermore, they reported a quantitative scaling of the correlation in the power law form, observed in numerous phenomena having a self-similar or fractal origin.

Different methods for the conversion of a nucleotide sequence into a one dimensional walk are possible, and some are reviewed in [9] (for example we could distinguish between nucleotides with strong vs weak hydrogen bonding - C and G vs T and A). Higher dimensional and more complex walks are also possible: for example in [9] nucleotides are mapped into the four cardinal points $(+1, -1, +j, -j)$ of the complex plane, which correspond respectively to the presence of A, G, T and C nucleotides in the sequence. In this representation purines are limited to values on the real axis, while pyrimidines are restricted to the imaginary axis. An example of such a walk, computed for a noncoding region of the *Helicobacter pylori* bateria sequence, is shown in Figure 2.6. The authors emphasise how this walk can help locating periodicities and nucleotide structures, and they also incorporate Gaussian-based wavelet analysis to distinguish between high and low complexity regions in genomic sequences.

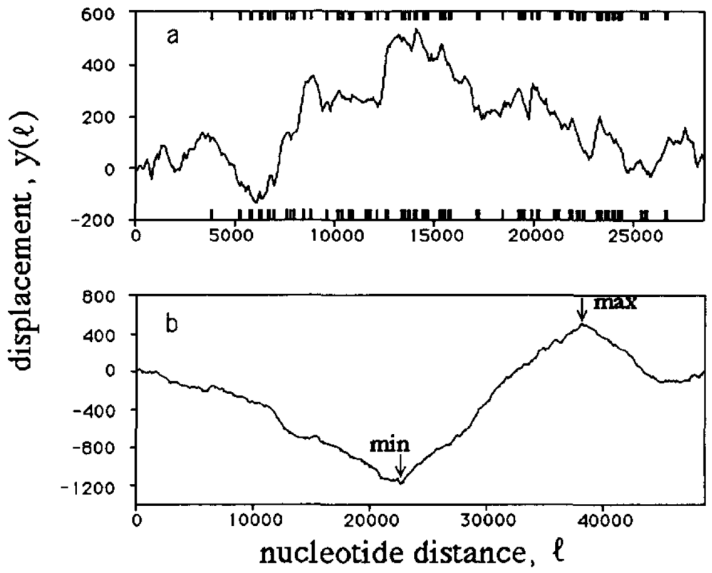


Figure 2.5: DNA walk representations of (a) intron-rich human β-cardiac myosin heavy chain gene sequence and (b) the intron-less bacteriophage λ DNA sequence. Heavy bars correspond to the coding regions of the gene. Image from [33].

In this work, we employ yet a different conversion of the nucleotide sequence into a DNA walk. Our interest here is mostly focused on dinucleotides, therefore our "walker" will:

- go up when the dinucleotide of interest is encountered (i.e. CpG);
- go down otherwise.

This method provides another possible one-dimensional walk representation of the DNA sequence under analysis.

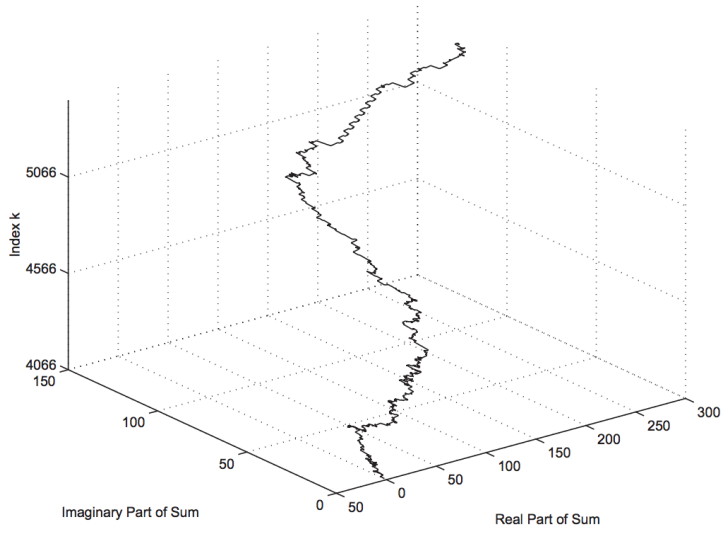


Figure 2.6: Complex DNA walk. This graphical representation highlights the nucleotide evolution of the DNA sequence and exposes trends in nucleotide composition. From this graphical representation, we can identify a region with an interesting nucleotide staircase structure occurring near bp 5150–5290. Image from [9].

CHAPTER 3

Characterisation of dinucleotide statistics

In this chapter we present the results of our genome-wide analysis of dinucleotide statistics. In the first section we focus on the abundance of the different dinucleotides in human DNA and compare it with the one observed for two different types of randomly generated sequences. The second section is devoted to the inter-dinucleotide distances distributions (as obtained with the method described in Chapter 2), which are characterised and compared with the corresponding random models (mainly the geometric distribution).

3.1 Relative frequencies of dinucleotides

In this section we analyze the different dinucleotides abundances in the human genome and compare them with the results obtained in the case of a random sequence. We consider two random sequences, generated respectively using a zeroth-order and a first-oder Markov chain (see the Methods chapter for details). In Table 3.1, the relative frequencies of the 16 dinucleotides are listed for chromosome 1 (taken here as an example) and for two synthetic random sequences. The data are also plotted in Figures 3.1 and 3.3 where they can be easily inspected visually. In Table 3.2 and Figures 3.2 and 3.4 we report the percentage differences in dinucleotide abundance between the chromosome 1 sequence and the two different random sequences.

By comparing the relative frequencies values for the different dinucleotides we immediately notice that CpGs are strongly depleted in the human genome, as expected due to the methylation and mutation processes that act on the cytosines in the couple. It is interesting to compare the two random models: the zeroth-order Markov model does reproduce a few dinucleotides in the abundance observed in the real sequence, but the predicted amount of CpGs is much higher than the observed one. The first-oder model, which takes into account the transition probability from a cytosine to a thymine, has the capability to reproduce this feature much better, as can be seen in Figure 3.3 and 3.4. Overall, the first-order Markov chain model generates sequences with a dinucleotide content extremely similar to the biological sequence: for this reason we believe that this model is appropriate for our analysis, as we wish to compare our results with sequences that contain approximately the same amount of dinucleotides, but where these are distributed randomly.

Dinucleotide	Rel freq observed	Rel freq Markov 0	Rel freq Markov 1
AA	0.095045182	0.069341454	0.095050808
AC	0.050228217	0.06418978	0.050213334
AG	0.07127676	0.06420494	0.071267626
AT	0.074513048	0.089614639	0.074527307
CA	0.072725744	0.064205948	0.07272302
CC	0.054485162	0.038075984	0.054500134
CG	0.010140554	0.046002714	0.010139958
CT	0.071385504	0.064255427	0.07139137
GA	0.059773523	0.064185349	0.059762304
GC	0.044171326	0.046032733	0.044165721
GG	0.054439241	0.038058532	0.054427571
GT	0.050318136	0.064241298	0.050309188
TA	0.063518745	0.089618062	0.063522939
TC	0.05985238	0.064241575	0.059875297
TG	0.07284555	0.064251727	0.072829629
TT	0.095280928	0.069479837	0.095293794

Table 3.1: Comparison of the relative frequencies of the 16 dinucleotides in the Chromosome 1 sequence and in two different reference random sequences (generated by a zeroth and first order Markov chain model).

Dinucleotide	Percentage difference Markov 0	Percentage difference Markov 1
AA	6.4162	-0.0059
AC	-18.1161	0.0296
AG	16.7448	0.0128
AT	-11.1572	-0.0191
CA	18.4022	0.0037
CC	18.5459	-0.0275
CG	-319.2884	0.0059
CT	16.8062	-0.0082
GA	0.7528	0.0188
GC	3.6798	0.0127
GG	18.5344	0.0214
GT	-17.9997	0.0178
TA	-30.4021	-0.0066
TC	0.7968	-0.0383
TG	18.4784	0.0219
TT	6.4407	-0.0135

Table 3.2: Comparison of the percentage differences in dinucleotide frequencies between the Chromosome 1 sequence and two different reference random sequences (generated by a zeroth and first order Markov chain model).

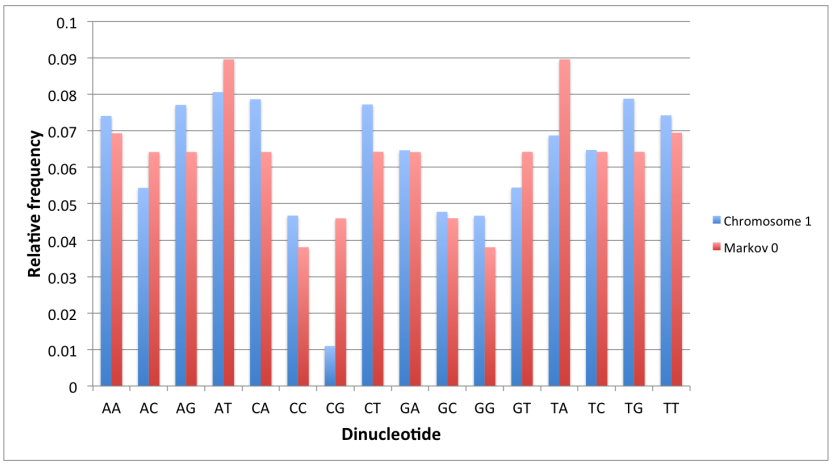


Figure 3.1: Comparison of relative frequencies of dinucleotides in chromosome 1 and in a random sequence generated with a zeroth-order Markov chain.

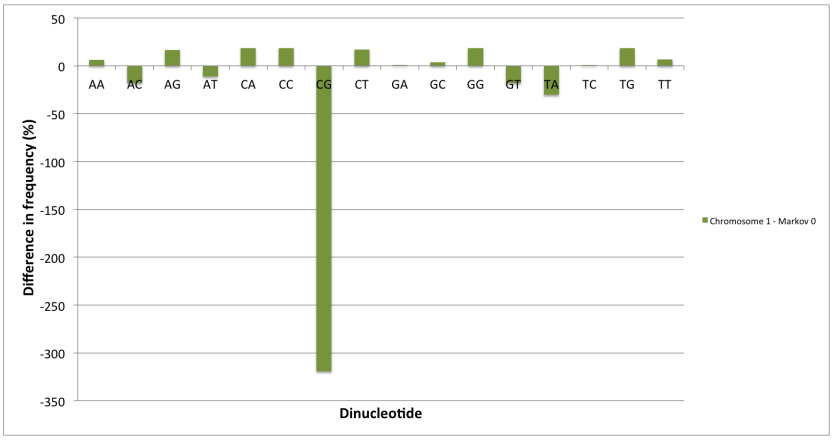


Figure 3.2: Percentage difference in dinucleotide content between the chromosome 1 sequence and a random sequence generated with a zeroth-order Markov chain.

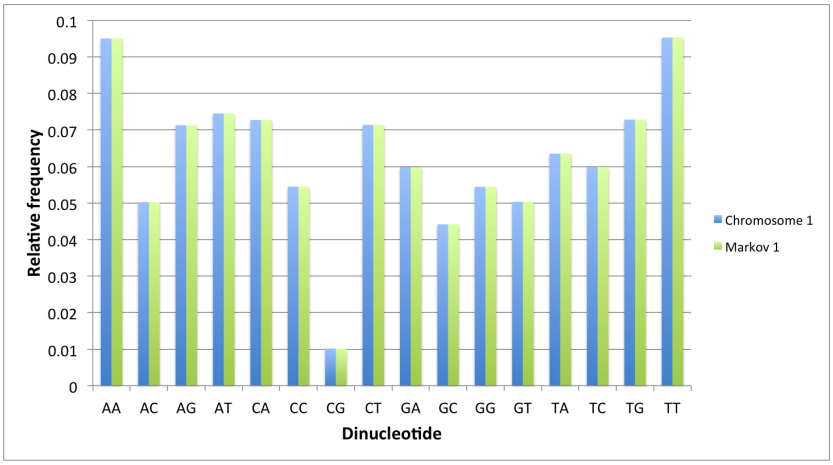


Figure 3.3: Comparison of relative frequencies of dinucleotides in chromosome 1 and in a random sequence generated with a first-order Markov chain.

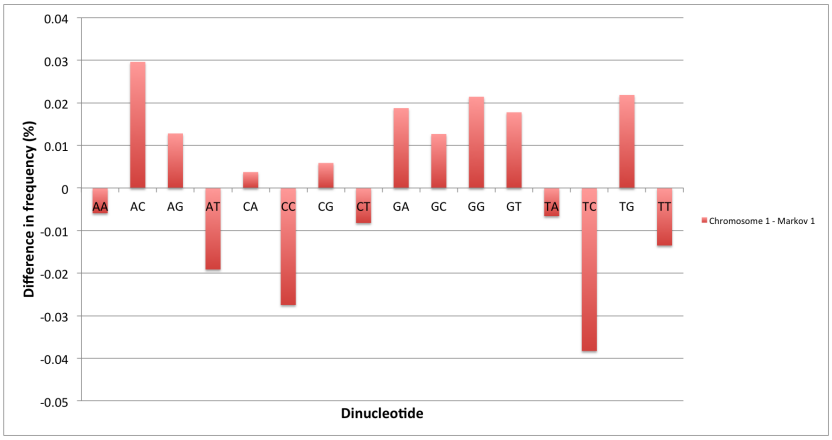


Figure 3.4: Percentage difference in dinucleotide content between the chromosome 1 sequence and a random sequence generated with a first-order Markov chain.

3.2 Comparison of dinucleotide distance distributions

In this section, we present and describe the results of our distance-based characterisation of dinucleotide positioning in the human genome. The distance distributions for the different dinucleotides are first compared with each other and then with the random models.

The histogram of the inter-dinucleotide distances for the CpGs is highly positive skewed and it spans a wide range of length values. The overall shape of the distance distribution for the other dinucleotides is quite similar to this one, so we choose to employ log-log plots that enable a better analysis of the distributions behaviour in the tails. In Figure 3.5 the distance histograms for all 16 dinucleotides are plotted on a log-log scale. In order to reduce noise in the tails of the plots, partial logarithmic binning was applied for distances over a fixed threshold, substantially improving the quality and readability of the plots.

We can see in this figure that the red curve, corresponding to CpG dinucleotides, clearly stands out from the other 15 curves, corresponding to the remaining dinucleotides. In order to quantitatively characterise the distributions we report in Table 3.3 some descriptive values for the different dinucleotides. All distances values throughout this work are intended in bp (base pairs). Comparison of the mean and median values for the different dinucleotides shows that both of them are significantly higher for CpGs: this fact, combined with the inspection of the log log plots in Figure 3.5 indicates that the CpG distance distribution has heavier tails when compared to the other dinucleotides.

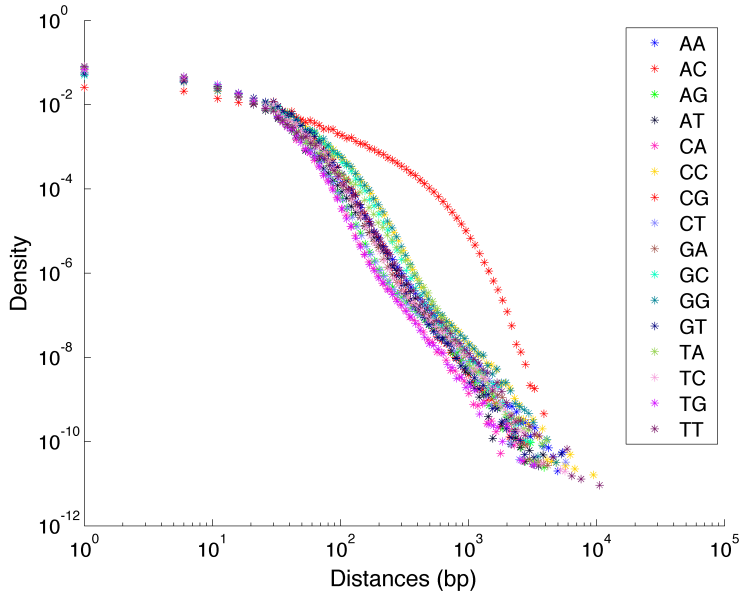


Figure 3.5: Histograms of the inter-dinucleotide distances distributions for all dinucleotides in log log scale.

Dinucleotide	Max distance	Mean distance	Median distance
AA	5874	13.24	7
AC	4039	18.86	12
AG	4238	13.30	8
AT	4154	11.94	7
CA	3427	12.79	9
CC	10266	23.11	13
CG	4210	100.40	41
CT	6245	13.29	8
GA	4611	15.85	10
GC	3274	22.44	13
GG	5256	23.10	13
GT	5763	18.82	12
TA	4743	14.23	8
TC	6172	15.85	10
TG	3562	12.76	9
TT	11516	13.20	7

Table 3.3: Properties of inter-dinucleotide distance distributions.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA		0.113	0.052	0.052	0.070	0.172	0.994	0.052	0.043	0.441	0.172	0.112	0.036	0.043	0.070	0.000
AC	0.113		0.098	0.168	0.125	0.047	0.678	0.099	0.030	0.139	0.046	0.000	0.094	0.03	0.126	0.114
AG	0.044	0.090		0.018	0.009	0.133	0.950	0.000	0.028	0.230	0.134	0.089	0.014	0.028	0.009	0.044
AT	0.046	0.145	0.017		0.017	0.189	1.046	0.017	0.057	0.304	0.190	0.143	0.019	0.057	0.017	0.046
CA	0.058	0.101	0.009	0.017		0.161	0.985	0.009	0.041	0.235	0.163	0.100	0.024	0.041	0.000	0.058
CC	0.193	0.058	0.205	0.276	0.275		0.539	0.205	0.097	0.094	0.006	0.059	0.142	0.097	0.277	0.195
CG	2.288	2.095	2.750	2.741	3.237	1.226		2.745	2.246	1.459	1.225	2.098	1.998	2.254	3.224	2.292
CT	0.044	0.090	0.000	0.018	0.009	0.133	0.951		0.028	0.231	0.134	0.089	0.014	0.028	0.009	0.044
GA	0.041	0.030	0.030	0.066	0.049	0.072	0.800	0.030		0.201	0.075	0.029	0.029	0.000	0.050	0.042
GC	0.285	0.099	0.232	0.310	0.279	0.071	0.603	0.233	0.149		0.075	0.099	0.193	0.149	0.281	0.288
GG	0.192	0.058	0.206	0.275	0.276	0.006	0.541	0.206	0.100	0.098		0.058	0.142	0.100	0.278	0.193
GT	0.112	0.000	0.097	0.166	0.124	0.047	0.679	0.098	0.030	0.139	0.046		0.093	0.030	0.125	0.113
TA	0.033	0.089	0.019	0.024	0.035	0.116	0.908	0.019	0.028	0.238	0.118	0.088		0.028	0.035	0.033
TC	0.041	0.030	0.030	0.066	0.049	0.072	0.801	0.030	0.000	0.202	0.075	0.029	0.029		0.050	0.042
TG	0.058	0.102	0.009	0.017	0.000	0.162	0.987	0.009	0.041	0.236	0.164	0.101	0.024	0.041		0.058
TT	0.000	0.114	0.053	0.052	0.070	0.174	0.996	0.053	0.044	0.444	0.173	0.113	0.037	0.044	0.070	

Table 3.4: Kullback-Leibler divergence of dinucleotide distance distributions.

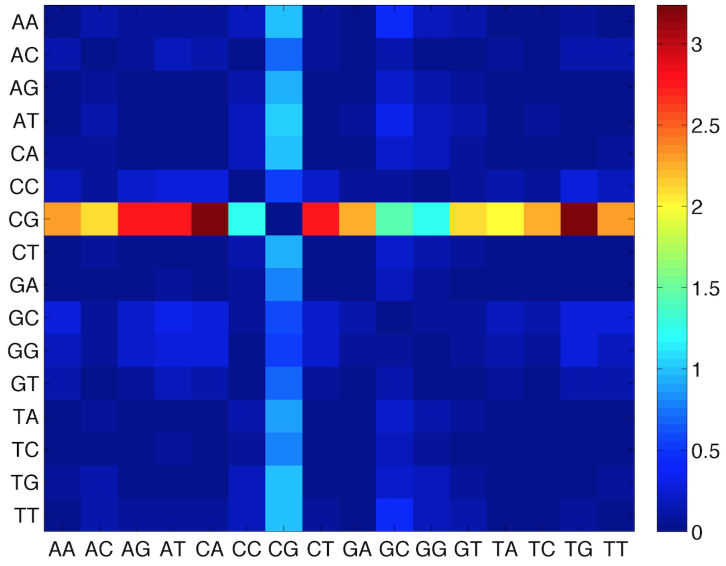


Figure 3.6: Heatmap of the KL-divergence of dinucleotide distance distributions.

Comparison with the random models

In the Methods chapter we introduced two different models for the generation of random references sequences, and we also described how the distance distribution in the random case can be modelled as a geometric distribution.

In Figure 3.14 we compare the distance frequencies for CpGs in the DNA sequence of chromosome 1 with all three references: first of all it is possible to notice how the green curve (Markov 0 chain sequence) is greatly enriched in short distances. This is due to the fact that CpG abundance in this random sequence is much greater than in the real biological sequence, as described in Section 3.1. Furthermore, this model also completely fails in describing the tails found in the real distance distribution: for these reason, we discard the Markov 0 sequence as a good random reference for the distance distributions. On the other hand, the geometric and Markov 1 curves correspond quite well and they both reflect the properties of a sequence with a similar amount of dinucleotides, but which are positioned randomly. In Figures 3.8 to 3.22 we report the linear-log plots of distances distribution for the 16 dinucleotides, compared with the geometric distribution (which is much less computationally intensive to generate than the Markov 1 sequence).

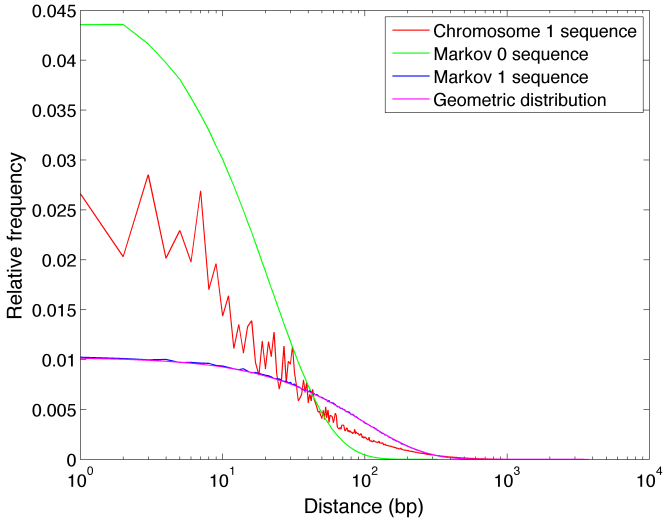


Figure 3.7: Comparison of CpGs distance frequencies in chromosome 1, in two random reference sequences (generated with a zeroth and first order Markov chain) and in the geometric distribution for CpGs.

We can notice how in many cases the real distributions differs from the one which represents a random positioning of dinucleotides along the sequence. For CpG dinucleotides we observe a sharp difference, which corresponds to an enrichment of short distance (up to approximately 50bp) followed by a depletion of medium distances, followed by an enrichment in very long distances (from 400bp onwards). This is consistent with what is expected from the biological insight into CpG distribution and with the presence of CpG islands (which could correspond to the short enriched distances).

In order to better understand the global differences between the "theoretical" and "experimental" distance distributions for the 16 dinucleotides, we performed a hierarchical clustering of the dinucleotides based on the difference between the two binned distribution. The metric employed was Spearman's rank correlation coefficient, and the results are shown in Figure 3.24, where the clustering is displayed by means of a heat map and a dendrogram.
We immediately see that CpGs are attributed to a separate cluster, and that the heat map reflects the observations made on the distance frequency plots, namely the enrichment in short distances and depletion in intermediate distances in the experimental distribution. Note however that we cannot observe here the enrichment in very long distances: this is a practical limitation due to the fact that distances generated according to the geometric distribution very rarely have high values, so this limits the maximum distance up to which we can perform our analysis. As far as the other dinucleotides are concerned, AA and TT are clustered in a separate group and show a depletion of short distances in the first two bins in the biological sequence, followed by an enrichment in distances up to the 8th bin. Most of the remaining dinucleotides are characterised by an enrichment of short distances in the biological sequence, followed by a depletion, but differences between the two distributions are significant only up to the 6th or 7th bin.

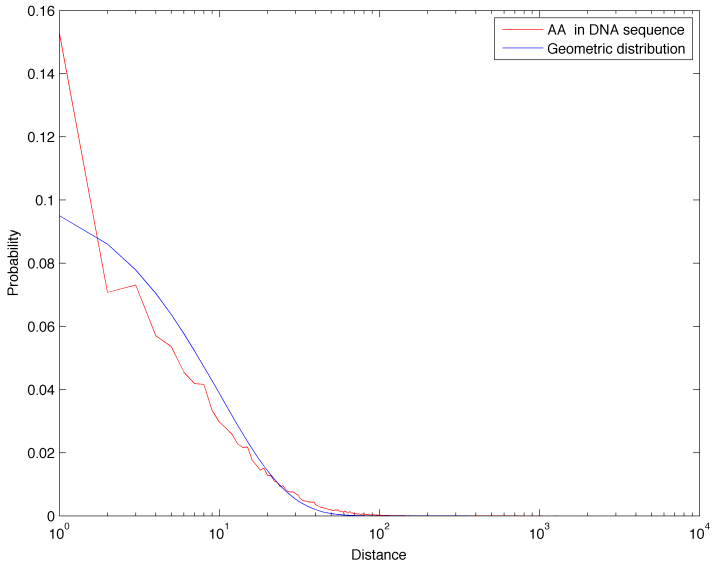


Figure 3.8: Comparison of AA distance distribution in chromosome 1 with the geometric distribution.

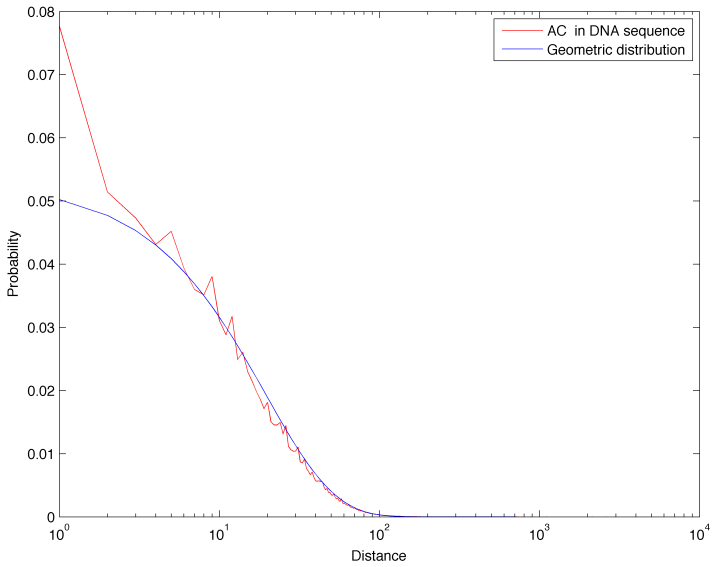


Figure 3.9: Comparison of AC distance distribution in chromosome 1 with the geometric distribution.

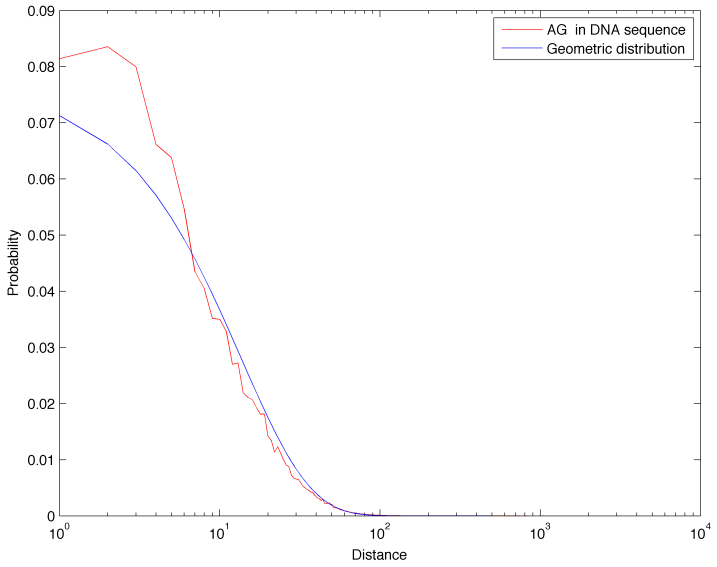


Figure 3.10: Comparison of AG distance distribution in chromosome 1 with the geometric distribution.

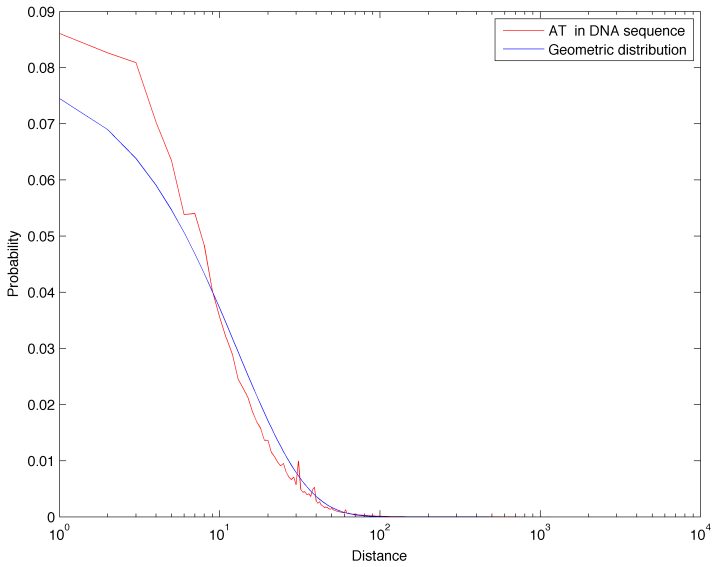


Figure 3.11: Comparison of AT distance distribution in chromosome 1 with the geometric distribution.

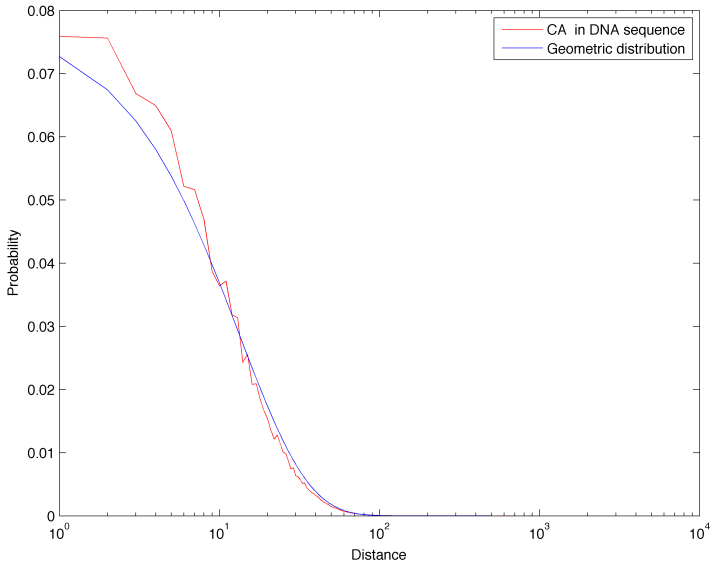


Figure 3.12: Comparison of CA distance distribution in chromosome 1 with the geometric distribution.

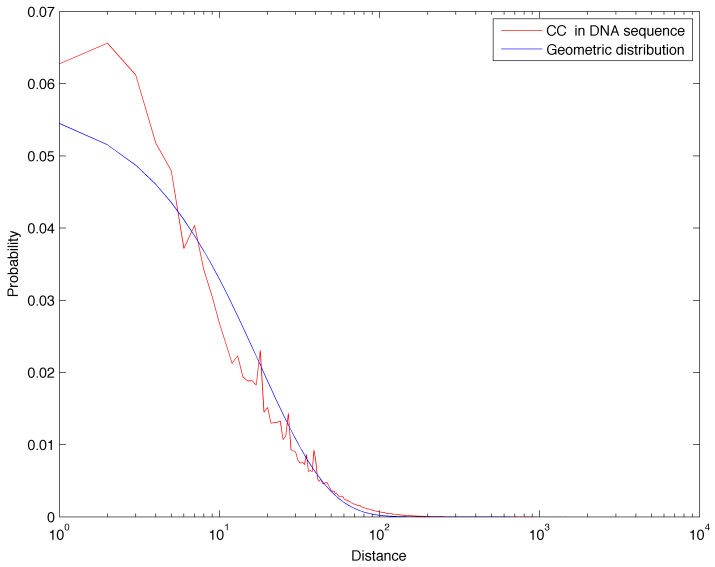


Figure 3.13: Comparison of CC distance distribution in chromosome 1 with the geometric distribution.

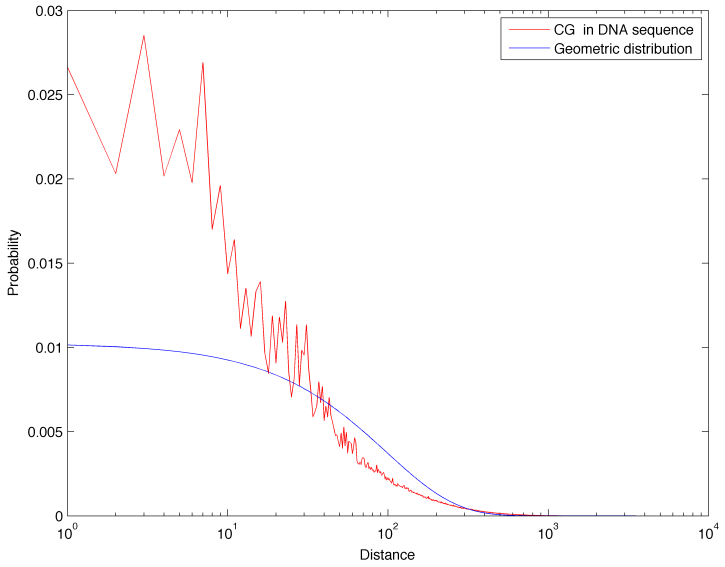


Figure 3.14: Comparison of CG distance distribution in chromosome 1 with the geometric distribution.

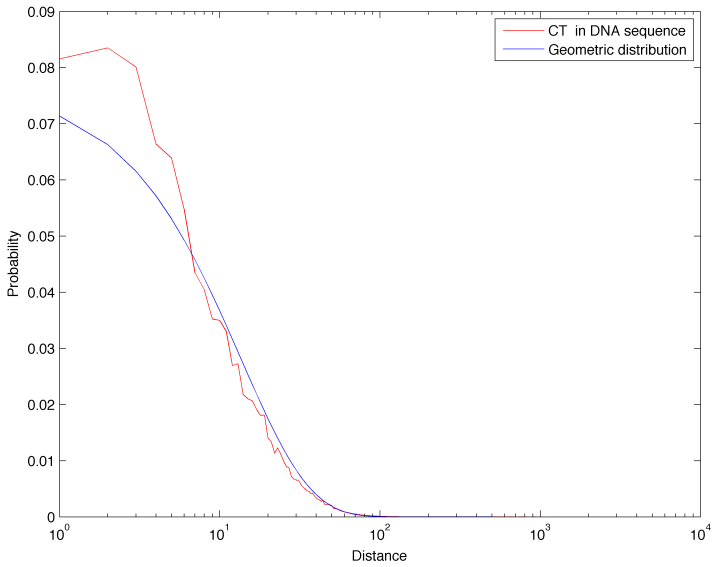


Figure 3.15: Comparison of CT distance distribution in chromosome 1 with the geometric distribution.

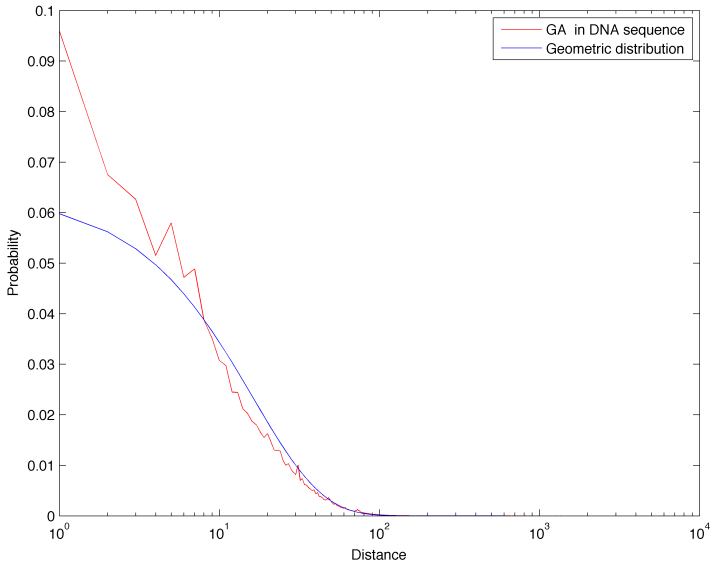


Figure 3.16: Comparison of GA distance distribution in chromosome 1 with the geometric distribution.

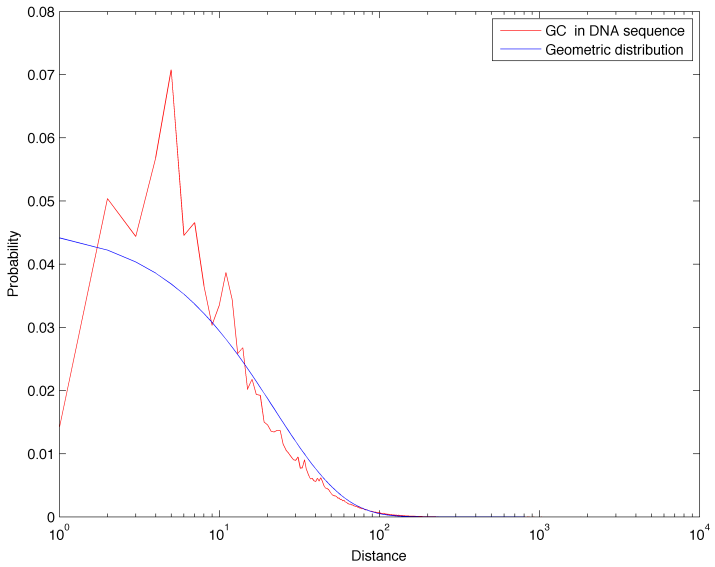


Figure 3.17: Comparison of GC distance distribution in chromosome 1 with the geometric distribution.

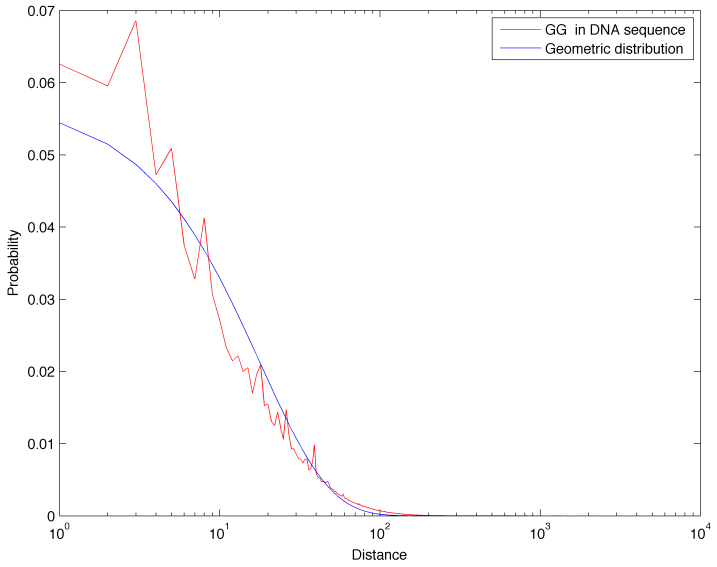


Figure 3.18: Comparison of GG distance distribution in chromosome 1 with the geometric distribution.

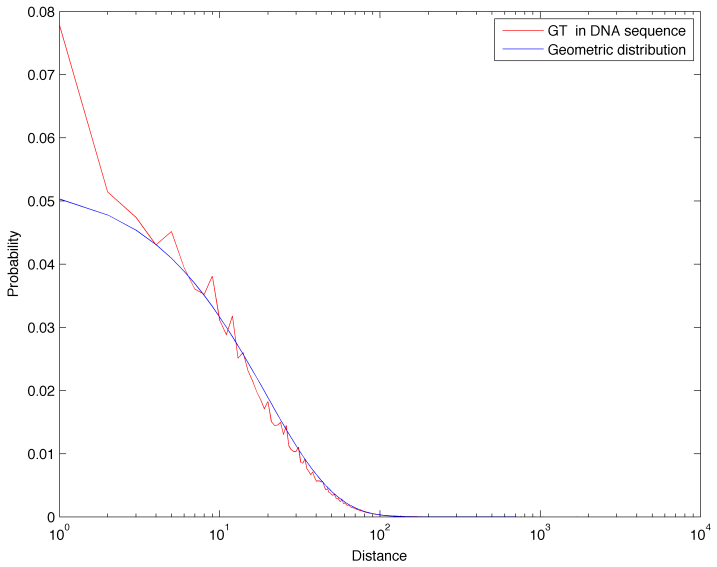


Figure 3.19: Comparison of GT distance distribution in chromosome 1 with the geometric distribution.

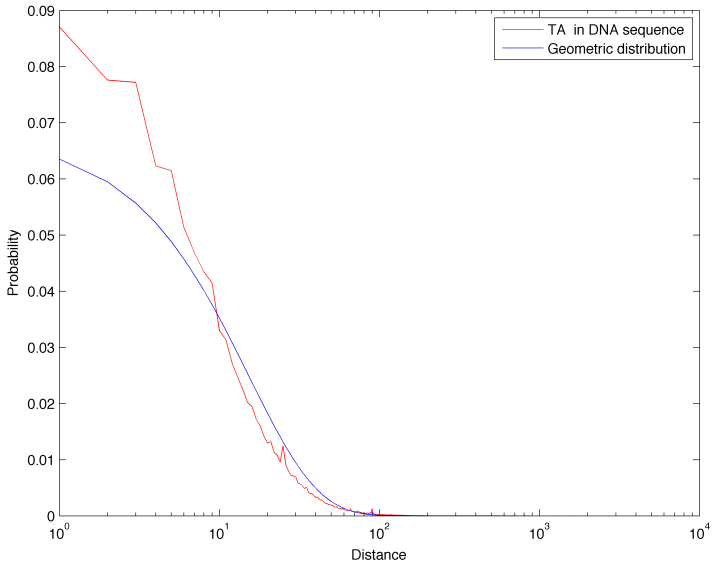


Figure 3.20: Comparison of TA distance distribution in chromosome 1 with the geometric distribution.

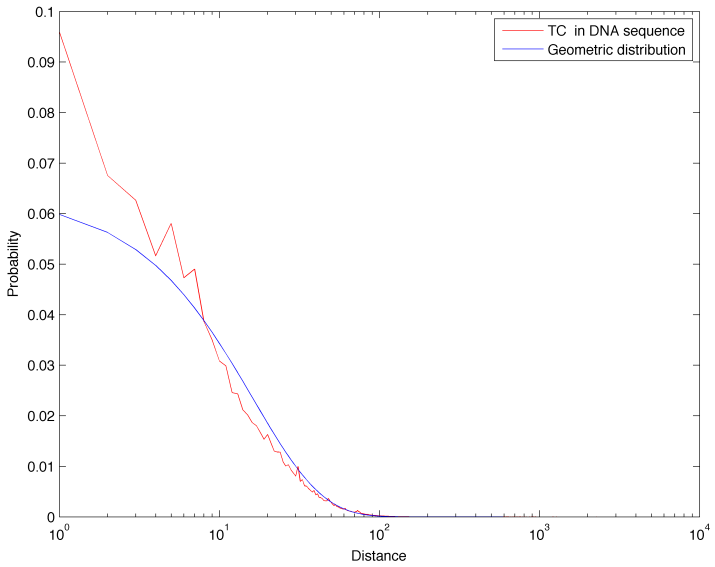


Figure 3.21: Comparison of TC distance distribution in chromosome 1 with the geometric distribution.

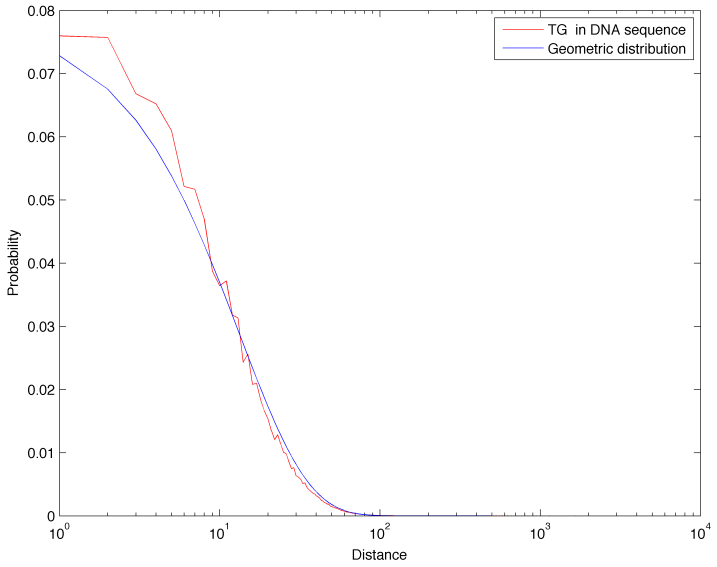


Figure 3.22: Comparison of TG distance distribution in chromosome 1 with the geometric distribution.

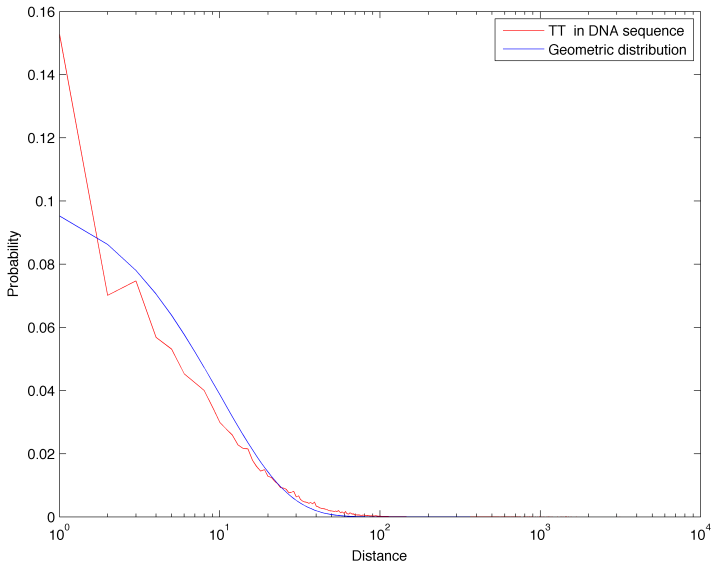


Figure 3.23: Comparison of TT distance distribution in chromosome 1 with the geometric distribution.

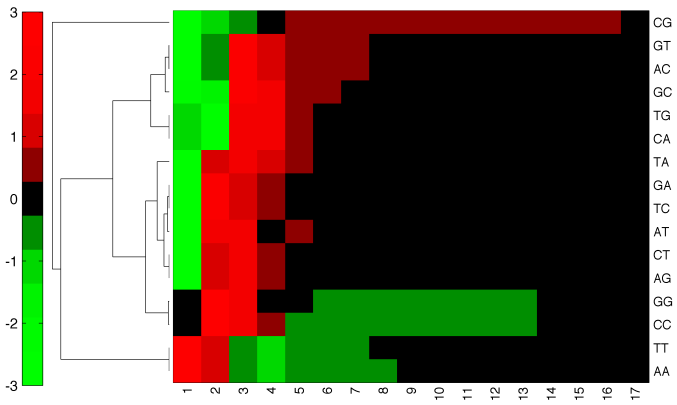


Figure 3.24: Hierarchical clustering of dinucleotides based on the difference between binned geometric distributed distances and sample distances (bin size=10 base pairs). The metric used here was the Spearman correlation coefficient.

CHAPTER 4

Analysis in genomic regions of interest

In this chapter we illustrate the results of our distance-based analysis of dinucleotides applied to genomic regions of interest that we expect to be related with CpG islands or CpG rich regions. These include CpG islands, as mapped by different algorithms, promoter regions and LADs (Lamina Associated Domains). In the last section we also describe the results of our DNA walk approach, especially in relation to CpG islands.

4.1 CpG islands

We now focus our analysis inside genomic regions annotated as CpG islands, comparing two different lists (the one by Irizarry and others and the one available on the Genome Browser - see the first chapter for details).

Dinucleotide frequencies inside CGIs

First we study dinucleotide abundances: the relative frequencies are listed in Table 4.1 and shown in Figure 4.1 for regions corresponding to annotated CGI and in the whole genome. The corresponding percentage difference values are listed in Table 4.2 and displayed in Figure 4.2. Notice how CpGs show the highest difference as expected, but also other dinucleotides containing either Cs or Gs are enriched in this region whereas the other ones are depleted.

Dinucleotide	Rel freq CGI Irizarry	Rel freq CGI GenBrows	Rel freq Whole gen
AA	0.036928231	0.031645865	0.09774721
AC	0.046059362	0.043084523	0.050339835
AG	0.068857378	0.065816349	0.069924165
AT	0.024965851	0.020040111	0.077258419
CA	0.062741974	0.056535687	0.072535448
CC	0.113627331	0.122211342	0.052096925
CG	0.077737625	0.095814051	0.009861512
CT	0.068985479	0.065650868	0.069961111
GA	0.059103342	0.056676477	0.059335043
GC	0.103364586	0.117742546	0.042660269
GG	0.114126457	0.121792202	0.052125953
GT	0.04639357	0.042702279	0.050454531
TA	0.018298294	0.015487839	0.065651918
TC	0.058892809	0.056683837	0.059358019
TG	0.062822783	0.056421869	0.072664114
TT	0.037094929	0.031694156	0.098025528

Table 4.1: Comparison of the relative frequencies of the 16 dinucleotides inside CpG islands corresponding to two different annotation (by Irizarry et.al. and from the Genome Browser) and in the whole genome.

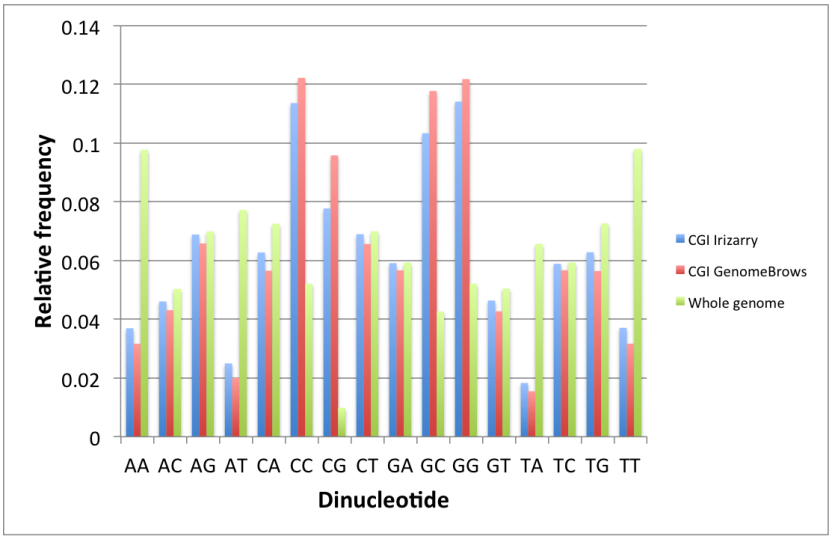


Figure 4.1: Comparison of the relative frequencies of the 16 dinucleotides inside CGIs corresponding to two different annotation (by Irizarry et.al. and from the Genome Browser) and in the whole genome.

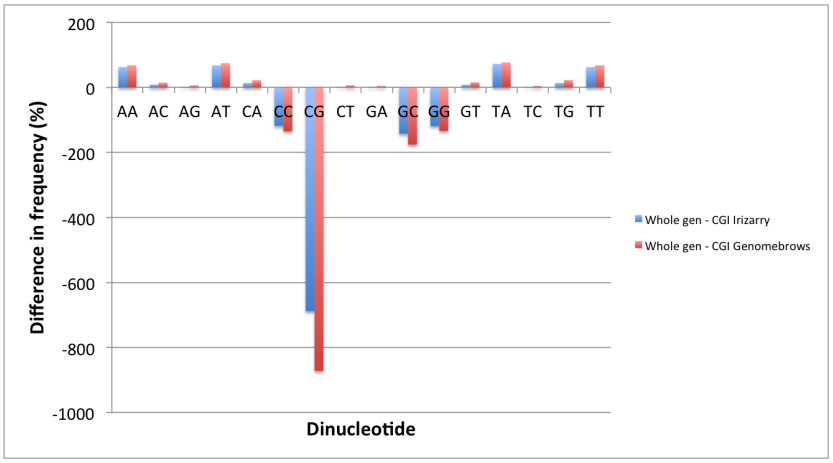


Figure 4.2: Percentage difference in dinucleotide content between the whole genome sequence and regions annotated as CGIs (respectively by Irizarry et.al. and from the Genome Browser).

Dinucleotide	Percentage Diff CGI Irizarry	Percentage Diff CGI GenomeBrows
AA	62.2207	67.6248
AC	8.5032	14.4127
AG	1.5256	5.8747
AT	67.6853	74.0609
CA	13.5016	22.0579
CC	-118.1076	-134.5846
CG	-688.2931	-871.5959
CT	1.3945	6.1609
GA	0.3905	4.4806
GC	-142.2971	-176.0005
GG	-118.9436	-133.6498
GT	8.0488	15.3648
TA	72.1283	76.4092
TC	0.7837	4.5052
TG	13.5436	22.3525
TT	62.1579	67.6674

Table 4.2: Comparison of the percentage differences in dinucleotide frequencies between the whole genome sequence and inside regions annotated as CGIs (by Irizarry et.al. and from the Genome Browser).

Inter-dinucleotide distance distributions inside CGIs

We then apply our distance-based analysis to dinucleotides inside regions mapped as CpG islands: the distributions in log-log scale are shown in Figure 4.3 and some distribution properties are listed in Table 4.3.

We can see that the distributions corresponding to CpG islands are shifted towards lower values, and naturally we see a cutoff for longer distances due to their finite size. If we now want to compare the distributions more quantitatively, recall that the mean and median CpG distance in the whole genome are respectively 100,4 and 41 base pairs: it is therefore evident that CpG inside annotated CpG islands have a very different distribution, as expected. Their mean and median values, in fact, are comparable to those of the other dinucleotides in the whole genome. A few dinucleotides show increased mean and long distances inside CpG islands: on one hand this could be due to the increased presence of CpG dinucleotides, or there could also be an effect of repetitive regions which are not included in annotated CpG islands.

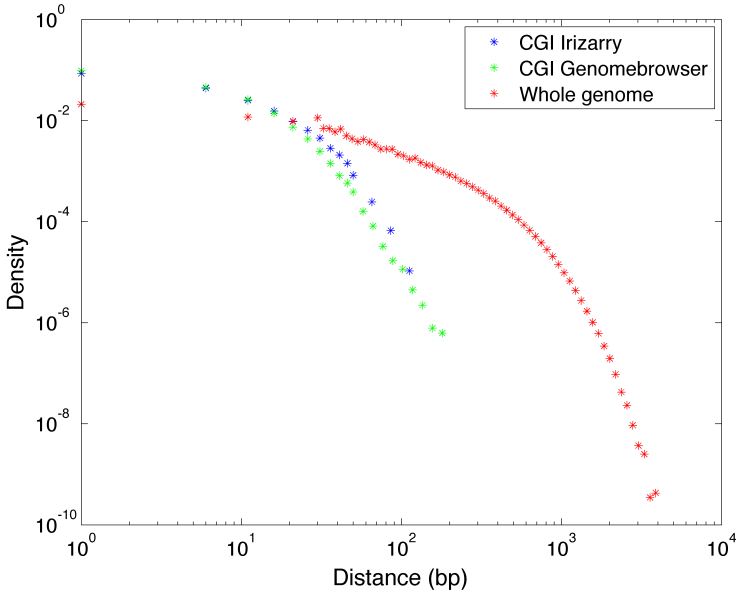


Figure 4.3: Log log plot of CpGs distances distribution in the whole genome and inside CpG islands corresponding to two different annotations.

	Inside HMM CGI		Inside GenomeBrowser CGI	
Dinucleotide	Mean dist	Median dist	Mean dist	Median dist
AA	34.30	19	42.48	25
AC	19.72	12	21.72	14
AG	13.30	9	14.09	9
AT	37.62	20	49.68	27
CA	14.31	9	16.13	11
CC	10.04	6	9.40	6
CG	11.15	7	8.82	6
CT	13.53	9	14.54	9
GA	15.65	10	16.54	11
GC	8.11	5	7.03	5
GG	10.15	6	9.42	6
GT	19.56	12	21.75	14
TA	51.73	26	65.57	37
TC	15.80	10	16.78	11
TG	14.36	9	16.55	11
TT	33.90	18	42.66	24

Table 4.3: Properties of inter-dinucleotide distance distributions computed inside CpG islands, according to two reference CGI annotations.

4.2 Promoters

Promoter regions 1000bp and 2000bp upstream of annotated genes were obtained from the Genome Browser database. As introduced in the first chapter, we expect promoters to be associated with CpG islands, therefore to find an enrichment in CpGs in these regions. In order to assess this we first compare dinucleotide abundances and then the inter-dinucleotide distance distributions, as done previously for CpG islands.

Dinucleotide frequencies inside promoter regions

First we study dinucleotide abundances in promoters: the relative frequencies are listed in Table 4.4 and shown in Figure 4.4 for regions 1000bp and 2000bp upstream of genes and in the whole genome. The corresponding percentage difference values are listed in Table 4.5 and displayed in Figure 4.5. Notice how CpGs show the highest difference.

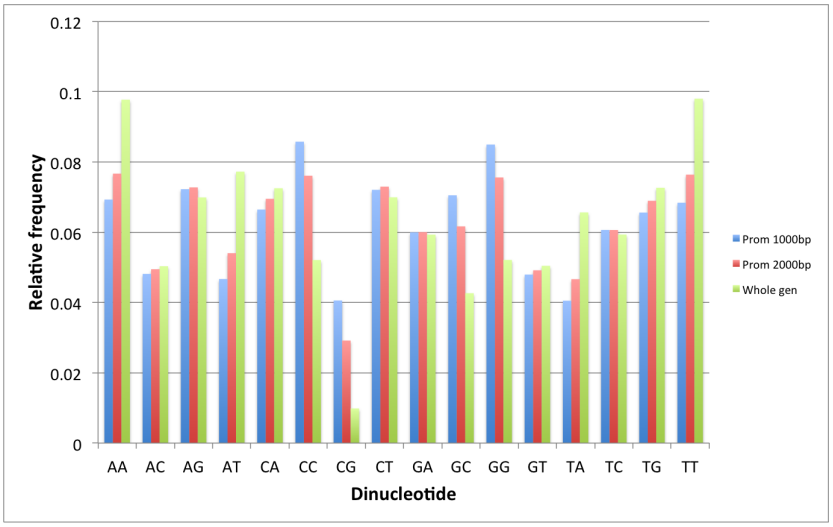


Figure 4.4: Comparison of the relative frequencies of the 16 dinucleotides inside promoter regions 1000 and 2000 bp upstream of genes and in the whole genome.

Dinucleotide	Rel freq 1000bp prom	Rel freq 2000bp prom	Rel freq Whole gen
AA	0.069313286	0.076693881	0.09774721
AC	0.048156293	0.049483474	0.050339835
AG	0.072263816	0.072772915	0.069924165
AT	0.046696086	0.054055506	0.077258419
CA	0.066472835	0.069523924	0.072535448
CC	0.085789358	0.076102362	0.052096925
CG	0.040585433	0.029185279	0.009861512
CT	0.072081175	0.072994215	0.069961111
GA	0.059982353	0.060057848	0.059335043
GC	0.070527912	0.061682248	0.042660269
GG	0.084967986	0.07560343	0.052125953
GT	0.04796255	0.049181298	0.050454531
TA	0.040516471	0.046652845	0.065651918
TC	0.060665188	0.060647377	0.059358019
TG	0.065607206	0.068959194	0.072664114
TT	0.068412052	0.076404206	0.098025528

Table 4.4: Comparison of the relative frequencies of the 16 dinucleotides inside promoter regions 1000 and 2000 bp upstream of genes and in the whole genome.

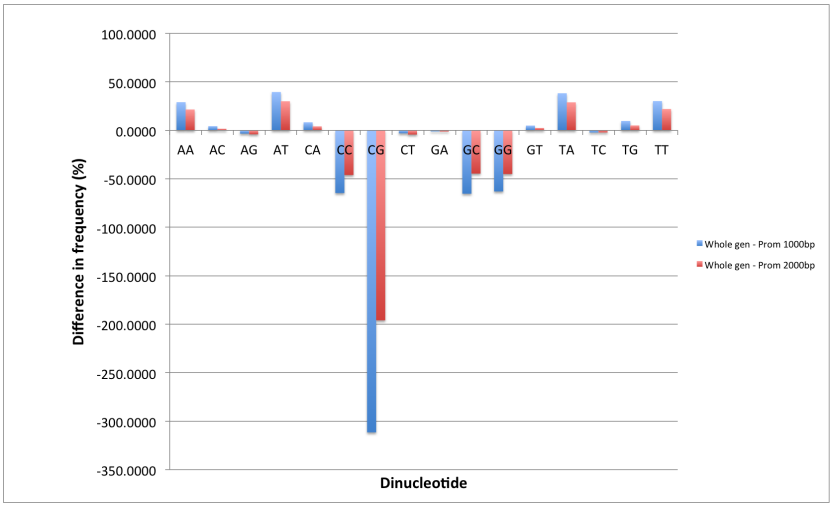


Figure 4.5: Percentage difference in dinucleotide content between the whole genome sequence and promoter regions 1000bp and 2000bp long.

Dinucleotide	Percentage Diff Prom 1000	Percentage Diff Prom 1000
AA	29.0892	21.5385
AC	4.3376	1.7012
AG	-3.3460	-4.0741
AT	39.5586	30.0329
CA	8.3581	4.1518
CC	-64.6726	-46.0784
CG	-311.5539	-195.9514
CT	-3.0303	-4.3354
GA	-1.0909	-1.2182
GC	-65.3246	-44.5894
GG	-63.0051	-45.0399
GT	4.9391	2.5235
TA	38.2859	28.9391
TC	-2.2022	-2.1722
TG	9.7117	5.0987
TT	30.2100	22.0568

Table 4.5: Comparison of the percentage differences in dinucleotide frequencies between the whole genome sequence and promoter regions 1000bp and 2000bp long.

Inter-dinucleotide distance distributions inside promoter regions

In Figure 4.6, the inter-dinucleotide distance distribution for CpG is shown on log-log scale for the whole genome and 1000bp and 2000bp long promoter regions. In Table 4.6 we also list distribution properties for the two distributions. Both from the log log plot and the mean and median values we can see evidence that CpGs are much more clustered inside promoter regions when compared to whole genome.

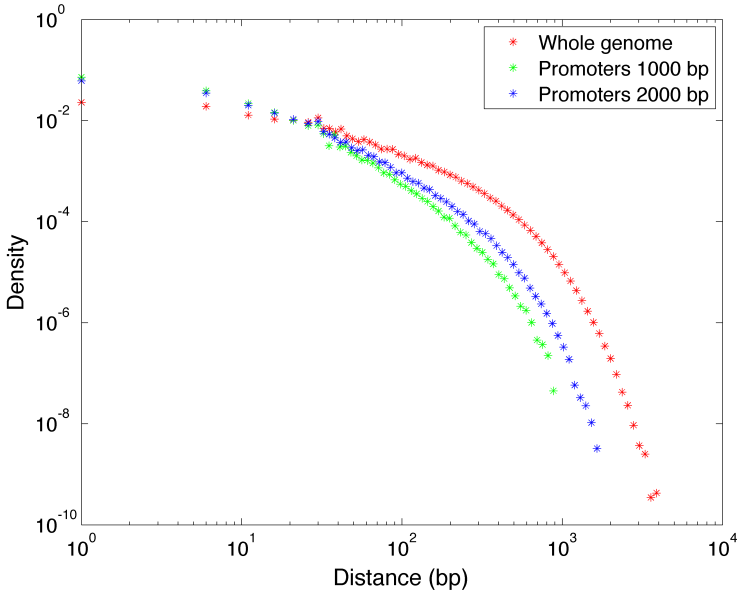


Figure 4.6: Log log plot of CpGs distances distribution in the whole genome and inside 1000bp and 2000bp promoter regions.

	Promoter regions 1000bp		Promoter regions 2000bp	
Dinucleotide	Mean dist	Median dist	Mean dist	Median dist
AA	17.65	9	16.45	9
AC	19.17	12	18.92	12
AG	12.60	8	12.62	8
AT	19.13	10	16.96	8
CA	13.75	9	13.24	9
CC	14.16	7	16.03	8
CG	21.33	9	31.30	11
CT	12.65	8	12.59	8
GA	15.27	9	15.45	9
GC	12.94	7	15.08	9
GG	14.31	8	16.15	8
GT	19.34	13	19.08	12
TA	21.80	11	19.64	10
TC	15.12	9	15.31	9
TG	13.97	9	13.37	9
TT	17.96	9	16.58	9

Table 4.6: Properties of inter-dinucleotide distance distributions in promoter regions 1000bp and 2000bp long.

4.3 LADs

LADs (Lamina Associated Domains) are portions of the genome which interact with the nuclear lamina, a dense fibrillar network found inside the nucleus which provides mechanical support and also regulates important cellular events such as DNA replication and cell division. These regions can be mapped experimentally using the DamID technique [45], which is based on targeted adenine methylation of DNA sequences that interact *in vivo* with the protein of interest. In [21], a high resolution map of nuclear lamina interactions was generated with this method for the whole human genome. In the same work the authors also report an enrichment of the number of CpG islands in proximity of LADs: we therefore decided to study CpGs abundance and distribution in these regions in order to gather insight into possible relationships between them and their CpG content.

In table 4.7, the total number and average length (in base pairs) of the LAD regions are listed for each chromosome. It is important to underline how LADs cover a significant portion of the human genome (approximately the 40%), however these regions are not expected to interact with the nuclear lamina simultaneously, as chromosome positioning is to be considered stochastic. Given the considerable size of these regions we decided to compare dinucleotide abundances and distributions as computed inside and outside of LADs, instead of inside LADs and in the whole genome (as was done previously for CGIs and promoters). The data in [21] were obtained for the human genome release hg18, so in order to use the LAD coordinate in our study (which is performed using the updated release hg19) they had to be converted using the LiftOver tool available at the Genome Browser website[1].

Following the same outline of the previous sections, we first compare dinucleotide frequencies inside and outside of LADs. We then move on to the results of our distance-based analysis applied to these regions. Finally we illustrate the findings of the same analysis repeated on smaller regions of interest located at the boundaries of LADs.

[1] http://genome.ucsc.edu/cgi-bin/hgLiftOver

Chromosome	Number of LADs	Mean Length (bp)
1	97	802792
2	97	989033
3	105	704726
4	92	1084255
5	91	880220
6	77	880076
7	71	928253
8	70	935566
9	48	902770
10	63	813983
11	70	797328
12	59	788448
13	44	796041
14	47	820431
15	35	688281
16	36	910199
17	24	678727
18	38	942753
19	19	405305
20	34	694245
21	16	568768
22	6	984732
X	59	1334770
Y	4	2397593

Table 4.7: Statistics of LADs: number and mean length for each chromosome.

Dinucleotide frequencies inside and outside LADs

In Table 4.8 and Figure 4.7 we compare the relative frequencies of dinucleotides in genomic areas inside and outside LADs. The corresponding percentage difference values are listed in Table 4.9 and displayed in Figure 4.8. CpGs show the highest difference, being less abundant inside than outside LADs. A similar behaviour is observed for CC, GC and GG whereas AA, AT ,TA and TT are more abundant inside LADs.

Dinucleotide	Rel freq inside LADs	Rel freq outside LADs
AA	0.103194	0.094100
AC	0.050136	0.050500
AG	0.068422	0.070900
AT	0.083351	0.073200
CA	0.071831	0.073000
CC	0.046911	0.055500
CG	0.007550	0.011400
CT	0.068447	0.071000
GA	0.059110	0.059500
GC	0.038546	0.045400
GG	0.046923	0.055600
GT	0.050193	0.050600
TA	0.070968	0.062100
TC	0.059147	0.059500
TG	0.071877	0.073200
TT	0.103394	0.094500

Table 4.8: Comparison of the relative frequencies of the 16 dinucleotides inside and outside LADs.

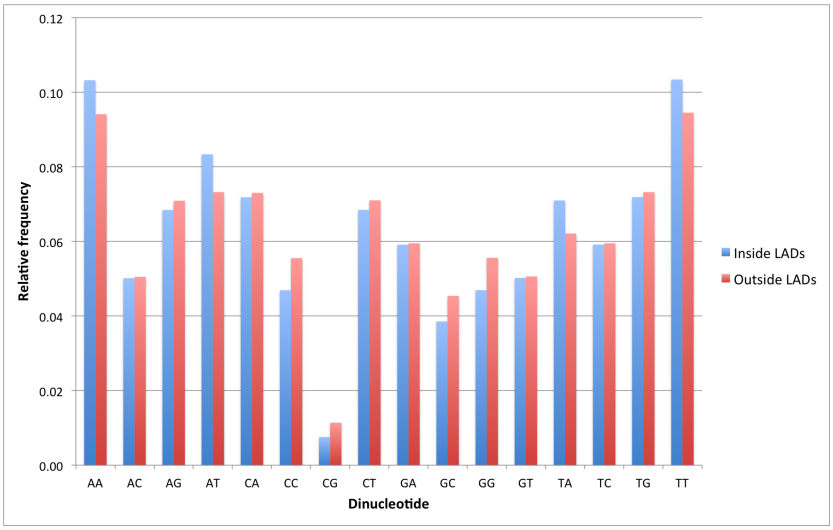


Figure 4.7: Comparison of the relative frequencies of dinucleotides inside and outside LADs.

Dinucleotide	Percentage Diff Inside-Outside LADs
AA	8.8123
AC	-0.7261
AG	-3.6217
AT	12.1786
CA	-1.6274
CC	-18.3084
CG	-50.9863
CT	-3.7299
GA	-0.6592
GC	-17.7823
GG	-18.4914
GT	-0.8109
TA	12.4952
TC	-0.5973
TG	-1.8411
TT	8.6024

Table 4.9: Comparison of the percentage differences in dinucleotide frequencies between genomic regions inside and outside of LADs.

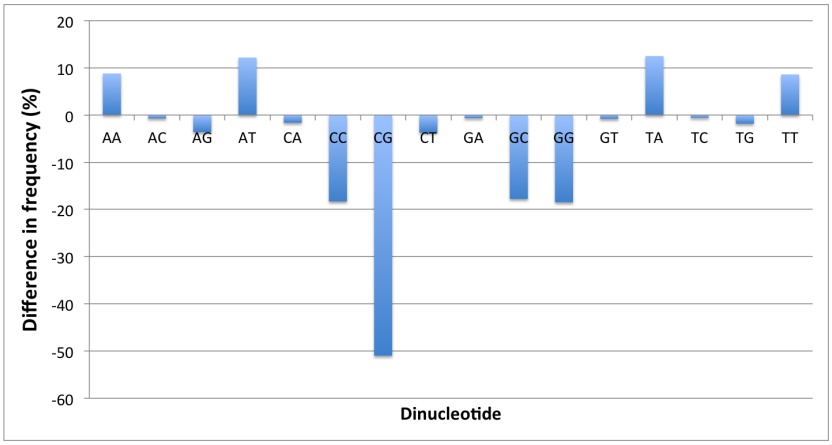


Figure 4.8: Percentage difference in dinucleotide content between genomic regions inside and outside LADs.

Inter-dinucleotide distance distributions inside and outside LADs

We now comment the results of our inter-dinucleotide distance analysis repeated inside and outside LADs. The log log plot for CpGs is shown in Figure 4.9. As done earlier we also list some properties of the two distributions in Table 4.10.

From these results and the one reported earlier we see that there is a difference in CpGs distribution inside and outside LADs: inside they tend to be more scarce and have higher mean and median distance values. However, this difference is not strong enough to be employed as a "marker" for candidate LADs in a genomic sequence.

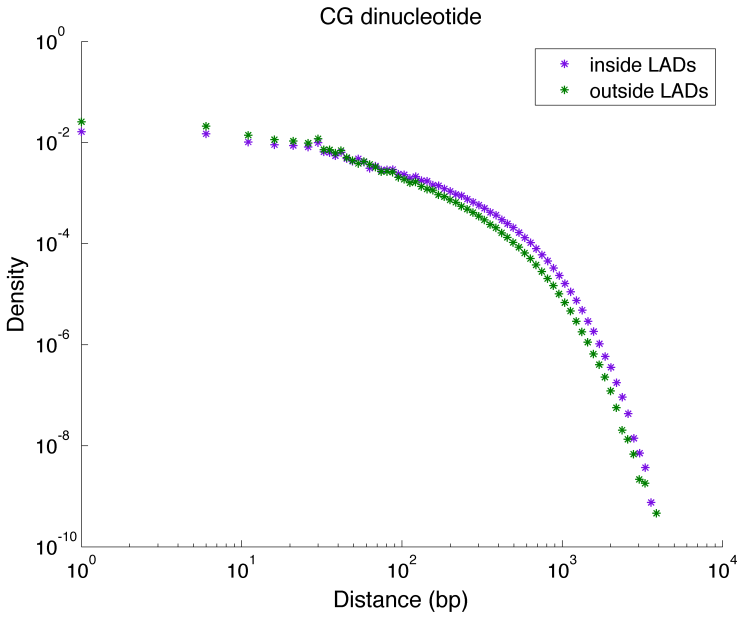


Figure 4.9: Distribution of inter-dinucleotide distances for CpGs, computed inside and outside LADs.

	Inside LADs		Outside LADs	
Dinucleotide	Mean dist	Median dist	Mean dist	Median dist
AA	12.49	7	13.78	8
AC	18.95	12	18.81	12
AG	13.61	9	13.10	8
AT	11.00	7	12.66	8
CA	12.92	9	12.70	9
CC	25.51	15	21.75	12
CG	131.41	61	86.77	35
CT	13.61	9	13.09	8
GA	15.92	10	15.81	9
GC	24.94	15	21.03	12
GG	25.51	15	21.73	12
GT	18.92	12	18.75	12
TA	13.09	8	15.09	8
TC	15.91	10	15.81	9
TG	12.91	9	12.66	9
TT	12.47	7	13.73	8

Table 4.10: Properties of inter-dinucleotide distance distributions in regions inside and outside LADs.

Analysis at LAD borders

After performing a comparison of CpG content and distribution inside and outside LADs, we concentrate our attention on LAD boundaries (see Figure 4.10): Guelen and others found in their work 4.7 that CpG islands occurred with a higher frequency in these areas relative to the remainder of inter-LAD regions. We therefore wanted to test whether our inter-dinucleotide distance method reflected this evidence and could give further insight.

We analysed inter-dinucleotide distances inside LADs boundary regions using windows of different sizes: 5,000, 10,000, 100,000 and 1,000,000 base pairs. The mean values of the distributions are listed in Table 4.11 for chromosome 1. We can see that on average CpGs are characterised by shorter distances in pre-LAD regions, however the difference is not very sharp. If we analyse the mean distance values for all LADs in chromosome 1 one by one we find a great variability: the smaller mean value found is 24 base pairs and the biggest is 300 base pairs. Overall in approximately the 30% of cases (30 LADs on a total of 97) we find a "promoter" region with a mean inter-CpG distance lower that both the values inside and outside LADs.

This result is consistent with the observation reported in [21], however it also suggests that CpG distribution cannot be exploited as a simple "marker" for LADs. On the other hand it is possible that a combination with other descriptors could be more effective or that LADs may be divided into different categories related to their CpG content and distributions.

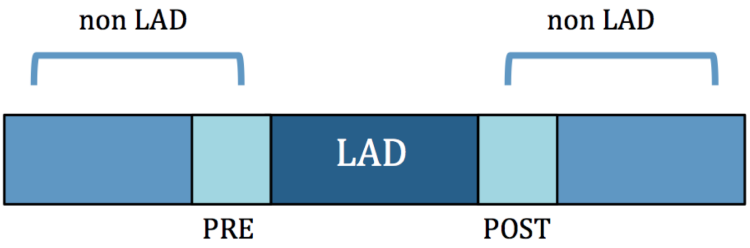


Figure 4.10: LAD regions of interest.

Dinucleotide	5.000 bp	10.000 bp	100.000 bp	1.000.000 bp
AA	13.88	14.06	13.46	13.39
AC	19.30	19.25	18.87	18.90
AG	13.21	13.16	13.11	13.02
AT	12.41	12.75	12.41	12.27
CA	13.07	13.06	12.89	12.79
CC	21.92	21.41	22.55	22.59
CG	89.83	84.86	97.85	104.74
CT	12.70	12.69	12.93	13.00
GA	15.84	15.97	15.92	15.73
GC	21.36	20.98	21.90	22.12
GG	21.92	21.83	22.61	22.60
GT	18.36	18.50	18.75	18.89
TA	14.64	14.93	14.46	14.47
TC	15.47	15.46	15.70	15.71
TG	12.40	12.58	12.81	12.79
TT	13.24	13.32	13.27	13.41

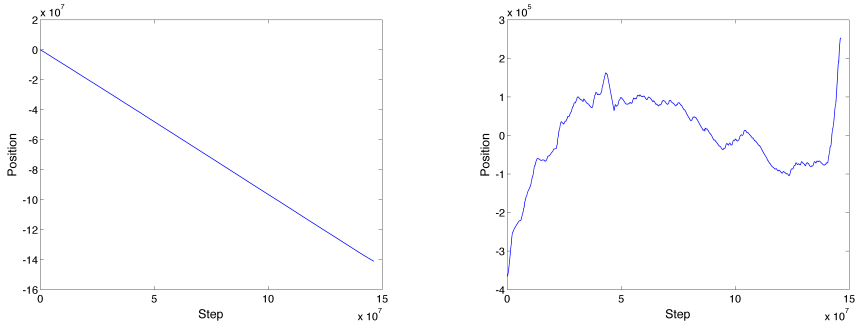
Table 4.11: Comparison of the mean distance computed with different window sizes in the pre-LAD area of chromosome 1 for all dinucleotides.

4.4 Random walk analysis of the DNA sequence

In this section we describe the results of our DNA walk analysis on dinucleotides in the human genome. In order to perform this kind of study, the nucleotide sequence has to be converted into a numerical sequence: as explained in the Methods chapter we choose a conversion into zeros and ones, where the ones correspond to the position of CG dinucleotides (or other nucleotide couples studied).

When we define a random walk on the DNA sequence such as the one described in the Methods chapter (go up when a CG dinucleotide is encountered, go down otherwise) we obtain for the CpG dinucleotide a plot of the walker position like the one displayed in Figure 4.11a for chromosome 1, taken here as an example. The plot is comparable to one expected for a random walk with drift, where the drift is caused here by the fact that it is much more likely to encounter a dinucleotide different from a CpG (and therefore go down) than a CpG. Given our knowledge of the values of dinucleotide frequencies in the chromosome, we can globally "detrend" the plot removing this effect in order to better visualise CpGs: the result of such an operation is shown in Figure 4.11b.

If we zoom in to observe a sample chromosomal region (as shown in Figure 4.12) we can notice "spikes" in which curve grows faster: if we superimpose CGI coordinates (i.e. the ones mapped by Irizarry) on this plot the two match as can be seen in the figure.



(a) DNA walk

(b) Detrended DNA walk

Figure 4.11: DNA walks for chromosome 1.

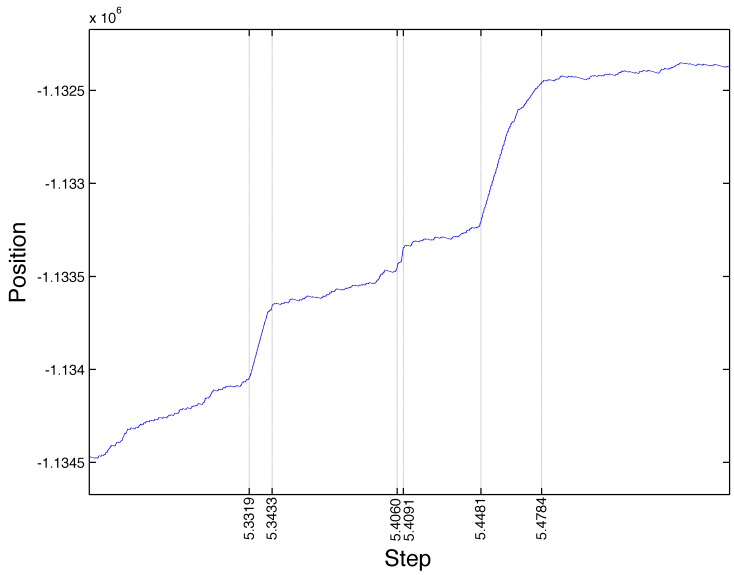


Figure 4.12: Detrended DNA walk for CpG dinucleotides in chromosome 1. The plot has been zoomed in and X coordinates correspond to start/end positions of CGI.

Prompted by this observation we then computed the slope of the walk inside and outside all genomic areas annotated as CpG islands in chromosome 1 (where 4898 of them are found) for the 16 walks constructed for the different dinucleotides. The histogram of slopes distribution for CpGs is shown in Figure 4.13, and some characteristic features of the distributions for the 16 walks are listed in Tables 4.12 and 4.13 . Note how the slope of the CpG walk inside CGI is approximately centered around a maximum between 0.2 and 0.3, and the distribution is slightly skewed to the right. In addition to this, the slope value for the CpG walk never becomes negative, as opposite to the other dinucleotides. As a quantitative measure of difference between the distributions we also list in Table 4.14 the values of the Kullback-Leibler divergence between the slope distribution inside and outside CpG island. Notice how CpGs are characterised by the highest difference, however other dinucleotides such as AC, CA, GC and TG have high values.

The above findings suggest that the slope of the DNA walk might be exploited, perhaps together with other methods, as an indicator for candidate CpG islands.

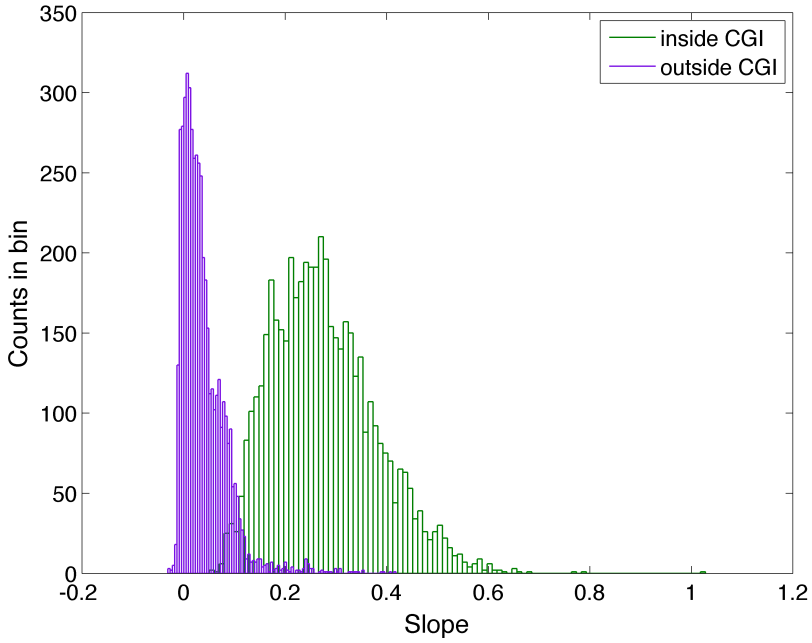


Figure 4.13: Distribution of slope values of the DNA walk, as computed inside and outside CpG islands in chromosome 1 (according to the annotation by Irizarry).

Dinucleotide	Min slope	Max slope	Mean slope	Median slope	Variance slope
AA	-0.2435	0.3787	-0.1257	-0.1394	0.0051
AC	-0.1728	1.2143	0.0138	0.0042	0.0073
AG	-0.2437	0.4828	0.0335	0.0324	0.0060
AT	-0.2667	0.3700	-0.1679	-0.1797	0.0037
CA	-0.2486	1.1546	0.0051	-0.0062	0.0085
CC	-0.1425	0.8828	0.1872	0.1812	0.0119
CG	0.0507	1.0277	0.2759	0.2648	0.0105
CT	-0.2442	0.5632	0.0359	0.0353	0.0057
GA	-0.2056	0.6118	0.0339	0.0297	0.0061
GC	-0.1452	0.7023	0.2690	0.2680	0.0140
GG	-0.1423	0.8577	0.1894	0.1870	0.0123
GT	-0.1734	0.9076	0.0167	0.0060	0.0070
TA	-0.2292	0.4680	-0.1571	-0.1678	0.0027
TC	-0.2058	0.6131	0.0312	0.0278	0.0058
TG	-0.2493	1.0440	0.0077	-0.0033	0.0082
TT	-0.2443	0.2962	-0.1271	-0.1383	0.0049

Table 4.12: Characteristics of the slope distribution for DNA walks corresponding to the different dinucleotides, as computed inside CpG islands in chromosome 1.

Dinucleotide	Min slope	Max slope	Mean slope	Median slope	Variance slope
AA	-0.2435	0.5706	-0.0145	-0.0082	0.0070
AC	-0.1728	0.7683	0.0276	0.0263	0.0022
AG	-0.2434	0.6431	0.0540	0.0527	0.0027
AT	-0.2667	0.6744	-0.0385	-0.0320	0.0070
CA	-0.2473	0.8573	0.0485	0.0490	0.0031
CC	-0.1425	0.8299	0.0773	0.0603	0.0065
CG	-0.0310	0.4185	0.0388	0.0279	0.0020
CT	-0.2431	0.8808	0.0543	0.0525	0.0030
GA	-0.2056	0.6435	0.0358	0.0350	0.0024
GC	-0.1452	0.4954	0.0762	0.0641	0.0050
GG	-0.1423	0.7524	0.0745	0.0584	0.0067
GT	-0.1734	0.7307	0.0275	0.0270	0.0017
TA	-0.2292	0.6336	-0.0418	-0.0386	0.0066
TC	-0.2058	0.9192	0.0385	0.0360	0.0028
TG	-0.2493	0.6776	0.0474	0.0482	0.0030
TT	-0.2443	0.5746	-0.0156	-0.0076	0.0067

Table 4.13: Characteristics of the slope distribution for DNA walks corresponding to the different dinucleotides, as computed outside CpG islands in chromosome 1.

Dinucleotide	KL divergence
AA	1.581
AC	3.780
AG	0.844
AT	2.761
CA	3.617
CC	1.793
CG	4.589
CT	2.588
GA	1.236
GC	3.549
GG	1.403
GT	2.378
TA	1.762
TC	2.312
TG	4.167
TT	2.697

Table 4.14: Values of Kullback-Leibler divergence between DNA walk slope values computed inside and outside regions mapped as CpG islands.

CHAPTER 5

Conclusions and future directions

CpG islands (CGIs) are functionally important regions of DNA which can show altered methylation patterns in many diseases, including cancer. Definitions of CGIs and algorithms for their annotation have been proposed since the 80s, and are still very debated.
In this work we approached this issue by first pursuing a more universal characterisation of CpGs in the human genome, studying their distribution and abundance and comparing them with the other dinucleotides and suitable random models.

We first analysed dinucleotide frequencies in the DNA sequence, finding a very low abundance for CpGs as expected from the biological processes of methylation and mutation that act on the cytosine in the couple. We then generated synthetic random reference sequences using a zeroth and first order Markov chain model, with transition probabilities estimated from dinucleotide and nucleotide abundances in the biological sequence. We find that the Markov 0 model greatly overestimates CpG content in the sequence: for example in chromosome 1 the percentage difference between CpGs relative frequency in the real sequence and in the synthetic one is over 300%. On the other hand, the Markov 1 model is capable of reproducing dinucleotide abundances very well, with a percentage difference that is less than 0.04% in the case of chromosome 1. From the analysis of inter-dinucleotide distances distribution, performed on the whole genome, we see that this method actually captures the intrinsic difference

in CpG positioning along the sequence. In fact, the histogram of the inter-CpG distance distribution clearly stands out from the other 15 plots, and the difference can also be emphasised quantitatively by looking at characteristic values of the distribution such as the mean or median, or by computing the Kullback-Leibler divergence between the distributions. As a random reference for the distance distributions we employ the geometric distribution: in the case of the CpG dinucleotide, we find that small distances (up to approximately 50bp) and very long distances (from approximately 400bp onwards) are enriched in the biological sequence; on the other hand intermediate distance values are more likely in the geometric reference distribution. The other dinucleotides also show differences from the random case: in order to better evaluate it we performed hierarchical clustering of the binned difference between the "theoretical" and "experimental" distributions, using Spearman correlation coefficient as metric. In this analysis we find that CpGs are attributed to a unique cluster separated from the other 15 dinucleotides: among these, AA and TT especially stand out but overall the distance between nested clusters is very small. We are currently not aware of possible biological motivations behind this difference, however in our opinion it should deserve further attention.

After this first characterisation of dinucleotide distribution in the whole genome, we moved on to focus on specific genomic regions of interest.
First we analysed dinucleotide abundances and distributions inside CpG islands, using two different annotations: as expected we find a high relative frequency of CpGs and much shorter inter-dinucleotide distances. We also study CpGs inside promoter regions, 1000 and 2000 base pairs upstream of genes, which are known to associate with CGIs. Our results confirm that CpGs (and the associated dinucleotides CC, GC and GG) are enriched in these areas. The last type of genomic regions we consider are LADs: Lamina Associated Domains, for which an increased frequency of CGI was reported at the boundaries. Our results show that CpGs are less abundant and on average more distant inside LADs than outside, however the difference is not extremely sharp. Analysis at boundaries found that approximately 30% of LADs in chromosome 1 are characterised by a "promoter" regions of clustered CpGs: this is consistent with previous observation, and suggests that CpGs alone cannot be exploited as a "marker" for LADs. However it might be possible to divide LADs in different categories according to their CpG association. Finally we employed a DNA walk representation based on dinucleotide positioning: for CpGs we obtain a heavily biased walk which, once detrended, shows sharp increases in slope in correspon-

dence with CpG islands. This is confirmed by a global analysis of slope values inside and outside regions annotated as islands, and we think that this framework could be exploited - perhaps in conjunction with different methods - in order to emphasise candidate CpG island regions.

In addition to this DNA walk approach applied to CpG islands annotation, other future directions could include for example the integration in our analyses of experimental methylation data. This would enable us to superimpose "markers" to the plain sequence data, indicating positions in which the cytosine is actually methylated. Another very interesting study which could shed light on the positioning and functional role of CpG sites would be to relate these genomic regions with the corresponding positions in the 3D chromosomal structure, which can now be obtained as a contact map by conformation captures techniques such as Hi-C.

Overall the results obtained confirm how powerful these quantitative methods can be for the study of DNA sequences: in fact, we see that these data analysis techniques are extremely effective at capturing biologically relevant features, such as the unique properties of CpGs distribution inside the human genome.

APPENDIX A

Practical details

This appendix covers practical details for the analysis of dinucleotides in the human genome sequence. In the first section we illustrate two important bioinformatics data formats, which were used for the analyses outlined in this thesis. Then, an overview of the databases and resources for DNA sequences analysis, which were employed in this work, is given. In the third section we describe the functions and scripts used to perform the different analyses, such as the computation of the inter-dinucleotide distance distributions in the sequence and the generation of random reference sequences using a zeroth and first order Markov chain model. The code was implemented in Matlab, and the Bioinformatics Tollbox provides some useful functions (for example to read FASTA format files). Finally we point at a few specific issues that need to be taken into consideration when performing these kinds of analyses.

A.1 Bioinformatics data formats

Two important bioinformatics data formats which are used in this analysis are now described: the FASTA format for the reference human genome sequence, and the GFF format for genomic regions of interest (CpG islands, LADs).

FASTA format

The FASTA file format is a widely used text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. A sequence in FASTA format begins with a single-line description, followed by lines of sequence data. The description line is distinguished from the sequence data by a greater-than ">" symbol in the first column. The word following the > symbol is the identifier of the sequence, and the rest of the line is the description (both are optional). As an example, the description line of human chromosome 14 is:

>gi|568336010|gb|CM000676.2|Homo sapiens chromosome 14, GRCh38 reference primary assembly.

GFF format

GFF (General Feature Format) is a format for describing genes and other features associated with DNA, RNA and Protein sequences. A GFF record is an extension of a basic (name,start,end) tuple that can be used to identify a substring of a biological sequence. The fields of a GFF file are:

- <seqname>, the name of the sequence. Normally it will be the identifier of the sequence in an accompanying FASTA format file, or the identifier for a sequence in a public database.
- <source>, the source of this feature (database or project name).
- <feature>, the feature type name (e.g. gene, variation...).
- <start>, start position of the feature (inclusive), with sequence numbering starting at 1.
- <end>, end position of the feature (inclusive).
- <score>, a floating point value.
- <strand>, defined as + (forward) or - (reverse).
- <frame>, one of '0', '1' or '2'. '0' indicates that the first base of the feature is the first base of a codon, '1' that the second base is the first base of a codon, and so on...
- [attributes], semicolon-separated list of tag-value pairs, providing additional information about each feature.

A.2 Useful resources

The main resources that were employed in this work in order to perform analyses on the human genome sequence are illustrated here, including genomic regions of interest (such as CpG islands, Lamina Associated Domains ...).

Human DNA sequence

The sequence of the human genome, which is continuously updated by the Human Genome Consortium, can be retrieved from publicly available databases. The download of sequence and annotation data can be made from the UCSC website: `http://hgdownload.cse.ucsc.edu/downloads.html#human`. Under the "Human Genome" section one can find the different releases, starting from the most recent one (currently hg38). It is possible to retrieve the full data set or the single chromosomes in a compressed FASTA format, which can then be viewed and modified using a simple text editor. Prepackaged promoters region sequences 1000bp, 2000bp and 5000 bp long can be downloaded from the same website in the bigZips downloads directory. Similarly, a complete copy of the entire known genes data set can be downloaded from the "Annotation" folder.

Genomic regions of interest

The genomic coordinates corresponding to many regions of interest can be extracted from the Genome Browser "Table Browser": `https://genome.ucsc.edu/cgi-bin/hgTables?org=human`. In this tool, the user has to select the genome release of interest, and the specific "track" that is to be retrieved: for example standard CpG islands annotations can be found under the"CpG islands" track. The "get output" button will produce a table containing the coordinates of the regions of interest and additional information (for example in this case the lengths of each CGI, or the O/E ratio), which can be easily copied into a spreadsheet file. The user interface of this tool is shown in Figure A.1. Other regions of interest which have been mapped experimentally can usually be found in the supplementary material of the corresponding paper, for example the list of LADs employed in this work can be retrieved at: `http://www.nature.com/nature/journal/v453/n7197/extref/nature06947-s2.txt` and HMM-based CpG islands coordinates can be found at `http://rafalab.jhsph.edu/CGI/`.

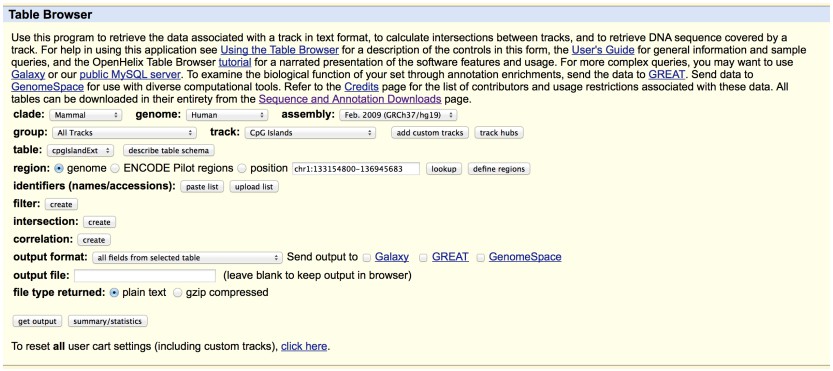


Figure A.1: User interface of the Table Browser tool.

A.3 Implementation of the analyses

The program for dinucleotide analysis in DNA sequences is described here in detail, along with the different studies that can be performed with it. The code was implemented in Matlab, and we also employ a few functions from the Bioinformatics Toolbox.

Inter-dinucleotide distance distribution

The first part of the program is dedicated to the analysis of inter-dinucleotide distance distributions in the sequence. These can be calculated using the calc_dist function, which has the nucleotide sequence and the dinucleotide of interest as inputs and gives the computed distance values as output in an array. In case no dinucleotides are found (this can happen when working with small genomic regions), the calc_dist function returns an empty array.
The results obtained can conveniently be stored in a 1x16 cell array for all the different dinucleotides. The distributions are then plotted on a log log scale, using partial logarithmic binning for the reduction of noise in the tails. The partial_log_bin function allows controlling of parameters such as the threshold from which logarithmic binning is applied and the bin width to apply in the other region. In order to quantitatively compare the distributions, characteristic values such as the maximum or mean are computed, as well as the Kullback-Leibler divergence between all dinucleotides couples. The KL-divergence values can also be conveniently displayed as an easy-to-read heat map.

Analysis in specific regions of interest

The above methods can be easily extended to analyse specific regions of interest: for example one may be interested in inter-dinucleotide distance distributions inside promoter regions or LADs. For promoter regions, one simply has to download the sequences (as described in the previous chapter) and use the same functions such as calc_dist to compute the desired distribution. In the case of LADs (or also CGI), the start and end coordinates are usually provided: to compute the distribution inside these regions one simply needs to give the corresponding fragment of sequence as an input to calc_dist, for example as in calc_dist(ch.Sequence(start:end), nuc). The results can then be compared with the one previously obtained for the whole chromosome sequence: however, if the regions of interest are very long (for example this is the case with LADs) it is better to repeat the computation inside and outside of it, and compare these two.

Random reference models

Two methods for the generation of random reference sequences are provided, which are based respectively on a zeroth and first order Markov chain model. The zeroth-order chain is generated using the randseq function, which takes as inputs the desired sequence length and the weights of the four nucleotides. These can be estimated from the biological sequence of interest, as the relative frequency of nucleotides computed with the basecount function. For the generation of the first order sequence the hmmgenerate function is employed, using an emission matrix equal to one and obtaining the state sequence, which is then converted into the four letters A, C, G, T. The choice to employ this function was motivated by efficiency reasons. The transition probabilities are calculated from the nucleotide and dinucleotide relative frequencies in the chromosome sequence, estimated with the basecount and dimercount functions (note that dinucleotide counts are "overlapping", i.e. AAA counts as two dinucleotides).

DNA walk

The DNA walk is constructed as follows: the walker steps up when the dinucleotide of interest (i.e. CG) is encountered, otherwise it steps down. As expected, a biased walk is obtained in this way, as it is much more likely to encounter a different dinucleotide rather than CG. After detrending the walk globally with the detrend function, it is interesting to plot it with the CGI start and end coordinates on the x axis: by zooming in different regions it will be

possible to notice how the slope changes inside CGIs. This observation is then quantified by computing the slopes both inside and outside all CGIs of a given chromosome, for all 16 dinucleotides. Finally, the histogram of the two slope distributions is shown and the minimum, mean, median and maximum values are computed for both of them.

A.4 Specific issues and additional considerations

Different issues can be encountered when performing the type of sequence analyses outlined in thesis: here some are pointed out and more general advice is given.

The choice of genome release should be considered carefully: the latest release should in general be the best option, but this also depends on the analyses that one intends to carry out. Genomic coordinates of regions of interest, in fact, are usually referred to specific releases and need to be converted in case the one employed is different from the one that was originally used. This problem was for example encountered here with LAD coordinates, which were relative to a previous release and needed to be converted using the LiftOver tool (`https://genome.ucsc.edu/cgi-bin/hgLiftOver`).

We assessed in initial tests (described in the Methods chapter) that the N letters present in the genomic sequences could be discarded without affecting the distance-based calculation. However, this is an issue that needs to be taken into consideration when working on genomic sequences, and treatment of Ns largely depends on the specific analyses that need to be carried out. For example here it was decided to keep Ns in the conversion of the chromosome in a DNA walk, in order to preserve the original chromosomal coordinates.

If one needs to perform the analyses on non-repeating portions of the genome only, the "masked" sequences can be retrieved from the same databases described in the first chapter. The same is also true if one wishes to study genomes of different organisms.

Bibliography

[1] Assemblathon, 2014. URL: http://assemblathon.org/

[2] Celera Assembler, 2014. URL: http://wgs-assembler.sourceforge.net

[3] Genome Reference Consortium, 2014. URL: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/

[4] Illuminia, 2014. URL: www.illuminia.com

[5] Sequence assembly teaching material by Ben Langmead, 2014. URL: http://www.langmead-lab.org/teaching-materials/

[6] Sanger sequencing, 2014. URL: http://en.wikipedia.org/wiki/Sanger_sequencing

[7] Human Genome Project declared finished: http://web.ornl.gov/sci/techresources/Human_Genome/project/press4_2003.shtml

[8] Carlos A. C. Bastos, Vera Afreixo et al. (2001). Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions, *Journal of Integrative Bioinformatics*.

[9] Berger, J. A., Mitra, S. K., Carli, M., & Neri, A. (2004). Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute*, 341(1), 37-53.

[10] Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, 321, 209-213.

[11] Bock, C., Walter, J., Paulsen, M., & Lengauer, T. (2007). CpG island mapping by epigenome prediction. *PLoS computational biology*, 3(6), e110.

[12] Burge, C., Campbell, A. M., & Karlin, S. (1992). Over-and under-representation of short oligonucleotides in DNA sequences. *Proceedings of the National Academy of Sciences*, 89(4), 1358-1362.

[13] Compeau, P. E., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11), 987-991.

[14] Durbin, R. (Ed.). (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

[15] Feil, R., & Fraga, M. F. (2012). Epigenetics and the environment: emerging patterns and implications. *Nature Reviews Genetics*, 13(2), 97-109.

[16] Feinberg, A. P. & Tycko, B. (2004). The history of cancer epigenetics. *Nature Rev. Cancer* 4, 1-11.

[17] Feinberg, A.P. (2007). Phenotypic plasticity and the epigenetics of human disease. *Nature* 447:433-440

[18] Gardiner-Garden M, Frommer M (1987). CpG islands in vertebrate genomes. *Journal of Molecular Biology* 196 (2): 261-82.

[19] Andrew J. Gentles and Samuel Karlin. (2001). Genome-scale compositional comparisons in eukaryotes, *Genome Research*.

[20] Glass, J. L., Thompson, R. F., Khulan, B., Figueroa, M. E., Olivier, E. N., Oakley, E. J., & Greally, J. M. (2007). CG dinucleotide clustering is a species-specific property of the genome. *Nucleic acids research*, 35(20), 6798-6807.

[21] Guelen, L., Pagie, L. et al (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions, *Nature Letters*.

[22] Hackenberg, M., Previti, C., Luque-Escamilla, P. L., Carpena, P., MartÌnez-Aroza, J., & Oliver, J. L. (2006). CpGcluster: a distance-based algorithm for CpG-island detection. *BMC bioinformatics*, 7(1), 446.

[23] Irizarry, R. A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., & Feinberg, A. P. (2009). The human colon cancer methylome shows similar hypo-and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics*, 41(2), 178-186.

[24] Irizarry, R. A., Wu, H., & Feinberg, A. P. (2009). A species-generalized probabilistic model-based definition of CpG islands. *Mammalian Genome*, 20(9-10), 674-680.

[25] Kapun, E., & Tsarev, F. (2013). De Bruijn superwalk with multiplicities problem is NP-hard. *BMC bioinformatics*, 14(Suppl 5), S7.

[26] Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) CpG islands as gene markers in the human genome. *Genomics* 13, 1095-1107.

[27] Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., ... & Venter, J. C. (2007). The diploid genome sequence of an individual human. *PLoS biology*, 5(10), e254.

[28] Metzker, M. L. (2009). Sequencing technologies - the next generation. *Nature Reviews Genetics*, 11(1), 31-46.

[29] Moore, L. D., Le, T., & Fan, G. (2012). DNA methylation and its basic function. *Neuropsychopharmacology*, 38(1), 23-38.

[30] Mortusewicz O, Schermelleh L, Walter J, Cardoso MC, Leonhardt H (2005). Recruitment of DNA methyltransferase I to DNA repair sites. *Proc Natl Acad Sci USA* 102: 8905-8909.

[31] Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., ... & Venter, J. C. (2000). A whole-genome assembly of Drosophila. *Science*, 287(5461), 2196-2204.

[32] Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., & Stanley, H. E. (1992). Long-range correlations in nucleotide sequences. *Nature*, 356(6365), 168-170.

[33] Peng, C. K., Buldyrev, S. V., Goldberger, A. L., Havlin, S., Sciortino, F., Simons, M., & Stanley, H. E. (1992). Fractal landscape analysis of DNA walks. *Physica A: Statistical Mechanics and its Applications*, 191(1), 25-29.

[34] Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences*, 98(17), 9748-9753.

[35] Pevzner, P. A. (1989). 1-Tuple DNA sequencing: computer analysis. *Journal of Biomolecular structure and dynamics*, 7(1), 63-73.

[36] Pop, M., Salzberg, S. L., & Shumway, M. (2002). Genome sequence assembly: Algorithms and issues. *Computer*, 35(7), 47-54.

[37] Robertson, K. D. (2005). DNA methylation and human disease. *Nature Reviews Genetics*, 6(8), 597-610.

[38] Russo, V. E., Martienssen, R. A., & Riggs, A. D. (1996). Epigenetic mechanisms of gene regulation. Cold Spring Harbor Laboratory Press.

[39] Sanger, F., Nicklen, S., Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12), 5463-5467.

[40] Saxonov, S., Berg, P., & Brutlag, D. L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America*, 103(5), 1412-1417.

[41] Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10), 1135-1145.

[42] Takai, D., & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*, 99(6), 3740-3745.

[43] L.A. Urry, M.L. Cain, S.A. Wasserman, and P.V. Minorsky. *Campbell biology - ninth edition*. Pearson Benjamin Cummings, 2011.

[44] Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*. Elsevier.

[45] Vogel, M. J., Peric-Hupkes, D. & van Steensel, B. (2007). Detection of in vivo protein–DNA interactions using DamID in mammalian cells. *Nature Protocols* 2, pp. 1467–1478.

[46] Yamada, Y., Watanabe, H., Miura, F., Soejima, H., Uchiyama, M., Iwasaka, T., & Ito, T. (2004). A comprehensive analysis of allelic methylation status of CpG islands on human chromosome 21q. *Genome research*, 14(2), 247-266.

[47] Wu, H., Caffo, B., Jaffee, H. A., Irizarry, R. A., & Feinberg, A. P. (2010). Redefining CpG islands using hidden Markov models. *Biostatistics*, kxq005.