Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea Magistrale in Matematica

CHARGAFF SYMMETRIC STOCHASTIC PROCESSES

Tesi di Laurea in Sistemi Dinamici

Relatore:
Chiar.mo Prof.
Mirko Degli Esposti

Presentata da:
Lucia Gagliardini

Correlatore:
Dott.
Giampaolo Cristadoro

Sessione III
2013-2014

A mamma e babbo

Contents

Introduction **iv**

1 Preliminary notions about DNA **1**
1.1 The cell 1
1.2 DNA structure 4
1.3 Coding 6
1.4 Changes and errors 9
1.5 Related Mathematical problems 10
1.6 Symmetries of DNA structure: the four Chargaff's rules 11
1.6.1 Chargaff's second parity rule for k-words 14

2 A stochastic model of bacteria DNA **15**
2.0.2 Model 15
2.0.3 Evaluating $\mathcal{N}$ 20
2.0.4 Model reliability 22
2.0.5 Observations 27

3 Simple stationary processes and concatenations of stationary processes **30**
3.1 Simple stationary processes 31
3.1.1 Preliminary definitions 32
3.1.2 Properties of Chargaff-processes 34
3.1.3 Bernoulli-scheme 35
3.1.4 1−Markov process 37
3.2 Concatenation of stationary processes 42
3.2.1 The concatenation 42
3.2.2 α−Chargaff processes 45
3.2.3 Bernoulli scheme 48
3.2.4 1-Markov chains 54
3.2.5 Mixed processes 55

Introduzione

La ricerca nel campo della genetica e in particolare nello studio del DNA ha radici piuttosto lontane nel tempo. Infatti già nel 1869 grazie a Miescher si scoprí la presenza del DNA, allora chiamato "nucleina", nelle cellule, anche se fu solo con un esperimento condotto nel 1944 da Avery-MacLoad-McCarty che si evidenziò come l'informazione genetica fosse contenuta nel DNA.

Da quel momento in poi, lo sviluppo delle tecniche di ricerca scientifica e la singergia del lavoro di studiosi di molti campi di ricerca diversi fecero sí che il DNA divenne un argomento in continuo sviluppo e oggetto di un interesse che dura tuttora.

Tra le numerose aree di interesse, un ruolo importante è ricoperto dalla modellizzazione delle stringhe di DNA. Lo scopo di tale settore di studio è la formulazione di modelli matematici che generano sequenze di basi azotate compatibili con il genoma esistente. Studi sulla composizione del DNA hanno infatti rivelato che la distribuzione delle basi azotate nei filamenti del DNA non può essere governata da un processo del tutto casuale. Carpire la natura del processo che ha portato al genoma attuale potrebbe darci una chiave di lettura della sua funzionalità e creare nuove opportunità in innumerevoli applicazioni.

La letteratura propone diversi modelli matematici per stringhe di DNA, ciascuno dei quali è costruito a partire da ipotesi e obiettivi diversi. In [6], per esempio, si prendono in considerazione solo porzioni codificanti di DNA, mentre numerosi articoli analizzano sia la parte codificante che quella non codificante, il cosídetto junk DNA, che rappresenta nella maggioranza dei casi la più alta percentuale del DNA di molti organismi.

Uno dei criteri su cui si sceglie di improntare lo studio della modellizzazione del DNA è l'ipotesi di aderenza alla seconda regola di Chargaff.

Nei primi anni ′50, il biochimico Erwin Chargaff, affascinato dai risultati ottenuti da Avery pochi anni prima, si imbattè in importanti regolarità nella composizione del DNA (see [12]). In particolare, trovò che le basi azotate erano presenti in proporzioni uguali sia considerando il doppio filamento di DNA, che il filamento singolo. La prima proprietà prese il nome di prima

regola di Chargaff, mentre l'altra si definí seconda regola di Chargaff.

Mentre la prima regola ha trovato una spiegazione grazie al modello di Watson e Crick, che idearono la doppia elica a partire anche dall'importante scoperta di Chargaff, la seconda risulta ancora parzialmente irrisolta. Sono ancora ignoti infatti i fattori che hanno generato questa simmetria. Restano inoltre sconosciuti tutti i possibili effetti di questa proprietà sul DNA, che potrebbe influire su molte funzioni, dalla trascrizione alla configurazione spaziale del DNA. I modelli matematici che tengono in conto le simmetrie di Chargaff si dividono principalmente in due filoni: uno la ritiene un risultato dell'evoluzione sul genoma, mentre l'altro la ipotizza peculiare di un genoma primitivo e non intaccata dalle modifiche apportate dall'evoluzione.

Questa tesi si propone di analizzare un modello del secondo tipo. In particolare ci siamo ispirati al modello definito da [13] da Sobottka e Hart. Dopo un'analisi critica e lo studio del lavoro degli autori, abbiamo esteso il modello ad un più ampio insieme di casi. Abbiamo utilizzato processi stocastici come Bernoulli-scheme e catene di Markov per costruire una possibile generalizzazione della struttura proposta in [13], analizzando le condizioni che implicano la validità della regola di Chargaff. I modelli esaminati sono costituiti da semplici processi stazionari o concatenazioni di processi stazionari.

Nel primo capitolo vengono introdotte alcune nozioni di biologia che rappresentano le basi del lavoro affrontato nelle pagine successive. Dopo una breve descrizione della cellula, si approfondiscono la struttura, il funzionamento e le caratteristiche del DNA.

Nel secondo capitolo si fa una descrizione critica e prospettica del modello proposto da Sobottka e Hart, introducendo le definizioni formali per il caso generale presentato nel terzo capitolo, dove si sviluppa l'apparato teorico del modello generale.

Sarebbe interessante proseguire il lavoro con l'analisi delle simulazioni pratiche dei processi definiti in modo teorico in questa tesi. In particolare, si potrebbero analizzare le realizzazioni dei processi definiti e studiare il confronto con dei veri filamenti di DNA.

Sebbene molti passi siano stati fatti dalla scoperta nel lontano 1944, molto ancora resta da scoprire circa il funzionamento e la struttura del DNA e questo lo rende uno dei campi di studio più affascinanti, anche in ambito matematico.

Introduction

Research in genetics and more in particular the study of DNA, have started long time ago. Indeed, thanks to Miescher, it was known back in 1869 that cells contain what was then called "nuclein" and that it was present in chromosomes, which lead Miescher to think that it could somehow be related to genetical inheritance. However, only in 1944 Avery-MacLoad-McCarty experiment with two different bacteria strains highlighted that the genetic information was probably contained in DNA (see [16]).

From that moment on, the development of scientific methods of investigation together with an increasing attention on the topic, allowed a more in-depth research on the field of genetics and made DNA an object of interest that lasts until nowadays.

Among all the numerous areas of interests, a very important role is represented by the modeling of DNA strings. This area of research aims to formulate mathematical models that generate sequence of nucleobasis such that they could be compared to real genome. Indeed, observations on the actual genome have shown that the distribution of nucleobasis can't be the result of a completely random mechanism. Succeeding in grasping the DNA structure could let us gain the key of its operation and open the door to countless applications.

Literature offers many models for DNA as symbolic sequences, each of them is defined from different hypothesis and purposes. Some of them take under consideration just the coding portion of the genome (see [6]), while the rest tries to analyze both coding and non-coding segments, that constitute the major percentage of the whole genomes in most of the organisms. One of the parameter that one may choose to shape the mathematical model is the compliance with a very important property of (almost) all kind of genome, that is Chargaff's second parity rule.

In the earliest 50s, the Austrian biochemist Erwin Chargaff, fascinated by Avery's work, made some experiments on animal genome and bumped into very important regularities in the DNA composition (see [12]). More in detail, he discovered that the amount of nucleotides were in a particular

equal percentages that ware the same both for double and singular strands of DNA.

The former was called "Chargaff's first parity rule", while the latter "Chargaff's second parity rule". While the first rule was totally explained with the famous model by Watson and Crick, which though the double helix structure of DNA starting from the important discovery made by Chargaff, the second parity rule is still partially unsolved. Indeed, the factors that bring to this symmetry at the intra-strand level remain unknown. Moreover, literature gives a non univocal explanation of the effects of the symmetry on DNA functions.

The mathematical models that comply with Chargaff's second parity rule can be divided in those taking under consideration evolution and those aiming to model a primitive DNA. The former consider the second parity rule as a consequence of evolution, while the latter hypothesizes that this property is peculiar to all primitive genomes and have not been destroyed by evolution.

This thesis' aim is to present a model of the second type. In particular, we took inspiration by the model defined in [13] by Sobottka and Hart. After a critical analysis and study of Sobottka and Hart's model, we extend the model to a wide range of cases. We use stochastic processes as Bernoulli and Markov chain to construct a possible generalization the structure proposed in [13]. We analyze stationary processes and simple composition of stationary processes and study conditions that enable Chargaff's second parity rule on the resulting strings.

In the first chapter we introduce some biology notions that represent the background of the processes analyzed. After a brief description of the cell, we focus our attention on DNA. We describe its structure, the main characteristics and the functions.

In the second chapter we make a perspective description of the model proposed by [13], setting the formal foundation to the general case presented in Chapter 3.

It would be interesting to extend this work studying the simulation of the processes defined in the last chapter and comparing the resulting realizations with actual sequences of DNA. It would also be interesting to add conditions and hypothesis to the model, such as variation of CG percentage, that represents a remarkable object of research, or constraints of different DNA species.

While much has been done about DNA and its activity, there is a need for further progress. Scientists still have to the fill the lack of knowledge that concerns the function and the nature of many structure of DNA, and this makes this field of study one of the most interesting and stimulating of nowadays science research.

Chapter 1

Preliminary notions about DNA

All organisms on this planet, as different as they can appear, share similarities. One of the most important analogy they have in common is that the genetic information is encoded by DNA.

In this chapter we introduce basilar notions about DNA. You can find most of the content below in ([4]).

We start with a brief description of the cell, underling the main differences and similarities between procaryote and eukaryote cells. We secondly focus our attention on DNA structure and coding function, and illustrate some errors and changes that modify DNA during evolution. In the end we rough out possible application and studies related with DNA.

1.1 The cell

Except for viruses, all life on this planet is based upon cells, small units that sequester biochemical reactions from the environment, maintain biochemical components at high concentrations and sequester genetic information. Cells are organized into a number of components and compartments, each one addressed to a specific task. Processes such as replication, DNA repair and glycolisys (extraction of energy by fermentation of glucose) are present and mechanistically similar in most organisms and broader insights into these functions can be obtained by studying simpler organisms such as yeast and bacteria. This allows biologists to focus on model organisms that conveniently embody and illustrate the phenomena under investigation. Model organisms are then chosen for convenience, economic importance, or medical relevance.

The more evolved is the organism, greater is the number of features involved in the cell. Indeed, a first organisms classification can be done looking

at the organization of the cell:

- Prokaryote : they have cell membranes and cytoplasm, but the DNA it is condensed into the nucleoid

- Eukaryotes: their cells have true nucleus and membrane bound organelles (ex. mitochondria and chloroplasts)

Fungi, insects, mammals are examples of eukaryotes, while bacteria are prokaryote organisms. Evidence supports the idea that eukaryotic cells are actually the descendents of separate prokaryotic cells. The presence of organelles in eukaryote cells itself, indeed, could be explained as the result of an engulfment of a bacteria by another bacteria. Organelles are DNA-containing, having originated from formerly autonomous microscopic organisms acquired via endosymbiosis. This is an important feature, as organenelle's DNA is more primitive then nuclear DNA and, as we will see later, doesn't comply with some properties of all other genome. Main functions of the organelles are the energy production from the oxidation of glucose substances and the release of adenosine triphosphate (mithocondria) or photosynthesis (chloroplast).
Because of their more complicated structure, eukaryotic cells have a more complex spatial partitioning of different biochemical reactions than do prokaryotes.

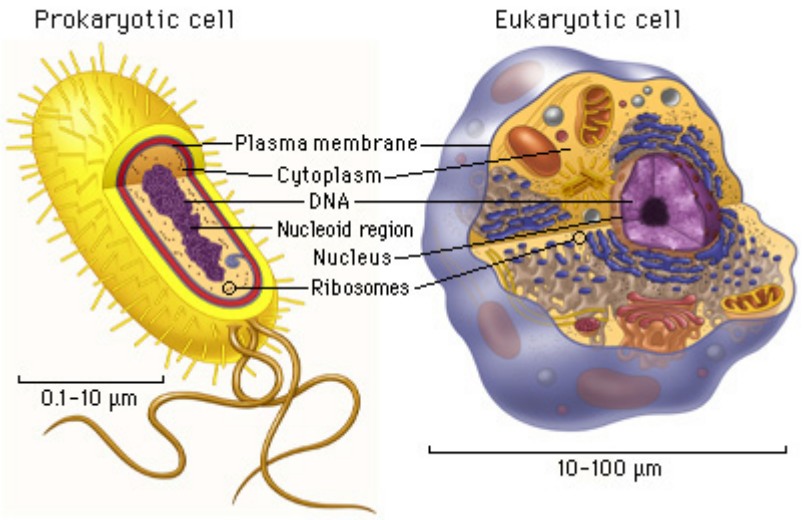

Figure 1.1: Some major components of a prokaryotic cell and an eukaryotic cell. It is easy to see the main differences and similarities listed above.

Prokaryote and eukaryote differ also in DNA storage, that is more organized more evolved the organism is. For example, while the former do not separate the DNA from the cytoplasm by a nuclear membrane, the latter provide a nucleus to contain it. Furthermore, Eukaryotes package their DNA in highly ordered structure, called chromosomes, which are condensed linear DNA molecules wrapped around octamers of proteins (histones). Eukaryotes are usually diploid, which means they contain N pairs of homologous chromosomes, differently from prokaryotes that are typically haploid and often have a single circular chromosomal DNA. Chromosomes are usually represented in the X-shape that is visible only during replication processes. The central part of the "X" is called *centromere* and links the two *sister chromatids* together. The *non-sister chromatids* are the halves of two homologue chromosomes.

Organization in chromosomes makes the process of replication cell more diversified for eukaryotes than prokaryotes, that just replicate their cell by binary fission. Otherwise, eukaryotic cells operate two different cell divisions: mitosis, which aims to replicate the cell with one identical to the original, and meiosis, which produces four gametes that are different from the original cell and from each other.

As we said before, DNA organization is more efficient in Eukaryote then in Prokaryote, thanks to sexual reproduction that reinforce evolution within changes and recombination in DNA.

An evident example of DNA recombination appears during meiosis.

At the beginning there is one mother cell, with $2N$ chromosomes, each one composed of one chromatid. Later, the chromatids duplicate to form the $X-$scructured chromosome. At that stage (Prophase I) the cell has $2N$ chromosomes, each one composed of two sister chromatids (see fig. 1.1). Before dividing in two haploid cells containing N chromosomes, the mother cells recombine its genome. More in detail, in Metaphase I, homologous chromosomes exchange genetic material. This recombination between non sister chromatids is called *crossing over*.

As shown in fig. 1.1, the resulting haploid cells contain each one a homolog chromosome of the original cell, but different in one chromatid. A further division yields to four cells containing chromatids all distinct from each other.

Recombination in meiosis supports evolution, since it generate organisms that can differ from parents in order to better adapt species to the environment.

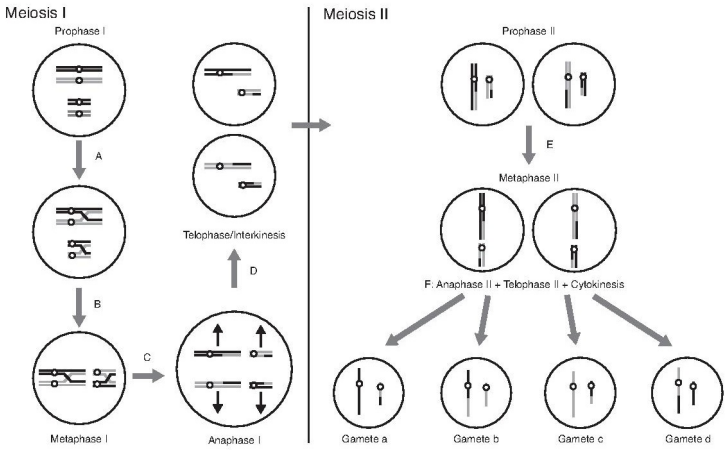

Figure 1.2: Schematic summary of steps in meiosis. In this example $N = 2$. Black chromosomes came from one parent, and grey chromosomes came from the other. The resulting gametes are composed of two chromatids, as $N = 2$. In metaphase they will duplicate to form the classic $X-$structure of the chromosome.

1.2 DNA structure

DNA (Deoxyribonucleic acid) is a long polymer made from repeating units called nucleotides. In most living organisms, DNA does not usually exist as a single molecule, but instead as a pair of molecules that are held tightly together in the shape of a double helix. The nucleotide repeats contain both the segment of the backbone of the molecule, which holds the chain together, and a nucleobase, which interacts with the other DNA strand in the shape of a double helix. A nucleobase linked to a sugar is called a nucleoside and a base linked to a sugar and one or more phosphate groups is called a nucleotide. A polymer comprising multiple linked nucleotides (as in DNA) is called a polynucleotide.

The subunits (nucleotides) of the DNA macromoleculas are deoxybonucleotides of four types: deoxyadenosine 5'-phosphate (A), deoxycytidine 5'-phosphate (C), deoxyguanosine 5'-phosphate (G) and thymidine 5'-phosphate (T). Nucleotides are more commonly identified with the nitrogen-containing nucleobase, i.e. adenine (A), cytosine (C), guanine (G) or thymine (T). The 5' position of the sugar of each nucleotide is connected via a phosphate group to the 3' position on the sugar of the immediately preceding nucleotide. Each DNA strand has a 5' end, corresponding to the phosphate group attached to

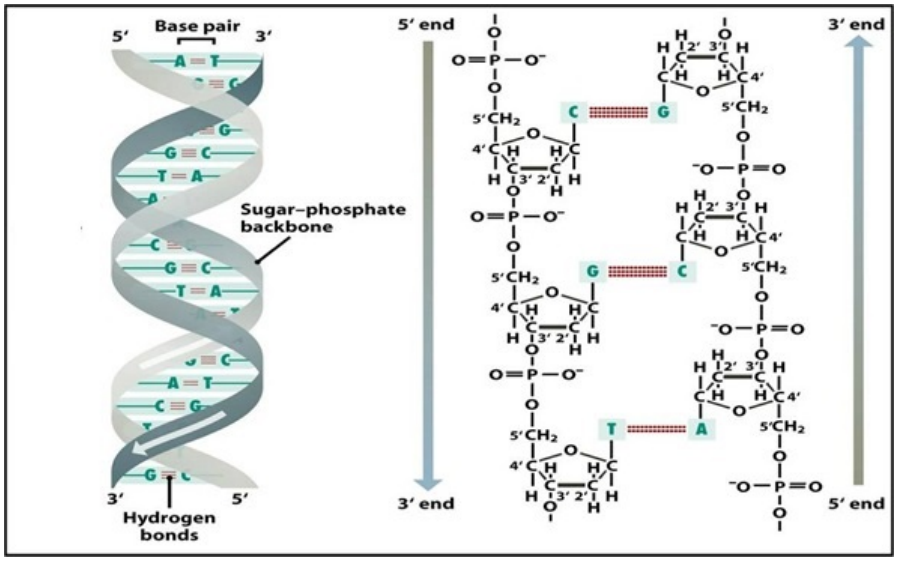

Figure 1.3: On the left, the helix structure of DNA. The picture on the right shows how nucleotides are attached to form the strand. The orientation is given by the position of the phosphate in the carbon ring.

the 5' position on the sugar molecule of the first nucleotide, and a 3' end, corresponding to the $-OH$ group at the 3' position on the sugar of the last nucleotide.

The fifth or third position on the sugar are used to denote direction of the complementary strands. Indeed, in the helix structured DNA, one strand is read in direction $5' - 3'$, while the other is read in direction $3' - 5'$.

Bases A and G are said to be purine (denoted with R), while bases C and T are called pyrimidine (denoted with Y). This classification is made on the chemical analogies between the couples, as shown in fig 1.2.

In addition, bases are classified (IUPAC) in weak (A,T) or strong (C,G) denoted with W and S respectively, and in keto (T,G) or amino (A,C) denoted with K and M respectively. In conclusion duplex DNA molecule can be represented by a string of letters drawn from $\{A, C, G, T\}$, with the left-to-right orientation of the string corresponding to the 5' to 3' polarity.

Note that a word read in $5 - 3$ direction is different from the same word read in the opposite direction $3-5$. This is due to the orientation of molecule to which bases are attached, as it can be seen in fig.1.2.

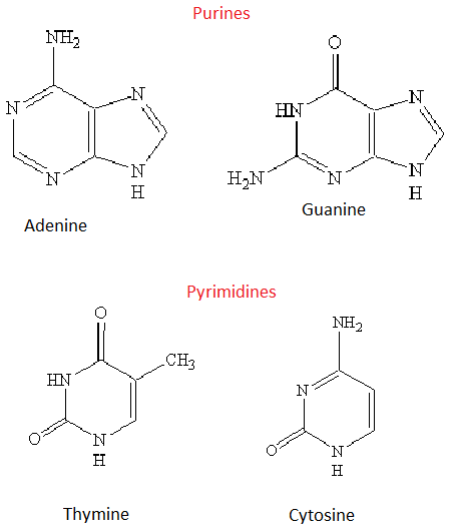

Figure 1.4: The common classification of nucleotides of similar chemical composition.

1.3 Coding

As we said before, DNA function is to store genomic information. In this section we briefly describe how information is encoded in the genome and what processes are put to use to decode it.

In order to do that, we need to introduce RNA, that is quite similar to DNA in its composition and differs from it in two primary ways: the residues contain hydroxyl groups (and thus are not "deoxy") and uracil (U) replaces the thymine base.

In most cases, RNA is encountered as a single strand, but often it will form intrastrand base pairs to form secondary structures that may be functionally important. RNA takes an important role in the readout process, as it may be used as a temporary copy of the information corresponding to genes or may play a role in the translational apparatus. In fact, the information flow in the cells can be summarized in four processes:

1. DNA replication, where a DNA sequence is copied to yield a molecule nearly identical to the starting molecule, during cellular division

2. Transcription, where a portion of DNA sequence is converted to the corresponding RNA sequence

3. Translation, where the polypeptide sequence corresponding to the mRNA sequence is synthesized

4. Reverse transcription, where the RNA sequence is used as a template for the synthesis of DNA, as in retrovirus replication, pseudogene formation, and certain types of transposition

Since during replication one strand is read continuously while its complement nascent strand is synthesized in discontinuous segments (due to the replication fork, that is the growing separation of the DNA strands), biologists use to call the former strand *leading strand* and the latter one *lagging strand*.

The processes involved in the decryption of the DNA sequence are (2) and (3), as shown in fig 1.3. First, in transcription, a temporary mRNA

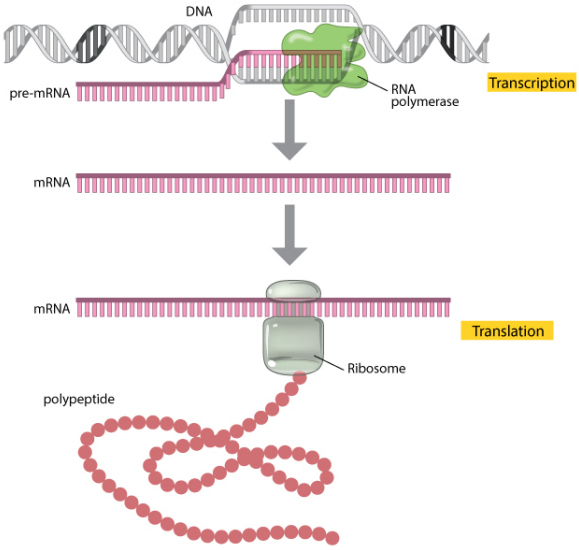

Figure 1.5: The process of deconing DNA sequences. First, the temporary mRNA is copied from the DNA strand and processed to form a mature mRNA molecule. This can be translated to build the protein molecule (polypeptide) encoded by the original gene.

[1]

(messenger RNA) is copied from portion of DNA sequence. Later, pre-mRNA is modified to remove certain stretches of non-coding sequences (i.e. portion of DNA that do not encode proteins) called introns; the stretches that remain include protein-coding sequences and are called exons. The sequences of mRNA are then translated thanks to the synergic action of ribosomes and tRNA. The process of translations consists of many steps. The ribosome assembles around the target mRNA allowing the first tRNA to attach at the start codon. The tRNA transfers an amino acid to the tRNA corresponding to the next codon. The ribosome then moves (translocates) to the next

mRNA codon to continue the process. When a stop codon is reached, the ribosome releases the complete chain.

Thus, the final product of the decoding of a DNA sequence is a polypeptide, i.e. a chain of amino acids attached together. The polypeptide, later, folds to form an active protein and performs its functions in the cell.

The DNA alphabet contains four letters but must specify polypeptide chains with an alphabet of 20 letters, that are all possible amino acids. This means that combinations of nucleotide are needed to code for each amino acid. There are 4^2 possible dinucleotides, that are still lesser then number of amino acid. Thus, the genetic code is a triplet code, and the code triplets in mRNA are called codons. Since all possible trinucleotides are 4^3, and there are three stop codons out of 64 triplets, there are 61 left triplets coding for the 20 amino acid. Codons are not used with equal frequencies in various genes and organisms, and the statistic of codon usage is a characteristic that can sometimes be used to distinguish between organisms. This phenomena is known as codon bias and the statistic that can describe each protein-coding gene for any given organism is the CAI (codon adaptation index), that compares the distribution of codons actually used in a particular protein with the preferred codons for highly expressed genes.

Even if the principal function of DNA is the production of proteins which are encoded by codons, i.e. words of length 3, larger words are also important for the strand organization. In particular words of length $k = 4, 5, 6$ or 8 are distributed in a way that may interfere with the action of some enzymes addressed to manipulate DNA strands. Furthermore, $4-$words are useful for analyzing particular genomic subsequences. In addition, $k-$tuple frequencies can assist in classifying DNA sequences by content, such as predicting whether a given sequence is coding or non-coding. In fact, because coding sequences commonly specify amino acid strings that are functionally constrained, the distribution of $k-$tuple frequencies differ from that of non-coding sequences.

From this perspective, DNA sequence organization is much more complicated than a simple list of proteins. Evolution has manipulated genome yielding to a powerful device which inner structure is yet unknown.

Information storage has indeed considerably different range in eukaryote and prokaryote. In fact, while the average prokaryotic gene is around 1000 bp (base pairs), the average human gene is about 27000 bp. Moreover, non-conding sequences take an important role in the evolution. In fact while approximately 90% of a typical prokaryotic genome codes for gene products, the percentage of coding sequence dramatically decrease for eukaryotes. For example, only the $1,2\%$ of human gene is coding , while the rest corresponds to extensive control regions, untranslated regions and intronic regions, i.e.

non-coding DNA segments that separate the exons and that are not included in the final mRNA. Other properties that interfere with DNA activities are G+C content, GC-skew (i.e. the quantity $(G-C)/(G+C)$) and AT skew, that may have a role in replication orientation and gene orientation (see [9]).

For a long time non-coding regions hadn't catch the attention of biologists, that used to referred to it as a "junk DNA". Nowadays it is known that non-coding sequences in DNA do have a very important role, and the research of reasons of its existence and functions is still an open and fascinating problem. It is with a good reason that more evolved organisms have more high percentage of non-coding DNA in their genome than primitive organisms have. Since non-coding sequences appear to accumulate mutations more rapidly than coding sequences due to a loss of selective pressure, Non-coding could serve as a raw material for evolution. Indeed, improvement of species is strongly connected with mutations of DNA. Those changes of sequences are mostly accidental, as we will describe in the next section.

1.4 Changes and errors

DNA is not immutable. Indeed, the sequence of bases contained on chromosomal DNA molecules is the result of a set of evolutionary processes that have occurred over time. These changes are intimately connected with many processes described above as chromosomes recombine and DNA replication. In fact, even if there were no recombination, the DNA of gametes would differ from the DNA of the parent cells because of errors that may occur at low frequency during DNA replication.

Principal types of changes that may occur to DNA sequences are:

- Deletion: removal of one or more contiguous bases
- Insertion: insertion of one or more contiguous bases between adjacent nucleotides in a DNA sequence
- Segmental duplication: appearance of two or more copies of the same extended portion of the genome in different locations in the DNA sequence
- Inversion: reversal of the order of genes or other DNA markers in a subsequence relative to flanking markers in a longer sequence. Within a longer sequence, inversion replaces one strand of the subsequence with its complement, mantaining $5'$ to $3'$ polarity
- Recombination: in vivo joining of one DNA sequence to another

- Point mutation: substitution of the base usually found at a position in the DNA by another as a result of an error in base insertion by DNA polymerase or misrepair after chemical modification of a base

If the errors occur within a gene, the result may be a recognizable mutation (alteration of the base sequence in a normal gene or its control elements). Anyway, base changes at the DNA sequence level do not always lead to recognizable phenotypes [2], particularly if they affect the third position of a codon [3]

Occurrence of errors in DNA replication may be a reason for the insertion of a large portion of non-coding region in evolved organisms. Indeed, high percentage of non-coding sequences would prevent mutations in meaningful regions and, at the same time, relax evolutionary constraints on the genome (see [3]).

1.5 Related Mathematical problems

Plenty of mathematical problems can be formulated in relation to genome and DNA sequences, ranging from statistical to computational problems.

For example, one can study processes occurring in a large number of interbreeding individuals, i.e. genetic variation There are two related statistical and computational problems in dealing with populations. First, characterization of genetic variation within and between populations in terms of allele frequencies or nucleotide sequence variation, and second the analysis of the trajectory of population parameters over time, that invokes evolutionary models to describe molecular data in parsimonious manner.

Other interesting studies deal with analysis of storage and readout information necessary to the function and reproduction of cells. In particular, codon usage and codon bias can be critical in classifying species and determine evolutionary mechanism.

DNA computing, in addition, aim in solving maths problem using DNA (see for example [7]).

Given a sequence of DNA, there are a number of questions one might ask. For instance, one can investigate if it represent a coding or non-coding

[2] With the word *phenotype* biologists refer to organism's actual observed properties. The full hereditary information is instead represented in what the call *genotype*. Genotype is a major influencing factor in the development of the phenotype of an organism, but still it is not the only one.

[3] Observations on the usage of nucleotides in the third position of codons lead to evaluation on evolutionary theory. Contrary to what Darwin's theory states, in particular, preferential codon usage suggests that origin of life was a plural form (see [15]).

sequence and can infer the sort of sequence that might be: could it be a protein coding sequence or a centromere or a control sequence?

In addition, by analyzing codon usage and codon bias, one can determine what sort of organism this sequence came from based on sequence content.

In the end, one may ask what sort of statistics should be used to describe this sequence. In the next chapter we will deal in such a problem, given an example of statistic model that can partially describe DNA properties.

1.6 Symmetries of DNA structure: the four Chargaff's rules

In the '50s Erwin Chargaff and his colleagues found some "regularities" in the base composition of DNA, that reveal the multiple levels of information in genomes (see [?],[?]).

Chargaff's results are summarized in four rules.

First and second Chargaff's parity rules affect the ratio of pyrimidine and purine bases on a DNA string, while the other two are about the content of some bases and how the nucleotides are distributed along DNA sequences.

The four rules can be summarized as follow

1. Chargaff's first parity rule: first parity rule states that the amount of guanine is equal to the amount of cytosine and the amount of adenine is equal to that of thymine. This property is species invariant.

2. Cluster rule: individual bases are clustered to a greater extent than expected on a random basis.

3. Chargaff's second parity rule: to a close approximation, the Chargaff's first parity rule holds to single stranded DNA. In other words, if the individual strands of DNA are isolated and their base composition determined, then $\#A \approx \#T$ and $\#C \approx \#G$ for each strand.

4. CG rule: The ratio of $(C+G)$content to the total bases content $A+C+G+T$ tends to be constant in a particular species, but varies between species.

These characteristics are shared by almost all genomes ([10])and most of them still don't have a biological unambiguous explanation.

The existence of these symmetries in the genome suggests that the inner process shaping DNA sequence is not completely random and is affected by rules that are invariant between species. Rule 2 and 4 do not find a great range in literature.

Furthermore, in spite of the importance of all of them, only one found a biological relevant role, being the prerequisite of the Watson and Crick model discovered in 1953. The double-helix structure of DNA, indeed, implies that the number of adenine and the number of thymine is the same, and similarly the number of cytosine equals the number of guanine, as every A and C on a strand match a T or G respectively on the complementary strand.

For what concerns the rule (2), it has been discovered that clustering in microorganisms often relates to transcription direction (see [5]).

Notice that the observation of base clustering did not necessarily imply a local conflict with Chargaff's second parity rule. For example, a run of T residues, might be accompanied by a corresponding number of dispersed A residues, so that $\#A \approx \#T$. However, there are distinct local deviations from the second parity rule, and they may correlate with transcription direction and gene location (see [2]). Furthermore, it could be interpreted in terms of stem-loop configurations. Indeed, Chargaff differences (i.e. relative richness of a region for a particular W base or S base)[4] would be reflected in the composition of loops in the stem-loops structures which might be extruded from supercoiled DNA under biological conditions ([2]).[5].

Chargaff's second parity rule imposes some form of evolutionary restraint on the double stranded genomes (see [10]). Nevertheless, the genomes which doesn't comply with this property, as organelles or one stranded genomes, seems to obey to a more relaxed imposition: $A + G = C + T$ (see [10]).

Althought many hypothesis on the genome and its origin are studied (see [5] and [14]), Chargaff's second parity rule still not has a confirmed and unique explanation. Sorimachi (2009) proposed a solution to Chargaff's second parity rule by analyzing the nucleotide contents in double stranded DNA as the union of ORF [6] (open reading frame)and NORF (non open reading frame).

---

[4]We remind that W denotes weak bases (A,T), while S denotes strong bases (C,G).

[5]Stem-loop is an intermolecular base-pairing. It can occur in single-stranded DNA or, more commonly, in RNA. The resulting structure is a key building block of many RNA secondary structure, i.e. the capability of assuming a regular spatial ripetitive structure. It appeared that the tendency of arrange the order of bases to support mRNA structure sometimes beats the coding function. Since in stems Chargaff differences tend to be zero (by definition), then overall Chargaff differences should be reflective of the base composition of loops.

[6]Transcription, that is the process that lead to the RNA synthesis and later to traslation into protein, affect portions of DNA (open reading frame) and stop each time the sequence run into a particular sequence of nucleotides, called stop codons. In molecular genetics an *open reading frame* (ORF) is the part of reading frame that contains no stop codons. The transcription termination pause site is located after the ORF. The presence of a ORF does not necessarily mean that the region is ever translated.

Even if each gene has a different nucleotide sequence, the genome is homogenously constructed from putative small units consisting of various genes displaying almost the same codon usages and amino acid compositions ([15]). Since a complete gene is assumed consisting of two huge molecules which represent a coding and a non coding sequence, Sorimachi operates identifying the ORF both on forward and reverse strand (s_1 and s_2 respectively). ORFs_1 and ORFs_2 denote the coding regions in s_1 and s_2 respectively, while NORFs_1 and NORFs_2 denote non-coding regions. Since they belong to the

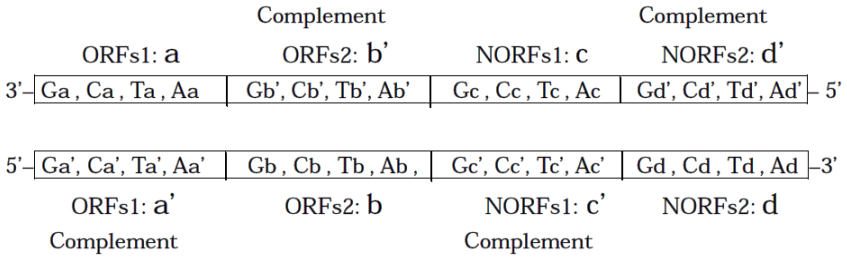

Figure 1.6: The double strand DNA is divided considering ORF and NORF. Segment a represents the ORF on strand s_1, while a' is its complement on strand s_2. Similarly, b is the coding region in s_2 and b' is the complement on the complement string. The same happens with non coding region c and d.

same genome ORFs_1 and ORFs_2 have almost the same size. Thus, nucleotide contents of ORF and NORF are related as follows. We denote with I_j the content of nucleotide $I \in \mathcal{I} = \{A, C, G, T\}$ in the portion j of the strand.
Then

$$\#A_b \approx \#A_a,\ \#C_b \approx \#C_a,\ \#G_b \approx \#G_a,\ \#T_b \approx \#T_a \tag{1.1}$$

for the coding segments. Similarly happens to the non coding sequence, so that

$$\#A_d \approx \#A_c,\ \#C_d \approx \#C_c,\ \#G_d \approx \#G_c,\ \#T_d \approx \#T_c \tag{1.2}$$

Nucleotide contents for a', b', c', d' depend on nucleotide contents on corresponding complementary segments, obeying Chargaff's first parity rule.
In particular it is

$$\#A_i = \#T_{i'},\ \#C_i = \#G_{i'},\ \#G_i = \#C_{i'},\ \#T_i = \#A_{i'},\ \ i = a, b, c, d \tag{1.3}$$

It follows that for each strand the content of a nucletide approximately equals the content of its complement.
For example, from (1.2) and (1.3) G and C content for s_1 can be written as follows

$$\#C_a + \#C_{b'} + \#C_c + \#C'_d \approx \#G_a + \#G_{b'} + \#G_c + \#G_{d'} \tag{1.4}$$

and similarly happens for A, T content.

1.6.1 Chargaff's second parity rule for k-words

A natural extension of Chargaff's second parity rule is that, in each DNA strand, the number of occurrences of a given word should match that of its reversed complement. In order to verify the extension of the parity rule to words of length k, $k-$mer, Afreixo and others (2013) investigated the distributions of symmetric pairs focusing on complete human genome, on each chromosome and on the transcriptome. They have found that, in the human genome, symmetry phenomenon is statistically significant at least for words of length up to 6 nucleotides.

More in general, the analysis of their results shows that, globally, Chargaff's second parity rule only holds for small oligonucleotides, in the human genome, even if there are some large oligonucleotides for which the extension of the rule to k-mer holds. The deviations from perfect symmetry are more pronuncced for large word lengths, for which the sample size limit might became the actual issue.

There are two different approaches to explainig Chargaff's second parity rule. It can either be supposed to arise from evolutionary convergence caused by mutation and selection (for example, see [1]) or it can be supposed to be a characteristic of the primordial genome. (see [13]).

In the next chapter we analyze a model that assumes the latter approach to explain the Chargaff's regularity.

Chapter 2

A stochastic model of bacteria DNA

In this chapter we present the model proposed by Sobottka and Hart (2011). This model aims to produce sequences of genome consistently with Chargaff's second parity rule for nucleotides and dinucleotides.

2.0.2 Model

This model is based on occurrence of random joins of nucleotides in a sequence. In particular, two half-strands extend in opposite directions by adding letters to a initial nucleotide on leading strand attached to its complement on the lagging strand, as shown in fig.2.0.2. The two half strings are thus generated by two processes, that we denote with $\mathcal{X}, \mathcal{Y}$ The remaining

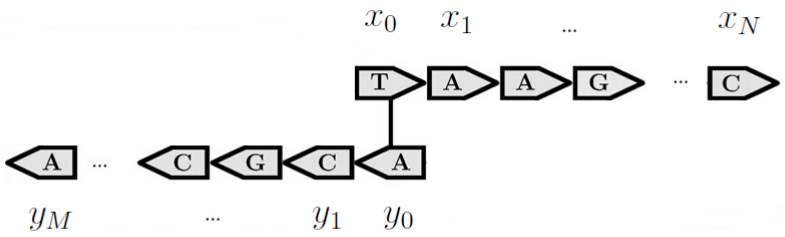

Figure 2.1: The double stranded DNA is generated by two half-strands growing in opposite directions. The successions of nucleotides belong to different strand.

halves of the strings generated above are filled with complementary bases, consistently with Watson and Crick model of DNA.

Given the initial nucleotide x_0 at the upper strand, and calling y_0 its complementary nucleotide on the second strand, we denote with $(x_l)_{l=0}^N$ and $(y_l)_{l=0}^M$ the sequences generated by processes $\mathcal{X}$ and $\mathcal{Y}$ respectively. Note that we are not supposing $M = N$, even if the model naturally brings to the equality, as we will see later. Halves $(y_l)_{l=-N}^0$ and $(x_l)_{-M}^0$ are obtained abiding by paring rule from sequences $(x_l)_{l=0}^N$ and $(y_l)_{l=0}^M$ respectively, as shown in Fig.2.0.2. We remind that paring rule states that each A (or T) in one strand matches a T (or A) in the complementary strand, and every C (or G) matches a G (or C).

For the model, Sobottka and others suppose $\mathcal{X} = \mathcal{Y} = \mathcal{Q}$, i.e. the two half-strands are supposed to be the resulting sequences of a same process, so that the final sequences $(x_l)_{l=0}^N$ and $(y_l)_{l=0}^M$ are statistically equivalent. In particular, the model taken into consideration is a Markov chain of transition matrix W and equilibrium distribution ν.

In the paper, the authors introduce the model in a different way, that better catches biological restraints on the construction of a DNA string. In this case, the probabilities defining processes are

1. probability vector $\mu = (\mu(A), \mu(C), \mu(G), \mu(T))$, that represents the availability of each nucleotide type

2. matrix $\mathcal{N} = (a_{ij})_{i,j=A,C,G,T}$, whose elements are the probabilities for nucleotides j of being accepted after nucleotide i

Let now $x_0 x_1 \dots x_l$ be a realization of $\mathcal{Q}$. Then a nucleotide x_{l+1} is randomly selected with probability $\mu(x_{l+1})$ and it is attached to the string $x_0 x_1 \dots x_l$ with probability $a_{x_l x_{l+1}}$ or it is rejected with probability $1 - a_{x_l x_{l+1}}$. Because of rejections, it might occur more than N random selections of nucleotides to construct a final string of length N.

The process $\mathcal{Q}$ defined by the new vector μ and matrix $\mathcal{N}$ can be seen as Markov chain. In other words, if we look at the final half-string simply considering each join and omitting the rejections, the sequences can be considered as two realizations of a Markov chain. In order to show that, we calculate the transition matrix defining the corresponding Markov chain.

We remind that given a Markov process defined on a state space $\mathcal{I} = \{A, C, G, T\}$, the transition matrix $T = (T_{ij})_{i,j \in \mathcal{I}}$ defining it has to satisfy the following:

1. $T_{ij} \geq 0, \quad \forall ij \in \mathcal{I}$

2. $\sum_{j \in S} T_{ij} = 1, \quad \forall i \in \mathcal{I}$

where T_{ij} is the probability of the state j to occur after the state i, with $i, j \in S$.

Thus, we can construct the transition matrix as follow. First, we calculate all the transition probabilities. If we look at $\mathcal{I}$ as the state space of the chain, then the probability of going from state i to state j is given by $\mu_j a_{ij}$, i.e. the product of the probability of the nucleotide j of being selected and the probability of the nucleotide selected of being accepted in the string. The resulting matrix will be of the form $T = (\mu_j a_{ij})_{i,j=A,C,G,T}$. Each element of T results as the product of two probabilities, so that hypothesis 1 is satisfied. On the contrary, equations 2 are unattended. Thus, we need to normalize the rows of T to get the transition matrix of the process.

The elements of the final matrix $T' = (T'_{ij})_{i,j=A,C,G,T}$ will be then of the form

$$T'_{ij} = \frac{T_{ij}}{\sum_{j \in \mathcal{I}} T_{ij}} \tag{2.1}$$

Initial nucleotides x_0 and y_0 are given according to probability vector ν. In this case, the initial probabilities are given by the stationary distribution ν of matrix T', which existence and uniqueness is guaranteed by Ergodic Theorem for Markov chains (see [8]). In fact, elements $(T'_{ij})_{i,j \in \mathcal{I}}$ are all non null, since we suppose that any letter can be attached after a given nucleotide. Thus, matrix T' is ergodic and there exist and is unique a vector ν such that $\nu T' = \nu$.

In conclusion, the construction of $(x_l)_{0 \leq l \leq N}$ and $(y_l)_{0 \leq l \leq M}$ proceeds according to a Markov chain of transition matrix $W = T'$ and stationary distribution $\nu = (\nu(A), \nu(C), \nu(G), \nu(T))$. Given a dinucleotide $\omega_1\omega_2$ with letters in $\mathcal{I}$, the probability of the dinucleotide is expressed by

$$P(\omega_1\omega_2) = \nu(\omega_1) T'_{\omega_1\omega_2}$$

Sobottka and Hart (2011) make some assumptions on vector μ and matrix $\mathcal{N}$ defined above, in order to create a model generating sequences that comply with Chargaff's second parity rule. More in detail, the following assumptions on the vector μ and the matrix $\mathcal{N}$ are taken:

- probability vector $\mu = (\mu(A), \mu(C), \mu(G), \mu(T))$ is constant, that is the probability of a base to be selected as candidate is constant throughout the construction of each strand

- the probabilities a_{ij} are supposed to be positive, constant and invariant for all primitive DNA sequences (and could be thought of a resulting from chemical and physical properties of bases themselves)

Moreover, they noticed that, due to the helix structure of DNA, each time a letter is attached in strand its complement has to join the complement strand. For example if nucleotide A is attached at position 2 in the upper strand, then a T as to join the lower strand at position -2, according to Chargaff's first parity rule (see Fig.2.0.2). This means that the probability of randomly selecting A has to be equal to the probability of selecting T. More in general, the probability of one nucleotide to be chosen has to be the same for its complement.

In addition, referring to example of picture 2.0.2, if the base G succeeds to join position 3 after A in the upper strand, then C has to succeed attaching after the T on the complement strand. In other words, the probabilities for a letter of being accepted after a nucleotide have to be the same as the probability of the complement letter of being accepted after the complement nucleotide.

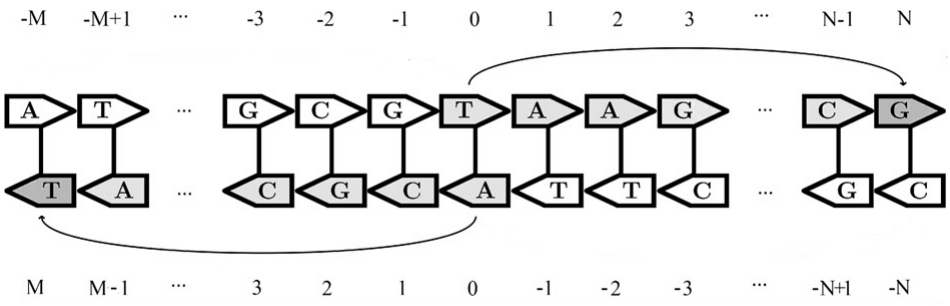

Figure 2.2: The picture shows an example of the realization of the model. The arrows identify the directions of the processes on the upper "half" and lower "half" on complementary strand. The starting nucleotides T and A are placed at initial time 0.

For convenience, we will denote the complementary base of a character $x_i \in \mathcal{I}$ with $\overline{x}_i$. For example, if $x_3 = G$, then $\overline{x}_3 = C$. Basing on how the model is constructed, we will have that $x_i = \overline{y}_{-i}$ and $\overline{x}_i = y_{-i}$, see fig.(2.0.2). Hence, he makes the following hypothesis

- H1: $\mu_A = \mu_T, \ \ \mu_C = \mu_G$.
  Thus, the probability vector μ takes the form

$$\mu = (m, 0.5 - m, 0.5 - m, m), \ \ 0 \leq m \leq 0.5$$

- H2: $a_{ij} = a_{\overline{j}\overline{i}}$

These hypothesis are speculated from Chargaff's first parity rule, i.e. they are an immediate consequence of the bases being paired in the double stranded DNA. Indeed, one may think at the set of all possible bases available in couple, since an abundance of one bases over its complement wouldn't give a higher probability of the former to join the string. For exemple, if quantity of A exceed that of T, the abound of A couldn't be used on the building of the string. For this reason, each time a nucleotide is pick up in order to be

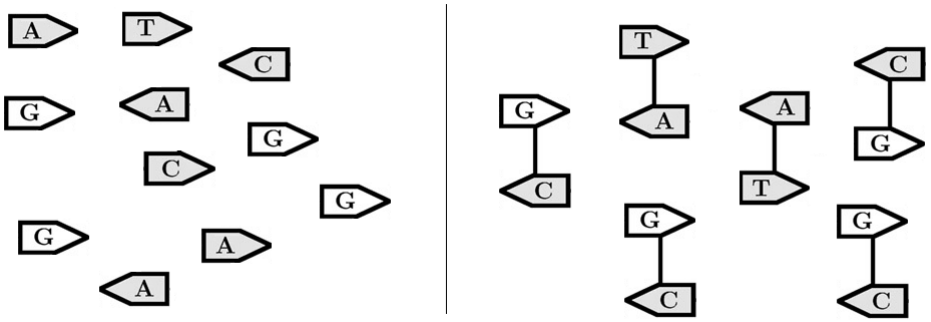

Figure 2.3: On the left, a set of nucleotides without assumption H1. On the right, the set of nucleotides thought under hypothesis H1.

attached (or rejected) to a strand, its complement has to be catch for the complement string. This means that $\mu(i)$ has to be equal to $\mu(\bar{i})$ for every $i \in \mathcal{I}$.

Similarly, if a nucleotide j is attached to a nucleotide i in a strand, complement nucleotide $\bar{j}$ is attached in the complement string. This would lead to a word ij on the top strand and a dinucleotide $\bar{j}\bar{i}$ on the bottom strand, according to $5-3$ orientation. In other words, for every $i, j \in \mathcal{I}$ it should be $a_{ij} = a_{\bar{j}\bar{i}}$.

Furthermore, the authors remark that $(x_l)_{l=1}^N$, generated by $\mathcal{X}$ and $(y_l)_{l=1}^M$, generated by $\mathcal{Y}$, are statistically equivalent, since they assume $\mathcal{X} = \mathcal{Y} = \mathcal{Q}$. For how the model is defined, $(x_l)_{l=-M}^0$ and $(y_l)_{l=-N}^0$ are also statistically equivalent, since they result by complementarity from $(x_l)_{l=1}^N$ and $(y_l)_{l=1}^M$ respectively.

Thus, one may look at the double stranded DNA as the result of a couple of processes $\bar{\mathcal{X}}$, $\mathcal{Y}$ defined similarly as above, and then complete the strands by making the complement, in the same way as before. By doing this, the processes would be described by complementary halves comparing to the model represented in fig.(2.0.2), and the matrices describing the transition probabilities are not the same, indeed they are complement matrices (see Chapter 3).

The model generates double stranded DNA. We remind that has been observed that Chargaff's second parity rule holds for double stranded DNA, while it fails to hold for organellar DNA and other types of genome (see [10]), this would support the effectiveness of the model provided.

We remind that Sobottka and Hart propose a model in order to produce a primitive sequence of DNA, assuming that all possible changes and errors that may occur over time slightly modify the main structure of a primordial genome. For this reason, they do not consider mutations for the model.

In addition, note that the process described by Sobottka can be seen as a concatenation of Markov chain. A simply Markov chain couldn't explain the long range correlation present in genome sequences (see [11]). However, this paper shows that the Markovian construction of primitive DNA sequences succeeds in capturing the gross structure at the level of mono and dinucleotide frequences.

The structure of the process that grounds this model is a first step towards investigation, and represents a keystone for the investigation we will introduce in Chapter 3.

2.0.3 Evaluating $\mathcal{N}$

In order to find the matrix of probabilities that could better suite all genomes, Sobottka and Hart (2011) made approximations and optimizations of actual frequencies of mononucleotides and dinucleotides in 1049 genome sequences.

Consistently with the notation used by Sobottka (2011), we denote with $(\pi(n), P(n))$ and $(\rho(n), R(n))$ the vectors and matrices containing mononucleotide and dinucleotide frequencies estimated for the primary and complementary strands respectively of the $n-$th bacterium, and they observed that $\pi(n) \approx \rho(n)$ and $P(n) \approx R(n)$ as expected.

In addition, if we look at the frequencies of mononucleotides and dinucleotides of each constructed sub-string, we will see that those of $(x_l)_{0\leq l\leq N}$ and $(y_l)_{0\leq l\leq M}$ are equal, due to the statistical similarity of the two sequences. Thus, if $\nu = (\nu_A, \nu_C, \nu_G, \nu_T)$ and $Q = (Q_{ij})_{i,j=A,C,G,T}$ are respectively the mononucleotide and dinucleotide frequencies in $(x_l)_{0\leq l\leq N}$, they are also the frequencies in $(y_l)_{0\leq l\leq M}$. Furthermore, as $(x_l)_{-M\leq l\leq 0}$ and $(y_l)_{-N\leq l\leq 0}$ are complementary strands, their observable frequencies are given by $\bar{\nu} = (\bar{\nu}_A, \bar{\nu}_C, \bar{\nu}_G, \bar{\nu}_T)$ and $\bar{Q} = (\bar{Q}_{ij})_{i,j=A,C,G,T}$, where $\bar{\nu}_i = \nu_{\bar{i}}$ and $\bar{Q}_{ij} = Q_{\bar{j}\bar{i}}$, according to Chargaff's first parity rule (see Fig.2.0.3).

As the length of the string $L = M + N + 1$ increases, $t = N/L$ tends to the proportion of the primary strand whose mononucleotide and dinucleotide frequencies are ν and Q respectively and similarly $1 - t$ approaches the proportion of the primary strand whose frequencies are given by $\bar{\nu}$ and $\bar{Q}$.

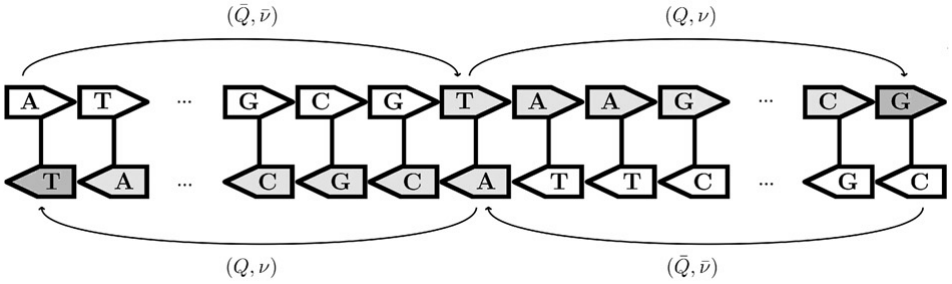

Figure 2.4: Mononucleotide and dinucleotide frequencies of each sub-string of a double stranded DNA obtained with the model.(original figure from [13])

Hence, the mononucleotide and dinucleotide frequencies estimated for each strand are approximated by

$$\begin{aligned} P_{ij} &= tQ_{ij} + (1-t)Q_{\bar{i}\bar{j}}, \\ \pi_i &= t\nu_i + (1-t)\nu_{\bar{i}} \end{aligned} \tag{2.2}$$

for the first strand $(x_l)_{-M\leq l\leq N}$, and

$$\begin{aligned} R_{ij} &= (1-t)Q_{ij} + tQ_{\bar{i}\bar{j}}, \\ \rho_i &= (1-t)\nu_i + t\nu_{\bar{i}} \end{aligned} \tag{2.3}$$

for its complementary strand $(y_l)_{-N\leq l\leq M}$, when L is large.

Since

$$Q_j = \nu_j W_{ij} \tag{2.4}$$

where the matrix W stands for the transition matrix obtained by normalization of the unknown matrix $\mathcal{N} = (a_{ij})$

$$W_{ij} = \frac{a_{ij}\mu_j}{\sum_{k\in\mathcal{I}} a_{ik}\mu_k} \tag{2.5}$$

the first equation of 2.2 can be written as

$$P_{ij} = t\nu_i \frac{a_{ij}\mu_j}{\sum_{k\in\mathcal{I}} a_{ik}\mu_k} + (1-t)\nu_{\bar{j}} \frac{a_{\bar{j}\bar{i}}\mu_{\bar{i}}}{\sum_{k\in\mathcal{I}} a_{\bar{j}k}\mu_k} \tag{2.6}$$

Thus, the matrix of the dinucleotides real frequencies can be written as function of vectors μ, ν, parameter t and matrix $\mathcal{N}$. However, since ν is the equilibrium distribution of the transition matrix W, it can be calculated from it [1].

[1] Sobottka and others ([13]) used a Matlab function to do evaluate equilibrium vector ν from transition matrix W in the optimization problem.

Hence, we have

$$P_{ij} = P_{ij}(t, (a_{ij})_{i,j\in\mathcal{I}}, \mu) \tag{2.7}$$

Similarly,

$$Q_{ij} = Q_{ij}(t, (a_{ij})_{i,j\in\mathcal{I}}, \mu) \tag{2.8}$$

In the same way, vectors π and ρ containing real nucleotide frequencies, depend on t and ν. As we have seen above, $\nu = \nu((a_{ij})_{i,j\in\mathcal{I}}, \mu)$, thus

$$\begin{aligned} \pi_i &= \pi_i(t, (a_{ij})_{i,j\in\mathcal{I}}, \mu), \\ \rho_i &= \rho_i(t, (a_{ij})_{i,j\in\mathcal{I}}, \mu) \end{aligned} \tag{2.9}$$

As we can see from 2.7, P_{ij} depends on a total number of 20 parameters, that are t, 16 elements of $\mathcal{N}$ and 3 elements of μ (since the vector has to sum to one).

Imposing hypothesis H1 and H2 makes the total number of parameters decrease to 12, since matrix $\mathcal{N}$ becomes antisymmetric and vector μ results of the form $(m, 0.5 - m, 0.5 - m, m)$.

Formulas 2.2 and 2.3 hold $\forall i, j \in \{A, C, G, T\}$, for every $n-$th bacteria analyzed, and were used to construct estimators of the matrices $\mathcal{N}(n)$ (and, consequently, elements of the corresponding equilibrium distribution $\nu(n)$), vectors $\mu(n)$ and values of $t(n)$ by determining the parameters for which the right side of equations most closely approximates $P(n), \pi(n), R(n), \rho(n)$ respectively. Afterward, the final matrix was calculate as the average of the resulting matrices $\mathcal{N}(n)$ obtained from the optimizations.

A first evaluation $\bar{\mathcal{N}}$ was made without the assumptions of H1 and H2. The vectors $\mu(n)$ estimated generally satisfied property of symmetry.

On the other hand, the average matrix obtained from the optimization under hypothesis H2 and H2 is approximately antisymmetric as expected.

$$\bar{\bar{\mathcal{N}}} = \begin{pmatrix} 0.7515 & 0.4807 & 0.5583 & 0.6785 \\ 0.6942 & 0.5584 & 0.6141 & 0.5583 \\ 0.6722 & 0.7407 & 0.5584 & 0.4807 \\ 0.5361 & 0.6722 & 0.6942 & 0.7515 \end{pmatrix}$$

2.0.4 Model reliability

Sobottka and Hart advance mainly three reasons to support the model supplied in [13].

First, they observe that simulations of the model for many distinct vectors μ and matrices $\mathcal{N}$ and values of t showed that, if the entries on the rows of $\mathcal{N}$ are very different from each other and the value of t is far from 0.5 then

the sequences produced by the model in general did not satisfy Chargaff's second parity rule. This would support the theory that in the construction of the sequences no strand is favored over the other, and could explain why, unlike single-stranded DNA (see [10]), many double-stranded genome sequences comply with Chargaff's second parity rule. For the same reason, they always imposed $t = 0.5$ as initial value, when computing matrices $\mathcal{N}(n)$, so that $N \approx M$ and the substrings generating the double strand DNA result to be half part of the final strand.

Secondly, they notice a consistent similarity between the sequences produced by the model and the bacteria genomes analyzed. In particular, they inferred equivalences on distribution of nucleotide and dinucleotide frequencies in the four parts of the double stranded DNA obtained with the process.

From now on, we will refer to the four parts of the double stranded realization of the model coherently with notation used in [13] (see Fig.2.0.4), calling first, second, third and fourth part $(x_l)_{-M\leq l\leq 0}$, $(y_l)_{0\leq l\leq M}$, $(x_l)_{0\leq l\leq N}$ and $(y_l)_{-N\leq l\leq 0}$ respectively.

Now, the occurrence of $t \approx 0.5$ for the model with mutations distributed uniformly throughout the sequence would imply that the mononucleotide and dinucleotide frequencies for the first (second) part of each strand (read in the process direction) are closer to each other than those over any other part. In fact frequencies only depends on probabilities of dinucleotide and second (first) and third (fourth) parts are statistically similar, as said in the previous section.

The authors found the same property in the 1049 bacteria genomes analyzed. After splicing in two (since $t \approx 0.5$) each genome sequence we observe nucleotide and dinucleotide frequencies in the four parts as shown in Fig2.0.4, we denote them with P_i and π_i respectively.

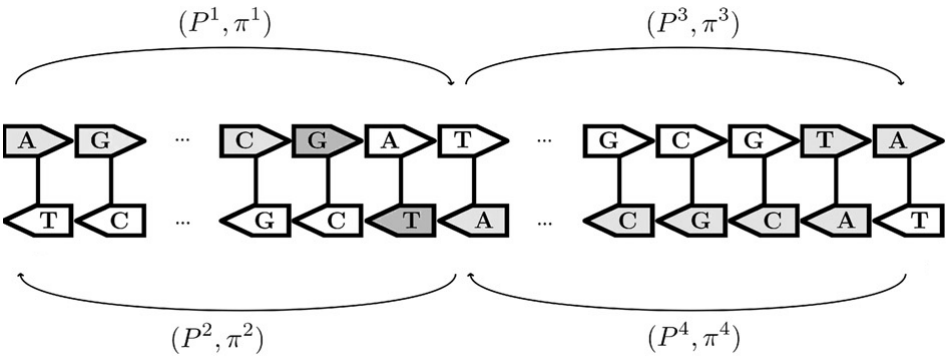

Figure 2.5: Matrices P_i and vectors π_i, $i = 1, 2, 3, 4$, contain the dinucleotide frequencies of each corresponding half of $n-$bacteria genome

All the genomes taken in exam satisfied the property which was predicted by the model, that is, the mononucleotide and dinucleotide frequencies were found to match most closely between half 1 and half 4 and between halves 2 and 3, with exception of dinucleotides AT,CG,TA,GC. This is easily explained by Chargaff's first parity rule, as $\overline{\omega} = \omega$, when $\omega \in \{AT, TA, CG, GC\}$, so that for those special dinucleotide frequencies on part 3 are exactly the same of frequencies on part 4, and frequencies on part 1 are equals to those of part 2.

Finally, they inferred that the model eventually respects an observed property of genomes. Indeed, a relation between $C + G$ content and dinucleotide frequencies is evident. In particular, if we plot the couples $(CG(n), P_{ij}(n))$, where, they appear to be systematically distributed around some curve.

They use matrices $\bar{\mathcal{N}}$ and $\overline{\overline{\mathcal{N}}}$ to produce mononucleotide and dinucleotide frequencies $(\overline{\pi}(m), \overline{P}(m))$ and $(\overline{\overline{\pi}}(m), \overline{\overline{P}}(m))$, with distinct values of $m \in (0, 0.5)$. We remind that different values of m give different probability vectors $\mu = (\mu_A, \mu_C, \mu_G, \mu_T)$. Then they calculate the $C+G$ content as function of m, i.e. points $(\overline{CG}(m), \overline{P}_{ij})$ and $(\overline{\overline{CG}}(m), \overline{\overline{P}}_{ij})$.

Plotting the point obtained as above in the same plot, they observed that not only the curves of the frequencies generated by $\overline{\mathcal{N}}$ and $\overline{\overline{\mathcal{N}}}$ were very close to each other, but also the majority of the points of the actual $C+G$ content and dinucleotide frequencies of the 1049 bacteria were distributed around those curves, as shown in Fig. 2.0.4. This suggests that the construction process defined by $\bar{\mathcal{N}}$ naturally lead to a process satisfying hypothesis of symmetries in the entries H2.

The construction of the model, produces double stranded DNA according to Chargaff's second parity rule even without assuming the Markovian hypothesis. In fact, it derives naturally from any stochastic construction around an initial nucleotide pair analogously to the way described above.

Note 2.1. *Considering the results of simulations of the model, it is shown that it works for $t \approx 0.5$ so that if we consider the given genome as the result of the model, we can suppose the process started approximately at the center of the double stranded DNA. However, if we suppose bacteria's genome as the product of the model too, establishing where the initial point is in circular DNA is not possible*

Since bacteria DNA is generally circular, finding the initial base x_0 and the corresponding base y_0 is not trivial, because there is no "half" to look for. For linear DNA the starting point is presumed to be in the middle.

Thus, an additional step it is needed in order to linearized the circular double stranded DNA. Bacteria DNA is then cut at some arbitrary point. However, this does not affect what was predicted by the model, in other words

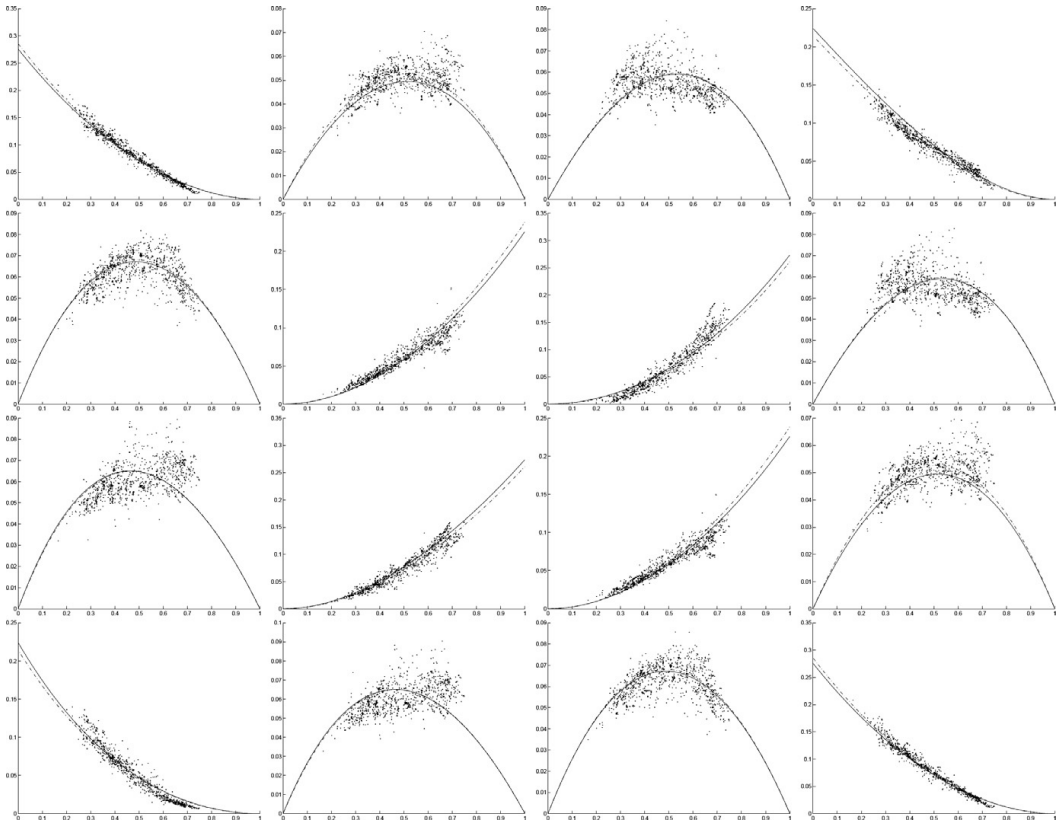

Figure 2.6: Plot in row i and column i shows points $(CG(n), P_{ij}(n))$, for the $n-$th bacteria genome examinated and the curves obteined from $(\overline{CG}(m), \overline{P}_{ij}(m))$ and $(\overline{\overline{CG}}(m), \overline{\overline{P}}_{ij}(m))$.

the linearized DNA still follows the property of having similar frequencies for parts $1-4$ *and* $2-3$*. In fact, whereas the slice is situated, the DNA will result as a transition of some sequence produced by the model. Each of four sub-strings contains both sequences with frequencies* (Q,ν) *and* $(\overline{Q},\overline{\nu})$*, in particular first and fourth parts have the same portion of sequences with frequencies* (Q,ν) *and* $(\overline{Q},\overline{\nu})$*, and similarly happens for second and third parts. More formally, referring to the example of fig.2.1, we have*

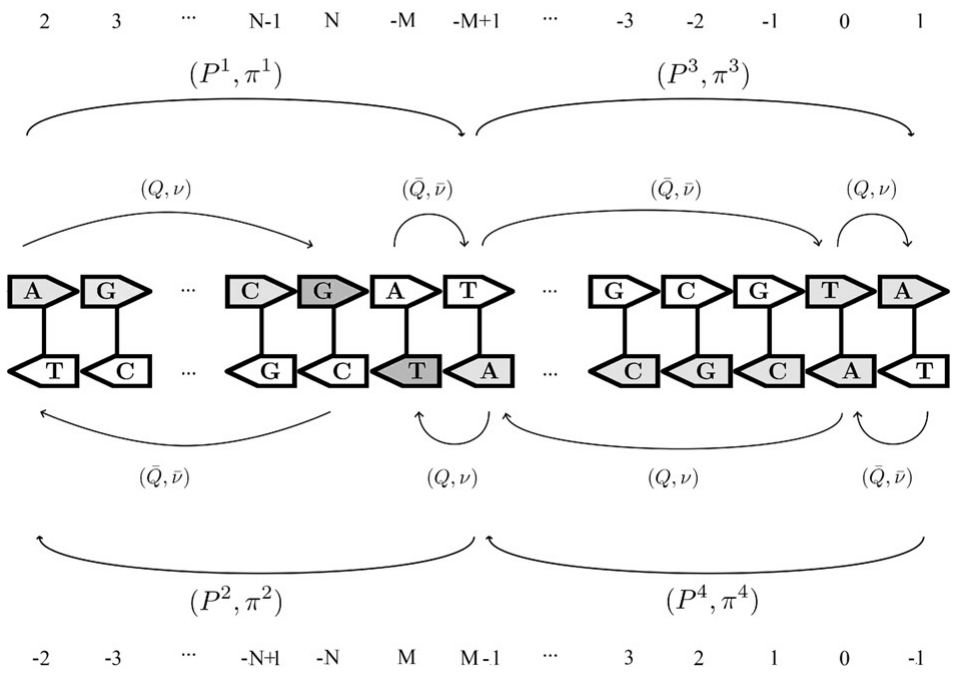

Figure 2.7: The figure shows a linearized string obtained by cutting a circular DNA sequence first generated by joining the extremities of a realization of the model. Naming halves $1,2,3,4$ as before, frequencies of nucleotide and dinucleotide result distributed in the same way as the linearized DNA. For example, half 1 has dinucleotide sequences.(original figure from [13])

$$f^1_{mono} = 2\overline{\nu} + (N-2)\nu$$
$$f^1_{bi} = 2\overline{Q} + (N-2)Q$$
$$f^2_{mono} = (N-2)\overline{\nu} + 2\nu$$
$$f^2_{bi} = (N-2)\overline{Q} + 2Q$$
$$f^3_{mono} = (M-2)\overline{\nu} + 2\nu$$

$$f^3_{bi} = (M-2)\overline{Q} + 2Q$$
$$f^4_{mono} = (M-2)\overline{\nu} + 2\nu$$
$$f^4_{bi} = (M-2)\overline{Q} + 2Q$$

where f^i_{mono} *are mononucleotide frequencies of part* i *and* f^i_{bi} *are dinucleotide frequencies of part* i*. It is evident that the frequencies are more close in parts* $1-2$ *and* $3-4$*, as predicted.*

2.0.5 Observations

In conclusion we want to make some remarks.

First, the model is studied to comply with Chargaff's second parity rule only for mononucleotides and dinucleotides. Anyway, this rule may hold for word lengths up to 10 nucleotides for bacteria and some eukariotic genomes and 6 nucleotides for human genome (see [**?**]). It would be interesting investigate on an extension of the model that could predict a symmetry for words of length greater then two.

Moreover, in [13] it is shown a correlation between CG content and elements of the frequencies matrix P. We saw that this is still valid if we consider Bernoulli processes instead of Markov chains generating the two half strands. One may want to verify this property for different processes applied to the model. In the end, one may be interested in analyzing the

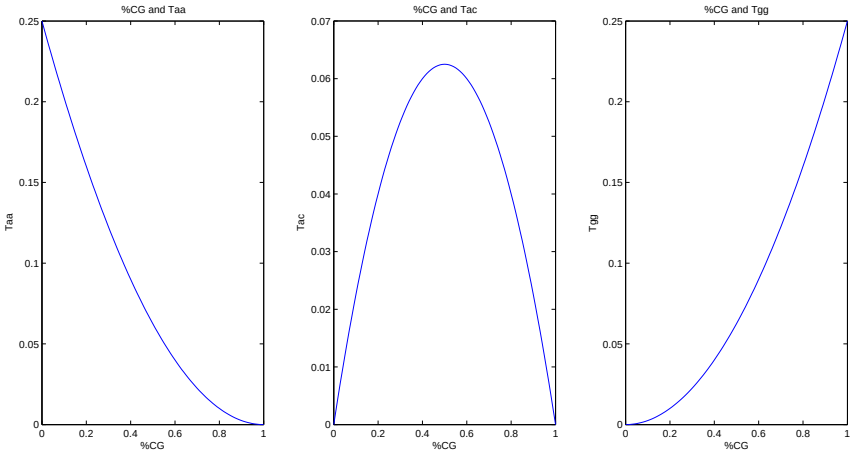

Figure 2.8: Three plots of couples $(CG(m), P_{AA}(m))$,$(CG(m), P_{AC})$ and $(CG(m), P_{GG}(m))$ generated with Bernoulli process instead of Markov chains show that the curves are close to the graphics generated by the model.

process by taking into consideration the time, i.e. looking at each rejection.

In fact, the resulting string analyzed in the previous sections do not shows all steps that brought to it. Let s be the beginning portion of a realization of the model, for example $s = (x_l)_{l=0}^{5} = ACCGTA$. This succession doesn't give information about bases that have been rejected over time. Some changes could be done to the model, so that the final string could reveal information about the process of joins of bases. We aim to briefly describe how such a process would be.

A first step to do is then to extend the alphabet of the string with a symbol assigned to the occurrence of a rejection. Let denote it with $*$, then the process that we want to describe is a Markov chain in the alphabet $\mathcal{I} \cup \{*\} = \{A, C, G, T, *\}$. In this way, the string counts both when a nucleotide is accepted and when it is not. If we see at the previous example, the string could be of the form $A**CCG*TA$ and it uses 9 steps instead of 6 to reach the final 6−bases long strand.

Anyway, the process using alphabet $\mathcal{I} \cup \{*\}$ is not a Markov chain. Indeed, the probability of a letter $x \in \mathcal{I} \cup \{*\}$ in a sequence that ends with n characters of kind $*$ do not depends only on the anterior letter. In fact it will depend on the previous $(n+1)-$nucleotides. Going back to the example, in the string $(x_l)_{l=0}^{8} = A**CCG*TA$, base $x_3 = C$ depends on A at position x_0.

In order to solve that, we add 4 different characters to the original alphabet $\mathcal{I}$, each one denoting a rejection that store the last base attached to a string. Denoting with i^* the rejection after a last base $i \in \mathcal{I}$, the alphabet describing the new process will be $\mathcal{I}^* = \{A, C, G, T, A^*, C^*, G^*, T^*\}$.

The transition matrix T^* is the 8×8 matrix with all possible transition probability. We can divide T^* in four blocks, equals in pairs.

In fact, we observe that if more than one rejection occurs, the stored final base remains fixed, so that $T_{ij^*} = 0$, $\forall i \neq j$. Moreover, the transition probability from the state i^* to the state j, $i, j \in \mathcal{I}$, is exactly T_{ij}, j depending only on the last letter attached to the string, that is i, for how we defined letters i^*.

Thus, matrix T^* can be represented as follow

$$T^* = \left(\begin{array}{c|c} T & D \\ \hline T & D \end{array}\right)$$

where $T = (T_{ij})_{i,j \in \mathcal{I}}$ is the matrix of probabilities of dinucleotides.

$$T_{ij} = \mu(j) a_{ij} \quad \forall i, j \in \mathcal{I}$$

On the other hand, D is the diagonal matrix of all possible rejections

$$D = \begin{pmatrix} T_{AA^*} & 0 & 0 & 0 \\ 0 & T_{CC^*} & 0 & 0 \\ 0 & 0 & T_{GG^*} & 0 \\ 0 & 0 & 0 & T_{TT^*} \end{pmatrix} \tag{2.10}$$

Each transition probability T_{ii^*} of matrix D is the probability of having a rejection of any base of alphabet $\mathcal{I}$.

Thus, recalling notation of the previous section, we have

$$T_{ii^*} = \sum_{j \in \mathcal{I}} \mu(j)(1 - a_{ij}), \forall i \in \mathcal{I} \tag{2.11}$$

Furthermore, T^* is stochastic. Indeed, elements of each row sum to the unit.

$$\begin{aligned} \sum_{j \in \mathcal{I}^*} T^*_{ij} &= \sum_{i \in \mathcal{I}} \mu(i) a_{ij} + \sum_{i \in \mathcal{I}} \mu(i)(1 - a_{ij}) = \\ &= \sum_{i \in \mathcal{I}} \mu(i)(a_{ij} + 1 - a_{ij}) = \sum_{i \in \mathcal{I}} \mu(i) = \\ &= 1 \end{aligned} \tag{2.12}$$

Matrix T^* is the transition matrix of the process describing both joins and rejections.

Chapter 3

Simple stationary processes and concatenations of stationary processes

Our goal is to define processes that reproduce symmetries found in DNA. Among all genome's properties, we choose to study processes that comply with Chargaff's second parity rule, in particular Chargaff's second parity rule extended to $k-$words. Since in a double stranded DNA one strand follows by the other, complying Chargaff's first parity rule, we provide a model for a unique strand, without taking under consideration its complement.

After giving preliminary notions of stochastic processes and stationary processes, we introduce the simplest case of a single stationary process such as Bernoulli process and Markov chain. In the end, we analyze the case of the concatenation of two stationary processes.

We study conditions on the probabilities such that the processes defined above can abide by Chargaff's second parity rule. We refer to [8] to recall definitions of stochastic processes and stationary processes.

Def 1. Let X_i be a family of random variables in a probability space $(\Omega, \mathcal{F}, P)$, indexed by a parameter $i \in T$, such that T is a subset of real line. Then X_t is called a *stochastic process.*

The set of all possible values of X_i is called *state set* and it is noted with $\mathcal{I}$.

In this Chapter, we consider the time set T as a discrete set that indicates the position of a specific state among the realization.

Def 2. We define the space of one sided sequences in the given alphabet $\mathcal{I}$

as

$$\sum_{\mathcal{I}}^{+} = \{(x_i)_{i=1}^{\infty} \mid x_i \in \mathcal{I}\} \tag{3.1}$$

that is the set of all possible sequences with elements in $\mathcal{I}$.

Similarly, the space of bi-sided sequences in $\mathcal{I}$ is

$$\sum_{\mathcal{I}} = \{(x_i)_{i=1 \in \mathbb{Z} \setminus \{0\}} \mid x_i \in \mathcal{I}\} \tag{3.2}$$

and includes all possible bi-infinite sequences in $\mathcal{I}$.

In our work, random variables take values in the alphabet of all possible nucleotides, i.e. $\mathcal{I} = \{A, C, G, T\}$. Moreover, the space Ω will be the set ${\sum_{\mathcal{I}}}^{+}$ for simple stationary process and $\sum_{\mathcal{I}}$ for the case of concatenation of stationary processes. The $\sigma-$algebra $\mathcal{F}$ is the $\sigma-$algebra generated by the cylinders, that we will define later. Probability measure will change consistently with the process generating the sequences.

We start studying stationary processes, that are processes such that any sequence of n consecutive points has the same distribution as any other sequence of n consecutive points.

More formally, a stochastic process X_i is said to be *stationary* if for every $i_1, \dots i_k$ for every sequence $x_1 \dots x_k$ and for every $n \in \mathbb{N}$

$$P(X_{i_1} = x_1 \dots X_{i_k} = x_k) = P(X_{i_1+n} = x_1 \dots X_{i_k+n} = x_k) \tag{3.3}$$

where $P(X_{i_1} = x_1 \dots X_{i_k} = x_k)$ is the probability of having sequence $x_1 \dots x_k$ at time $i_1 \dots i_k$.

From now on, we will denote with $\mathcal{X}$ a stationary stochastic process that generates sequences s in ${\sum_{\mathcal{I}}}^{+}$ or $\sum_{\mathcal{I}}$, with random variables that vary in the state space $\mathcal{I} = \{A, C, G, T\}$.

3.1 Simple stationary processes

In this section we study two cases of simple stationary processes. The first one is a generalization of the Bernoulli process, the second is a Markov chain.

The space of the stochastic process is ${\sum_{\mathcal{I}}}^{+}$, that is the collection of the infinite sequences $(x_i)_{i=1}^{\infty}$, with x_i in the state space $\mathcal{I}$.

Note 3.1. *For convenience index i varies in $\mathbb{N} \setminus \{0\}$, as in the case of the combination of simple stationary process will be useful to remove 0-term.*

We may refer to the realization of $\mathcal{X}$ in ${\sum_{\mathcal{I}}}^{+}$ with s during the work.

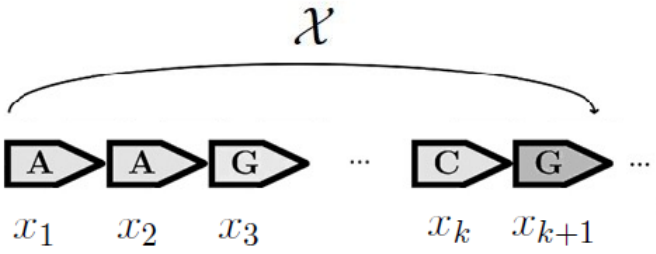

Figure 3.1: A possible realization of a given stochastic process $\mathcal{X}$. The arrow shows the direction of the process. Each x_i is a letter in the alphabet of the nucleotides $\mathcal{I} = \{A, C, G, T\}$.

3.1.1 Preliminary definitions

In this section we give definitions of cylinders, probability of cylinders and Chargaff processes that are valid for both Bernoulli and $1-$Markov processes.

A finite succession of length k of letters $x_i \in \mathcal{I}$ is a *word*, and it is denoted with ω. Hence, $\omega = \omega_1\omega_2 \ldots \omega_k$ is an element of $\mathcal{I}^k$, and $\omega_i \in \mathcal{I}$.

We denote with $\bar{\omega}_i$ the complement of the $i-$term of a word ω, according to the parity rule given by Chargaff's first parity rule. In other words,

$$\begin{aligned} \bar{A} &= T \qquad \bar{T} = A \\ \bar{C} &= G \qquad \bar{G} = C \end{aligned} \tag{3.4}$$

Thus, the reverse complement of a word $\omega = \omega_1\omega_2 \ldots \omega_k$ is the word composed of the complements of every letter written in the opposite order. Denoting with $\hat{\omega}$ and we have that

$$\hat{\omega} = \bar{\omega}_k \ldots \bar{\omega}_2\bar{\omega}_1$$

Example 3.2. *Let* $\omega = ACGGTGAAG$ *a* $9-$*word. Then* $\hat{\omega} = CTTCACCGT$, $\omega_2 = C$, $\bar{\omega}_2 = G$ *and* $\hat{\omega}_2 = T$.

Since we want to make hypothesis on the process to enable Chargaff parity rule, we need to introduce definitions of cylinders, and then impose restrictions on cylinders' probabilities. Thus, we need to recall the following

Def 3. A *cylinder* is a subset of $\sum_{\mathcal{I}}^{+}$

$$\mathcal{S}_{j,j+k-1}(\omega_1 \ldots \omega_k) = \{(x_i)_{i=1}^{+\infty},\ x_i \in \mathcal{I} : x_i = \omega_{i-j+1}\ \forall j \leq i \leq j+k-1\} \tag{3.5}$$

A cylinder of the form $\mathcal{S}_{1,k}$ is called a *simple cylinder*.

We will use only simple cylinders and we will use the notation $\mathcal{S}_k$ instead of $\mathcal{S}_{1,k}$ for convenience.

Since the process is stationary, the probability of a letter (or a word) doesn't depend on the position it occupies in the string. This means that, given $\omega = \omega_1\omega_2 \dots \omega_k$ a word of length k, the probability of finding the word at the beginning point x_1 of a string $s = x_1x_2 \dots\dots \in \sum_{\mathcal{I}}^{+}$ is equal to the probability of find the word at the starting point x_i.

So that we have

$$P(x_1 = \omega_1, x_2 = \omega_2 \dots x_k = \omega_k) = P(x_i = \omega_1, x_{i+1} = \omega_2 \dots x_{i+k-1} = \omega_k) \tag{3.6}$$

Thus, as

$$\mathcal{S}_k = \{(x_i)_{i=1}^{+\infty},\ x_i \in \mathcal{I}\ \ s.t.\ \ x_i = \omega_i\ \forall 1 \leq i \leq k\}$$

from (3.6) we have

$$P(\mathcal{S}_k) = P(\mathcal{S}_{i,i+k-1}),\ \ \forall i = 1, 2, \dots$$

From now on we will consider the cylinders centered at the initial point of the string, as it doesn't affect the hypothesis.

Similarly as we have done before with words and letters, we can define the reverse complement of a given cylinder.

Def 4. Given a simple cylinder

$$\mathcal{S}_k(\omega_1 \dots \omega_k) = \{(x_i)_{i=1}^{+\infty},\ x_i \in \mathcal{I}\ \ s.t.\ \ x_i = \omega_i\ \forall 1 \leq i \leq k\}$$

its *reverse complement* $\hat{\mathcal{S}}_k$ is the cylinder generated by the reverse complement of the word generating $\mathcal{S}$, i.e.

$$\hat{\mathcal{S}}_k(\omega) = \mathcal{S}_k(\hat{\omega}) = \{(x_i)_{i=1}^{+\infty},\ x_i \in \mathcal{I}\ \ s.t.\ \ x_i = \bar{\omega}_{k+n-i}\ \forall 1 \leq i \leq k\} \tag{3.7}$$

Now we can define a process that enable Chargaff's second parity rule in the resulting sequences. This can be done naturally by imposing the probability of a cylinder and the probability of its reverse complement to be the same.

Def 5 (*Chargaff process*)**.** Let $\mathcal{X}$ be a stochastic stationary process with sequences$(x_t)_{t\in\mathbb{N}}$ that take values in the alphabet $\mathcal{I} = \{A, C, G, T\}$ in a probability space $(\mathcal{I}, P)$. We call $\mathcal{X}$ a *Chargaff process* if and only if, for every $k > 0$, for every word $\omega \in \mathcal{I}^k$ the following equality is attended

$$P(\mathcal{S}_k(\omega)) = P(\hat{\mathcal{S}}_k(\omega)) \tag{3.8}$$

We will say that a process is $\mathcal{C}^{II}$ to denote that $\mathcal{X}$ is a Chargaff processes.

Since in some cases it is better to consider the property expressed by 3.8 for words of a fixed lengh k, it is convenient to define a more relaxed Chargaff process.

Def 6. Let k be fixed in $\mathbb{N}$ and let $\mathcal{X}$ be a process defined as in (5). Then $\mathcal{X}$ is said to be a $k-$*Chargaff process* if for every word ω in $\mathcal{I}^k$ it is

$$P(\mathcal{S}_k(\omega)) = P(\hat{\mathcal{S}}_k(\omega)) \tag{3.9}$$

and we say $\mathcal{X}$ is $\mathcal{C}_k^{II}$.

3.1.2 Properties of Chargaff-processes

Here we study the main properties of Chargaff processes. As we will see, the following observations hold for both simple stochastic processes and concatenations of stochastic processes.

A question one might ask is how k-Chargaff processes are related in function of k.

Theorem 3.3. *Let $\mathcal{X}$ be a process as defined above. If the process is $k-$Chargaff for a $k \in \mathbb{N}$, then the process is $(k-1)-$Chargaff.*

More formally

$$\mathcal{C}_k^{II} \Longrightarrow \mathcal{C}_{k-1}^{II} \tag{3.10}$$

Proof. Let $\mathcal{S}_{k-1}$ be a cylinder in $\sum_{\mathcal{I}}^{+}$. We have to prove that $\forall\ \mathcal{S}_{k-1} \in \sum_{\mathcal{I}}^{+},\ \forall k \in \mathbb{N}$

$$P(\mathcal{S}_{k-1}) = P(\hat{\mathcal{S}}_{k-1}) \tag{3.11}$$

Given $\omega = \omega_1\omega_2...\omega_{k-1}$ word in the alphabet $\mathcal{I} = \{A, C, G, T\}$, the probability of the cylinder centered in ω is

$$P(\mathcal{S}_{k-1}) = P(x_1 = \omega_1, x_2 = \omega_2, \ldots, x_{k-1} = \omega_{k-1})$$

One can define the probability of the cylinder centered in ω of length $k-1$ as function of the probability of a word of length k and apply $\mathcal{C}_k^{II}$, so that

$$
\begin{aligned}
P(\mathcal{S}_{k-1}) &= P(x_1 = \omega_1, x_2 = \omega_2, \ldots, x_{k-1} = \omega_{k-1}) = \\
&= \sum_{x \in \mathcal{I}} P(x_1 = \omega_1, x_2 = \omega_2, \ldots, x_{k-1} = \omega_{k-1}, x_k = x) = (3.12) \\
&= \sum_{x \in \mathcal{I}} P(\mathcal{S}_k(\omega x)) = \sum_{x \in \mathcal{I}} P(\hat{\mathcal{S}}_k(\omega x)) = \sum_{\bar{x} \in \mathcal{I}} P(\mathcal{S}_k(\bar{x}\hat{\omega})) = (3.13) \\
&= \sum_{x \in \mathcal{I}} P(x_1 = \bar{x}, x_1 = \bar{\omega}_{k-1}, \ldots, x_k = \bar{\omega}_1) = \\
&= P(\hat{\mathcal{S}}_{k-1}(\omega)) \qquad (3.14) \\
& \qquad (3.15)
\end{aligned}
$$

that proves the theorem. □

3.1.3 Bernoulli-scheme

In this section we analyze the conditions that guarantee the process to be $\mathcal{C}^{II}$ when $\mathcal{X}$ is a Bernoulli scheme. In particular, we will see that asking the Bernoulli scheme to be a Chargaff process, in the sense of the definition (5), is sufficient for the validity of the Chargaff rule on words of every length. Let us first introduce some definitions and notations.

A homogeneous sequence of independent trials is called a sequence of Bernoulli trials if the state space S consists of two elements (see [8]). In our case the state space consists of all possible nucleotides, i.e. four elements. However, every sequence result to be made of independent trials, so that the process can be compared to a Bernoulli one.

Def 7. Let $(x_i)_{i=1}^{+\infty}$ be a stationary stochastic process with $x_i \in \mathcal{I} = \{A, C, G, T\}$. Let $\pi = (\pi(A), \pi(C), \pi(G), \pi(T))$ be a vector such that

- $\pi(x) \leq 1, \quad \forall x \in \{A, C, G, T\}$
- $\sum_{x \in \mathcal{I}} \pi(x) = 1$

Then the process is called *Bernoulli scheme* if for every $k \in \mathbb{N}$ and for every word $\omega \in \mathcal{I}^k$ the probability is defined as

$$
P(\mathcal{S}_k(\omega)) = \prod_{i=1}^{k} \pi(\omega_i) \qquad (3.16)
$$

We recall that in a Bernoulli process the probabilities are independent and so it is for a Bernoulli scheme. Therefore, since the conditions on $\mathcal{X}$ that make the process a Chargaff process are limitations on the probabilities of cylinders, we have to make restrictions on π. As we will see, a condition on π will provide $\mathcal{C}_k^{II}$ for every k.

Note 3.4. *Let $\mathcal{X}$ be a Bernoulli scheme.*

By definition of Chargaff process, $\mathcal{X}$ is a $1-$Chargaff process if and only if for every $1-$word $\omega = x$ in $\mathcal{I}$ it is

$$P(\mathcal{S}_1(x)) = P(\hat{\mathcal{S}}_1(x)) \tag{3.17}$$

Since probabilities are defined by (3.16), we have

$$P(\mathcal{S}_1(x)) = \pi(x) \qquad P(\hat{\mathcal{S}}_1(x)) = \pi(\bar{x}) \tag{3.18}$$

Thus 3.17 become

$$\pi(x) = \pi(\bar{x}), \quad \forall x \in \{A, C, G, T\} \tag{3.19}$$

3.19 gives the restriction on probability vector π that enable $\mathcal{X}$ to be $\mathcal{C}_1^{II}$.

Condition (3.19) on the probability vector is sufficient to guarantee the Chargaff property for any $k-$word, $\forall k \in \mathbb{N}$.

More formally, we have the following result:

Proposition 3.5. *Let $\mathcal{X}$ be a Bernoulli scheme defined by the probability vector π. If $\mathcal{X}$ is a $1-$Chargaff process, then it is Chargaff.*

In other words, if (3.19) are true, then

$$P(\omega_1 \dots \omega_k) = P(\bar{\omega}_k \dots \bar{\omega}_1) \forall k \in \mathbb{N}, \forall \omega_i \in \mathcal{I}, i = 1, \dots k \tag{3.20}$$

that is the second Chargaff parity rule is valid for any word of any length.

Proof. Let us fix an arbitrary$k$ in $\mathbb{N}$ and let $\omega = \omega_1\omega_2 \dots \omega_k$ be a word in $\mathcal{I}^k$. Suppose $\mathcal{X}$ is Chargaff, we want to prove that

$$P(\mathcal{S}_k(\omega)) = P(\hat{S}_k(\omega)) \tag{3.21}$$

Because the process is a Bernoulli scheme, the probabilities of the cylinders are

$$P(\mathcal{S}_k(\omega)) = \prod_{i=1}^{k} \pi(\omega_i), \quad P(\hat{\mathcal{S}}) = \prod_{i=1}^{k} \pi(\bar{\omega}_i) \tag{3.22}$$

Consequently, (3.21) is true if and only if

$$\prod_{i=1}^{k} \pi(\omega_i) = \prod_{i=1}^{k} \pi(\bar{\omega}_i) \tag{3.23}$$

But from the hypothesis $\mathcal{P}$ is $\mathcal{C}_1^{II}$, therefore

$$\pi(\omega_i) = \pi(\bar{\omega}_i) \quad \forall i = 1, \dots k \tag{3.24}$$

Hence equation (3.21) is satisfied and it ends the proof. □

3.1.4 1−Markov process

In this section we study the case of $\mathcal{X}$ Markov chain. We give definitions od $k-$Chargaff and Chargaff process and we analyze the conditions on the process in order to be $\mathcal{C}^{II}$. Before we define a $1-$Markov process, we recall the definition of a stochastic matrix and a distribution vector (see ([8])).

Def 8. Let π on S be a vector $\pi = (\pi_i)_{i\in\mathcal{I}}$ s.t.

1. $\pi(i) \geq 0, \quad \forall i \in \mathcal{I}$
2. $\sum_{i\in\{A,C,G,T\}} \pi(i) = 1$

Then π is called a *distribution vector*.

Def 9. A matrix $T = (T_{ij})_{i,j=1,\dots n}$ is said to be *stochastic* if

1. $T_{ij} \geq 0 \ \ \forall i, j = 1, \dots n$
2. $\sum_{j=1}^{n} t_{ij} = 1, \quad \forall i = 1, \dots n$

Let now $\mathcal{I} = l_1 \dots _n$ be the state space of a stochastic stationary process $(x_t)_{t\in\mathbb{N}}$, $\pi = (\pi_i)_{i\in\mathcal{I}}$, and $T = (t_{ij})_{i,j\in\mathcal{I}}$ be a stochastic matrix of dimensions $n * n$.

Def 10. The *Markov chain* with the state space $\mathcal{I}$ generated by the distribution π on $\mathcal{I}$ and the stochastic matrix T is the probability measure P on ${\sum_{\mathcal{I}}}^{+\infty}$ s.t.

$$P(x_1 = \omega_1 \dots x_k = \omega_k) = \pi(\omega_k) T_{\omega_1\omega_2} \dots T_{\omega_{k-1}\omega_k} \tag{3.25}$$

T is called the *transition matrix* and elements $(t_{ij})_{i,j\in\mathcal{I}}$ are the transition probabilities of going from the state i to the state j.

In other words, in a Markov chain, each state only depends on the previous state in the sequence (similarly, in a $k-$Markov process a state depends on the k previous states in the sequence).

In our case $T = (T_{ij})_{i,j\in\{A,C,G,T\}}$ is the transition matrix defining the probabilities of the letter j to follow letter i, while $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ is the initial probability vector. For convenience of notation, we will may use $\pi(i)$ instead of π_i. For example, the element T_{AG} is the probability that the nuclotide G join the realization after the Adenine.

According to the definition (5) given in the section above, the process $\mathcal{P}$ is said to be Chargaff if $\forall k \in \mathbb{N}, \quad \forall \omega \in \mathcal{I}^k$

$$P(\mathcal{S}_k(\omega)) = P(\hat{\mathcal{S}}_k(\omega)) \tag{3.26}$$

while it is $k-$Chargaff if

$$P(S_k(\omega)) = P(\hat{S}_k(\omega)) \forall \omega \in \mathcal{I}^k$$

As before, because the process is stationary, we choose $\mathcal{S}_k$ simple cylinder without loss of generality.

The probability of a given cylinder $\mathcal{S}_k$ will thus be

$$\begin{aligned} P(\mathcal{S}_k) &= P(x_1 = \omega_1, \dots x_k = \omega_k) = \\ &= P(x_k = \omega_k \mid x_1 = \omega_1 \dots x_{k-1} = \omega_{k-1}) P(x_1 = \omega_1, \dots x_{k-1} = \omega_{k-1}) = \\ &= P(x_k = \omega_k \mid x_{k-1} = \omega_{k-1}) P(x_1 = \omega_1, \dots x_{k-1} = \omega_{k-1}) \\ &= T_{\omega_{k-1}\omega_k} P(x_1 = \omega_1, \dots x_{k-1} = \omega_{k-1}) \end{aligned} \tag{3.27}$$

Proceeding by iteration we have that

$$P(\mathcal{S}_k) = \pi(\omega_1) \prod_{i=2}^{k} T_{\omega_{i-1}\omega_i} \tag{3.28}$$

First, let us analyze the simplest case. When $k = 1$ the conditions are the same as the Bernoulli process.

Proposition 3.6. *Let $\mathcal{X}$ be a $1-$Markov process. Let $T = (t_{ij})_{i,j \in A,C,G,T}$ be the transition matrix and the stationary distribution.*
Then, $\mathcal{X}$ is $1-$Chargaff if and only if

$$\pi(x) = \pi(\bar{x}),\ \forall x \in \{A, C, G, T\} \tag{3.29}$$

Proof. Given $x \in \mathcal{I} = \{A, C, G, T\}$, let $\mathcal{S}_1(x) = \mathcal{S}(x)$ a simple cylinder centered on x. The probability of the cylinder is the probability of the letter x under π

$$P(\mathcal{S}(x)) = P(x_1 = x) = \pi(x)$$

So

$$P(\mathcal{S}(x)) = P(\hat{\mathcal{S}}(x)) \iff \pi(x) = \pi(\bar{x}),\ \forall x \in \mathcal{I} \tag{3.30}$$

□

Proposition 3.7. *The Markov process $\mathcal{X}$ defined by the probability vector $\pi = (\pi(A), \pi(C), \pi(G), \pi(T))$ and the transition matrix $T = (T_{i,j})_{i,j \in \mathcal{I}}$ is $2-$Chargaff if and only if*

$$T_{\omega_1\omega_2} = \frac{\pi(\bar{\omega}_2)}{\pi(\omega_1)} T_{\bar{\omega}_2\bar{\omega}_1} \quad \forall\, \omega_1, \omega_2 \in \mathcal{I} \tag{3.31}$$

Proof. First we prove ($\longrightarrow$).

Let $\omega = \omega_1\omega_2$ be a word in the alphabet $\mathcal{I}$. Since the process is $\mathcal{C}_2^{II}$ we have that

$$\mathcal{S}_2(\omega) = \hat{\mathcal{S}}_2(\omega) \tag{3.32}$$

that is

$$\begin{aligned} P(x_1 = \omega_1, x_2 = \omega_2) &= P(x_1 = \bar{\omega}_2, x_2 = \bar{\omega}_1) \\ \pi(\omega_1)T_{\omega_1\omega_2} &= \pi(\bar{\omega}_2)T_{\bar{\omega}_2\bar{\omega}_1} \qquad (3.33) \\ & \qquad (3.34) \end{aligned}$$

From (3.33) we have the thesis (2).

We prove now ($\longleftarrow$).

$$\begin{aligned} P(\mathcal{S}_2(\omega_1\omega_2)) &= \pi(\omega_1)T_{\omega_1\omega_2} = \qquad (3.35) \\ &= \pi(\omega_1)\frac{\pi(\bar{\omega}_2)}{\omega_1}T_{\bar{\omega}_2\bar{\omega}_1} = \\ &= \pi(\bar{\omega_2})T_{\bar{\omega}_2\bar{\omega}_1} = P(\hat{\mathcal{S}}_2(\omega_1\omega_2)) \qquad (3.36) \end{aligned}$$

It ends the proof. □

We remind that if a process is $k-$Chargaff, then it is l-Chargaff $\forall l \neq k$, see (3.10). Thus, the condition (3.31)) guarantee that the process is $\mathcal{C}_k^{II}$ for every k.

Proposition 3.8. *If $\mathcal{X}$ is a $2-$Chargaff process, then it is a $k-$Chargaff process $\forall k \geq 2$. In other words, if Chargaff second parity rule holds for every word of length two, then it holds for every word.*

More formally

$$\mathcal{C}_2^{II} \Longrightarrow \mathcal{C}_k^{II}, \quad \forall k \geq 2 \tag{3.37}$$

Proof. We will prove the assertion by induction.

We first prove that

$$\mathcal{C}_2^{II} \Longrightarrow \mathcal{C}_3^{II} \tag{3.38}$$

We have to see that

$$P(\omega_1\omega_2\omega_3) = P(\bar{\omega}_3\bar{\omega}_2\bar{\omega}_1), \quad \forall \omega_i \in \mathcal{I} = \{A, C, G, T\}, i = 1, 2, 3$$

Because the process is $1-$Markov, we can write the probabilities above as

$$\pi(\omega_1)T_{\omega_1\omega_2}T_{\omega_2\omega_3} = \pi(\bar{\omega}_3)T_{\bar{\omega}_3\bar{\omega}_2}T_{\bar{\omega}_2\bar{\omega}_1} \tag{3.39}$$

As the process is $\mathcal{C}_2^{II}$ we have that

$$\pi(\omega_1)T_{\omega_1\omega_2} = \pi(\bar{\omega}_2)T_{\bar{\omega}_2\bar{\omega}_1}$$

Replacing the first member in 3.39 we obtain

$$\pi(\omega_1)T_{\omega_1\omega_2}T_{\omega_2\omega_3} = \pi(\omega_2)T_{\bar{\omega}_2\bar{\omega}_1}T_{\omega_2\omega_3} = \pi(\bar{\omega}_2)T_{\bar{\omega}_2\bar{\omega}_1}T_{\omega_2\omega_3}$$

because $\pi(\omega) = \pi(\bar{\omega})$ for every $\omega \in \mathcal{I} = A, C, G, T$ as the process is Chargaff for the 2−words.

Replacing $\pi(\bar{\omega}_2)T_{\omega_2\omega_3}$ with $\pi(\bar{\omega}_3)T_{\bar{\omega}_3\bar{\omega}_2}$, similarly as above we have

$$P(\omega_1\omega_2\omega_3) = P(\bar{\omega}_3\bar{\omega}_2\bar{\omega}_1)$$

We suppose that the process is k−Chargaff and we prove that

$$\mathcal{C}_k^{II} \Longrightarrow \mathcal{C}_{k+1}^{II} \tag{3.40}$$

As above, we will prove that the probabilities are equal, i.e.

$$P(\omega_1\omega_2\ldots\omega_{k+1}) = P(\bar{\omega}_{k+1}\bar{\omega}_k\ldots\bar{\omega}_1), \tag{3.41}$$

for every $\omega_i \in \mathcal{I} = \{A, C, G, T\}, i = 1, \ldots, k{+}1$. Let now calculate $P(\omega_1\omega_2\ldots\omega_k)$. Because the process is 1−Markov the probabilities are

$$P(\omega_1\omega_2\ldots\omega_{k+1}) = \pi(\omega_1)T_{\omega_1\omega_2}\ldots T_{\omega_k\omega_{k+1}} \tag{3.42}$$

Since $P(\omega_1\omega_2\ldots\omega_k) = \pi(\omega_1)T_{\omega_1\omega_2}\ldots T_{\omega_k\omega_{k+1}}$, 3.42 can be written as

$$P(\omega_1\omega_2\ldots\omega_{k+1}) = P(\omega_1\omega_2\ldots\omega_k)T_{\omega_k\omega_{k+1}} \tag{3.43}$$

Moreover $\mathcal{C}_k^{II}$ holds, so we have

$$\begin{aligned} P(\omega_1\omega_2\ldots\omega_{k+1}) &= T_{\omega_k\omega_{k+1}}P(\bar{\omega}_k\bar{\omega}_{k-1}\ldots\bar{\omega}_1) \\ &= \pi(\bar{\omega}_k)T_{\bar{\omega}_k\bar{\omega}_{k-1}}\ldots T_{\bar{\omega}_2\bar{\omega}_1}T_{\omega_k\omega_{k+1}} \end{aligned} \tag{3.44}$$

and because, in particular, $\mathcal{C}_k^{II}$ implies $\mathcal{C}_{k-1}^{II}$, we have that the process is $\mathcal{C}_1^{II}$ and $\mathcal{C}_2^{II}$.

Hence

$$\begin{aligned} \pi(\bar{\omega}_k)T_{\omega_k\omega_{k-1}} &= \pi(\omega_k)T_{\omega_k\omega_{k-1}} = \\ &= P(\omega_k\omega_{k-1}) = P(\bar{\omega}_{k-1}\bar{\omega}_k) = \\ &= \pi(\bar{\omega}_{k+1})T_{\bar{\omega}_{k+1}\bar{\omega}_k} \end{aligned} \tag{3.45}$$

By substitution we obtain

$$\begin{aligned} P(\omega_1\omega_2\ldots\omega_{k+1}) &= \pi(\bar{\omega}_{k+1})T_{\bar{\omega}_{k+1}\bar{\omega}_k}\ldots T_{\bar{\omega}_2\bar{\omega}_1} = \\ &= P(\bar{\omega}_{k+1}\ldots\bar{\omega}_2\bar{\omega}_1) \end{aligned} \tag{3.46}$$

That ends the proof. □

In conclusion, for a $1-$Markov process of transition matrix $T = (T_{ij})_{ij\in\mathcal{I}}$ and probability vector π is Chargaff the following equations are true:

1. $\pi(\omega) = \pi(\bar{\omega}), \quad \forall\omega \in \mathcal{I}$

2. $T_{\omega_i\omega_j} = \frac{\pi(\bar{\omega}_j)}{\pi(\omega_i)} T_{\bar{\omega}_j\bar{\omega}_i}, \ \ \forall\omega_i, \omega_j \in \mathcal{I}$

In addition, since T is a transition matrix, we have that the arrows sum to the unit, that is

$$\sum_{\omega_j\in\mathcal{I}} T_{\omega_i\omega_j} = 1, \ \ \forall\omega_i \in \mathcal{I} \tag{3.47}$$

Thus, from (2) we have

$$\begin{aligned} T_{AG} &= \frac{\pi(C)}{\pi(A)} T_{CT}, \qquad T_{GA} = \frac{\pi(T)}{\pi(G)} T_{TC} \\ T_{CA} &= \frac{\pi(T)}{\pi(C)} T_{TG}, \qquad T_{AC} = \frac{\pi(G)}{\pi(A)} T_{GT} \end{aligned} \tag{3.48}$$

In addition, from (1) it is

$$\begin{aligned} T_{AT} &= 1 - T_{AA} - T_{AC} - T_{AG}, \qquad T_{CG} = 1 - T_{CA} - T_{CC} - T_{CT} \\ T_{GC} &= 1 - T_{GA} - T_{GG} - T_{GT}, \qquad T_{TA} = 1 - T_{TC} - T_{TG} - T_{TT} \end{aligned} \tag{3.49}$$

Hence, from (1), 3.48 and 3.49 we have that the transition matrix T of a $\mathcal{C}^{II}$ Markov chain can be expressed as

$$\begin{pmatrix} T_{AA} & T_{AC} & T_{AG} & 1 - T_{AA} - T_{AC} - T_{AG} \\ \lambda T_{TG} & T_{CC} & 1 - \lambda(T_{TG} + T_{AG}) - T_{CC} & \lambda T_{AG} \\ \lambda T_{TC} & 1 - \lambda(T_{TC} + T_{AC}) - T_{CC} & T_{CC} & \lambda T_{AC} \\ 1 - T_{AA} - T_{TG} - T_{TC} & T_{TC} & T_{TG} & T_{AA} \end{pmatrix}$$

where

$$\lambda = \frac{\pi(A)}{\pi(C)} \tag{3.50}$$

We remind that

$$\lambda = \frac{\pi(T)}{\pi(C)} = \frac{\pi(A)}{\pi(G)} = \frac{\pi(T)}{\pi(G)} = \frac{\pi(A)}{\pi(C)} = \frac{\pi(T)}{\pi(G)} \tag{3.51}$$

3.2 Concatenation of stationary processes

In this section we describe a very particular case of non stationary processes.

We begin by introducing the model and defining cylinders, probabilities and $\alpha-$Chargaff processes. Secondly, we study properties of $\alpha-$Chargaff processes and conditions on probability vectors that enable second parity rule, similarly to how we did for simple stationary processes.

This model is a generalization of the one proposed in [13] and described in Chapter 2.

3.2.1 The concatenation

We define the model $\mathcal{Z}$ as the composition of two different processes $\mathcal{X}, \mathcal{Y}$, that we will assume stationary.

The two processes generate sequences $(x_i)_1^N, (y_i)_1^M$, that grows in opposite directions and lie in different strands. Remaining sequences $(x_i)_{-1}^{-M}, (y_i)_{-1}^{-N}$ are constructed as complements of the corresponding successions.

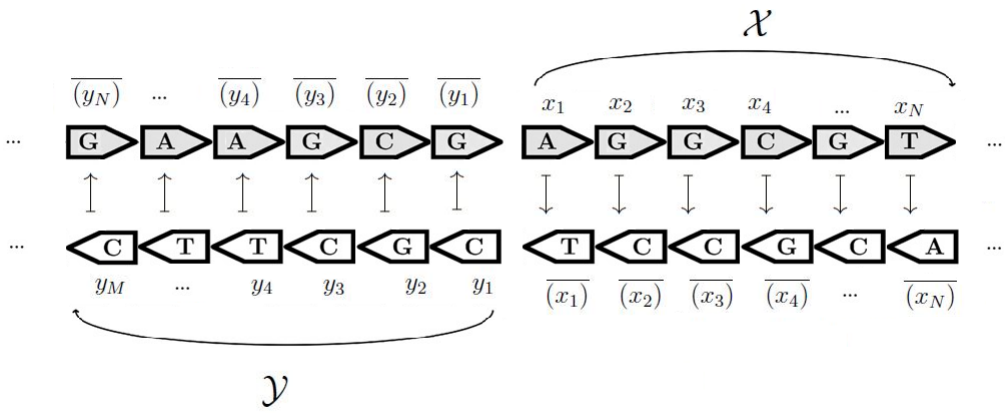

Figure 3.2: The model is made of two different processes $\mathcal{X}, \mathcal{Y}$ that are supposed stationary. The sequence produced by the process is the concatenation of the realizations of the two processes and their complements. The arrows show the directions of progression of main sub-strings $(x_i)_1^N$ and $(y_i)_1^M$.

Then, $\mathcal{Z}$ produces two strands. However, since they are complement of each other, it is sufficient to study one. Indeed, the other acts similarly for complementarity.

From now on, we will work with the upper strand. We denote with

$$(z_i)_{i=-M}^{N} = (y_i)_{i=1}^{M} \sqcup (x_i)_{i=1}^{N} \tag{3.52}$$

the concatenation of two sequences constituent the first strand expressed as function of succession generated by process $\mathcal{X}, \mathcal{Y}$ Plus, since

$$z_i = \bar{y}_{-i}, \quad -M \leq i \leq -1 \tag{3.53}$$

the notation 3.52 can be written also as

$$(z_i)_{i=-M}^{N} = (\bar{y}_i)_{i=-M}^{-1} \sqcup (x_i)_{i=1}^{N} \tag{3.54}$$

The latter notation is the one we mostly use during this work, as it express the final string as union of the realization of original processes $\mathcal{X}, \mathcal{Y}$.

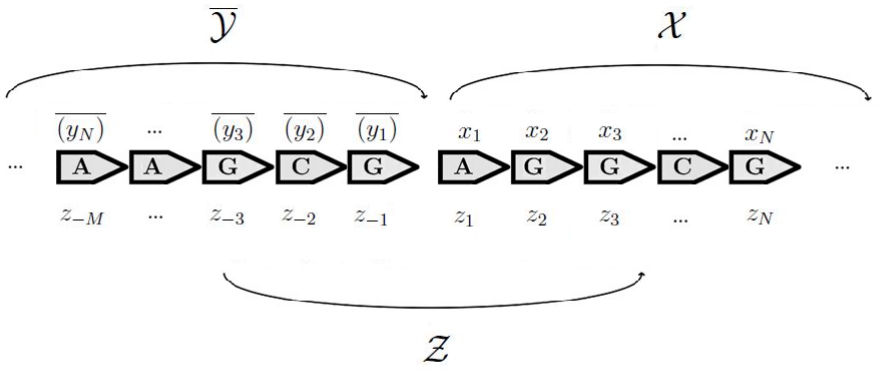

Figure 3.3: If we look at one string, for example the upper one, we can consider it as the resulting succession $(z_i)_{i\in\mathbb{Z}}$.

Def 11. Given a word $\omega = \omega_{-m} \dots \omega_n$ in $\mathcal{I}^{m+n}$, $n, m \in \mathbb{N}$, we call *cylinder* a subset of $\sum_{\mathcal{I}}$ of the form

$$\mathcal{S}_{-m,n}(\omega_{-m} \dots \omega_n) = \{(z_i)_{i\in\mathbb{Z}\setminus\{0\}},\ z_i; \in \mathcal{I}\ s.t.\ z_i = \omega_i\ \forall -m \leq i \leq n\} \tag{3.55}$$

We remark that sequences in $\sum_{\mathcal{I}}$ do not contain 0−term. This is due to the fact that each sequence can result from the union of two sequences in $\sum_{\mathcal{I}}{}^{+}$. In fact, from how we constructed the model, each cylinder of the space $\sum_{\mathcal{I}}$ can be written as the concatenation of two cylinders in $\sum_{\mathcal{I}}{}^{+}$.

Given a word $\omega = \omega_{-m} \dots \omega_n$ in $\mathcal{I}^{m+n}$, it can be spliced in two words $\omega^m = \omega_{-m} \dots \omega_{-1}$ and $\omega^n = \omega_1 \dots \omega_n$ in $\sum_{\mathcal{I}}{}^{m}$ and $\sum_{\mathcal{I}}{}^{n}$ respectively.

Then, we have that

$$\omega = \omega_{-m} \dots \omega_{-1}\omega_1 \dots \omega_n$$

We say that ω is the concatenation of the two words $\omega^n \in \mathcal{I}^n, \omega^m \in \mathcal{I}^m$ and we denote it with

$$\omega = \omega^m \sqcup \omega^n \tag{3.56}$$

Remark 3.9. *Note that words ω are centered. Indeed, there is not a unique way of splicing a word ω in two words ω^n, ω^m. Plus, the division affects the cylinders $S_n(\omega^n), S_m(\omega^m)$.*

We can now describe formally the process of the new model.

Def 12 (*Concatenation* of processes)**.** Let $\mathcal{X}, \mathcal{Y}$ be two stationary processes on the probability spaces $(\sum_{\mathcal{I}}{}^+, \mathcal{F}, P_X)$ and $(\sum_{\mathcal{I}}{}^+, \mathcal{F}, P_Y)$ respectively.

The *concatenation of* $\mathcal{X}, \mathcal{Y}$ is stochastic process $\mathcal{Z}$ defined on the probability space $(\sum_{\mathcal{I}} \mathcal{F}, P_{\mathcal{Z}})$, which realizations are all possible sequences $s = (z_i)_{i \in \mathbb{Z} \setminus \{0\}}$ of the form

$$(z_i)_{i \in \mathbb{Z} \setminus \{0\}} = (\bar{y}_{-i})_{i=1}^{\infty} \sqcup (x_i)_{i=1}^{\infty} \tag{3.57}$$

where $(x_i)_{i=1}^{\infty}, (y_i)_{i=1}^{\infty} \in \sum_{\mathcal{I}}{}^+$ and which probability measure $P_{\mathcal{Z}}$ is defined as the product of the probabilities $P_{\mathcal{X}}, P_{\mathcal{Y}}$ as defined below

$$P_{\mathcal{Z}}(\mathcal{S}_{-m,n}(\omega)) = P_{\mathcal{X}}(\mathcal{S}_n(\omega_{\mathcal{X}})) P_{\mathcal{Y}}(\mathcal{S}_m(\bar{\omega}_{\mathcal{Y}})), \quad \forall \omega \in \mathcal{I}^{n+m} \tag{3.58}$$

Calling $s_{\mathcal{X}}, s_{\mathcal{Y}}$ the sequences generated by the stationary processes $\mathcal{X}, \mathcal{Y}$ respectively, we denote the sequence $s \in \sum_{\mathcal{I}}$ as follow

$$s = s_{\mathcal{Y}} \sqcup s_{\mathcal{X}} \tag{3.59}$$

Notation $s_{\mathcal{X}}$ and $s_{\mathcal{Y}}$ will be used to denote also finite portion of the realization of the processes. It will be specified to avoid ambiguity.

Similarly to the simple stochastic case, the $\sigma-$algebra $\mathcal{F}$ of the probability space $(\sum_{\mathcal{I}}, \mathcal{F}, P_{\mathcal{Z}})$ is the $\sigma-$algebra generated by all possible cylinders $\mathcal{S}_{-m,n}(\omega)$, with $\omega \in \mathcal{I}^{m+n}$.

Although $\mathcal{X}, \mathcal{Y}$ are both stationary, the process $\mathcal{Z}$ is not, so that we have to give a completely different definition for the new process $\mathcal{Z}$ to be Chargaff.

Since a stochastic process is defined by the probabilities of all cylinders, we first need to give the probabilities on every cylinder. The probability of a cylinder generated by $\mathcal{Z}$ will depend both on $\mathcal{X}$ and on $\mathcal{Y}$.

From now on, $\mathcal{X}, \mathcal{Y}$ are the two stationary processes generating $s_{\mathcal{X}}$ and $s_{\mathcal{Y}}$ with probabilities $P_{\mathcal{X}}, P_{\mathcal{Y}}$. The final string s is the union of the infinite strings $s_{\mathcal{X}}$ and $s_{\mathcal{Y}}$, and is a bi-infinite sequence in the alphabet $\mathcal{I}$.

Example 3.10. *The following string*

$$s = \ldots x_{-m-2}x_{-m-1}\omega_{-m}\ldots\omega_{-1}\omega_1\ldots\omega_n x_{n+1}x_{n+2}\ldots$$

belongs to the cylinder $C_{-m,n}(\omega_{-m}\ldots\omega_{-1}\omega_1\ldots\omega_n)$. *and results as the union of* s_2 *and* s_1 *where*

$$s_{\mathcal{X}} = \omega_1\ldots\omega_n x_{n+1}\ldots$$

$$s_{\mathcal{Y}} = \omega_{-1}\ldots\omega_{-m}x_{-m-1}\ldots$$

It is easy to see that the $P_{\mathcal{Z}}$ defined on the measured space $\sum_{\mathcal{I}}$ as the product of probabilities $P_{\mathcal{X}}, P_{\mathcal{Y}}$ is a probability.

Indeed, we have to prove that

1. $P(\mathcal{S}_k(\omega)) \geq 0 \;\; \forall\omega \in \mathcal{I}^k$
2. $P(\sum_{\mathcal{I}}) = 1$

We remind that for every cylinder $\mathcal{S} \in \Omega$ there exist two cylinders $\mathcal{S}_{\mathcal{X}}, \mathcal{S}_{\mathcal{Y}} \in \sum_{\mathcal{I}^+}$ such that

$$P(\mathcal{S}) = P_{\mathcal{X}}(\mathcal{S}_{\mathcal{X}})P_{\mathcal{Y}}(\mathcal{S}_{\mathcal{Y}}) \tag{3.60}$$

where $P_{\mathcal{X}}, P_{\mathcal{Y}}$ are the probabilities in ${\sum_{\mathcal{I}}}^+$. Since $P_{\mathcal{X}}(\mathcal{S}_{\mathcal{X}}) \geq 0, P_{\mathcal{Y}}(\mathcal{S}_{\mathcal{X}} \geq 0$ since $P_{\mathcal{X}}$ and $P_{\mathcal{Y}}$ are probabilities, we have that $P_{\mathcal{Z}}(C) \geq 0$ and (1) is proved.

In order to prove (2), we notice that we can write the total space $\sum\mathcal{I}$ as the union of the total spaces of processes $\mathcal{X}, \mathcal{Y}$, both equal to ${\sum_{\mathcal{I}}}^+$. Indeed, as we saw above, every bi-infinite sequence can be written as the union of two infinite sequences. Then, since $\mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}}$ are processes of probabilities $P_{\mathcal{X}}, P_{\mathcal{Y}}$, it results that $P(\sum_{\mathcal{I}}^+) = 1$ and so (2) is satisfied.

3.2.2 $\alpha-$Chargaff processes

In this section we give definition of Chargaff processes that take under consideration the window frame used to read the final string s to check its compliance with the rule. In fact, being $\mathcal{Z}$ a non-stationary process, definitions given in the previous chapter are not still valid. While for the stationary process the only significant element for counting the number of a given word ω was its length k, in the case of concatenation of stationary processes not only it is necessary to consider k but also we need to know in what proportion the sub-strings that are read. For exemple, counting the number of words ω scanning the sequence $(z_i)_{i\in\infty}$ in equal portions of $(x_i))_{i\in\infty}$ and $(\bar{y}_i))_{i\in\infty}$ is not the same as considering the sequence generated by $\mathcal{X}$ in double the length of that generated by $\mathcal{Y}$.

In order to analyze the conditions on the processes that enable Chargaff's second parity rule, we use a new probability P_α, that allows us to relax the condition on the words ω. In fact we define the probability as function of cylinders in $\sum_{\mathcal{I}}^{+}$.

Def 13 ($P_\alpha-$)**.** Let $\omega = \omega_1 \dots \omega_k$ with $\omega_i \in \mathcal{I}, i = 1, \dots k$. Given $\alpha \in [0,1]$, we use the following probability and we indicate it with P_α

$$P_\alpha(\omega) = \alpha P_{\mathcal{X}}(S_k(\omega)) + (1-\alpha) P_{\mathcal{Y}}(S_k(\bar{\omega})) \tag{3.61}$$

with $P_{\mathcal{X}}, P_{\mathcal{Y}}$ probabilities of the processes $\mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}}$ respectively, and $S_k(\omega) \in \sum_{\mathcal{I}}^{+}$.

Since process $\mathcal{P}_{\bar{\mathcal{Y}}}$ is constructed as complement of process $\mathcal{P}_{\mathcal{Y}}$, we have that

$$P_{\mathcal{Y}}(S_k(\bar{\omega})) = P_{\bar{\mathcal{Y}}}(S_k(\omega)) \tag{3.62}$$

so that it is equivalent defining P_α for process $\mathcal{Y}$ or process $\bar{\mathcal{Y}}$.

In particular, definition is equivalent 3.61 to

$$P_\alpha(\omega) = \alpha P_{\mathcal{X}}(S_k(\omega)) + (1-\alpha) P_{\bar{\mathcal{Y}}}(S_k(\omega)) \tag{3.63}$$

We will use both equations equally during this work.

Def 14 ($\alpha-$Chargaff process)**.** Let $\mathcal{Z}$ be the concatenation of stationary processes $\mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}}$ defined on spaces $(\sum_{\mathcal{I}}^{+}, P_{\mathcal{X}})$,$(\sum_{\mathcal{I}}^{+}, P_{\mathcal{Y}})$.

Given $\alpha \in [0,1]$, we say that $\mathcal{Z}$ is a *$\alpha-$Chargaff process* if and only if, $\forall k \in \mathbb{N}$, for every word $\omega = \omega_1 \dots \omega_k$ in $\mathcal{I}$

$$P_\alpha(\omega) = P_\alpha(\bar{\omega}) \tag{3.64}$$

It will be necessary to use a more relaxed definition of $\alpha-$Chargaff process. In particular, if (3.64) holds only for the words of a given length k, we say the process is $k-\alpha-$Chargaff.

Def 15 (Birkhoff sums)**.** Let $\mathcal{Z}$ be the concatenation process generated by $\mathcal{P}_{\mathcal{X}}, \mathcal{P}_{\mathcal{Y}}$ and let ω be a word of length k. Let $(z_i)_{i\in\mathbb{Z}}$ be a realization of $\mathcal{Z}$. Given $m, n \in \mathbb{N}$, we call *(-m,n)-Birkhoff sums* and we indicate it as $B_{-m,n}(\omega)$.

$$B_{-m,n}(\omega) = \sum_{j=-m}^{n} \chi_j(\omega) \tag{3.65}$$

where χ *chi*-function defined as

$$\chi_j(\omega) = \begin{cases} 1, & \text{if } z_j \dots z_{j+k-1} = \omega_1 \dots \omega_k \\ 0, & \text{otherwise} \end{cases}$$

Note that definition of $B_{-m,n}(\omega)$ depends not only on word $\omega \in \mathcal{I}^k$, but also on the proportion $n : m$ of scanning windows reading successions (x_i) and $(\bar{y}_{-i})$.

Theorem 3.11 (Ergodic therorem for concatenation of processes)**.** *Let $\mathcal{Z}$ be concatenation process generated by the stationary processes $\mathcal{P}_\mathcal{X}, \mathcal{P}_\mathcal{Y}$, associated with probability P_α, with $\alpha \in [0,1]$.*

Then, $\forall k \in \mathbb{N}$, for every $\omega \in \mathcal{I}^k$

$$\lim_{n\to\infty} \frac{B_{-m_l,n_l}(\omega)}{m_l + n_l} = P_\alpha(\omega) \tag{3.66}$$

with $(m_l)_{l\in\mathbb{N}}, (n_l)_{l\in\mathbb{N}}$ successions s.t.

$$\frac{n_l}{n_l + m_l} \longrightarrow_{l\to\infty} \alpha$$

Proof. Let be ω a word in $\mathcal{I}^k$ and let $(z_i) = (\bar{y}_i) \sqcup (x_i)$ be a realization of the process $\mathcal{Z}$. Given $(m_l)_{l\in\mathbb{N}}, (n_l)_{l\in\mathbb{N}}$ increasing successions s.t. $\frac{n_l}{n_l+m_l} \longrightarrow \alpha$ for $l \to \infty$ one can express the Birkhoff sums B_{-m_l,n_l} among s as

$$B_{-m_l,n_l}(\omega) = \sum_{j=-m_l}^{n_l-k} \chi_j(\omega) = \sum_{j=-m_l}^{-k} \chi_j(\omega) + \sum_{j=-k+1}^{-1} \chi_j(\omega) + \sum_{j=1}^{n_l-k} \chi_j(\omega) \tag{3.67}$$

In other words, the total number of words ω in the string $(z_i)_{i=-m_l}^{n_l}$ results as the sum of the number of ω in $(y_i)_{i=-m_l}^{1}$ and $(x_i)_{i=1}^{n_l}$ plus the number of those generated by the overlap of the union of the two substrings.

As usual, we denote with $s_\mathcal{X}$ and $s_{\bar{\mathcal{Y}}}$ the two substrings $(x_i)_{i=1}^{n_l}, (y_i)_{i=-m_l}^{-1}$ respectively.

Then, the first and third addends in equation (3.67) can be written as function of the empirical frequencies $f_{s_\mathcal{X}}(\omega)$ and $f_{s_{\bar{\mathcal{Y}}}}(\omega)$ calculated on the portions of the two realizations

$$\sum_{j=-m_l}^{-k} \chi_j(\omega) = f_{s_\mathcal{X}} m_l$$

$$\sum_{j=1}^{n_l-k} \chi_j(\omega) = f_{s_{\bar{\mathcal{Y}}}} n_l$$

Thus, eq.3.67 became

$$B_{-m_l,n_l}(\omega) = m_l f_{s_{\bar{\mathcal{Y}}}}(\omega) + \sum_{j=-k+1}^{-1} \chi_j(\omega) + n_l f_{s_\mathcal{X}} \tag{3.68}$$

and diving by n_l we obtain

$$\frac{B_{-m_l,n_l}(\omega)}{n_l+m_l} = \frac{m_l}{n_l+m_l} f_{s_{\bar{\mathcal{Y}}}}(\omega) + \frac{\sum_{j=-k+1}^{-1} \chi_j(\omega)}{n_i+m_i} + \frac{n_i}{n_i+m_i} f_{s_{\mathcal{X}}}(\omega) \tag{3.69}$$

that is

$$\frac{B_{-m_l,n_l}(\omega)}{n_l+m_l} = (1-\alpha) f_{s_{\bar{\mathcal{Y}}}}(\omega) + \frac{\sum_{j=-k+1}^{-1} \chi_j(\omega)}{n_l+m_l} + \alpha f_{s_{\mathcal{X}}}(\omega) \tag{3.70}$$

Moreover, the number of words ω counted in the overlap are at most $k-1$.

Thus, we can bound $\frac{B_{-m_l,n_l}(\omega)}{n_l+m_l}$ above and write

$$\frac{B_{-m_l,n_l}(\omega)}{n_l+m_l} \leq (1-\alpha) f_{s_{\bar{\mathcal{Y}}}}(\omega) + \frac{k-1}{n_l+m_l} + \alpha f_{s_{\mathcal{X}}}(\omega) \tag{3.71}$$

Since the processes generating the two substrings are stationary, we can use the Ergodic Theorem to prove that each frequency tents to the measure as the length of $s_{\mathcal{X}}$, $s_{\bar{\mathcal{Y}}}$ tents to infinity.

Formally we have that

$$f_{s_{\mathcal{X}}}(\omega) \xrightarrow[l\to\infty]{P}_{\mathcal{X}} (\omega)$$

where $P_{\mathcal{X}}(\omega)$ is the probability of seeing the word ω among $s_{\mathcal{X}}$ according to the process $\mathcal{P}_{\mathcal{X}}$.

In fact, we have that m_l tents to ∞ when $l \to \infty$, so that frequencies of ω tents to probability of word ω to appear among $(x_i)_{i\in\mathbb{N}}$. Similarly

$$f_{s_{\bar{\mathcal{Y}}}}(\omega) \xrightarrow[l\to\infty]{} P_{\bar{\mathcal{Y}}}(\omega)$$

Hence, for l that tends to infinity we got

$$\lim_{l\longrightarrow\infty} \frac{B_{-m_l,n_l}(\omega)}{n_l+m_l} = (1-\alpha) P_{\bar{\mathcal{Y}}}(\omega) + \alpha P_{\mathcal{X}}(\omega)$$

that ends the proof. □

3.2.3 Bernoulli scheme

In the previous section we analyzed some of the main properties of a Chargaff process. In particular, we saw that if the Chargaff's second parity rule holds for every word of length k, then it holds for every word of length $k-1$. This means that it is sufficient for a string to satisfy Chargaff for a k word to

ensure the validity of Chargaff second parity rule for every word at most of length k. This implication is valid whether the process is stationary or not.

Here we study the reverse implication. In other words, we aim to analyze the conditions that ensure the validity of Chargaff second parity rule for k-words, knowing that it holds for every $(k-1)$-word.

Let $\pi_{\mathcal{X}}, \pi_{\mathcal{Y}}$ the probability vectors defining Bernoulli-like processes $\mathcal{X}, \mathcal{Y}$ respectively.

By the definition (14), a process $\mathcal{Z}$ composition of two stationary processes $\mathcal{X}, \mathcal{Y}$ is $\alpha-$Chargaff if and only if

$$\alpha P_{\mathcal{X}}(S_k(\omega)) + (1-\alpha)P_{\mathcal{Y}}(S_k(\bar{\omega})) = \alpha P_{\mathcal{X}}(S_k(\bar{\omega})) + (1-\alpha)P_{\mathcal{Y}}(S_k(\omega)) \quad (3.72)$$

for every word ω in $\{A, C, G, T\}$ of every length k.

We remind that

$$P_{\mathcal{X}}(\omega) = P_{\mathcal{X}}(x_j \dots x_{j+k} = \omega_1 \dots \omega_k), \quad x_j \dots x_{j+k} = (s_{\mathcal{X}})_j^{j+k}$$

Plus, since process $P_{\mathcal{X}}$ is stationary

$$P_{\mathcal{X}}(\omega) = P_{\mathcal{X}}(x_1 \dots x_k = \omega_1 \dots \omega_k), \quad x_1 \dots x_k$$

Moreover, we recall that for how we construct the model, the probability $P_{\bar{\mathcal{Y}}}$ of process $\bar{\mathcal{Y}}$ is defined from probability $P_{\mathcal{Y}}$ for complementarity. The probability vector corresponding to process $\bar{\mathcal{Y}}$ will then be

$$\pi_{\bar{\mathcal{Y}}}(x) = \pi_{\mathcal{Y}}(\bar{x}), \quad \forall x \in \mathcal{I} \quad (3.73)$$

As we saw earlier, the condition that enable $\mathcal{Z}$ to be $\alpha-$Chargaff can be written as

$$\alpha P_{\mathcal{X}}(S_k(\omega)) + (1-\alpha)P_{\bar{\mathcal{Y}}}(S_k(\omega)) = \alpha P_{\mathcal{X}}(S_k(\bar{\omega})) + (1-\alpha)P_{\bar{\mathcal{Y}}}(S_k(\bar{\omega})) \quad (3.74)$$

In particular, because we assume $\mathcal{X}, \mathcal{Y}$ Bernoulli-like, we have that

$$\alpha \prod_{x \in \omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x \in \omega} \pi_{\bar{\mathcal{Y}}}(x) = \alpha \prod_{\bar{x} \in \bar{\omega}} \pi_{\mathcal{X}}(\bar{x}) + (1-\alpha) \prod_{\bar{x} \in \bar{\omega}} \pi_{\bar{\mathcal{Y}}}(\bar{x}) \quad (3.75)$$

Let us considerate for example $k = 1$. The conditions on the word ω become conditions on $\omega = x \in \mathcal{I}$[1]. More formally

$$\alpha P_{\mathcal{X}}(x) + (1-\alpha)P_{\mathcal{Y}}(\bar{x}) = \alpha P_{\mathcal{X}}(\bar{x}) + (1-\alpha)P_{\mathcal{Y}}(x) \quad (3.76)$$

[1]Note that here x denote a character, that can indifferently belong to both sub-strings $(x_i)_{i \in \mathbb{N}}$ and $(y_i)_{i \in \mathbb{N}}$. Notation x was merely assumed for semplicity

Since $\mathcal{X}, \mathcal{Y}$ are Bernoulli-like processes of probability vectors $\pi_{\mathcal{X}}, \pi_{\mathcal{Y}}$, (3.76) it is equivalent to

$$\alpha\pi_{\mathcal{X}}(x) + (1-\alpha)\pi_{\mathcal{Y}}(\bar{x}) = \alpha\pi_{\mathcal{X}}(\bar{x}) + (1-\alpha)\pi_{\mathcal{Y}}(x) \tag{3.77}$$

or equally

$$\alpha\pi_{\mathcal{X}}(x) + (1-\alpha)\pi_{\bar{\mathcal{Y}}}(x) = \alpha\pi_{\mathcal{X}}(\bar{x}) + (1-\alpha)\pi_{\bar{\mathcal{Y}}}(\bar{x}) \tag{3.78}$$

First we will study the conditions on a concatenation of stationary processes, $\mathcal{Z}$ composed of two Bernoulli processes $\mathcal{X}$,$\mathcal{Y}$.

The probabilities of the $1-$words are totally independent, so that a $k-$word $\omega = \omega_1 \dots \omega_k$ depends only on the probabilities of the single characters $\omega_i, i = 1, \dots k$.

We remind that a Bernoulli-like process is given by a probability vector.

Let $\mathcal{X}, \mathcal{Y}$ be two bernoulli processes defined by two probability vectors $\pi_{\mathcal{X}} = (\pi_{\mathcal{X}}(A), \pi_{\mathcal{X}}(C), \pi_{\mathcal{X}}(G), \pi_{\mathcal{X}}(T)), \quad \pi_{\mathcal{Y}} = (\pi_{\mathcal{Y}}(A), \pi_{\mathcal{Y}}(C), \pi_{\mathcal{Y}}(G), \pi_{\mathcal{Y}}(T))$. Let also be $M, N \in \mathbb{N}$ the lengths of the realizations $s_{\mathcal{X}}, s_{\mathcal{Y}}$ of $\mathcal{X}$ and $\mathcal{Y}$ respectively.

Our goal is to find the restrictions on $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ such that 3.64 holds. In other words, being s a string generated by the process $\mathcal{Z}$, we want to study under which conditions the number of a word ω in a string s generated by $\mathcal{Z}$ is the same of the number of the reverse complement of ω.

Simple cases of Bernoulli Chargaff-processes

We aim to study the conditions on $\pi_{\mathcal{X}}, \pi_{\mathcal{Y}}$ (or similarly $\pi_{\bar{\mathcal{Y}}}$) that make $\mathcal{Z}$ a $\alpha-$Chargaff process, that is such that $P_\alpha(\omega) = P_\alpha(\bar{\omega})$ for every word ω of every length k. Let be $k = 1$. We ask the equality (3.77) is sufficient to guarantee the process to be $\alpha-$Chargaff.

Firstly, we notice that there are some particular probability vectors that satisfy (3.77). For example, let be $\pi_{\mathcal{X}} = \pi_{\bar{\mathcal{Y}}} = \pi$ and consequently $\mathcal{X} = \mathcal{Y}$. If we suppose $\mathcal{X}$ $1-$Chargaff, i.e. $\pi(x) = \pi(\bar{x})\ \forall x \in \mathcal{I}$, we have that $\mathcal{Z}$ is $\alpha-$Chargafffollowing

$$\begin{aligned} \alpha\pi_{\mathcal{X}}(x) + (1-\alpha)\pi_{\bar{\mathcal{Y}}}(x) &= \alpha\pi_{\mathcal{X}}(\bar{x}) + (1-\alpha)\pi_{\mathcal{X}}(\bar{x}) = \\ &= \alpha\pi(\bar{x}) + (1-\alpha)\pi(\bar{x}) \\ &= \alpha\pi_{\mathcal{X}}(\bar{x}) + (1-\alpha)\pi_{\bar{\mathcal{Y}}}(\bar{x}) \end{aligned} \tag{3.79}$$

The process $\mathcal{Z}$ is naturally $1-\alpha-$Chargaff also if the processes generating it are both self-Chargaff. In other words, if

$$\pi_{\mathcal{X}}(x) = \pi_{\mathcal{X}}(\bar{x}), \forall x \in \{A, C, G, T\}$$

$$\pi_{\mathcal{Y}}(x) = \pi_{\mathcal{Y}}(\bar{x}), \forall x \in \{A, C, G, T\}$$

then (3.77) are true.

The last particular couple of vectors which enable (3.77) are

$$\pi_{\mathcal{X}}(x) = \pi_{\bar{\mathcal{Y}}}(\bar{x}), \forall x \in \{A, C, G, T\} \tag{3.80}$$

In fact, replacing (3.80) two times in (3.77) we obtain

$$\alpha\pi_{\mathcal{X}}(x) + (1-\alpha)\pi_{\bar{\mathcal{Y}}}(x) \;\; = \;\; \alpha\pi_{\bar{\mathcal{Y}}}(\bar{x}) + (1-\alpha)\pi_{\mathcal{X}}(\bar{x}) \tag{3.81}$$

The equality holds for $\pi_{\mathcal{X}} = \pi_{\bar{\mathcal{Y}}}$ self-Chargaff, which is the case above, or $\alpha = \frac{1}{2}$.

As we will see in the next Proposition, these are the only cases which allow the implication

$$\mathcal{C}_1^{II} \Longrightarrow \mathcal{C}_k^{II}, \;\; \forall k \in \mathbb{N}$$

for $\alpha-$Chargaff processes.

Proposition 3.12. *Let $\mathcal{Z}$ be a process concatenation of two Bernoulli processes $\mathcal{P}_1, \mathcal{P}_2$ and let $\pi_{\mathcal{X}}, \pi_{\mathcal{Y}}$ be the respective probability vectors.*

If

1. $\pi_{\mathcal{X}} = \pi_{\bar{\mathcal{Y}}}$, *and* $\mathcal{X}, \mathcal{Y}$ *are* $1-$*Chargaff,* $\alpha \in [0, 1]$
2. $\pi_{\mathcal{X}} \neq \pi_{\bar{\mathcal{Y}}}$, *and* $\mathcal{X}, \bar{\mathcal{Y}}$ *are* $1-$*Chargaff,* $\alpha \in [0, 1]$
3. $\pi_{\mathcal{X}} = \pi_{\mathcal{Y}}, \;\; \alpha = \frac{1}{2}$

then the process $\mathcal{Z}$ is $\alpha-$Chargaff. Less formally, Chargaff parity rule holds for every word ω in $\mathcal{I} = \{A, C, G, T\}$ of length $k, \forall k \in \mathbb{N}$.

Proof. We have to prove that, for every k

$$P_\alpha(\omega) = P_\alpha(\bar{\omega}), \;\; \forall \omega \in \mathcal{I}^k$$

from the definition of $\alpha-$Chargaff.

Because $\mathcal{X}, \mathcal{Y}$ are Bernoulli-like processes, this is equivalent to prove

$$\alpha \prod_{x \in \omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x \in \omega} \pi_{\bar{\mathcal{Y}}}(x) = \alpha \prod_{\bar{x} \in \bar{\omega}} \pi_{\mathcal{X}}(\bar{x}) + (1-\alpha) \prod_{\bar{x} \in \bar{\omega}} \pi_{\bar{\mathcal{Y}}}(\bar{x}) \tag{3.82}$$

(case (1)) Consider now the first couple of vectors, $\pi_{\mathcal{X}} = \pi_{\mathcal{Y}}$. Because both processes are Chargaff, $\pi_{\mathcal{X}}(x) = \pi_{\mathcal{X}}(\bar{x}), \;\; \pi_{\bar{\mathcal{Y}}}(x) = \pi_{\bar{\mathcal{X}}}(\bar{x}) \;\; \forall x \in \mathcal{I}$.

Then we have that (3.82) is the same as

$$\alpha \prod_{x \in \omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x \in \omega} \pi_{\mathcal{X}}(x) = \alpha \prod_{\bar{x} \in \bar{\omega}} \pi_{\mathcal{X}}(\bar{x}) + (1-\alpha) \prod_{\bar{x} \in \bar{\omega}} \pi_{\mathcal{X}}(\bar{x}) \tag{3.83}$$

As the process $\mathcal{X}$ is $1-$Chargaff, we have

$$\alpha \prod_{x\in\omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x\in\omega} \pi_{\mathcal{X}}(x) = \alpha \prod_{x\in\omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x\in\omega} \pi_{\mathcal{X}}(x) \quad (3.84)$$

that prove the theorem for the first case.

(case (2)) Let now be $\pi_{\mathcal{X}}, \pi_{\bar{\mathcal{Y}}}$ such that processes $\mathcal{X}, \bar{\mathcal{Y}}$ are $1-$Chargaff, that is

$$\begin{aligned} \pi_{\mathcal{X}}(x) &= \pi_{\mathcal{X}}(\bar{x}) \\ \pi_{\bar{\mathcal{Y}}}(x) &= \pi_{\bar{\mathcal{Y}}}(\bar{x}) \end{aligned} \quad (3.85)$$

for every $x \in \mathcal{I}$. By substitution in (3.82) we obtain

$$\alpha \prod_{x\in\omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x\in\omega} \pi_{\bar{\mathcal{Y}}}(x) = \alpha \prod_{x\in\omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{x\in\omega} \pi_{\bar{\mathcal{Y}}}(x) \quad (3.86)$$

so that the equality is satisfied.

(case 3) We end the proof with the third case. Let $\pi_{\mathcal{X}}, \pi_{\bar{\mathcal{Y}}}$ be the probability vectors of $\mathcal{X}, \mathcal{Y}$ such that $\pi_{\mathcal{X}} = \pi_{\mathcal{Y}}$. It has to be

$$\pi_{\mathcal{X}}(x) = \pi_{\bar{\mathcal{Y}}}(\bar{x}), \quad \forall x \in \mathcal{I} \quad (3.87)$$

Hence, we have that (3.82) becomes

$$\alpha \prod_{x\in\omega} \pi_{\mathcal{X}}(x) + (1-\alpha) \prod_{\bar{x}\in bar\omega} \pi_{\bar{\mathcal{Y}}}(\bar{x}) = \alpha \prod_{x\in\omega} \pi_{\bar{\mathcal{Y}}}(x) + (1-\alpha) \prod_{x\in\omega} \pi_{\mathcal{X}}(x) \quad (3.88)$$

The equality holds if and only if $\alpha = \frac{1}{2}$, and it ends the proof. □

General case

In the previous section we analyzed the conditions on three simple cases of processes $\mathcal{X}, \mathcal{Y}$ that generate a concatenation process $\mathcal{Z}$ $\alpha-$Chargaff. Here we study the general case of a process $\mathcal{Z}$ that is $\alpha-$Chargaff for the letters, but not for words ω of length k.

For simplicity, from now on we suppose $\alpha = 1/2$. We will see that in general it is not true that a $\alpha - \mathcal{C}_1^{II}$ is consequently $\alpha - \mathcal{C}_k^{II}$ for every $k \in \mathbb{N}$.

First of all, as the vectors $\pi_{\mathcal{X}}, \pi_{\mathcal{Y}}$ defining the processes are probability vectors, they have to sum to 1. Thus it must be

$$\begin{aligned} \pi_{\mathcal{X}}(A) + \pi_{\mathcal{X}}(C) + \pi_{\mathcal{X}}(G) + \pi_{\mathcal{X}}(T) &= 1 \\ \pi_{\mathcal{Y}}(A) + \pi_{\mathcal{Y}}(C) + \pi_{\mathcal{Y}}(G) + \pi_{\mathcal{Y}}(T) &= 1 \end{aligned} \quad (3.89)$$

Also, as we are supposing that the process $\mathcal{Z}$ is $\alpha-$Chargaff for $k=1$, the equations (3.77) must be satisfied. In other words, it has to be

$$\begin{aligned}\pi_{\mathcal{X}}(A)-\pi_{\mathcal{Y}}(A)&=\pi_{\mathcal{X}}(T)-\pi_{\mathcal{Y}}(T)\\ \pi_{\mathcal{X}}(C)+\pi_{\mathcal{Y}}(C)&=\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(G)\end{aligned} \tag{3.90}$$

Hence we are looking for $\pi_{\mathcal{X}},\pi_{\mathcal{Y}}$ such that (3.89) and (3.90) are both true. In other words, $\pi_{\mathcal{X}},\pi_{\mathcal{Y}}$ are solutions of the system:

$$\begin{cases}\pi_{\mathcal{X}}(A)+\pi_{\mathcal{X}}(C)+\pi_{\mathcal{X}}(G)+\pi_{\mathcal{X}}(T)=1\\ \pi_{\mathcal{Y}}(A)+\pi_{\mathcal{Y}}(C)+\pi_{\mathcal{Y}}(G)+\pi_{\mathcal{Y}}(T)=1\\ \pi_{\mathcal{X}}(A)-\pi_{\mathcal{Y}}(A)=\pi_{\mathcal{X}}(T)-\pi_{\mathcal{Y}}(T)\\ \pi_{\mathcal{X}}(C)+\pi_{\mathcal{Y}}(C)=\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(G)\end{cases} \tag{3.91}$$

Let us fix $\pi_{\mathcal{X}}(G),\pi_{\mathcal{Y}}(A),\pi_{\mathcal{Y}}(C),\pi_{\mathcal{Y}}(G)$, we write the other components as function of them

$$\begin{cases}\pi_{\mathcal{X}}(T)=1-\pi_{\mathcal{X}}(A)-\pi_{\mathcal{X}}(C)-\pi_{\mathcal{X}}(G)\\ \pi_{\mathcal{Y}}(T)=1-\pi_{\mathcal{Y}}(A)-\pi_{\mathcal{Y}}(C)-\pi_{\mathcal{Y}}(G)\\ \pi_{\mathcal{X}}(A)=\pi_{\mathcal{Y}}(A)+\pi_{\mathcal{Y}}(G)-\pi_{\mathcal{X}}(G)\\ \pi_{\mathcal{X}}(C)=\pi_{\mathcal{Y}}(C)+\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(G)\end{cases} \tag{3.92}$$

Replacing $\pi_{\mathcal{X}}(G),\pi_{\mathcal{X}}(G)$ and $\mathcal{X}(G)$ in the first equation, we obtain

$$\begin{cases}\pi_{\mathcal{X}}(T)=1-\pi_{\mathcal{Y}}(A)-\pi_{\mathcal{Y}}(G)+\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(C)-\pi_{\mathcal{X}}(G)+\pi_{\mathcal{Y}}(G)-\pi_{\mathcal{X}}(G)\\ \pi_{\mathcal{Y}}(T)=1-\pi_{\mathcal{Y}}(A)-\pi_{\mathcal{Y}}(C)-\pi_{\mathcal{Y}}(G)\\ \pi_{\mathcal{X}}(A)=\pi_{\mathcal{Y}}(A)+\pi_{\mathcal{Y}}(G)-\pi_{\mathcal{X}}(G)\\ \pi_{\mathcal{X}}(C)=\pi_{\mathcal{Y}}(C)+\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(G)\end{cases}$$

that is equivalent to

$$\begin{cases}\pi_{\mathcal{X}}(T)=1-\pi_{\mathcal{Y}}(A)-\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(C)\\ \pi_{\mathcal{Y}}(T)=1-\pi_{\mathcal{Y}}(A)-\pi_{\mathcal{Y}}(C)-\pi_{\mathcal{Y}}(G)\\ \pi_{\mathcal{X}}(A)=\pi_{\mathcal{Y}}(A)+\pi_{\mathcal{Y}}(G)-\pi_{\mathcal{X}}(G)\\ \pi_{\mathcal{X}}(C)=\pi_{\mathcal{Y}}(C)+\pi_{\mathcal{X}}(G)-\pi_{\mathcal{Y}}(G)\end{cases}$$

The solutions of this system are all and the only probability vectors $\pi_{\mathcal{X}},\pi_{\mathcal{Y}}$ defining the Bernoulli-like processes $\mathcal{X},\mathcal{Y}$ that produce a concatenation process $\mathcal{Z}$ which is $1-$Chargaff. Anyway, supposing the vectors satisfying (3.2.3) does not guarantee that the process is Chargaff also for words of length k.

Indeed we give the following example

Example 3.13. *Let $\omega = \omega_1\omega_2$ be a word in $\mathcal{I}^2$. In order that the process is $2-$Chargaff, it has to be*

$$P_\alpha(\omega_1\omega_2) = P_\alpha(\bar{\omega}_2\bar{\omega}_1), \forall \omega_1, omega_2 \in \{A, C, G, T\}$$

Let consider the word $\omega = GG$. Reminding that we are taking under consideration the case $\alpha = \frac{1}{2}$, we have to verify that

$$P_\alpha(GG) = P_\alpha(CC) \iff P(\omega_2 = A \mid \omega_1 = A) = P(\omega_2 = T \mid \omega_1 = T)$$

i.e.

$$\pi_{\mathcal{X}}(G)\pi_{\mathcal{X}}(G) + \pi_{\mathcal{Y}}(C)\pi_{\mathcal{Y}}(C) = \pi_{\mathcal{X}}(C)\pi_{\mathcal{X}}(C) + \pi_{\mathcal{Y}}(G)\pi_{\mathcal{Y}}(G)$$

that in respect to (3.2.3) become

$$\pi_{\mathcal{X}}(G)^2 + \pi_{\mathcal{Y}}(C)^2 = \pi_{\mathcal{X}}(G)^2 + (\pi_{\mathcal{Y}}(C) + \pi_{\mathcal{X}}(G) - \pi_{\mathcal{Y}}(G))^2$$

It is easy to see that if we choose $\pi_{\mathcal{X}}$ and $\pi_{\mathcal{Y}}$ as

$$\pi_{\mathcal{X}}(G) = 0 \quad \pi_{\mathcal{Y}}(C), \pi_{\mathcal{Y}}(G) \neq 0$$

the equality is not satisfied, thus the process is not $2-$Chargaff.

3.2.4 1-Markov chains

Here we introduce the case of a process concatenation of two Markov chains.

Let $\mathcal{X}, \mathcal{Y}$ be two Markov chains of probabilities $(\pi, T), (\rho, Q)$ respectively. Let $\mathcal{Z}$ be the concatenation process of $\mathcal{X}$ and $\mathcal{Y}$.

We want to describe constrictions on probability vectors π, ρ and transition matrices T, Q, such that process $\mathcal{Z}$ is $\alpha-$Chargaff for some $k \in \mathbb{N}$.

Firstly, we observe that for $k = 1$ the conditions only affect vectors π and ρ, similarly to the simpler case of concatenation of Bernoulli-like processes. Indeed, given a letter $x \in \mathcal{I}$, we have that

$$\begin{aligned} P_{\mathcal{X}}(x) &= \pi(x), \\ P_{\mathcal{Y}}(x) &= \rho(x) \end{aligned} \tag{3.93}$$

so that

$$P_\alpha(x) = \alpha\pi(x) + (1-\alpha)\rho(\bar{x}) \tag{3.94}$$

Thus, the process $\mathcal{Z}$ is $\alpha-$Chargaff for $1-$words if and only if

$$\alpha\pi(x) + (1-\alpha)\rho(\bar{x}) = \alpha\pi(\bar{x}) + (1-\alpha)\rho(x) \tag{3.95}$$

for every x in $\mathcal{I}$, that is the same of equation (3.88) for Bernoulli-like case.

Let $\omega = xy$ be a word in $\mathcal{I}$. The condition on the probabilities that enable $\alpha-$Chargaff property are

$$\alpha\pi(x)T_{xy} + (1-\alpha)\rho(\bar{y})Q_{\bar{y}\bar{x}} = \alpha\pi(\bar{y})T_{\bar{y}\bar{x}}(1+\alpha)\rho(x)Q_{xy} \tag{3.96}$$

that can be written as

$$\alpha[\pi(x)T_{xy} - \pi(\bar{y})T_{\bar{y}}\bar{x}] = (1-\alpha)[\rho(x)Q_{xy} - \rho(\bar{y})Q_{\bar{y}\bar{x}}] \tag{3.97}$$

In order to simplify the notation, we suppose $\alpha = \frac{1}{2}$ from now on.

$$\begin{cases} \sum_{x\in\mathcal{I}} \pi(x) = 1 \\ \sum_{x\in\mathcal{I}} \rho(x) = 1 \\ \sum_{i,j\in\mathcal{I}} T_{ij} = 1, \forall i \in \mathcal{I} \\ \sum_{i,j\in\mathcal{I}} Q_{ij} = 1, \forall i \in \mathcal{I} \\ \pi(x) - \pi(\bar{x}) = \rho(x) - \rho(\bar{x}) \\ \pi(x)T_{xy} - \pi(\bar{y})T_{\bar{y}\bar{x}} = \rho(x)Q_{xy} - \rho(\bar{y})Q_{\bar{y}\bar{x}} \end{cases} \tag{3.98}$$

Equations third and fourth of system 3.98 give each one four equations. Plus, conditions on the letters give two equations. In the end, conditions on two letters give 6 equations

$$\begin{aligned} \pi(A)T_{AA} - \pi(T)T_{TT} &= \rho(A)Q_{AA} - \rho(T)Q_{TT} \\ \pi(C)T_{CC} - \pi(G)T_{GG} &= \rho(C)Q_{CC} - \rho(G)Q_{GG} \\ \pi(A)T_{AC} - \pi(G)T_{GT} &= \rho(A)Q_{AC} - \rho(G)Q_{GT} \\ \pi(C)T_{CA} - \pi(T)T_{TG} &= \rho(C)Q_{CA} - \rho(T)Q_{TG} \\ \pi(C)T_{CT} - \pi(A)T_{AG} &= \rho(C)Q_{CT} - \rho(A)Q_{AG} \\ \pi(T)T_{TC} - \pi(G)T_{GA} &= \rho(T)Q_{TC} - \rho(G)Q_{GA} \end{aligned} \tag{3.99}$$

Plus, if we analyze words such that $xy = \bar{y}\bar{x}$ we obtain

$$\begin{aligned} T_{AT} = Q_{AT}, \;\; T_{TA} = Q_{TA} \\ T_{CG} = Q_{GC}, \;\; T_{GC} = Q_{GC} \end{aligned} \tag{3.100}$$

In conclusion we have 40 parameters and 18 conditions, for concatenation of Markov chains.

3.2.5 Mixed processes

In the previous sections we studied cases of concatenations of processes of the same nature, i.e. $\mathcal{X}, \mathcal{Y}$ both Bernoulli-like or Markov-chains. Here we

construct final process $\mathcal{Z}$ as concatenation of stationary processes $\mathcal{X}, \mathcal{Y}$, that are Markov chain and Bernoulli-like respectively.

Consistently with the previous sections, we call π, T the probability vector and the transition matrix of process $\mathcal{Y}$, and ρ the probability vector of Bernoulli-process $\mathcal{X}$.

As usual, we want to describe conditions for what values of π, T and ρ the process $\mathcal{Z}$ is $\alpha-$Chargaff. We start analyzing probability on $1-$words.

Given $\alpha \in [0,1]$, since it has to be $P_\alpha(x) = P_\alpha(\bar{x})$ for every x in $\mathcal{I}$, we have

$$\alpha\pi(x) + (1-\alpha)\rho(\bar{x}) = \alpha\pi(\bar{x}) + (1-\alpha)\rho(x) \tag{3.101}$$

Let now be $\omega = xy$ a in $\mathcal{I}$. Then the process $\mathcal{Z}$ is $\alpha-$Chargaff for $2-$words if and only if

$$\alpha\pi(x)T_{xy} + (1-\alpha)\rho(\bar{y})\rho(\bar{x}) = \alpha\pi(\bar{y})T_{\bar{y}\bar{x}} + (1-\alpha)\rho(x)\rho(y) \tag{3.102}$$

that can be written as

$$(1-\alpha)[\rho(\bar{y})\rho(\bar{x}) - \rho(x)\rho(y)] = \alpha[\pi(\bar{y})T_{\bar{y}\bar{x}} - \pi(x)T_{xy}] \tag{3.103}$$

For $\alpha = \frac{1}{2}$ equality (3.104) become

$$\rho(\bar{y})\rho(\bar{x}) - \rho(x)\rho(y) = \pi(\bar{y})T_{\bar{y}\bar{x}} - \pi(x)T_{xy} \tag{3.104}$$

Thus, for $\alpha = \frac{1}{2}$, the only vectors π, ρ and matrices T that produce a process $\mathcal{Z}$ that is $\alpha-$Chargaff in respect to words of length two, are all and only that one which satisfy the following

$$\begin{cases} \sum_{x\in\mathcal{I}} \pi(x) = 1 \\ \sum_{x\in\mathcal{I}} \rho(x) = 1 \\ \sum_{i,j\in\mathcal{I}} T_{ij} = 1, \forall i \in \mathcal{I} \\ \pi(x)T_{xy} + \rho(\bar{y})\rho(\bar{x}) = \pi(\bar{y})T_{\bar{y}\bar{x}} + \rho(x)\rho(y) \\ \rho(\bar{y})\rho(\bar{x}) - \rho(x)\rho(y) = \pi(\bar{y})T_{\bar{y}\bar{x}} - \pi(x)T_{xy} \end{cases} \tag{3.105}$$

Bibliography

[1] Guenter Albrecht-Buehler. Asymptotically increasing compliance of genomes with chargaff's second parity rules through inversions and inverted transpositions. *Proceedings of the National Academy of Sciences*, 103(47):17828–17833, 2006.

[2] SJ Bell and DR Forsdyke. Deviations from chargaff's second parity rule correlate with direction of transcription. *Journal of theoretical biology*, 197(1):63–76, 1999.

[3] Cristian I Castillo-Davis. The evolution of noncoding dna: how much junk, how much func? *Trends in Genetics*, 21(10):533–536, 2005.

[4] Richard C Deonier, Simon Tavaré, and Michael Waterman. *Computational genome analysis: an introduction*. Springer Science & Business Media, 2005.

[5] Donald R Forsdyke and James R Mortimer. Chargaff's legacy. *Gene*, 261(1):127–137, 2000.

[6] Diego Luis Gonzalez, Simone Giannerini, and Rodolfo Rosa. On the origin of the mitochondrial genetic code: towards a unified mathematical framework for the management of genetic information. 2012.

[7] Lila Karl. Dna computing: arrival of biological mathematics. *The mathematical intelligencer*, 19(2):9–22, 1997.

[8] Leonid Koralov and Yakov G Sinai. *Theory of probability and random processes*. Springer Science & Business Media, 2007.

[9] Michael J McLean, Kenneth H Wolfe, and Kevin M Devine. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of molecular evolution*, 47(6):691–696, 1998.

[10] David Mitchell and Robert Bridge. A test of chargaff's second rule. *Biochemical and biophysical research communications*, 340(1):90–94, 2006.

[11] CK Peng, SV Buldyrev, AL Goldberger, S Havlin, F Sciortino, M Simons, HE Stanley, et al. Long-range correlations in nucleotide sequences. *Nature*, 356(6365):168–170, 1992.

[12] Rivka Rudner, John D Karkas, and Erwin Chargaff. Separation of b. subtilis dna into complementary strands. 3. direct analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 60(3):921, 1968.

[13] Marcelo Sobottka and Andrew G Hart. A model capturing novel strand symmetries in bacterial dna. *Biochemical and biophysical research communications*, 410(4):823–828, 2011.

[14] Kenji Sorimachi. A proposed solution to the historic puzzle of chargaff's second parity rule. *The Open Genomics Journal*, 2(3):12–14, 2009.

[15] Kenji Sorimachi and Teiji Okayasu. An evaluation of evolutionary theories based on genomic structures in saccharomyces cerevisiae and encephalitozoon cuniculi. *Mycoscience*, 45(5):345–350, 2004.

[16] James D Watson and Andrew Berry. *DNA: The secret of life*. Knopf, 2009.

Acknowledgements

I thank my supervisor Dr Mirko Degli Esposti and assistant supervisor Giampaolo Cristadoro for having introduced me to the fascinating study of DNA modeling, for the support and for the patience they revealed in working with me.

I also would like to thank Professor Sobottka for the immediate help he gave me during all the writing of this thesis, answering to my questions from the opposite part of the world.

My friends, you are so many that it is impossible to mention all of you here. Thank you for helping me, for listening to me, for being happy for me, for crying with me, for loving me, for saying "spegni Facebook e mettiti a studiare" to me. If you're reading these lines, it means that you decided to share this important moment with me, so hang on and I'll hug you in a few minutes.

You, thank you too.

And most important of all, I thank my family, that has supported me and took care of me all over these years (and it's quite a big amount of years) always with love.