# Stationarity and Reversibility in the Nucleotide Evolutionary Process

Federico Squartini

Dissertation zur Erlangung des Grades
eines Doktors der Naturwissenschaften (Dr. rer. nat.)
am Fachbereich Mathematik und Informatik
der Freien Universität Berlin

Gutachter:
Prof. Dr. Martin Vingron
Prof. Dr. Arndt von Haeseler

1. Referent: Prof. Dr. Martin Vingron
2. Referent: Prof. Dr. Arndt Von Haeseler

Tag der Promotion: 3.Mai 2010

# Acknowledgments

All the research contained in this work was carried out at the Max Planck Institute for Molecular Genetics in Berlin, department of Computational Molecular Biology. The years spent in the institute have been an enriching experience, which has greatly helped me bringing my scientific skills to maturity.

I specially thank Peter Arndt for choosing me as his PhD student and supervising my research, Martin Vingron for giving me the opportunity to work in his department, and Hannes Luz for all the invaluable help he has given me during my years in Berlin.

Many thanks also go to my friends and colleagues in the Max Planck Institute, too many to mention here, for the interesting discussions and the enduring encouragement.

Finally, I thank my parents, who never failed to support me during my studies.

Federico Squartini Berlin, May 2010

# Contents

# Chapter 1

# Introduction

*When studying a natural phenomenon it is a well established and fruitful practice to disregard some of its properties in order to get a simpler and neater mathematical description. In a first stage we can use physical and mathematical intuition to decide what to incorporate and what to eliminate from the description. But once a theory has been laid out it becomes important to go back to the assumptions previously made and to test in a rigorous way their validity in the phenomenon under study.*

*In computational evolutionary genomics one example of this simplification process can be found in the assumptions that are made in the various models of sequence evolution, the nucleotide substitution process which leads to the divergence of the DNA sequences of different species originating from a common ancestor.*

*It is the aim of this thesis to investigate two such assumptions, namely the assumption that nucleotide sequence is in equilibrium with respect to the substitution process and the assumption that the process is time reversible.*

## 1.1 DNA, the molecule of life

Ever since its iconic double helix structure was determined by Francis Crick and James Watson [58], the deoxyribonucleic acid (DNA) has been the most popular and most studied molecule in biology (Fig. 1.1).

The DNA molecule is a polymeric chain composed of a sugar backbone on which four monomers, called nucleotides or bases, are attached. These are Adenine (`A`), Thymine (`T`), Cytosine (`C`) and Guanine (`G`) (Fig. 1.2).

The nucleotides have the fundamental property of being able to couple with each other, `A` can bond with `T` and `C` with `G`, forming the so called Watson Crick pairs (Fig. 1.3). It is because of these bonds that two linear chains of nucleotides running in opposite directions pair with each other if they are complement symmetric, i.e. they can be

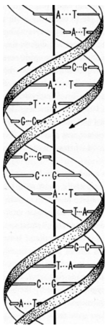

**Figure 1.1:** A DNA molecule. The double helix structure is due to the formation of bonds among complementary nucleotides.

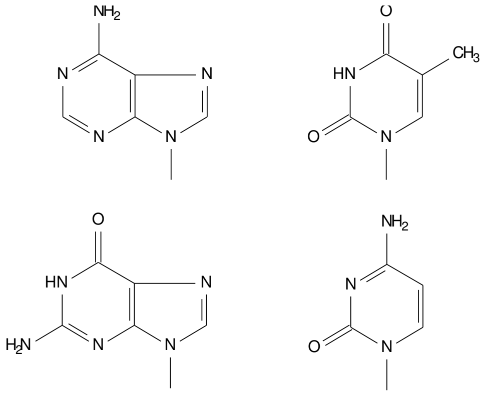

**Figure 1.2:** The four nucleotides, from left to right and top to bottom:Adenine, Thymine, Cytosine and Guanine.

obtained from each other by taking the Watson-Crick complementary nucleotide to each of their bases. The double polymeric chain so obtained further coils in the notorious double helix structure.

Each continuous DNA molecule present in a living cell is called a *chromosome*. One refers to the total number of chromosomes in a cell as the *genome*. Prokariotyc organisms, like bacteria, have only one chromosome, a circular molecule of DNA. Eukaryotic organism instead have a much more complex cellular structure. They have several chromosomes, each of which is not just arranged linearly like in prokaryotes, but it's folded in a highly packed structure called chromatin.

Chromatin is the product of the repeated folding of DNA on a backbone of special proteins know as histones. It has several benefits, first of all it allows much longer amount of DNA to occupy a small space. The human genome arranged linearly is a couple of meters long, quite an impressive length if we compare it with the size of a cell which is about $10^{-5}$ meters. The second important role of chromatin is in regulating the chemical activity of DNA by packing and unpacking portions of it, thus rendering them accessible to the action of proteins or not. Studying how this is explicated is one of the subjects of epigenetics [2], and has been one of the most active field of molecular biology research in recent years.

Eukaryotic cells are furthermore divided in two categories. Haploid cells have only one set of chromosomes, in a similar fashion to prokaryotic cells. On the other hand diploid cells, which are present in multicellular organisms with sexual reproduction, have two set of chromosomes, one set inherited from the father and one set inherited from the mother, so that chromosomes are in this case present in homologous pairs. Each set is very similar to the other, but the variation in the base composition between the two elements of each pair adds a further level of complexity and robustness. Even more important it allows through the mechanism of meiotic recombination the possibility that beneficial mutations present on homologous chromosomes come together on the same one.

Even Eukaryotic cells may have non homologous chromosomes though, the sexual chromosomes which determine the sex of individual. As an example in mammals there are two sexual chromosomes, the X and the Y. Males of the species have a non homologous XY couple in their cells, while females have an homologous XX couple.

## 1.2 The central dogma

Apart from DNA there are two others fundamental polymers in cells: the ribonucleic acid (RNA) and proteins. RNA is structurally very similar to DNA, the only difference being in the sugar backbone, which is in this case ribonucleic sugar, and in the use of the nucleotide uracil (U) in place of thymine. Furthermore RNA is only present in single stranded form. A linear RNA chain can fold on itself forming Watson Crick pairs that define its three-dimensional shape, which can be determined computationally with a good precision [66, 65]. RNAs are classified according to their function and there are many different varieties, the most relevant being messenger RNA, transfer RNA and ribosomal RNA.

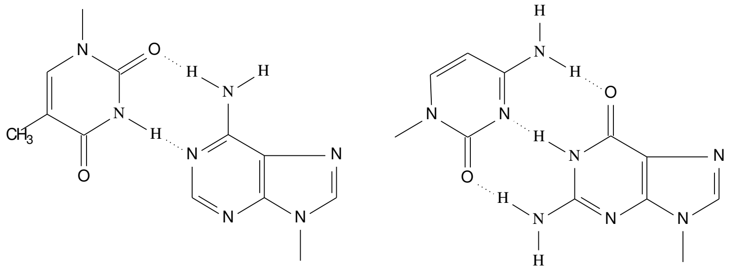

**Figure 1.3:** The Watson-Crick pairings. We can see the pairing of thymine and adenine on the left and that of cytosine and guanine on the right.

Proteins are also, like DNA and RNA, polymers. However their component monomers, amino-acids, are twenty and they can interact with each other in more ways than the simple pairing mechanism that shapes RNA structure. So a protein has not a simple structure, but instead it coils on itself forming a complex globular structure. The problem of predicting from first principles how a protein will coil, given a linear sequence of amino-acids, has not yet found a solution despite being 50 years old. With our present technology we can determine protein structure only with experimental methods like crystallography [59, 23] or nuclear magnetic resonance [45]. Experimentally known protein structures are stored in databases which can be used to infer new structures, using the assumption that proteins with similar sequences will coil in similar ways [37].

The second fundamental discovery of Francis Crick [12] was how DNA, RNA and proteins are related and functional to each other. His proposed mechanism, which is known as the *central dogma* of molecular biology [13], has two steps. First, the portion of DNA molecule which encodes for a protein is first transcribed into an RNA molecule. The RNA transcript is then processed in a specific cellular machine, the ribosome, where the linear chain of nucleotides is converted in a linear chain of amino-acids, converting three nucleotides (a codon) into one amino-acid. This process is called translation, and the conversion code used by the cell is universal across all organisms and is called genetic code (Tab. 1.1).

| | T | C | A | G |
|---|---|---|---|---|
| T | TTT Phe (F) | TCT Ser (S) | TAT Tyr (Y) | TGT Cys (C) |
| | TTC ” | TCC ” | TAC ” | TGC ” |
| | TTA Leu (L) | TCA ” | TAA **Stop** | TGA **Stop** |
| | TTG ” | TCG ” | TAG **Stop** | TGG Trp (W) |
| C | CTT Leu (L) | CCT Pro (P) | CAT His (H) | CGT Arg (R) |
| | CTC ” | CCC ” | CAC ” | CGC ” |
| | CTA ” | CCA ” | CAA Gln (Q) | CGA ” |
| | CTG ” | CCG ” | CAG ” | CGG ” |
| A | ATT Ile (I) | ACT Thr (T) | AAT Asn (N) | AGT Ser (S) |
| | ATC ” | ACC ” | AAC ” | AGC ” |
| | ATA ” | ACA ” | AAA Lys (K) | AGA Arg (R) |
| | ATG Met (M) | ACG ” | AAG ” | AGG ” |
| G | GTT Val (V) | GCT Ala (A) | GAT Asp (D) | GGT Gly (G) |
| | GTC ” | GCC ” | GAC ” | GGC ” |
| | GTA ” | GCA ” | GAA Glu (E) | GGA ” |
| | GTG ” | GCG ” | GAG ” | GGG ” |

**Table 1.1:** The genetic code is a dictionary which translates triplets of nucleotides (codons) to amino-acids.

As the sequence of amino-acids comes out of the processing ribosome, it starts coiling and forming the spatial structure which confers to each different protein its specific function in the cell.

An interesting fact is that the central dogma, in its orthodox formulation, states that the flow of information in the cell has a precise direction: out of DNA and into proteins. The key point here is that this is in perfect agreement with the Darwinian theory of evolution.

Darwin observed that individuals with beneficial traits will survive at the expense of less fit individuals, and pass their genomic set to future generations. He ruled out the possibility that beneficial traits acquired during the lifetime of and individual organism would be passed to the offspring and would thus contribute to evolution. According to his theory, and in contrast with the views of the french biologist Lamarck, beneficial traits acquired during the lifetime of and individual organism would not be passed to the offspring and would thus not contribute to the evolution of the species. Instead as an example, if Lamarck had been right and Darwin wrong, a giraffe who stretches its neck to be able to eat higher leaves of a tree would have had offspring with a longer neck too.

Darwin's view is in perfectly confirmed by the central dogma, according to which genotype, DNA, determines phenotype, proteins, and never the converse. However some recent studies (see [27] for a review) have found evidence that the central dogma (this being maybe the fate of any dogma) is in fact violated. That is, there are molecular mechanisms by which information can flow from the environment back into DNA, thus effectively suggesting a come back of Lamarckian kind of evolution on which Darwin seemed to have put a gravestone 150 years ago.

## 1.3 The genomic landscape

Not all portions of the genome of an organisms are coding for a protein or an RNA. The fraction of DNA with such a purpose varies greatly across different organisms. In higher eukaryotes only a very small portion has such functions, for example in the human genome only about 3% has such coding role. The rest is composed of different non coding sequences, like repetitive sequences, transposable elements, pseudo-genes and genomic desert with no known function [26].

Another peculiarity of eukaryotes is that proteins are not encoded in continuous stretches of DNA, but instead their coding sequence is split into chunks called exons, which are interspersed into much longer sequence stretches called introns. Intronic regions are spliced from the RNA transcript before translation begins. The usefulness of all this non coding elements is still debated, and they are usually referred to as “junk”.

## 1.4 Molecular replication and evolution

One prominent feature of living organism is the capability to generate offspring, so that like begets like. The process is of course very complex and involves a vast number of different chemical pathways in the cell. However there is a very simple concept at its heart, nature's stroke of genius one may say, a trait which must have been present in cells already at the dawn of life 3.5 billions of years ago.

This feature is the fact that the DNA double helix contains twice the genetic information, because as we said previously each of its two strands is a complementary copy of the other. Thus the key step the cell has to perform to replicate the information is unzipping the double stranded DNA in two single stranded chains. Each of these chains will then be complemented again with nucleotides, so that at the end there will be two new double helices in place of one.

Even though the idea is simple, it still involves lots of molecular machinery working on it in order to be accomplished successfully, the most important being a protein complex known as DNA polymerase. The replication process is quite accurate, as this is the fundamental requirement to preserve the information content and have "working" cellular offspring.

Nonetheless, however precise the replication process is, errors happen and there is no guarantee that the copied DNAs will be completely identical to their parent. Such errors are called *mutations*, and are the basis of biological evolution.

In order to have an evolutionary impact mutations have to be inheritable, in other words the individual in which they appear should have the possibility of passing the mutated DNA to its offspring. While this is always the case for unicellular organisms which replicate by mitosis, it is not a given for multicellular organisms where only mutations happening in the germ line will have a chance of being inherited. On the contrary mutations appearing in somatic cells will only have phenotypical effects, cancer being most well known and devastating result of such category of DNA alterations.

## 1.5 Mutation classes

The kind of possible mutations affecting DNA can be partitioned into two different classes, those which affect single nucleotides, and those which insert or delete whole portions of the genome.

**Point mutations** The first, and more important for the rest of this work, kind of mutations are the ones which exchange a nucleotide with another one, commonly called *single nucleotide mutations* or just *point mutations*.

There are many chemical process that can lead to point mutations. They can be partitioned in two kinds: exogenous ones, due two mutagenic agents (exposure to radioactive sources or mutagenic chemicals) and endogenous ones, where mutations are either induced by the thermal fluctuations of the environment, mainly tautomerism [56], or are the result of errors in the replication process.

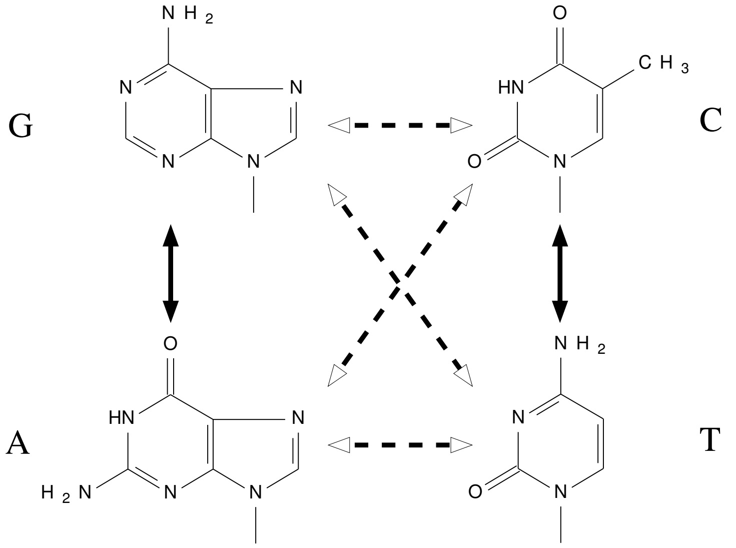

**Figure 1.4:** The twelve possible point mutational processes. Transitions are indicated by the continuous line and transversions by the dashed one.

There are twelve possible point mutations, from each of the four nucleotides to any of the other three. Mutations exchanging a purine with a purine or a pyrimidine with a pyrimidine are called *transitions*, while mutations from purine to pyrimidines and viceversa are known as *transversions* (Fig. 1.4). On the basis of chemical similarity a transition is more likely to happen than a transversion.

**Neighbor dependencies** Another fundamental point mutational process taking place mainly in vertebrates genomes, but first discovered in bacteria [11], originates from the interaction of nucleotides with the aqueous cellular environment. In fact, as can be seen in Fig. 1.5, a cytosine may interact with a molecule of water by an hydrolysis reaction, and mutate into an uracil according to the following stoichiometric equation:

$$\mathtt{C} + H_2O \rightarrow \mathtt{U} + NH_3 \tag{1.1}$$

But in this case repair enzymes recognize the uracil as an extraneous nucleotide, and correct the error excising it and replacing it with the correct complementary nucleotide.

However it is well known [11], and in fact one of the cornerstones of the already cited epigenetics, that in vertebrates cytosines adjacent to guanines, the so called `CG` pairs or

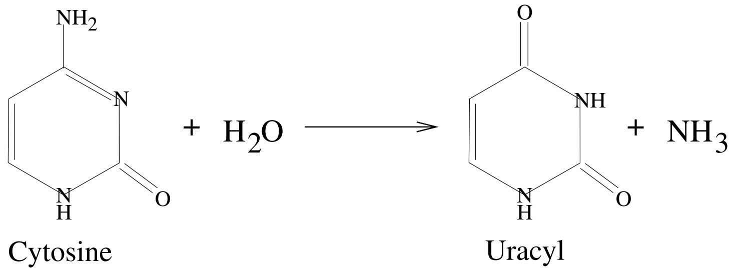

**Figure 1.5:** The hydrolysis reaction leading to the mutation of a Cytosine into a Uracil.

CpG's (the "p" referring to the phosphate backbone), are often present in a methylated state known as 5-methiylcytosine and indicated as C*. The hydrolysis reaction in this case is (Fig. 1.6):

$$\texttt{C}^* + H_2O \to \texttt{T} + NH_3 \tag{1.2}$$

In this case the cytosine mutates into a thymine and so it is not any longer possible for the repair enzymes to determine whether they should excise the newly created thymine or the original complementary base, a guanine. The repair mechanism has to make an arbitrary choice in this case, so this is a very effective mechanism to introduce new mutations in a genome. The net effect is then the mutation of a CpG pair into a TpG and it is called CpG deamination or CpG decay process.

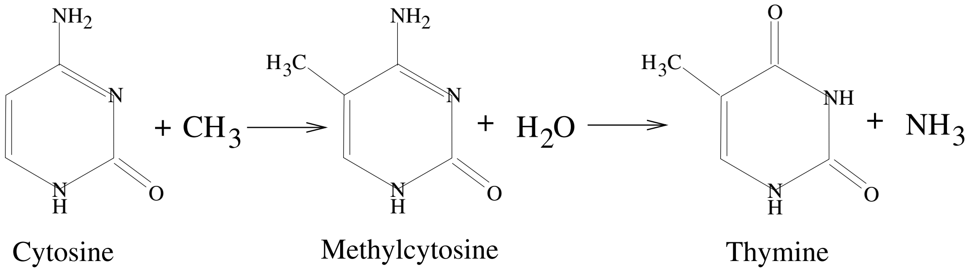

**Figure 1.6:** The CpG decay process is an hydrolysis reaction leading to the mutation of a methiylcytosine into a guanine.

The importance of this process can be easily understood by looking at (Tab. 1.2) from [3]. The table shows the ratios of dinucleotide frequencies in the human genome, where the numerator is obtained multiplying the single nucleotide frequencies, while the denominator is just the actual count of dinucleotides. If there were no neighbor dependencies the ratios should be all very close to 1, however what one can see that the value of the ratio for CGs is 0.2, meaning that this doublet is underrepresented in the human genome.

**Insertion and deletions** The second big group of mutational events are those which add or remove portions of DNA (indels). Unlike for point mutations, the spectra of events

|   | A | C | G | T |
|---|---|---|---|---|
| A | 1.10 | 0.87 | 1.11 | 0.91 |
| C | 1.20 | 1.21 | 0.20 | 1.11 |
| G | 0.99 | 1.05 | 1.22 | 0.87 |
| T | 0.80 | 0.99 | 1.21 | 1.10 |

**Table 1.2:** This table shows the ratio dinucleotide frequencies computed using the product of single nucleotide frequencies, and computed by counting the actual number of occurrences in the human genome.

is in this case much wider, and the characteristic much less understood.

According to recent studies [8] the vast majority of indels are actually the result of duplications of portion of genomes. There is a well established research line, pioneered by the Japanese geneticist Susumu Ohno [44], which investigates the importance of duplications in genome evolution.

The basic idea is that duplicating a portion of the genome, a gene for example, will leave one copy free to evolve new useful functions [36, 7]. Although initially the study of duplications had focused on gene duplications, recent studies have used triple alignments of closely related genomes to show that duplications happen not only at the level of single genes but at all length scales, from single nucleotides to long genomic stretches [40].

## 1.6 Motivations and aims of the thesis

In chapter 2 we will introduce the Markov model of molecular evolution, an ubiquitous framework which is one of the cornerstone of bioinformatics. The focus of this thesis will be the analysis of some particular properties of this model.

In fact, as we will see in later chapters, out of historical reasons and computational convenience, several simplifications are usually made inside this framework. The first Markov model, the Jukes-Cantor model or simply JC69, had only one free parameter [28]. The substitution rate from one nucleotide to any other different nucleotide was assumed to be the same regardless of the particular nucleotides.

A successive model was Kimura's two parameter model, also known as K80 [31]. This model breaks the complete symmetry present in the JC69, stating that nucleotide evolution has two different classes of events. One class is that of *transitions* in which a purine is exchanged with another purine (i.e. $\mathtt{A} \leftrightarrow \mathtt{G}$), or a pyrimidine with another pyrimidine (i.e. $\mathtt{T} \leftrightarrow \mathtt{C}$). The other class is the one of *transversions* in which a purine is exchanged with a pyrimidine or viceversa (eight possible events: $\mathtt{A} \leftrightarrow \mathtt{T}$, $\mathtt{A} \leftrightarrow \mathtt{C}$, $\mathtt{G} \leftrightarrow \mathtt{T}$ and $\mathtt{G} \leftrightarrow \mathtt{C}$). This reflects biochemical knowledge because as we have seen the two purines, as well as the two pyrimidines, have similar chemical structure so that transitions are more likely to happen than transversions.

Other models followed which broke more symmetries in the rate matrix: the F81 [18], the HKY85 [21], the T92 [53] and the TN93 [54]. Eventually, it was realized [33, 55] that all these models shared a particular symmetry, time reversibility. Time reversible Markov processes have two basic features. First, by looking at their realizations, it is not possible to decide whether the phenomenon we are observing is running forward or backward in time. Second and maybe even more important, the assumption of time reversibility also implies that the statistical properties of the system, DNA, do not change in time. In other words the process is stationary in time.

Later on, several other extension of these models were introduced, including those which also describe rate heterogeneities along the DNA sequence [63, 57], but they still assume the validity of the time reversibility assumption for the evolution of each single nucleotide.

In fact one of the fundamental problems of evolutionary genomics is how to estimate the parameters of the above mentioned models. In chapter 3 I will show how this can be reliably done using a procedure known as maximum likelihood estimation. However, it should be noted that models of nucleotide evolution were developed long before whole genome sequences were available. Researchers had at their disposal the sequences of only small portions of genomes, thus the scarcity of data forced them to use models with as few parameters as possible, in order to obtain reliable estimates.

In this context, assuming time reversibility and equilibrium in Markov models of nucleotide substitution was an elegant way of restricting the dimensionality of the parameter space. Furthermore, in maximum likelihood calculations, the possibility of rerooting the phylogenetic tree anywhere without affecting the resulting likelihood (the so called Felsenstein's "pulley principle"[18]), leads to an efficient algorithm for calculating the branch lengths of the tree. This speed up is extremely useful when searching the tree space for the maximum likelihood tree.

But is the evolutionary process of nucleotide substitutions really time reversible and in its stationary state? Making such assumptions could cause some important features of genome evolution to be overlooked. As an example, if the genome were always in its equilibrium state during evolution, quantities like the average GC content would not evolve in time. However, it was shown by [4] that, for example, the GC content in the human genome is not in equilibrium, and is still evolving. Similar results have also been found for the mouse genome [15].

Following [16, 50], I will show how it is possible, using at least three present day genomic sequences, to extend the maximum likelihood estimation procedure to the case where time reversibility and equilibrium are not assumed.

Using this methodology, in chapter 4 I will show how it is possible to measure deviations from time reversibility and equilibrium in the evolution of genomes. To this aim I will introduce two sets of indices, the stationarity indices and the irreversibility indices, STIs and IRIs for short, which can be calculated from the substitution frequencies along one branch in a phylogenetic tree and the nucleotide composition at the node at its end. When non-zero, the indices indicate violations of the basic assumptions mentioned above.

I will first derive the indices for Markov models describing the evolution of independent sites. However, in order to apply the indices to the analysis of the human genome we will have to face a complication, due to the `CpG` decay process. In this case it is not any longer possible to assume that nucleotide sites are evolving independently, since the process couples adjacent nucleotides. We will see how it is possible to extend the IRI to include neighbor dependencies and asses the deviations from time reversibility in the human case.

It is important to note that although other tests for stationarity and time reversibility have been proposed so far [49, 48, 17, 1] all of them operate on pairs of sequences, which limits their power. For example, situations where a sequence evolved under non-reversible conditions might go undetected as pointed out by [1].

The analysis based on STI and IRI, has the advantage that it tests for stationarity and time reversibility on just any single phylogenetic branch connecting an ancestral node with a more recent one (like for example the branch from the human-chimp common ancestor to present day human). To compute the indices the rate matrix has been estimated using the mentioned maximum likelihood procedure which does not assume either time reversibility or the stationarity of the process. In order to test the equilibrium and time reversibility properties in test cases, I will calculate the STIs and the IRIs for the evolutionary process of two different species, Drosophila simulans species and Homo sapiens.

# Chapter 2

# Models of Sequence Evolution

*The need to understand the features of the mutational process, in order to reconstruct phylogenetic trees and to apply general bioinformatics models, requires the development of ad hoc mathematical models.*

*In this chapter we will discuss the principles of sequence evolution, starting from the main mathematical tools, Markov chains and Markov processes. We will then discuss the evolutionary dynamics of new mutations appearing in a population, introducing Kimura's results on neutral evolution. Finally we will apply these models to the analysis of inter-species variations in the DNA sequence studying the features and the assumptions of different proposed mathematical models.*

## 2.1 Introduction to Markov processes

An appropriate probabilistic framework to model biological sequence evolution is that of *Markov processes*, in this and the following sections we will briefly introduce the basic mathematical concepts.

In what follows let $X(t)$ be a family of random variables (a stochastic process) with values $x$ in a discrete state space $\mathcal{S}$ and time $t$ belonging to a set, $\mathcal{T}$, which can in general be the set of integers (in which case we will speak of a *Markov chain*), or the set of real numbers (a *continuous time* or *proper Markov process*).

What we are interested in, is calculating the probabilities $P(X(t_0) = x_0, \ldots, X(t_n) = x_n)$ of trajectories. This can be accomplished using the product rule:

$$
\begin{aligned}
P(X(t_0) = x_0 \ldots X(t_n) = x_n) = {} & P(X(t_n) = x_n | X(t_{n-1}) = x_{n-1} \ldots X(t_0) = x_0) \\
& \times P(X(t_{n-1}) = x_{n-1} | X(t_{n-2}) = x_{n-2} \ldots X(t_0) = x_0) \ldots \\
& \ldots P(X(t_1) = x_1 | X(t_0) = x_0) P(X(t_0) = x_0)
\end{aligned}
\tag{2.1}
$$

Often when modeling phenomena we can make the assumption that the probability to be in a given state at time $t$ is only influenced by the previous state at time $t-1$. In other words the process only has a one step memory and in this case we speak of a Markov process:

**Definition 1.** *A Markov process is a stochastic process for which the following holds:* $P(X(t_n) = x_n | X(t_{n-1}) = x_{n-1} \dots X(t_0) = x_0) = P(X(t_n) = x_n | X(t_{n-1}) = x_{n-1})$ *for* $t_0 < t_1 < \dots < t_n$.

A further assumption that is commonly made is that the dynamical properties of the process do not vary in time. For example, in the case of neutral sequence evolution this would reflect the assumption that the mechanisms affecting the mutation rate do not change in time. In turn this would assume that the efficiency of the repair enzymes has not varied during the evolutionary time. This property of a Markov process is called *time homogeneity*:

**Definition 2.** *A Markov process is said to be time homogeneous if* $P(X(t+\tau) = x_1 | X(t) = x_0)$ *does not depend on* $t$. *In this case it is possible to use the following compact notation* $p(x_1, x_0, \tau) \equiv P(X(t+\tau) = x_1 | X(t) = x_0)$. *We call* $p(x, y, \tau)$ *the stochastic transition function.*

Alternatively time homogeneity can be defined in the following way, easily shown to be equivalent to the previous:

**Definition 3.** *A Markov process is said to be time homogeneous if the distribution of* $X(t_1), X(t_2), \dots, X(t_n)$ *is equal to the distribution of* $X(t_1+\tau), X(t_2+\tau), \dots, X(t_n+\tau)$ *for all* $\tau$ *and* $t_1, t_2, \dots \in \mathcal{T}$.

It follows immediately that for a homogeneous Markov chain, we can use the following short notation:

$$P_{i,j} \equiv P(X(t_n) = i | X(t_{n-1}) = j) \tag{2.2}$$

Equivalently for homogeneous Markov processes we can define:

$$P_{i,j}(t) \equiv P(X(t_0 + t) = i | X(t_0) = j) \tag{2.3}$$

One can also use an abstract matrix form. The $P$ operator is called the *transition semigroup*, and it can be shown to obey the Chapman-Kolmogorov equation (semi-group compositional property):

$$P(t_1)P(t_2) = P(t_1 + t_2) \tag{2.4}$$

In the case of continuous time Markov processes we can also define the concept of *transition rate*, as a time derivative of the transition probabilities:

**Definition 4.** *The transition rate from state $i$ to state $j$ is defined by the limit:*

$$q(i,j) \equiv \lim_{\tau \to 0^+} \frac{P(X(t+\tau)=i|X(t)=j)}{\tau} \tag{2.5}$$

As for the transition probability we can even in this case introduce an associated *transition rate matrix*:

$$Q_{i,j} \equiv q(i,j) \tag{2.6}$$

We can now define the equilibrium distribution, as the probability distribution that the process reaches asymptotically:

$$\lim_{t\to\infty} P(X(t)=k|X(0)=j) = \pi(k) \tag{2.7}$$

The equilibrium distribution has the property of being a fixed point of the dynamics. It can thus obtained solving the following eigenvalue problem:

$$\pi(j) = \sum_{k\in\mathcal{S}} P_{j,k}\pi(k) \tag{2.8}$$

In the case of Markov processes it is often more convenient to solve an equivalent problem, but for the transition matrix:

$$0 = \sum_{k\in\mathcal{S}} Q_{j,k}\pi(k) \tag{2.9}$$

Therefore one can either find an eigenvector with eigenvalue one of the transition probability, or an eigenvector with eigenvalue zero of the transition rate matrix.

## 2.2 The master equation

The transition matrix $Q$ is a first order approximation for the Markov process. From it we can derive the basic equation that describes the dynamic of a Markov process. We use definition (2.6):

$$Q = \lim_{t\to 0} \frac{P(t) - \mathbb{I}}{t} \tag{2.10}$$

Where $\mathbb{I}$ is the identity matrix. Applying Chapman-Kolmogorov we get:

$$\begin{aligned}\frac{dP(t)}{dt} &= \lim_{dt\to 0} \frac{P(t+dt)-P(t)}{t} \\ &= \lim_{dt\to 0} \frac{(P(dt)-\mathbb{I})}{t}P(t) \\ &= \lim_{dt\to 0} \frac{(P(dt)-\mathbb{I})}{t}P(t) \\ &= QP(t)\end{aligned} \tag{2.11}$$

This is known as Kolmogorov forward equation, or master equation, and determines the evolution in time of the probability density, $\rho(t) = P(X(t))$. In term of the probability density it can be written as:

$$\frac{d\rho(t)}{dt} = Q\rho(t) \tag{2.12}$$

The solution of a master equation can be obtained calculating the exponential of the rate matrix:

$$P(t) = e^{Qt} = \sum_{k=0}^{\infty} \frac{(Qt)^k}{k!} \tag{2.13}$$

Although here we define the exponential of the matrix in terms of its Taylor expansion, summing a finite number of terms of this series it's not the most efficient way of computing it. Efficient algorithms, are reviewed in [42].

## 2.3 Time reversibility

A relevant class of Markov processes, are those with the following property: inverting the arrow of time we obtain a new process that cannot be distinguished from the original one. Intuitively this means that a watching a movie of a time reversible phenomenon we should not be able to determine whether the tape is running forwards or backwards. Such a class of processes is called *time reversible.* Historically the origin of the concept of time reversibility comes from classical mechanics where, in absence of dissipative forces, the equations of motion are assumed to be invariant under a time reversal transformation. However, as we will see, the concept has been widely used sequence evolution too.

Formal aspects of time reversibility in Markov processes are introduced in the book of Kelly [29]. Here and in the following (when cited) we reproduce some proofs relevant for our work:

**Definition 5.** *A Markov process is time reversible if* $X(t_1), X(t_2), \dots, X(t_n)$ *has the same distribution as* $X(\tau - t_1), X(\tau - t_2), \dots, X(\tau - t_n)$ *for all* $\tau$ *and* $t_1, t_2, \dots \in \mathcal{T}$

The following is a fundamental property of reversible processes:

**Proposition 1.** *A time reversible Markov process is stationary.*

*Proof.* The proof follows immediately from the definition. Since $X(t_1 + \tau), X(t_2 + \tau), \dots, X(t_n+\tau)$ has the same distribution as $X(t_1), X(t_2), \dots, X(t_n)$ and $X(t_1), X(t_2), \dots,$ $X(t_n)$ has the same distribution as $X(\tau - t_1), X(\tau - t_2), \dots, X(\tau - t_n)$. So the process is stationary. □

A well known criterion, the *detailed balance* condition, can be used to test the reversibility of a Markov process [29]:

**Proposition 2.** *A stationary Markov chain is reversible iff there exist a probability distribution* $\pi_j$*, with* $j \in \mathcal{S}$ *such that*

$$P_{j,k}\pi_k = P_{k,j}\pi_j \tag{2.14}$$

*Proof.* Let's assume the process is reversible. Since the process is stationary $P(X(t) = j)$ is independent of $t$. Let's define $\pi(j) \equiv P(X(t) = j)$. Since the process is reversible:

$$P(X(t+1) = k, X(t) = j) = P(X(t+1) = j, X(t) = k) \tag{2.15}$$

from which :

$$P_{k,j}\pi_j = P_{j,k}\pi_k \tag{2.16}$$

Let's now suppose that the distribution $\pi_j$ exists then summing over both sides of Eq. (2.14):

$$\sum_j P_{k,j}\pi_j = \sum_j P_{j,k}\pi_k = \pi_k \tag{2.17}$$

From which we get the $\pi_j$ is the stationary distribution of the chain. Then let's calculate the probability of a trajectory:

$$\begin{aligned} P(X(t+m) = j_m, X(t+m-1) = j_{m-1}, \dots X(t+1) = j_1, X(t) = j_0) = \\ P_{j_m,j_{m-1}} \dots P_{j_1,j_0}\pi_{j_0} \end{aligned} \tag{2.18}$$

And:

$$P(X(t'+m)=j_0, X(t'+m-1)=j_1, \dots X(t'+1)=j_{m-1}, X(t')=j_m) = \\ P_{j_0,j_1}\dots P_{j_{m-1},j_m}\pi_{j_m} \tag{2.19}$$

Finally applying detailed balance (Eq. 2.14) we can check that the right hand sides are equal, this proves the theorem. □

**Proposition 3.** *A stationary Markov process is reversible iff there exist a probability distribution $\pi_j$, with $j \in \mathcal{S}$ such that*

$$Q_{k,j}\pi_j = Q_{j,k}\pi_k \tag{2.20}$$

*Proof.* Let's assume reversibility, in this case:

$$P(X(t+\tau)=j, X(t)=k) = P(X(t+\tau)=k, X(t)=j)$$

Multiplying both sides by the stationary distribution, we get:

$$P(X(t+\tau)=j|X(t)=k)\pi(k) = P(X(t+\tau)=k|X(t)=j)\pi(j)$$

Dividing by $\tau$ and taking the limit to zero:

$$\lim_{\tau\to 0}\frac{P(X(t+\tau)=j|X(t)=k)}{\tau}\pi(k) = \lim_{\tau\to 0}\frac{P(X(t+\tau)=k|X(t)=j)}{\tau}\pi(j)$$

Which gives the detailed balance condition, as desired.

Let's assume the converse, as in the previous proof summing over both sides of the detailed balance conditions (Eq. 2.20). We get

$$\sum_j Q_{k,j}\pi_j = \sum_j Q_{j,k}\pi_k = 0 \tag{2.21}$$

Which means that $\pi$ is a stationary distribution of the process.

Now, in order to complete the proof, I will use the associated Markov chain jump process. Let's assume the process visits states $i_1,\dots,i_n$ and sojourns in each of them a time $t_1,\dots,t_n$. The times have probability density:

$$q(i_1)e^{-q(i_1)t_1}$$

At each jump the process will go form state $i_1$ to state $i_n$ with probability $\frac{q(i_1)}{q(i_n)}$. So that

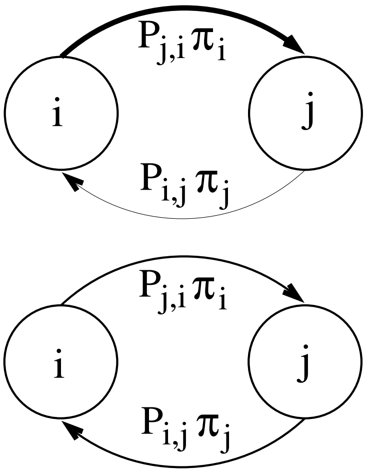

**Figure 2.1:** The detailed balance theorem asserts that in order to have time reversibility the total probability flux among any two states must be zero. We see in this figure two fluxes, in the upward one detailed detailed balance is not satisfied, in the lower one it is.

the probability density for the process is:

$$\begin{aligned} &\pi(i_1)e^{-q(i_1)t_1}\frac{q(i_1,i_2)}{q(i_2)}q(i_2)e^{-q(i_2)t_1}\dots\frac{q(i_{n-1},i_n)}{q(i_n)}q(i_n)e^{-q(i_n)t_n} \\ &= \pi(i_1)e^{-q(i_1)t_1}q(i_1,i_2)e^{-q(i_2)t_1}\dots q(i_{n-1},i_n)e^{-q(i_n)t_n} \\ &= \pi(i_n)e^{-q(i_n)t_n}q(i_{n-1},i_n)e^{-q(i_{n-1})t_{n-1}}\dots q(i_2,i_2)e^{-q(i_1)t_1} \end{aligned} \tag{2.22}$$

Where in the last line we have applied the hypothesis of detailed balance. We have thus found that, under the given assumptions, the density of the original process is equivalent to the probability density of the time reversed one, which proves the theorem. □

## 2.4 Principles of species evolution

We have seen in the introduction how the information content of a DNA molecule varies over generations due to the appearance of mutations. However by itself the appearance of a new mutation in a genome is not sufficient to have an evolutionary effect, as this mutation has also to spread from the single individual where it appears in, to the whole species. How this happens, is the subject of a branch of evolutionary science known as population genetics, whose origin can be dated back to the work of Gregor Mendel on plant hybrids [39].

The fundamental concept is that of *species*, a population of individuals having a common descent and, in case of sexual species, capable of mating with each other. All the indi-

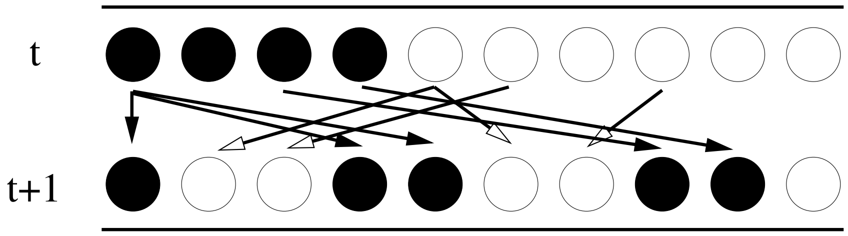

**Figure 2.2:** The wright fisher model of neutral evolution. Black a white circles represent two different kind of competing alleles.

vidual genomes in a species are derived from a common ancestral sequence, and share a high genomic similarity.

Each time a new mutation appears it propagates from the individual bearing it to its offspring. Eventually there can be two possibilities, either it spreads so much that it will be present in the whole population, or it decreases in frequency until it is not present anymore, according to a dynamic which is known as *fixational process*. In this thesis I will only be concerned with the molecular evolution of nonfunctional regions, which represent a special case of the general process of Darwinian evolution. The dynamic of neutral regions was studied by two of the leading geneticists of the last century, Sewall Wright and sir Ronald Fisher. Not surprisingly the model they developed is known as Wright-Fisher (WF) model [19, 61].

The WF model is quite simple to describe. If we restrict our study to a given locus on the genome, there will be in general a number $n$ variants of this locus in different individuals, this variants are named *alleles*. With no loss of generality, but just to avoid a more cumbersome notation, we restrict the analysis to a locus with two alleles ($A$ and $a$), an ancestral one and a new variant which has arisen due to some mutational event. What we are interested in, is the probability that the frequency of the allele $a$ in the population has a given value as a function of time, $p(t) \equiv P\left(\frac{n_a(t)}{2N}\right)$. We take discrete time, assuming non overlapping generations. For diploid organism, if the population has size $N$, when the mutation first appears we will have $p(0) = \frac{1}{2N}$, since of course it will appear on one of two homologous chromosomes.

Having no phenotypical effect the fixational process can be modeled as random sampling with replacement. This is because the amount of resources the environment can provide is limited, so that only a finite number $N$ of individuals will be able to coexist at the same time. The generation at time $t+1$ is created by sampling $N$ individuals from those composing the population at time $t$. An individual can have more than one offspring, hence we need to sample with replacement (Fig. 2.2). In other words in this scenario the Darwinian mechanism of the survival of the fittest, is substituted by a neutral mechanism,

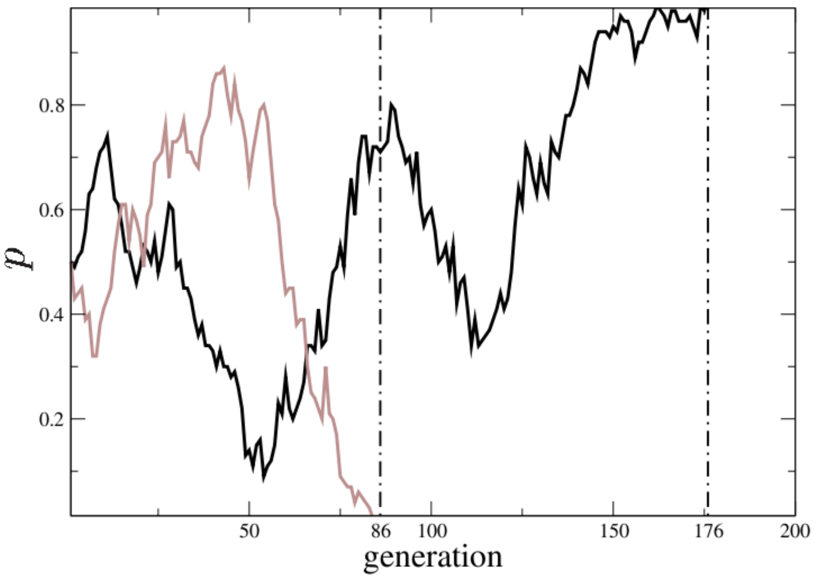

**Figure 2.3:** This figure (from [22]) shows two example trajectories of allele frequencies for two neutral loci. As a result a random sampling, due to the finite size of the population, one allele fixates while the other is purged from the population.

a *survival of the luckiest.*

$$p(j, t+1|i, t) = \binom{2N}{j} \left(\frac{i}{2N}\right)^{j} \left(1 - \frac{i}{2N}\right)^{(2N-j)} \tag{2.23}$$

The fundamental result, due to Kimura [30] is that in this case the probability of fixation depends only on the population size:

**Theorem 1.** *(Kimura). The probability of fixation of a neutral allele, is equal to its initial frequency $p_0$.*

Kimura's result is the connection between population genetics and species evolution. In a species with $N$ individuals and mutation rate $\mu$ there will be $N\mu$ mutations appearing per generation. Multiplying by the probability of fixation we get the substitution rate

$$u = 2N\mu \left(\frac{1}{2N}\right) = \mu \tag{2.24}$$

So, the result is that for neutral evolution mutation rate is equal to substitution rate.

## 2.5 Mathematical models of evolution

Computational molecular biology uses a very schematic representation of DNA, namely an ordered sequence of letters. As we have seen even for a single species there is nothing like a unique DNA, since individuals have similar, but not identical genomes. However, given that the differences among genomes of different individuals comprise a small fraction of the total genome, for analysis which aim to compare different species it is customary to use a given individual genome (a reference genome arbitrarily chosen) as a representative of all the different individual genomes.

We will use the following mathematical definition of a genome:

**Definition 6. Genomic Sequence** *A genomic sequence $S$ is an ordered list of nucleotides $S = (x_1, \ldots, x_n) \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$. One refers to nucleotide at position $k$ with the notation $S[k] \equiv x_k$.*

One other key concept for evolutionary studies is that of an *alignment* of sequences. An alignment is a way to recapitulate the evolutionary history of a genomic sequence with respect to the indel events which we have defined in the introduction. In fact as we have discussed, genomes have a common origin and because of this for any two genomes of different species there is an ancestral genome from which both have evolved. However as a consequence of insertions or deletions contemporary genomes may be longer or shorter than their ancestor, and so given two nucleotides we cannot, without further analysis, tell whether they both evolved from the same ancestral nucleotide or not. Aligning a group of sequences means extending them adding gaps (a special character) in such a way that they all have the same length, and that nucleotide having a common ancestor are in the same position in the new gap-extended representation.

Although we usually align present day sequence with one another, ideally we would align ancestral sequences with the present day descendants. The idea of alignment is exemplified graphically in Fig. 2.4. We won't elaborate further on the methodologies to generate alignments, but we will note that the production of reliable alignments is a crucial problem in computational biology, as a wrong set of alignments can drastically alter the result of a biological study [60].

As already seen, the nature of mutations is such that, given a succession of $n$ nucleotides at a given locus and time $X(t_0), \ldots, X(t_n)$ where $X(t) \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ and time is measured in appropriate units, there should be a dependence only on the state at time $t-1$. In general it is always possible to rescale time to 1 and we will often use this rescaling in the following.

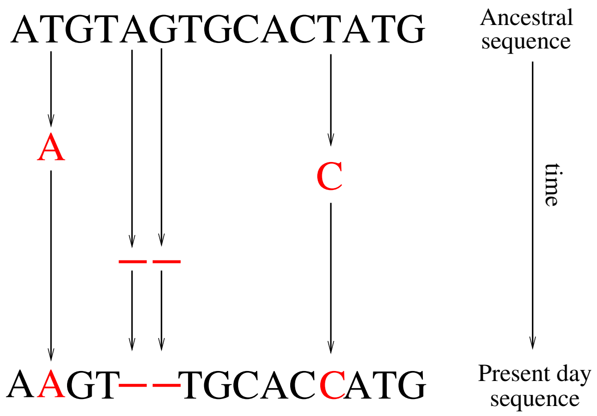

**Figure 2.4:** This figure illustrates the evolutionary process with indels. It shows two mutational events and a deletion event which removes simultaneously two adjacent nucleotides from the sequence. If we had wanted to represent an insertion event, we would have inserted gaps in the ancestral sequence corresponding to the loci were new nucleotides have been inserted.

## 2.6 Jukes Cantor and Kimura 2 parameter models

The use of probabilistic models in the study of sequence evolution started with the landmark papers [64, 28], where for the first time a Markovian model of evolution was introduced and used. The model is known as Jukes-Cantor, or JC69. The original focus of the papers was protein evolution, however what I present in this section is an equivalent model for nucleotide evolution.

Two key assumptions in the JC69 model, are shared by almost all successive probabilistic models of sequence evolution. The first is that each nucleotide position in the sequence evolves independently from all others. Probabilistically speaking this means that the probability distribution for the whole sets of sites is the product of identically distributed probability distributions at each site.

The second assumption is that the evolution of each single site is a Markov process, each nucleotide position evolution is completely determined by assigning a $4 \times 4$ rate matrix with 12 different transition probabilities.

Furthermore the JC69 model also assumes that the transition probabilities are all equal to the same value. Each nucleotide is equally likely to turn into each other nucleotide.

Thus for a single site the resulting rate matrix must have the following form:

$$Q_{\mathrm{JC69}} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{G} \\ \mathtt{T} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array} \\ \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \end{array} \tag{2.25}$$

The transition probabilities for the Jukes Cantor model can be explicitly calculated, computing the exponential of the matrix:

$$P_{\mathrm{JC69}}(t) = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{G} \\ \mathtt{T} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array} \\ \begin{pmatrix} 1-3a(t) & a(t) & a(t) & a(t) \\ a(t) & 1-3a(t) & a(t) & a(t) \\ a(t) & a(t) & 1-3a(t) & a(t) \\ a(t) & a(t) & a(t) & 1-3a(t) \end{pmatrix} \end{array} \tag{2.26}$$

Where the function $a(t)$ is the following:

$$a(t) = \frac{1-3e^{-4\alpha t}}{4} \tag{2.27}$$

It is very simple to check that this evolutionary model has a uniform equilibrium distribution:

$$\pi_{\mathrm{JC69}} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \tag{2.28}$$

So the Jukes-Cantor model, although very appealing in its simplicity, immediately shows its shortcomings, as a simple count of nucleotide frequencies in available genomes shows distributions different from the uniform one. Ideally a probabilistic model of evolution should incorporate as much biochemical information as possible. This was the reason for the successor model of JC69, Kimura's K80 model [31]. Based on the analysis of the similarity among the chemical structure of the four different nucleotides, Kimura came to the conclusion that a purine should be more likely to turn into another purine than into a pyrimidine , and equivalently pyrimidines should be more likely to mutate into one another than into purines. He thus proposed a model with two free parameters:

$$Q_{K80} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{G} \\ \mathtt{T} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array} \\ \begin{pmatrix} -\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha-2\beta & \beta & \beta \\ \beta & \beta & -\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha-2\beta \end{pmatrix} \end{array} \tag{2.29}$$

As for the Jukes Cantor model we can find an explicit formula for the transition probabilities:

$$P_{K80}(t) = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{G} \\ \mathtt{T} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \end{array} \\ \begin{pmatrix} 1-a(t)-2b(t) & a(t) & b(t) & b(t) \\ a(t) & 1-a(t)-2b(t) & b(t) & b(t) \\ b(t) & b(t) & 1-a(t)-2b(t) & a(t) \\ b(t) & b(t) & a(t) & 1-a(t)-2b(t) \end{pmatrix} \end{array} \tag{2.30}$$

Were the function $a(t)$ and $b(t)$ are the following ones:

$$\begin{aligned} a(t) &= \frac{2c(t)-d(t)}{4} & c(t) &= 1-e^{-2t\alpha+\beta} \\ b(t) &= \frac{c(t)}{4} & d(t) &= 1-e^{-4t\beta} \end{aligned} \tag{2.31}$$

As in the case of JK69, it is simple to check that the equilibrium distribution of the Kimura model is the uniform one:

$$\pi_{\mathrm{K80}} = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right) \tag{2.32}$$

## 2.7 The general time reversible model

Following the models of Jukes, Cantor and Kimura there was a plethora of other models. In fact the most general independent site evolution rate matrix, has twelve different parameters [46]:

$$
Q_{12} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \\ \hline \mathtt{A} & - & q_{\mathtt{GA}} & q_{\mathtt{TA}} & q_{\mathtt{CA}} \\ \mathtt{G} & q_{\mathtt{AG}} & - & q_{\mathtt{TG}} & q_{\mathtt{CG}} \\ \mathtt{T} & q_{\mathtt{AT}} & q_{\mathtt{GT}} & - & q_{\mathtt{CT}} \\ \mathtt{C} & q_{\mathtt{AC}} & q_{\mathtt{GC}} & q_{\mathtt{TC}} & - \end{array} \tag{2.33}
$$

However due to historical reasons, trying to reduce the number of free parameters people used less general models. In fact the vast majority of the models used are all nested into the *General Time Reversible model* (GTR) [33, 55]. In order to derive it one can use a particular representation of a rate matrix:

$$
Q = \mathcal{D}(\pi)\Pi = \begin{pmatrix} \pi_{\mathtt{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathtt{G}} & 0 & 0 \\ 0 & 0 & \pi_{\mathtt{T}} & 0 \\ 0 & 0 & 0 & \pi_{\mathtt{C}} \end{pmatrix} \begin{pmatrix} - & a & b & c \\ g & - & d & e \\ h & i & - & f \\ j & k & l & - \end{pmatrix} \tag{2.34}
$$

It can be checked that a matrix with such representation has $(\pi_{\mathtt{A}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}}, \pi_{\mathtt{C}})^t$ as equilibrium distribution. So it is possible to derive a time reversible model just by choosing a symmetric matrix as the second factor in 2.34:

$$
Q_{GTR} = \begin{pmatrix} \pi_{\mathtt{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathtt{G}} & 0 & 0 \\ 0 & 0 & \pi_{\mathtt{T}} & 0 \\ 0 & 0 & 0 & \pi_{\mathtt{C}} \end{pmatrix} \begin{pmatrix} - & a & b & c \\ a & - & d & e \\ b & d & - & f \\ c & e & f & - \end{pmatrix} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \\ \hline \mathtt{A} & - & a\pi_{\mathtt{A}} & b\pi_{\mathtt{A}} & c\pi_{\mathtt{A}} \\ \mathtt{G} & a\pi_{\mathtt{G}} & - & d\pi_{\mathtt{G}} & e\pi_{\mathtt{G}} \\ \mathtt{T} & b\pi_{\mathtt{T}} & d\pi_{\mathtt{T}} & - & f\pi_{\mathtt{T}} \\ \mathtt{C} & c\pi_{\mathtt{C}} & e\pi_{\mathtt{C}} & f\pi_{\mathtt{C}} & - \end{array} \tag{2.35}
$$

It is particularly simple to check the detailed balance using this representation:

$$
(\Pi Q_{GTR})^t = (\Pi\mathcal{D}(\pi)\Pi)^t = \Pi^t\mathcal{D}(\pi)^t\Pi^t = \Pi Q_{GTR} \tag{2.36}
$$

This is the most general time reversible model. By putting constraint on the parameters we can recover all the model proposed and used over the years.

With $a = b = c = d = e = f = 1$ we get the model proposed by Felsenstein [18] to improve over the Kimura two parameter and get a non uniform base distribution:

$$Q_{\text{F81}} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \\ \hline \mathtt{A} & - & \pi_{\mathtt{A}} & \pi_{\mathtt{A}} & \pi_{\mathtt{A}} \\ \mathtt{G} & \pi_{\mathtt{G}} & - & \pi_{\mathtt{G}} & \pi_{\mathtt{G}} \\ \mathtt{T} & \pi_{\mathtt{T}} & \pi_{\mathtt{T}} & - & \pi_{\mathtt{T}} \\ \mathtt{C} & \pi_{\mathtt{C}} & \pi_{\mathtt{C}} & \pi_{\mathtt{C}} & - \end{array} \tag{2.37}$$

With $a = f := \alpha$ and $c = d = e = g := \beta$ the HKY model [21]

$$Q_{\text{HKY85}} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \\ \hline \mathtt{A} & - & \alpha\pi_{\mathtt{A}} & \beta\pi_{\mathtt{A}} & \beta\pi_{\mathtt{A}} \\ \mathtt{G} & \alpha\pi_{\mathtt{G}} & - & \beta\pi_{\mathtt{G}} & \beta\pi_{\mathtt{G}} \\ \mathtt{T} & \beta\pi_{\mathtt{T}} & \beta\pi_{\mathtt{T}} & - & \alpha\pi_{\mathtt{T}} \\ \mathtt{C} & \beta\pi_{\mathtt{C}} & \beta\pi_{\mathtt{C}} & \alpha\pi_{\mathtt{C}} & - \end{array} \tag{2.38}$$

With $a = f := \alpha$, $c = d = e = g := \beta$ and $\pi_{\mathtt{A}} = \pi_{\mathtt{T}} := \frac{\pi_{\mathtt{AT}}}{2}$, $\pi_{\mathtt{C}} = \pi_{\mathtt{G}} = \frac{\pi_{\mathtt{GC}}}{2} = 1 - \frac{\pi_{\mathtt{AT}}}{2}$ the Tamura model [53]

$$Q_{\text{T92}} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \\ \hline \mathtt{A} & - & \alpha\pi_{\mathtt{AT}} & \beta\pi_{\mathtt{AT}} & \beta\pi_{\mathtt{AT}} \\ \mathtt{G} & \alpha\pi_{\mathtt{GC}} & - & \beta\pi_{\mathtt{GC}} & \beta\pi_{\mathtt{GC}} \\ \mathtt{T} & \beta\pi_{\mathtt{AT}} & \beta\pi_{\mathtt{AT}} & - & \alpha\pi_{\mathtt{AT}} \\ \mathtt{C} & \beta\pi_{\mathtt{GC}} & \beta\pi_{\mathtt{GC}} & \alpha\pi_{\mathtt{GC}} & - \end{array} \tag{2.39}$$

Finally with $a := \alpha$, $e := \beta$ and $c = d = e = g := \gamma$ the Tamura-Nei model [54]

$$Q_{\text{TN93}} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{G} & \mathtt{T} & \mathtt{C} \\ \hline \mathtt{A} & - & \alpha\pi_{\mathtt{A}} & \gamma\pi_{\mathtt{A}} & \gamma\pi_{\mathtt{A}} \\ \mathtt{G} & \alpha\pi_{\mathtt{G}} & - & \gamma\pi_{\mathtt{G}} & \gamma\pi_{\mathtt{G}} \\ \mathtt{T} & \gamma\pi_{\mathtt{T}} & \gamma\pi_{\mathtt{T}} & - & \beta\pi_{\mathtt{T}} \\ \mathtt{C} & \gamma\pi_{\mathtt{C}} & \gamma\pi_{\mathtt{C}} & \beta\pi_{\mathtt{C}} & - \end{array} \tag{2.40}$$

## 2.8 The reverse complement symmetric model

A different and lesser known form of an evolutionary matrix has his origin in some studies conducted during the years 50 and 60s by biochemist Erwin Chargaff. In his works [10, 47] he showed some remarkable symmetry properties in the nucleotide composition of the DNA molecule. Based on his observations he could formulate two rules, later called by Sueoka [52] *Chargaff's parity rules.* The first rule expresses a property of the whole DNA double strand:

**Law 1. First Parity Rule**. *In a DNA sequence, indicating with $N_x$ the number of occurences of nucleotide $x$, the following holds $N_{\mathtt{A}} = N_{\mathtt{T}}$ and $N_{\mathtt{G}} = N_{\mathtt{C}}$.*

The reason for the first parity rule was found shortly thereafter [58], it is the direct consequence of the double helix structure of DNA and of the Watson-Crick base pairing. The second rule, refining the first one, specifies a property of single strands:

**Law 2. Second Parity Rule**. *The first rule also holds, but only in an approximate sense, for single stranded DNA. In other words the following approximate equalities hold for a single strand: $N_{\mathtt{A}} \simeq N_{\mathtt{T}}$ and $N_{\mathtt{G}} \simeq N_{\mathtt{C}}$.*

The reason for this second rule turned out to be quite elusive, and the underlying mechanism was only found much later, and was proposed in the papers [62, 52, 34].

The essential point is that when describing the evolution of a DNA sequence we should try to incorporate as much information about the underlying biological processes as possible, as Kimura first showed when he reasoned about similarities in the chemical structure of nucleotides.

One fundamental fact that we should take into account is the double stranded nature of DNA, which has an immediate consequence: the two strands reciprocally influence their evolution. When we translate it in a probabilistic framework, the result of this interaction is summarized in the following (already suggested in [9]):

**Theorem 2.** *If the repair mechanism acts with equal efficiency on the leading and lagging strand, and if the mutation rate is equal on both strands, and if the probability of base $X$ turning into base $Y$ does not depend on the strand, then the most general single nucleotide mutation rate matrix has the following reverse complement symmetric form (RCS), with six parameters:*

$$Q_{RCS} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & r_{\mathtt{AC}} & r_{\mathtt{AG}} & r_{\mathtt{AT}} \\ r_{\mathtt{GT}} & \cdot & r_{\mathtt{CG}} & r_{\mathtt{CT}} \\ r_{\mathtt{CT}} & r_{\mathtt{CG}} & \cdot & r_{\mathtt{GT}} \\ r_{\mathtt{AT}} & r_{\mathtt{AG}} & r_{\mathtt{AC}} & \cdot \end{pmatrix} \end{array}. \tag{2.41}$$

*Proof.* In this proof I use the complementary operator, $*$, which exchanges a nucleotide with its Watson-Crick complement, e.g. $\mathtt{A}^* = \mathtt{T}$. It follows that "$*$" is a conjugation, so that $\alpha^{**} = \alpha$ where $\alpha \in \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$.

I will also use the following abbreviations:

$\alpha \to \beta$ = "We observe nucleotide $\alpha$ going into nucleotide $\beta$"
$(\alpha \to \beta)_{\mathbb{F}}$ = "There was a mutation on the forward strand"
$(\alpha \to \beta)_{\mathbb{R}}$ = "There was a mutation on the reverse strand"
$\mathbb{R}$ is repaired = "The repair mechanism repairs the base on the reverse strand"
$\mathbb{F}$ is repaired = "The repair mechanism repairs the base on the forward strand"

In order to get the rate matrix we need to calculate the following probability:

$$\begin{aligned}
p(\alpha \to \beta) &= p(((\alpha \to \beta)_{\mathbb{F}}, \mathbb{R}\text{is repaired}) \vee ((\alpha^* \to \beta^*)_{\mathbb{R}}, \mathbb{F}\text{is repaired})) \\
&= p((\alpha \to \beta)_{\mathbb{F}}, \mathbb{R} \text{ is repaired}) + p((\alpha^* \to \beta^*)_{\mathbb{R}}, \mathbb{F} \text{ is repaired})) \\
&= p(\mathbb{R} \text{ is repaired}|(\alpha \to \beta)_{\mathbb{F}})p((\alpha \to \beta)_{\mathbb{F}}) \\
&+ p(\mathbb{F} \text{ is repaired}|(\alpha^* \to \beta^*)_{\mathbb{R}})p((\alpha^* \to \beta^*)_{\mathbb{R}})
\end{aligned}$$

On the other hand the probability of the transition between the complementary nucleotides to the first two is given by:

$$\begin{aligned}
p(\alpha^* \to \beta^*) &= p(\mathbb{R} \text{ is repaired}|(\alpha^* \to \beta^*)_{\mathbb{F}})p((\alpha^* \to \beta^*)_{\mathbb{F}}) \\
&+ p((\mathbb{F} \text{ is repaired}|(\alpha^{**} \to \beta^{**})_{\mathbb{R}})p((\alpha^{**} \to \beta^{**})_{\mathbb{R}}) \\
&= p(\mathbb{R} \text{ is repaired}|(\alpha^* \to \beta^*)_{\mathbb{F}})p((\alpha^* \to \beta^*)_{\mathbb{F}}) \\
&+ p((\mathbb{F} \text{ is repaired}|(\alpha \to \beta)_{\mathbb{R}})p((\alpha \to \beta)_{\mathbb{R}})
\end{aligned}$$

Equating we get:

$$\begin{aligned}
p(\alpha \to \beta) &= p(\mathbb{R} \text{ is repaired}|(\alpha \to \beta)_{\mathbb{F}})p((\alpha \to \beta)_{\mathbb{F}}) \\
&+ p(\mathbb{F} \text{ is repaired}|(\alpha^* \to \beta^*)_{\mathbb{R}})p((\alpha^* \to \beta^*)_{\mathbb{R}}) \\
&= p((\mathbb{F} \text{ is repaired}|(\alpha \to \beta)_{\mathbb{R}})p((\alpha \to \beta)_{\mathbb{R}}) \\
&+ p(\mathbb{R} \text{ is repaired}|(\alpha^* \to \beta^*)_{\mathbb{F}})p((\alpha^* \to \beta^*)_{\mathbb{F}}) \\
&= p(\alpha^* \to \beta^*)
\end{aligned}$$

Now we can use our hypotheses. First if there is no mutational bias, then the probability of a mutation happening on the forward strand is equal to the same probability on the reverse strand $p((\alpha \to \beta)_{\mathbb{R}}) = p((\alpha \to \beta)_{\mathbb{F}})$. Second if there is no bias in the repair mechanism, with an analogous reasoning we get: $p(\mathbb{R} \text{ is repaired}|(\alpha \to \beta)_{\mathbb{F}}) = p((\mathbb{F} \text{ is repaired}|(\alpha \to \beta)_{\mathbb{R}})$.

And we get the hypothesis $p(\alpha \to \beta) = p(\alpha^* \to \beta^*)$. □

We now prove the extension of the theorem to substitution matrices:

**Proposition 4.** *For neutrally evolving region, if the assumptions of theorem (2) hold,* *then the most general substitution matrix is:*

$$Q_{RCS} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & r_{\mathtt{AC}} & r_{\mathtt{AG}} & r_{\mathtt{AT}} \\ r_{\mathtt{GT}} & \cdot & r_{\mathtt{CG}} & r_{\mathtt{CT}} \\ r_{\mathtt{CT}} & r_{\mathtt{CG}} & \cdot & r_{\mathtt{GT}} \\ r_{\mathtt{AT}} & r_{\mathtt{AG}} & r_{\mathtt{AC}} & \cdot \end{pmatrix} \end{array} \tag{2.42}$$

*Proof.* A substitution is then the result of a mutation and a fixation. As before we will use an abbreviation:

$\boldsymbol{\alpha} \rightarrow \boldsymbol{\beta}$ =“we observe a nucleotide substitution from $\alpha$ to $\beta$”

In this case we use a boldface to distinguish the substitution process, $\boldsymbol{\alpha} \rightarrow \boldsymbol{\beta}$, which takes place at a species level from the mutational process, $\alpha \rightarrow \beta$, which takes place at individual level.

We can then write for the probabilities:

$$\begin{aligned} p(\boldsymbol{\alpha} \rightarrow \boldsymbol{\beta}) &= p(\alpha \rightarrow \beta, \text{fix}) \\ &= p(\text{fix}|\alpha \rightarrow \beta)p(\alpha \rightarrow \beta) \end{aligned}$$

While for the complementary substitution:

$$\begin{aligned} p(\boldsymbol{\alpha}^* \rightarrow \boldsymbol{\beta}^*) &= p(\alpha^* \rightarrow \beta^*, \text{fix}) \\ &= p(\text{fix}|\alpha^* \rightarrow \beta^*)p(\alpha^* \rightarrow \beta^*) \\ &= p(\text{fix}|\alpha^* \rightarrow \beta^*)p(\alpha \rightarrow \beta) \end{aligned}$$

When the sequence is neutrally evolving the probability of fixation becomes independent of the particular nucleotide at the site:

$$p(\text{fix}|\alpha \rightarrow \beta) = p(\text{fix}|\alpha^* \rightarrow \beta^*)$$

□

It can be checked by substitution that the generator (Eq. 2.42) has the following equilibrium probabilities:

$$\pi_{\text{RCS}} = \begin{pmatrix} \pi_{\mathtt{AT}} \\ \pi_{\mathtt{GC}} \\ \pi_{\mathtt{GC}} \\ \pi_{\mathtt{AT}} \end{pmatrix} \tag{2.43}$$

With $\pi_{\mathtt{GC}} = \frac{r_{\mathtt{CT}}+r_{\mathtt{GT}}}{r_{\mathtt{AC}}+r_{\mathtt{AG}}+r_{\mathtt{GT}}+r_{\mathtt{CT}}}$, and $\pi_{\mathtt{AT}} + \pi_{\mathtt{GC}} = 1$. This proves the following proposition:

**Proposition 5.** *Under the hypothesis of proposition (4) at equilibrium the following equalities hold on a single strand:* $P(\mathtt{A}) = P(\mathtt{T})$ *and* $P(\mathtt{G}) = P(\mathtt{C})$.

This explains why the approximate equalities of Chargaff's second law hold, however the equalities do not hold if selection is present. Deviation from the second parity rule can be detected using AT and GC skews:

$$\begin{aligned} \mathtt{AT}_{\text{skew}} &= \frac{\mathtt{A}-\mathtt{T}}{\mathtt{A}+\mathtt{T}} \\ \mathtt{GC}_{\text{skew}} &= \frac{\mathtt{G}-\mathtt{C}}{\mathtt{G}+\mathtt{C}} \end{aligned} \tag{2.44}$$

## 2.9 The time reversible RCS model

One important feature of the RCS model, is that while incorporating more information on the structure of the DNA molecule, it does not assume time reversibility, which means it is not nested in the GTR model as can be seen in Fig. 2.5. There is however a subset of it's parameter space it has the property of time reversibility. This feature will be important when we will need to check whether the evolution of real genomes is time reversible or not.

The subclass of the time reversible RCS matrices can be obtained by imposing the detailed balance conditions on its parameters, and we

**Proposition 6.** *Under the assumptions of the RCS and of time reversibility, the most general rate matrix is the following:*

$$\begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} - & c\pi_{\mathtt{AT}} & a\pi_{\mathtt{AT}} & b\pi_{\mathtt{AT}} \\ c\pi_{\mathtt{GC}} & - & d\pi_{\mathtt{GC}} & a\pi_{\mathtt{GC}} \\ a\pi_{\mathtt{GC}} & d\pi_{\mathtt{GC}} & - & c\pi_{\mathtt{GC}} \\ b\pi_{\mathtt{AT}} & a\pi_{\mathtt{AT}} & c\pi_{\mathtt{AT}} & - \end{pmatrix} \end{array} \tag{2.45}$$

*Where* $a, b, c, d$ *are independent parameters and* $\pi_{\mathtt{GC}} + \pi_{\mathtt{AT}} = 1$.

*Proof.* One just need to inpose the equality of complementary rates in the detailed balance

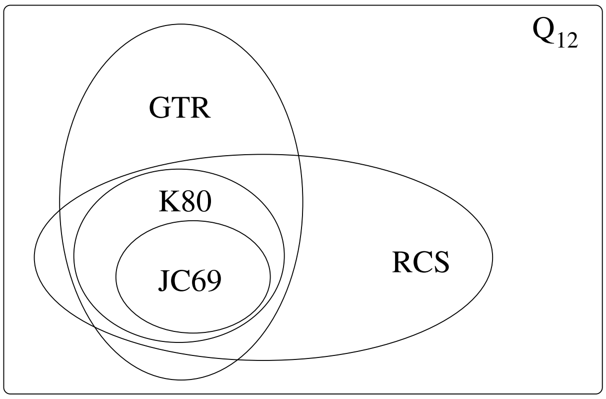

**Figure 2.5:** Here it's shown the hierarchical nesting of different evolutionary models. The GTR and the RCS overlap but none of them is nested into the other. They both include some of the simpler models and both are included in the most general 12 parameters model.

conditions:

$$\begin{array}{lll} q_{\mathtt{CA}} = q_{\mathtt{GT}} & \Rightarrow & c\pi_{\mathtt{A}} = d\pi_{\mathtt{T}} \\ q_{\mathtt{GA}} = q_{\mathtt{CT}} & \Rightarrow & a\pi_{\mathtt{A}} = f\pi_{\mathtt{T}} \\ q_{\mathtt{TA}} = q_{\mathtt{AT}} & \Rightarrow & b\pi_{\mathtt{A}} = b\pi_{\mathtt{T}} \\ q_{\mathtt{AC}} = q_{\mathtt{TG}} & \Rightarrow & c\pi_{\mathtt{C}} = d\pi_{\mathtt{G}} \\ q_{\mathtt{GC}} = q_{\mathtt{CG}} & \Rightarrow & e\pi_{\mathtt{C}} = e\pi_{\mathtt{G}} \\ q_{\mathtt{TC}} = q_{\mathtt{AG}} & \Rightarrow & f\pi_{\mathtt{C}} = a\pi_{\mathtt{G}} \end{array}$$

From which follows the following conditions on the parameters, $c = d := \gamma$, $a = f := \alpha$, $b := \beta$, $e := \delta$, $\pi_{\mathtt{A}} = \pi_{\mathtt{T}} := \pi_{\mathtt{AT}}$, $\pi_{\mathtt{G}} = \pi_{\mathtt{C}} := \pi_{\mathtt{GC}} = 1 - \pi_{\mathtt{AT}}$: □

So we find out that the time reversible RCS model is a 5 parameter subclass of the general RCS model.

## 2.10 Evolution with neighbor dependencies

So far we have always used the assumption first introduced in the JC69 model, that the evolution of each nucleotide can be treated independently from all the others. This is apparently a good approximation, especially for non-coding sequences, where we don't expect the presence of phenotypical effects which could introduce correlations in the evolution of different nucleotides. However even if not functionally relevant there may still be biochemical mechanisms which couple nucleotides.

In particular, as we have seen previously, in the evolution of vertebrates one cannot disregard the so called deamination process. This is a process that causes the depletion of CpG nucleotide doublets.

As a result neighbor dependencies play a significant role in the evolution of vertebrate genomes [6, 25]. In this case it is favorable to take into account the CpG decay process as shown in [4]. This is because in presence of methylation a CpG dinucleotide has an increased mutation rate to a CpA dinucleotide due to the reaction described in [11]. The formalism used for describing sequence evolution must then be appropriately generalized.

First, in general the configuration space of a nucleotide sequence of length $N$, is the Cartesian product of single nucleotide states having $4^N$ possible configurations:

$$\mathcal{C} = s_1 \times \ldots \times s_N \qquad s_i = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}. \tag{2.46}$$

The nucleotide substitution process in this space will then be described by a $4^N \times 4^N$ rate matrix. If we assume site independence the generator can be written in the following form:

$$\mathbb{Q} = \sum_{k=1}^{N} \mathcal{Q}_k \tag{2.47}$$

Where each of the generators in the right hand side is a $4^N \times 4^N$ matrix acting on one nucleotide:

$$\mathcal{Q}_k = \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{k-1} \otimes Q \otimes \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{N-k}. \tag{2.48}$$

Here $\mathbb{I}$ is the $4 \times 4$ identity matrix and $Q$ is given in Eq. (2.33). In the rest of this section we will use the RCS parameterization for $Q$ (Eq. (2.41)).

The tensor, or Kronecker, product of matrices, is defined in the following way:

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \tag{2.49}$$

That this abstract formulation is equivalent to the usual one, can be easily proved using the following identity:

$$e^{A \otimes \mathbb{I} + \mathbb{I} \otimes B} = e^A \otimes e^B. \tag{2.50}$$

We have thus a formulation equivalent to the one used in the previous sections. However in this new formalism it is simple to add generators that couple neighboring nucleotides:

$$\mathbb{Q}^{\mathtt{CpG}} = \sum_{k=1}^{N} \mathcal{Q}_k + \sum_{k=1}^{N-1} \mathcal{Q}^{\mathtt{CpG}}_{k,k+1}. \tag{2.51}$$

The second sum in Eq. (2.51) represents nearest neighbor dependencies and has the following form:

$$\mathcal{Q}^{\mathtt{CpG}}_{k,k+1} = \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{k-1} \otimes Q^{\mathtt{CpG}} \otimes \underbrace{\mathbb{I} \otimes \ldots \otimes \mathbb{I}}_{N-k-1}. \tag{2.52}$$

$Q^{\mathtt{CpG}}$ is a $16 \times 16$ matrix which models transitions on dinucleotides. In order to include the CpG decay in the model we parameterize it as follows:

$$Q^{\mathtt{CpG}}_{\alpha'\beta'\,\alpha\beta} = \begin{cases} r_{\mathtt{CpG}} & \text{if } (\alpha'\beta'\,\alpha\beta) = (\mathtt{CA}\,\mathtt{CG}) \text{ or } (\alpha'\beta'\,\alpha\beta) = (\mathtt{TG}\,\mathtt{CG}) \\ -2r_{\mathtt{CpG}} & \text{if } (\alpha'\beta'\,\alpha\beta) = (\mathtt{CG}\,\mathtt{CG}) \\ r^{\mathtt{rev}}_{\mathtt{CpG}} & \text{if } (\alpha'\beta'\,\alpha\beta) = (\mathtt{CG}\,\mathtt{CA}) \text{ or } (\alpha'\beta'\,\alpha\beta) = (\mathtt{CG}\,\mathtt{TG}) \\ -r^{\mathtt{rev}}_{\mathtt{CpG}} & \text{if } (\alpha'\beta'\,\alpha\beta) = (\mathtt{CA}\,\mathtt{CA}) \text{ or } (\alpha'\beta'\,\alpha\beta) = (\mathtt{TG}\,\mathtt{TG}) \\ 0 & \text{otherwise,} \end{cases} \tag{2.53}$$

where $r_{\mathtt{CpG}}$ is the rate of CpG decay substitutions CG → CA and CG → TG, and $r^{\mathtt{rev}}_{\mathtt{CpG}}$ is the rate of the corresponding back substitutions. This way we constructed a $4^N \times 4^N$ rate matrix $\mathcal{Q}$, while the corresponding transition probability matrix $\mathcal{P} = \exp \mathcal{Q}$ is computed by matrix exponentiation.

However the size of the resulting matrix is too big, and this renders the computation of the exponential unfeasible. We thus applied the cluster approximation described in [4]:

$$P(S_1 \to S_2|\mathcal{Q}) \simeq \prod_k P(S_1[k-1]S_1[k]S_1[k+1] \to *S_2[k] * |\mathcal{Q}) \tag{2.54}$$

Where we have used the following notation:

$$P(S_1[k-1]S_1[k]S_1[k+1] \to *S_2[k]*) = \sum_{i,j\in\{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}} [\exp \mathcal{Q}]_{S_1[k-1]S_1[k]S_1[k+1],i\;S_2[k]\;j} \tag{2.55}$$

After the cluster approximation we have to calculate the exponential of matrices of size $4^3 = 64$, and the problem is tractable again.

# Chapter 3

# Parameters Estimation Methods

*We have seen in detail how the theory of Markov processes provides us a powerful formalism to describe the evolution in time of genomic sequences. Using this formalism, given a sequence and set of evolutionary rates we can predict in a probabilistic sense the evolution of different quantities, like for example the average* GC *content of a sequence [6].*

*However in most cases we face the opposite problem. This is because we have no direct observations of how genomes evolve, for this would only be possible having historic series of fossil genomes at our disposal. Unfortunately the genome degrades rapidly once an organism dies, and so the acquisition of a sequence of even just one extinct genome proves to be an extraordinary task. Only recently it has been possible to obtain the sequence of two extinct animal species [20, 43, 41]. However interesting such an achievement is, this is not a viable way of approaching general evolutionary problems.*

*The kind of data we have at our disposal are instead sets of genomic sequences of different present day species. One of the problem we face, the one we will focus in this thesis, is how to reliably infer the evolutionary rates. In this chapter we will show how one may use a method know as maximum likelihood estimation (MLE) to accomplish this task. We will first introduce it as a general procedure, and then we will show a formulation specialized to the evolutionary case.*

*We will also show the differences of the maximum likelihood methodology between the time reversible and the non time reversible case.*

## 3.1 Markov processes on trees

First of all, it is important to note that so far we have been discussing a mathematical model apt to describe the evolution of the genomic sequence of single species lineage, as for example could be if we were studying evolution of the genome of the Homo Sapiens species from its last common ancestor with the Chimpanzee to the present day.

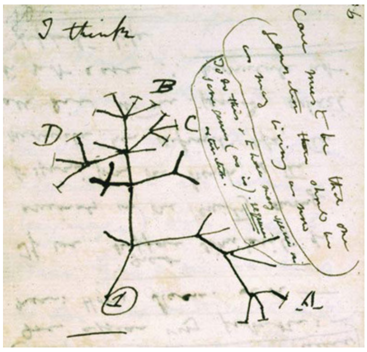

**Figure 3.1:** The famous phylogenetic tree sketched by Darwin in his notebook.

However it is a known fact, since Darwin first published the origin of species [14], that animal species, and thus genomes, do have evolutionary histories independent from one another. Rather the concept of *common descent* assumes that the evolution of species proceeds with a tree like structure, marked by specific events, known as *speciations*, which give rise to new species. The number of species increases exponentially by successive bifurcations of the existing lineages. A speciation event takes place when, a population of individuals is in a state in which it is composed of different subpopulations which because of environmental or geographical or social reasons stop sharing genetic material with each other. We have to note that broadly speaking this not a completely rigorous definition, as no species is ever completely reproductively isolated from other species. However if used with the necessary care this is a good definition for most practical cases.

So in order to translate mathematically the concepts of common descent we need to extend our framework, to include speciation events. We will then introduce the concept of phylogenetic tree and define an evolutionary Markov process on its branches.

Following [32] we give a the following recursive definition of tree:

**Definition 7.** *A Tree $\mathcal{T}$ on a set $A$ is a set of elements that can be one of the following:*

- *A leaf, containing an element $a \in A$;*
- *A branch, containing an element $a \in A$, called internal node and two trees, on the same set $A$, called left and right subtrees.*

In the evolutionary case one uses the following specialized definition:

**Definition 8.** *A phylogenetic tree $\mathcal{T}$ is a set of elements that can be one of the following:*

- *A leaf, containing a present day genomic sequence.*
- *A branch, containing an ancestral sequence, two trees, and two values corresponding to the evolutionary distance from the ancestral sequence to the subtrees.*

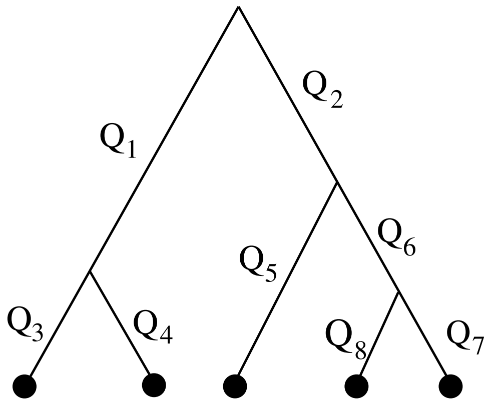

**Figure 3.2:** An example of phylogenetic Markov process. The tree represents 5 present day species, and 4 extinct ones, the common ancestors. Along its branches there are 8 different Markov processes acting, each representing the sequence evolutionary dynamics.

The recursive nature of trees makes very convenient to use a nested parenthesis format (also known as Newick format) to describe their structure. For example if we wanted to write down the tree describing the evolutionary relationships between Human (H), chimpanzee (C) and macaque (M), we could use the following notation:

$$(\text{"}HCM\text{"}(\text{"}HC\text{"}(\text{"}H\text{"})(\text{"}C\text{"}))(\text{"}M\text{"})) \tag{3.1}$$

In this case “HCM” is the root node of the three, the one at the top of the recursive hierarchy. In evolutionary terms it is named the *last common ancestor* of the three species.

The tree structure describes the history of successive speciations leading from extinct species to existing ones. Assuming an evolutionary tree is known, we can superimpose on each of its branches a Markov process, like the one we introduced in the first chapter. In this case we talk of a Phylogenetic Markov Process (Fig. 3.2):

**Definition 9.** *A phylogenetic Markov process is a tuple $(\mathcal{T}, Q_1, \ldots, Q_n)$ of a phylogenetic Tree $\mathcal{T}$, and $n$ Markov processes $Q_1, \ldots, Q_n$.*

We construct a phylogenetic Markov Process associating to each of the branches of a phylogenetic tree a rate matrix which recapitulates the evolutionary dynamic along that branch, according to the observations made in the previous chapter. In the majority of studies the same $Q$ matrix is used for all branches. However this is not in general a safe assumption, because it is equivalent to stating that the same kind of mutational processes and repair mechanism have been acting along the evolutionary history of different species. In general there will be a trade off, if one chooses to have only one rate matrix for all branches there will be less free parameters in the model, at the expense of a more realistic representation of the evolutionary processes.

## 3.2 The maximum likelihood approach

Maximum likelihood is a powerful paradigm, which can be used to infer parameters in probabilistic models. It has a straightforward derivation from Bayes theorem, which in turn is just a simple application of the law of product probabilities.

Let's suppose we have an hypothesis $\mathcal{H}$ and some experimental data $\mathcal{D}$, Bayes theorem is the following equality for probability of the hypothesis given the data:

$$P(\mathcal{H}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{H})P(\mathcal{H})}{P(\mathcal{D})} \tag{3.2}$$

Each of the factors on the right hand side has a standard name. $P(D|\mathcal{H})$, the probability of getting the data given that the hypothesis is true, is known as the *likelihood of the data*. $P(\mathcal{H})$ is called the *prior*, and encodes our belief about the validity of the hypothesis before having the possibility of looking at the data. Lastly, $P(D)$ is named *evidence* plays a fundamental role in Bayesian model comparison. In what follows we will only be concerned only in the first of these three factors, but should be noticed, that the other two are fundamental in Bayesian statistical theory.

The maximum likelihood approach follows from the observation that when the prior probabilities for different hypothesis have the same value, having no reason to favor one hypothesis over another, the hypothesis with the highest probability is the one which maximizes the likelihood of the data. That is, maximizing the likelihood we can find which hypothesis we should favor among the possible ones.

To make the concept clearer, I will show how Bayes theorem can be used to estimate the parameters of a simple toy example. Let's assume we toss a coin $n$ times, therefore having two possible outcomes, head and tails which we indicate with $H$ and $T$. The toss of a fair coin would have probability 0.5 for any of these two outcomes. If however, we had no information regarding the fairness of the coin, how could we quantify it, given a series of coin tosses?

In the most general case we may indicate with $0 \leq p \leq 1$ the probability of getting a tail on a coin toss, having a continuum of possible hypotheses. This example has as boundary cases $p = 0$, a coin that will always give a head, $p = 1$, one which will always give a tail, while $p = 0.5$ is the fair one.

We could try to apply maximum likelihood to see what is the probability $p$ of having $H$ as result. Let $N$ be the number of tosses, and $N_H$ the number of heads we have observed. Then the likelihood of the data is the binomial distribution:

$$P(N_H|p) \propto p^N (1-p)^{N-N_H} \tag{3.3}$$

Calculations are usualy done using the logarithm of the likelihood, the so called log-likelihood, which in this case is:

$$\log P(N_H|p) \propto N_H \log p + (N - N_H) \log(1 - p) \tag{3.4}$$

The maximum of the log likelihood is:

$$\frac{N_H}{p} - \frac{N - N_H}{1 - p} = 0 \Rightarrow p = \frac{N_H}{N} \tag{3.5}$$

This is what we expect intuitively, that the best estimate should be the fraction of heads observed, but we derived it using a general formalism applicable to any probabilistic problem.

However in general, if there is more than one variable, it is not possible to find the maximum of likelihood function analytically, and one has to resort to numerical maximization algorithms.

## 3.3 Maximum likelihood on a tree

Now that we have introduced this procedure, we can go back to solving the general problem stated in the introduction. Given a phylogenetic Markov process and an alignment of $n$ sequences $S_1 \ldots S_n$ how can we infer the parameters of the rate matrices $Q_1 \ldots Q_m$?

The solution [18] can be found in two steps, first we need to calculate the likelihood of the tree given the sequences with given rate matrices, then we will have to find which parameters of the rate matrices maximize it.

As first step, let's write down the probability assuming that we know the ancestral sequences. We assume that the tree has $n$ leaves, $m$ internal nodes and a root node that we label with 0. We also indicate with $Q_{i,j}$ the rate matrix which describes evolution on the tree branch connecting $i$ with $j$, while $\{Q_{i,j}\}$ is the set of all such matrices for every pair of nodes $(i, j)$ in the tree. Then we have

$$P(S_1, \ldots, S_n, S_{n+1}, \ldots, S_{n+m}|\{Q_{i,j}\}, \mathcal{T}) = \rho(S_0) \prod_{(i,j)} P(S_i \to S_j|Q_{i,j}) \tag{3.6}$$

To get the likelihood we now have to sum over all the internal states representing the unknown ancestral sequences.

$$\mathcal{L}(S_1, \ldots, S_n|\{Q_{i,j}\}, \mathcal{T}) = \sum_{S_0, S_{n+1}, \ldots, S_{n+m}} \rho(S_0) \prod_{(i,j)} P(S_i \to S_j|Q_{i,j}) \tag{3.7}$$

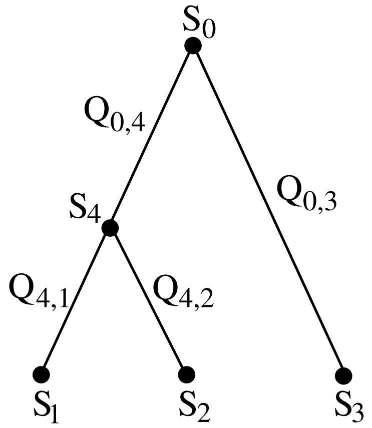

**Figure 3.3:** The three species phylogenetic three used in the text to illustrate Felsenstein's pruning algorithm.

In order to be clearer we also show the form of the likelihood, for a simple three species alignment:

$$
\begin{aligned}
\mathcal{L}(S_1, S_2, S_3|\{Q_{i,j}\}, \mathcal{T}) = \sum_{S_0,S_4} & \rho(S_0)P(S_0 \to S_4|Q_{0,4})P(S_0 \to S_3|Q_{0,3}) \\
& P(S_4 \to S_1|Q_{4,1})P(S_4 \to S_2|Q_{4,2})
\end{aligned}
\tag{3.8}
$$

## 3.4 The independent sites case: pruning algorithm

Now, if we make the further assumption that the sites in the sequence are evolving independently, it follows that the likelihood for an alignment, of length $l$, of $n$ sequences is the product of the likelihood at each single site:

$$
\mathcal{L} = \prod_{k=1}^{l} \mathcal{L}(\mathcal{S}_1[k], \ldots, S_n[k]|\{Q_{i,j}\}, \mathcal{T}) = \prod_{k=1}^{l} \mathcal{L}_k \tag{3.9}
$$

We are then left with the problem of calculating the likelihood of a tree at a single site. This can be accomplished starting from the root node and summing over all possible internal unknown states.

To avoid cumbersome notation, I will show the calculation in the case of a three species alignment, with the corresponding phylogenetic tree shown in Fig. 3.3.

$$\mathcal{L}_k = \mathcal{L}(\mathcal{S}_1[k], \mathcal{S}_2[k], \mathcal{S}_3[k] | \{Q_{i,j}\}, \mathcal{T}) = \sum_{j_0, j_4 \in \{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}} \rho(j_0) P(j_0 \to j_4 | Q_{0,4}) P(j_0 \to S_3[k] | Q_{0,3}) \\ P(j_4 \to S_1[k] | Q_{4,1}) P(j_4 \to S_2[k] | Q_{4,2}) \tag{3.10}$$

It is not difficult to realize that written as it is this summation is redundant, as several terms contains factors which are equal and should then be only computed once, and memorized for later reoccurences. Even in our simple case the term $P(j_4 \to S_1[k]|Q_{4,1})P(j_4 \to S_2[k]|Q_{4,2})$ as we sum over the index $j_0$ four different times. It is evident that as the number of species, and thus internal nodes, increases as a power of two, the number of such recomputations grows exponentially thus rendering a naive approach to the calculation of the likelihood impracticable.

In order to eliminate redundant summation steps, Felsenstein [18] devised a dynamic programming algorithm, the so called *pruning* algorithm. The idea is simple and consists just in shifting the summations over nucleotides at the internal nodes as far to the right as possible. This way after the pruning Eq. (3.10) has the form:

$$\mathcal{L}_k = \sum_{j_0 \in \{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}} \rho(j_0) P(j_0 \to j_4 | Q_{0,4}) P(j_0 \to S_3[k] | Q_{0,3}) \\ \sum_{j_4 \in \{\mathtt{A},\mathtt{G},\mathtt{C},\mathtt{T}\}} P(j_4 \to S_1[k] | Q_{4,1}) P(j_4 \to S_2[k] | Q_{4,2}) \tag{3.11}$$

The procedure can be also formulated more formally, introducing the notion of postorder traversal of a tree, a procedure which visits all the nodes in the tree and execute some specific action on them:

```
postorder (root)
for all $k$ child of root do
  postorder($k$)
  action($k$)
end for
```

The characteristic of a postorder traversal is that it executes the required action starting from the leaves. Using this particular traversal, we can reformulate Felsenstein pruning algorithm introducing the notion of conditional likelihood for a given sub-tree, which is simply the likelihood of the sub-tree when we fix the nucleotide at its root node. Indicating with $i \in \{\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T}\}$ the nucleotide present at the root node of the sub-tree we are considering, we define:

$$\mathcal{L}_k(i) = \prod_{k=1}^{l} \mathcal{L}(\mathcal{S}_1[k], \ldots, S_n[k] | \{Q_{i,j}\}, i, \mathcal{T}) \tag{3.12}$$

Then Felsenstein pruning algorithm can be formulated as follows:

1. The conditional likelihood of a leaf $j$ is:

$$\mathcal{L}_k(i) = \delta_{(i,S_j[k])} \tag{3.13}$$

2. The conditional likelihood of an internal node is:

$$\mathcal{L}_k(i) = \sum_{j \in \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} \prod_{j \in \text{children} i} P(i \to j | Q_{i,j}) \mathcal{L}_k(j) \tag{3.14}$$

3. The total likelihood of the tree is:

$$\mathcal{L}_k = \sum_{j_0 \in \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} \prod_{j \in \text{children } j_0} \rho(j_0) P(j_0 \to j | Q_{j_0,j}) \mathcal{L}_k(j_0) \tag{3.15}$$

The algorithm can then be implemented with a post-order traversal of the tree which at each step of the traversal computes the conditional likelihood according to the rules given above.

## 3.5 Equilibrium and time reversibility in the maximum likelihood procedure

As already anticipated traditionally maximum likelihood methods used in molecular evolution studies rely on the assumptions of time reversibility and equilibrium. We will clarify these two concepts one after the other.

First of all let's assume that the process has the same $Q$ matrix on all the branches, and also let's assume that the nucleotide distribution is at equilibrium, with respect to $Q$ in every point of the phylogeny. This allows us to rewrite Eq. (3.7) using as probability of the root sequence, a product of equilibrium frequencies, $\rho(S_0) = \prod_k \pi_{S_0[k]}$, where each frequency computed by calculating how often the given nucleotide appears in the present day sequences $S1, \ldots, S_n$. We have then:

$$\mathcal{L}(S_1, \ldots, S_n | Q, \mathcal{T}) = \sum_{S_0, S_{n+1}, \ldots, S_{n+m}} (\prod_k \pi_{S_0[k]}) \prod_{(i,j)} P(S_i \to S_j | Q) \tag{3.16}$$

Similarly Eq. (3.15) becomes:

$$\mathcal{L}_k = \prod_{j \in \text{children } j_0} \sum_{j_0 \in \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} \pi_{S_0[j]} P(j_0 \to j|Q) \mathcal{L}_k(j_0) \tag{3.17}$$

The second fundamental simplification is that, as intuitively obvious, in order to calculate the likelihood of a tree under a time reversible model it is no longer necessary to start from the root proceeding forward in time to calculate the likelihood of the branches representing new species.

It is as well possible to start the computation from any leaf proceeding backward and forward in time until the likelihood of the whole tree has been computed . In other words supposing as an example that we wanted to calculate the likelihood of an homologous sequences of human and chimp. Instead of having the sequence evolving forward in time from the common ancestor to human on one branch and to chimp on the other, we would obtain the same result having the sequence evolving backward in time

This can be formulated more precisely mathematically:

$$\begin{aligned}
\mathcal{L}_k = \mathcal{L}(\mathcal{S}_1[k], \mathcal{S}_2[k], \mathcal{S}_3[k]|Q, \mathcal{T}) = & \sum_{j_0, j_4 \in \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} \rho(j_0) P(j_0 \to j_4|Q) P(j_0 \to S_3[k]|Q) \\
& P(j_4 \to S_1[k]|Q) P(j_4 \to S_2[k]|Q) \\
= & \sum_{j_0, j_4 \in \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} \rho(j_4) P(j_4 \to j_0|Q) P(j_0 \to S_3[k]|Q) \\
& P(j_4 \to S_1[k]|Q) P(j_4 \to S_2[k]|Q) \\
= & \sum_{j_4 \in \{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} \rho(j_4) P(j_4 \to S_3[k]|Q) \\
& P(j_4 \to S_1[k]|Q) P(j_4 \to S_2[k]|Q)
\end{aligned} \tag{3.18}$$

Where we have used detailed balance condition and the Kolmogorov property of a Markov process. Analogously one can shift the root of the tree on other nodes.

## 3.6 Maximum Likelihood with Neighbor Dependencies

As we have see in section 2.10 the `CpG` decay process in vertebrates requires the introduction of neighbor dependent Markov models of sequence evolution. Unfortunately in this case we cannot rely anymore on the factorization in Eq. (3.9). We could still in principle use the pruning algorithm, but the number of intermediate states at internal nodes over which we should sum over would be $4^l$, making the computation infeasible.

In the following we will assume that the time dynamics is given by neighbor independent nucleotide substitutions and nearest neighbor dependent substitutions only. The corresponding generator is given in Eq. (2.51). The transition probability matrix is then $\mathcal{P}^{ji} = \exp(t\mathcal{Q}^{ji})$. Without loss of generality we again set $t = 1$.

To maximize the likelihood in Eq. (3.7) we introduce a mixed Monte-Carlo Maximum-Likelihood (MCML) approach, which combines elements of the two methods in a very efficient way: In an iterative fashion we will first (M-step) estimate substitution frequencies for a given ancestral sequence at internal nodes (using a maximum likelihood approach) and then (E-step) get a new estimate for the sequence at internal nodes for given substitution frequencies (using a Monte Carlo approach). This algorithm actually falls into the class of stochastic Expectation Maximization (EM) algorithms [38].

The iteration is initialized setting the sequences at the internal nodes to be the consensus of all its descendant sequences. If nucleotides at one position are not equal in all descendant sequences one of them is chosen at random. Initializing with a random sequence prolongs but not prevents the convergence of the algorithm to the maximum.

In the *M-step*, for each branch of the phylogeny the substitution frequencies (including those for neighbor dependent processes)are estimated from comparisons of ancestral and daughter sequences as described in [4]. In practice the method relies on the cluster decomposition to compute, exponentiating the rate matrix, the probability of going from the ancestral to the daughter sequence:

$$\begin{aligned} P(S_i \to S_j | S_i, S_j, \mathbb{Q}_{i,j}) &= \prod_k P(S_1[k-1]S_1[k]S_1[k+1] \to *S_2[k]* |\mathbb{Q}) \\ P(S_1[k-1]S_1[k]S_1[k+1] \to *S_2[k]*) &= \sum_{i,j\in\{\texttt{A},\texttt{G},\texttt{C},\texttt{T}\}} [\exp \mathbb{Q}]_{(S_1[k-1]S_1[k]S_1[k+1],i\ S_2[k]\ j)} \end{aligned} \tag{3.19}$$

The rate matrix is then varied until the maximum of the likelihood is found.

In the *E-step* then, we update the ancestral sequences at the internal nodes. To do this we make use of a Monte Carlo procedure. We first consider the internal sequence $S_4$. For each position $k = 1, \dots, l$ we propose to update the nucleotide $S_4[k]$ by another nucleotide $S_4[k]'$. The newly proposed nucleotide is accepted with some probability, which is computed using a four nodes likelihood which gives the probability of finding a given nucleotide at a given position of the ancestral sequence given the sequences at the parent node and at the two children nodes.

$$\begin{aligned} \mathbb{L}^4_k(S_4[k-1]S_4[k], S_4[k]) = \quad & P(S_0[k-1]S_0[k]S_0[k] \to S_4[k-1]S_4[k]S_4[k]|S_0, S_4, \mathbb{Q}_{0,4})\times \\ & P(S_1[k-1]S_1[k]S_1[k] \to S_4[k-1]S_4[k]S_4[k]|S_1, S_4, \mathbb{Q}_{1,4})\times \\ & P(S_2[k-1]S_2[k]S_2[k] \to S_4[k-1]S_4[k]S_4[k]|S_2, S_4, \mathbb{Q}_{2,4}) \end{aligned} \tag{3.20}$$

where the probabilities $P(\alpha_1\alpha_2\alpha_3 \rightarrow \beta_1\beta_2\beta_3\,|\alpha, \beta, \mathbb{Q}_{\alpha,\beta})$ of substitutions of three consecutive nucleotides $\alpha_1\alpha_2\alpha_3$ on node $i$ to $\beta_1\beta_2\beta_3$ on node $j$ are given as matrix element of the $4^3 \times 4^3$ dimensional transition probability matrix $\mathcal{P}^{ji} = \exp \mathcal{Q}^{ji}$ describing the time evolution on $N = 3$ sites with $\mathcal{Q}^{ji}$ given by Eq. (2.51). The substitution frequencies along each branch, which fix the corresponding matrices $\mathcal{Q}^{ji}$, are taken from the estimates in the previous M-step. An update $S_4[k] \rightarrow S_4[k]'$ is always accepted if the likelihood increases, i.e. if the likelihood ratio

$$\lambda = \mathbb{L}^4_k(S_4[k-1]S_4[k], S_4[k])/\mathbb{L}^4_k(S_4[k-1]S_4[k]', S_4[k]) \tag{3.21}$$

is larger than one. If this ratio is smaller than one the substitution is accepted with probability $\lambda$. In this case the (local) likelihood is decreased in order to increase the (global) likelihood in the following M-step.

After the entire internal sequence $S_4$ is updated, the sequence on the root node $S_0$ is updated in a similar fashion. Only the definition of the local likelihood differs and now involves the trinucleotide distribution $\rho(S_0[k-1]S_0[k], S_0[k])$ of the ancestral sequence $S_0$:

$$\begin{aligned}\mathbb{L}^0_k(S_0[k-1]S_0[k], S_0[k]) = \rho(S_0[k-1]S_0[k], S_0[k])&\\ \times P(S_0[k-1]S_0[k]S_0[k] \rightarrow S_4[k-1]S_4[k]S_4[k]|S_0, S_4, \mathbb{Q}_{0,4})&\\ \times P(S_0[k-1]S_0[k]S_0[k] \rightarrow S_3[k-1]S_3[k]S_3[k]|S_0, S_3, \mathbb{Q}_{0,3})&\end{aligned} \tag{3.22}$$

The trinucleotide distribution is assumed to be homogeneous along the sequence and is estimated from $S_0$ right before starting with the E-step. The transition probabilities are defined as above; substitution frequencies are given from the estimates in M-step.

This two E- & M-step iteration is performed several times until convergence of all the substitution frequencies and of the trinucleotide distribution $\rho(S_0[k-1]S_0[k], S_0[k])$ is established. In our applications this happens after about 40 iterations.

By the virtue of the Monte Carlo step, we allow that ancestral sites might not be in their *most* likely ancestral state. This is done by intention since such situations can actually be observed for sufficiently long sequences. The Monte Carlo step introduces such configurations into the ancestral sequence in as much as they are expected to occur with regard to the substitution model. This is crucial for the accurate estimation of substitution frequencies and ancestral single and di-nucleotide frequencies. Note that while the number of those sites that are not in their most likely state is given by the substitution models, their positions are not uniquely defined. Therefore, the ancestral sequence is one representative out of the set of sequences that maximize the likelihood. While for a general EM algorithm one would require to take the expectation over all possible ancestral sequences (or a sample of those for a Monte Carlo EM algorithm), we rely here on only one representative ancestral sequence. This is possible since the average over all positions along the sequence offer an implicit equivalent of the expectation. If only little amounts of sequence data is available a sampling over different realization of ancestral sequences can easily be incorporated into the MCML approach.

As mentioned above for the neighbor independent case, the substitution frequencies of edges connected to the root and the trinucleotide distribution of the ancestral sequence $S_0$ cannot be reconstructed. However, substitution frequencies in the two branches for the two sister species as well as the nucleotide distribution in the last common ancestor of the two sister species are not affected by this ambiguity. For more details and numerical verification of this approach see [16].

After maximizing the likelihood of a model for given data, the value of the likelihood can also be used to judge whether the use of particular parameterizations is indicated. We performed such a study for the fly data set. A comparison of the Reverse Complement Symmetric (RCS) model and the General Time Reversible (GTR) model, both of which have 6 free parameters along each branch, came out in favor of the RCS model. Models with more parameters (like the one in Eq. (2.33) with 12 independent parameters) or less parameters (like the HKY85 or JC69 model) compared less favorable to the RCS model when taking into account the total numbers of parameters using the Akaike information criterion (see Tab. 3.1).

| model | $\log L$ | AIC |
|---:|---|---|
| RCS | -88579.26(2) | 177212.52(1) |
| 12-parameter | -88570.44(1) | 177242.88(2) |
| HKY85 | -88612.49(4) | 177246.97(3) |
| GTR | -88602.97(3) | 177259.93(4) |
| K80 | -88647.21(5) | 177316.43(5) |
| JC69 | -88776.99(6) | 177567.97(6) |

**Table 3.1:** Comparison of different models of nucleotide substitutions for *D. simulans*. For each model we report the mean log likelihood,$\log L$, as well as the mean value of the Akaike information criterion AIC $= 2p - 2\log L$, where $p$ is the number of parameters of the respective models on the phylogenetic tree. Means are taken over the 539 windows used in the main text. Numbers in brackets report the rank of the corresponding model when sorted by decreasing $\log L$ or increasing AIC.

# Chapter 4

# Testing Reversibility and Equilibrium

*In the previous chapters we have presented Markov models of evolution that assume stationarity and time reversibility, and shown that it is possible to remove this assumption and still be capable of estimating the evolutionary rates. The question is how far the evolutionary properties of organisms are from the stationarity-reversibility assumption? In this chapter we will derive a set of statistical indices, in order to see how it is possible to answer such a question and we will apply them to the analysis of some animal genomes.*

## 4.1 Equilibrium conditions: the stationarity indices

The stationary, or equilibrium, state of the process is the probability distribution which does not evolve in time under the evolution defined in Eq. (2.11). It is usually denoted as $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})^t$, and it can be calculated solving the following system of linear equations:

$$Q\pi = 0. \tag{4.1}$$

The easiest way to check whether the process is stationary or not is to define the following indices, which quantify deviations of the present day nucleotide composition, $\rho$, from the equilibrium one, $\pi$:

$$\Delta_\alpha = \rho_\alpha - \pi_\alpha. \tag{4.2}$$

Due to the normalization constraint only three of these 4 equations are independent. If all of them are equal to zero, i.e. $\Delta_\alpha = 0 \, \forall \alpha$, then the process is in its stationary state.

It is important to note that checking for the equality of the nucleotide distribution at different leaf nodes is not a sufficient condition for equilibrium. As an example, all the sequences in the tree could be evolving from a `GC` rich state to a `GC` poor one with the same rate, in which case they would show the same nucleotide composition even if they are not in equilibrium. Our method does not have such inconveniences and quantifies equilibrium in the most precise way.

We can recast the conditions in a more insightful form if we take independent linear combinations of the $\Delta_\alpha$ in Eq. (4.2) and define:

$$\begin{aligned} \mathrm{STI}_1 &= \Delta_{\mathtt{C}} + \Delta_{\mathtt{G}} = \rho_{\mathtt{GC}} - \pi_{\mathtt{GC}} \\ \mathrm{STI}_2 &= \Delta_{\mathtt{A}} - \Delta_{\mathtt{T}} \\ \mathrm{STI}_3 &= \Delta_{\mathtt{C}} - \Delta_{\mathtt{G}}, \end{aligned} \tag{4.3}$$

which we call Stationarity Indices (STIs). The first index is just the difference between the actual GC content , $\rho_{\mathtt{GC}}$ (i.e. the frequencies of Gs and Cs present on one strand), and the equilibrium GC content, $\pi_{\mathtt{GC}}$. The second and third equations, are reminiscent of the AT skew and GC skew indices. A system is in its stationary state if all STIs vanish.

We further want to quantify whether deviations from zero of the three indices are statistically significant when only a finite amount of sequence data is available to measure the present day nucleotide distribution. To achieve this we compare the distribution of nucleotides, $\rho_\alpha$, of a sequence of length $N$ to the stationary distribution, $\pi_\alpha$, using a $\chi^2$-test with

$$\chi^2 = N \sum_\alpha \frac{(\rho_\alpha - \pi_\alpha)^2}{\pi_\alpha}\,. \tag{4.4}$$

This quantity follows a $\chi^2$ distribution with 3 degrees of freedom. Deviations from stationarity are significant (with 95% confidence) if $\chi^2 > 7.8147$.

## 4.2 Kolmogorov cycle conditions

We have seen in the first chapter a criterion to test whether a Markov process is reversible or not, we will now show an alternative formulation which we will use to develop a test for reversibility.

We will first show a proof for Markov chains (taken from the book of Kelly [29]). We define *irreducible*, a Markov chain or process where each state can be reached from any other state.

**Proposition 7.** *An irreducible and stationary Markov chain is reversible if and only if the transition probabilities satisfy the following conditions.*

$$P_{j_1,j_2}P_{j_2,j_3}\dots P_{j_{n-1},j_n}P_{j_n,j_1} = P_{j_1,j_n}P_{j_n,j_{n-1}}P_{j_{n-2},j_{n-3}}\dots P_{j_3,j_3}P_{j_2,j_1} \tag{4.5}$$

*For any possible choice of the indices $j_1 \dots j_n$*

*Proof.* Let's assume that the process is time reversible, then it must satisfy the detailed balance conditions:

$$\begin{aligned} P_{j_2,j_1}\pi_{j_1} &= P_{j_1,j_2}\pi_{j_2} \\ P_{j_3,j_2}\pi_{j_2} &= P_{j_2,j_3}\pi_{j_3} \\ &\vdots \\ P_{j_n,j_{n-1}}\pi_{j_{n-1}} &= P_{j_{n-1},j_n}\pi_{j_n} \\ P_{j_1,j_n}\pi_{j_n} &= P_{j_n,j_1}\pi_{j_1} \end{aligned}$$

And if we multiply all of the first sides and all of the second sides we get the Kolmogorov conditions.

Let's now assume the converse, that the process satisfies the Kolmogorov conditions. Then since the process is irreducible there is a sequence of states $(j_0, j_1, j_2, \dots, j)$ connecting any two states $j_0$ and $j$. Then let $B$ be a positive constant, and $\pi_j$ be defined by:

$$\pi_j = B\frac{P_{j_0,j_1}P_{j_1,j_2}\dots P_{j_n,j}}{P_{j,j_n}P_{j_n,j_{n-1}}\dots P_{j_1,j_0}}$$

$\pi_j$ is independent of the sequence chosen for the right hand side, as we can see choosing another sequence $(j_0, j_1', j_2', \dots, j)$ and applying Kolmogorov criteria:

$$\frac{P_{j_0,j_1}P_{j_1,j_2}\dots P_{j_n,j}}{P_{j,j_n}P_{j_n,j_{n-1}}\dots P_{j_1,j_0}} = \frac{P_{j_0,j_1'}P_{j_1',j_2'}\dots P_{j_n',j}}{P_{j,j_n'}P_{j_n',j_{n-1}'}\dots P_{j_1',j_0}} \tag{4.6}$$

Furthermore defining:

$$\pi_k = B\frac{P_{j_0,j_1}P_{j_1,j_2}\dots P_{j_n,k}}{P_{k,j_n}P_{j_n,j_{n-1}}\dots P_{j_1,j_0}}$$

We get:

$$P_{k,j}\pi_k = P_{j,k}\pi_j \tag{4.7}$$

So that we prove that $\pi$ is in fact the equilibrium distribution of the chain and that the detailed balance conditions hold, which proves the proposition. □

An equivalent proposition can be proven equivalently for Markov processes.

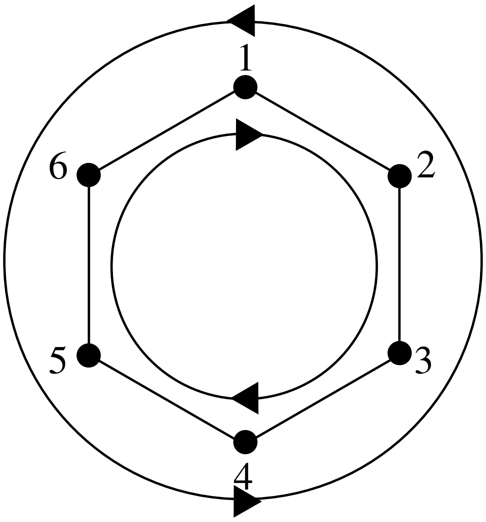

**Figure 4.1:** An example of a Kolmogorov cycle.

**Proposition 8.** *An irreducible and stationary Markov process is reversible if and only if the transition rates satisfy the following conditions.*

$$Q_{j_1,j_2}Q_{j_2,j_3}\dots Q_{j_{n-1},j_n}Q_{j_n,j_1} = Q_{j_1,j_n}Q_{j_n,j_{n-1}}\dots Q_{j_3,j_2}Q_{j_2,j_1} \tag{4.8}$$

*For any possible choice of the indices $j_1 \dots j_n$*

Intuitively the Kolmogorov criterion states that for a reversible system if we pick an arbitrary state $i$ of the chain and follow a path which eventually closes on $i$, then the probability of the path is the same regardless of the direction which we follow. One can say probability flux shows forms no vorticity in state space.

The usefulness of Kolmogorov criterion comes from thee fact that often, it is not necessary to test the reversibility of every possible cycle, but it is possible to show analytically that testing a subset of the cycles will test all of them.

## 4.3 Kolmogorov conditions for a four state process

**Proposition 9.** *If the off–diagonal coefficients of the rate matrix are strictly positive and if Kolmogorov conditions hold for 3–cycles then they hold for cycles of arbitrary length.*

*Proof.* From the positivity of the off–diagonal rate matrix coefficients, we can deduce the ergodicity of the process.

The rest of the proposition follows by induction. It holds trivially for two cycles and it holds by hypothesis for 3–cycles. Then let's show that if it holds for $n$-cycles then it is also valid for $(n+1)$-cycles. Let's assume we want to test whether the equality still holds for a chain which has element $i_{n+1}$ inserted between element $i_n$ and element $i_1$. We multiply both sides by the following factor $(Q_{i_n i_1} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n})$, obtaining:

$$\begin{aligned}&(Q_{i_n i_1} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n}) Q_{i_1 i_n} Q_{i_n i_{n-1}} \dots Q_{i_2 i_1} = \\ &\quad Q_{i_1 i_2} \dots Q_{i_{n-1} i_n} Q_{i_n i_1} (Q_{i_n i_1} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n}),\end{aligned} \tag{4.9}$$

which after applying Kolmogorov condition for 3–cycles and simplifying leads to:

$$Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n} Q_{i_n i_{n-1}} \dots Q_{i_2 i_1} = Q_{i_1 i_2} \dots Q_{i_{n-1} i_n} Q_{i_1 i_{n+1}} Q_{i_{n+1} i_n}. \tag{4.10}$$

So that the equality holds for $(n+1)$-cycles and the proposition is proven. □

We now restrict ourselves to a Markov process with only four states $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ and prove the following:

**Proposition 10.** *Given a four states Markov process with strictly positive rate matrix coefficients, if the conditions:*

$$Q_{\alpha\delta} Q_{\delta\gamma} Q_{\gamma\beta} Q_{\beta\alpha} = Q_{\alpha\beta} Q_{\beta\gamma} Q_{\gamma\delta} Q_{\delta\alpha}, \tag{4.11}$$

*hold for $(\alpha, \beta, \gamma, \delta)$ equal to $(\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T})$, $(\mathtt{A}, \mathtt{G}, \mathtt{T}, \mathtt{C})$ and $(\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T})$, (Fig. 4.2b), then Kolmogorov conditions hold for 3–cycles.*

*Proof.* It suffices to multiply the generators for the 4-cycles:

$$\begin{aligned}&(Q_{\mathtt{AT}} Q_{\mathtt{TC}} Q_{\mathtt{CG}} Q_{\mathtt{GA}})(Q_{\mathtt{AT}} Q_{\mathtt{TG}} Q_{\mathtt{GC}} Q_{\mathtt{CA}})(Q_{\mathtt{AC}} Q_{\mathtt{CT}} Q_{\mathtt{TG}} Q_{\mathtt{GA}}) = \\ &\quad (Q_{\mathtt{AG}} Q_{\mathtt{GC}} Q_{\mathtt{CT}} Q_{\mathtt{TA}})(Q_{\mathtt{AC}} Q_{\mathtt{CG}} Q_{\mathtt{GT}} Q_{\mathtt{TA}})(Q_{\mathtt{AG}} Q_{\mathtt{GT}} Q_{\mathtt{TC}} Q_{\mathtt{CA}}).\end{aligned} \tag{4.12}$$

Simplifying both sides and squaring we get the equivalence for one of the 3–cycles:

$$Q_{\mathtt{GA}} Q_{\mathtt{AT}} Q_{\mathtt{TG}} = Q_{\mathtt{GT}} Q_{\mathtt{TA}} Q_{\mathtt{AG}}. \tag{4.13}$$

It can be easily seen that exchanging factors between left and right hand side the remaining 3–cycles can be obtained. □

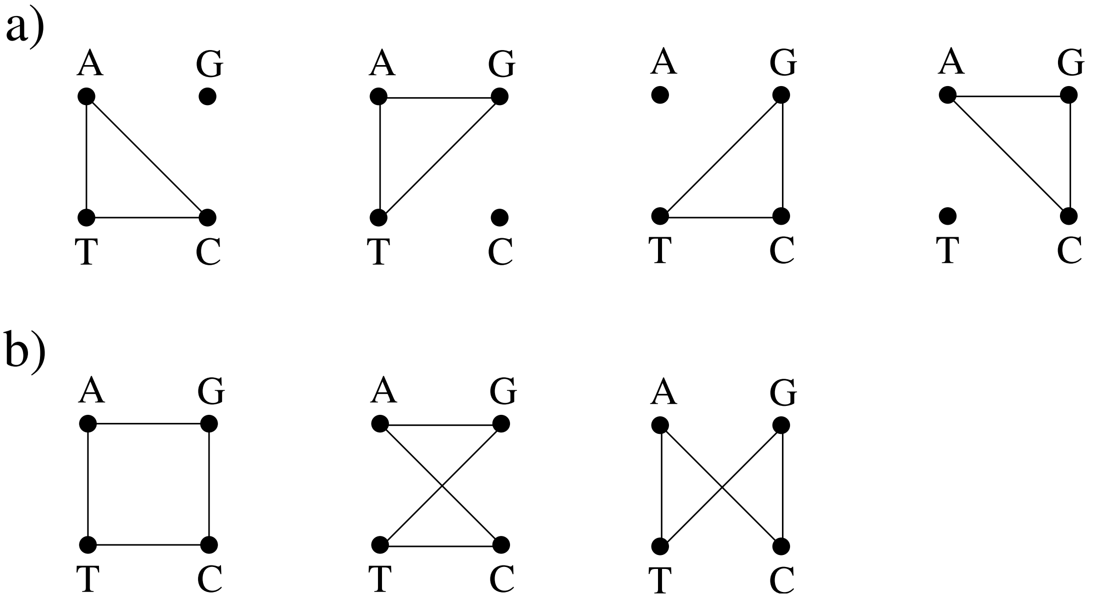

**Figure 4.2:** All the possible 3–cycles (a) and 4–cycles (b) for a Markov model with four states.

## 4.4 Kolmogorov conditions for the nucleotide evolution process

As we have seen, the importance of the Kolmogorov's conditions comes from the fact that if it holds and if the process has strictly positive rates, as is the case in the evolutionary process, then the process is time reversible. Also notable is the fact that, unlike detailed balance, the Kolmogorov's condition does not make use of the equilibrium distribution of the process.

We will now apply the theory to the case of nucleotide evolution, first for a model with independently evolving sites and then for a model with `CpG` neighbor dependencies.

### The Independent sites case

In order to check in what case the Markov model defined by Eq. (2.33) is also time reversible we have to consider equalities for the four 3–cycles conditions shown in Fig. 4.2a. However, substituting the rate matrix into Kolmogorov conditions one can immediately check that if any three of the four conditions are fulfilled then the fourth holds. That is, there are only three independent 3–cycles, so in order to derive an IRI we could single out three of the four possible 3–cycles. Instead we decided to check the equalities on 4–cycles, as there are only three non-trivial 4–cycles (Fig. 4.2b) and they are all independent. This approach is equivalent to the previous one as proven in proposition 9. The process is time reversible if the following conditions

$$Q_{\alpha\delta}Q_{\delta\gamma}Q_{\gamma\beta}Q_{\beta\alpha} = Q_{\alpha\beta}Q_{\beta\gamma}Q_{\gamma\delta}Q_{\delta\alpha}, \tag{4.14}$$

hold for $(\alpha, \beta, \gamma, \delta)$ equal to $(\mathtt{A}, \mathtt{G}, \mathtt{C}, \mathtt{T})$, $(\mathtt{A}, \mathtt{G}, \mathtt{T}, \mathtt{C})$ and $(\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T})$.

It is then straightforward to construct indices out of such equations just by taking the difference of both sides and normalizing it by the sum. We end up with three IRIs:

$$\begin{aligned}
\mathrm{IRI}_1 &:= \frac{Q_{\mathtt{AG}}Q_{\mathtt{GT}}Q_{\mathtt{TC}}Q_{\mathtt{CA}} - Q_{\mathtt{AC}}Q_{\mathtt{CT}}Q_{\mathtt{TG}}Q_{\mathtt{GA}}}{Q_{\mathtt{AG}}Q_{\mathtt{GT}}Q_{\mathtt{TC}}Q_{\mathtt{CA}} + Q_{\mathtt{AC}}Q_{\mathtt{CT}}Q_{\mathtt{TG}}Q_{\mathtt{GA}}} \\
\mathrm{IRI}_2 &:= \frac{Q_{\mathtt{AT}}Q_{\mathtt{TG}}Q_{\mathtt{GC}}Q_{\mathtt{CA}} - Q_{\mathtt{AC}}Q_{\mathtt{CG}}Q_{\mathtt{GT}}Q_{\mathtt{TA}}}{Q_{\mathtt{AT}}Q_{\mathtt{TG}}Q_{\mathtt{GC}}Q_{\mathtt{CA}} + Q_{\mathtt{AC}}Q_{\mathtt{CG}}Q_{\mathtt{GT}}Q_{\mathtt{TA}}} \\
\mathrm{IRI}_3 &:= \frac{Q_{\mathtt{AT}}Q_{\mathtt{TC}}Q_{\mathtt{CG}}Q_{\mathtt{GA}} - Q_{\mathtt{AG}}Q_{\mathtt{GC}}Q_{\mathtt{CT}}Q_{\mathtt{TA}}}{Q_{\mathtt{AT}}Q_{\mathtt{TC}}Q_{\mathtt{CG}}Q_{\mathtt{GA}} + Q_{\mathtt{AG}}Q_{\mathtt{GC}}Q_{\mathtt{CT}}Q_{\mathtt{TA}}}.
\end{aligned} \tag{4.15}$$

The three IRIs will thus be comprised in the interval $[-1, 1]$ and will be simultaneously zero if and only if the system under study evolves time reversibly.

We conclude this section noting that, as we already pointed out in the first chapter, evolutionary models traditionally used in the literature, belong to a family of nested models which originate from the GTR model [33, 55], which assumes the following parameterization of the rate matrix:

$$Q_{\mathrm{GTR}} = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} \cdot & a\pi_{\mathtt{A}} & b\pi_{\mathtt{A}} & c\pi_{\mathtt{A}} \\ a\pi_{\mathtt{C}} & \cdot & d\pi_{\mathtt{C}} & e\pi_{\mathtt{C}} \\ b\pi_{\mathtt{G}} & d\pi_{\mathtt{G}} & \cdot & f\pi_{\mathtt{G}} \\ c\pi_{\mathtt{T}} & e\pi_{\mathtt{T}} & f\pi_{\mathtt{T}} & \cdot \end{pmatrix} \end{array}. \tag{4.16}$$

The four $\pi$'s appearing in this matrix define the equilibrium distribution of nucleotides; only three of them are independent because they are assumed to be normalized. It can easily be checked by substitution of the parameterization of Eq. (4.16) in Eq. (4.15) that all three IRIs vanish for the GTR model, which therefore is indeed time reversible. The same is true for all its nested sub-models, which are mentioned in the introduction. As expected the GTR model has 9 free parameters. The 12-dimensional parameter space of the most general model Eq. (2.33) is reduced by 3 dimensions since equating the three IRI indices to zero yields 3 conditions on the 12 parameters.

## The reverse complement symmetric case

We now specialize the theory to the reverse complement symmetric model, defined by Eq. (2.42). In general, this model is not time reversible and in this case the Stationarity Indices have the following simple form:

$$\begin{aligned} \mathrm{STI}_1 &= \rho_{\mathtt{GC}} - \pi_{\mathtt{GC}} \\ \mathrm{STI}_2 &= \rho_{\mathtt{A}} - \rho_{\mathtt{T}} \\ \mathrm{STI}_3 &= \rho_{\mathtt{C}} - \rho_{\mathtt{G}}. \end{aligned} \tag{4.17}$$

It is worth noting that in this case $\mathrm{STI}_2$ and $\mathrm{STI}_3$ are the unnormalized AT and GC skews. They depend only on the nucleotide composition of the sequence, and not on the evolutionary rates. For reverse complement symmetric processes, it can be proven that once these indices or skews vanish they will stay stationary even if the rate matrix $Q_{\mathrm{RCS}}$ changes in time [35]. Therefore the skews can equilibrate even in the presence of reverse complement symmetric rate variations.

To derive an IRI for the RCS model we substitute the reverse complement symmetric parameterization in Eq. (4.15). We find that in this case we can check time reversibility with just one index:

$$\mathrm{IRI}_1 := \frac{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 - r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2}{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 + r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2}, \tag{4.18}$$

because $\mathrm{IRI}_2$ and $\mathrm{IRI}_3$ are equal to zero.

## The case with neighbor dependencies

To check for the time-reversibility of this model of evolution we should in principle check the Kolmogorov conditions for cycles with vertices in $\mathcal{C}$, the big configuration space introduced in Eq. (2.46). However, the generator of the dynamics (Eq. 2.53) permits only single nucleotide changes at a time and any cycle factorizes and can be decomposed into cycles changing only one site. Therefore, it is sufficient to check Kolmogorov conditions on single nucleotide 3–cycles like we did before, leading to the $\mathrm{IRI}_1$ for the RCS model. In addition to that one has to consider the particular configuration in which a C is followed by a G in the sequence. One example is the 3–cycle CG → CA → CT → CG. In this case the factorization is still possible but it is necessary to add to the total rate the contribution which comes from the CpG deamination process. In summary, there are then two IRIs for a process with neighbor dependencies:

$$\mathrm{IRI}_1 := \frac{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 - r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2}{r_{\mathtt{AG}}^2 r_{\mathtt{GT}}^2 + r_{\mathtt{AC}}^2 r_{\mathtt{CT}}^2} \tag{4.19}$$

$$\mathrm{IRI}_{\mathtt{CpG}} := \frac{r_{\mathtt{GT}}^2 (r_{\mathtt{AG}} + r_{\mathtt{CpG}})^2 - (r_{\mathtt{CT}} + r_{\mathtt{CpG}}^{\mathtt{rev}})^2 r_{\mathtt{AC}}^2}{r_{\mathtt{GT}}^2 (r_{\mathtt{AG}} + r_{\mathtt{CpG}})^2 + (r_{\mathtt{CT}} + r_{\mathtt{CpG}}^{\mathtt{rev}})^2 r_{\mathtt{AC}}^2} \tag{4.20}$$

Note that, as expected, in the absence of neighbor dependent processes we have $\mathrm{IRI}_1 = \mathrm{IRI}_{\mathtt{CpG}}$.

## 4.5 Measurements of STI and IRI in Drosophila

We first measure the STIs and $\mathrm{IRI}_1$ for the *Drosophila simulans* lineage from the time of the split with *Drosophila sechellia* until the current time, using *Drosophila melanogaster* as the outgroup. Whole genome alignments of the species are freely available on the Internet [51]. The genomic sequences have been split into 539 tiles corresponding to 50 Kbp long non-overlapping windows along the Drosophila chromosomes. We disregarded all gaps and masked the regions that were annotated as coding sequence in the Ensembl database [24]. The remaining nucleotides can be regarded to evolve independently from each other and without any significant contribution from the CpG decay process [4].

We estimate in each of the 50 Kbp windows all 6 free parameters of the RCS model in the *D. simulans* branch. From the inferred substitution rates in each fragment we have calculated the values of the STIs, thus obtaining the statistical distribution of the indices along the *D. simulans* genome (Fig. 4.3). The finite variance in the distribution of the indices arises as a statistical effect, since we are analyzing finite length sequences in each window and each of them is a realization of a Markov process. To count how many window can be assumed to be out of equilibrium we use the $\chi^2$-test mentioned in the Methods section. Since multiple independent tests are performed we have applied an appropriate Bonferroni correction, dividing the statistical significance level by the total number of windows. The test does not reject the hypothesis of stationarity in only 82 windows while it rejects it in 457.

Since the majority of tiles is not in the stationary state we also analyzed the distribution of the $\mathrm{IRI}_1$ index. The results are summarized in Fig. 4.4.

We do not present a closed form for the distribution of the $\mathrm{IRI}_1$ for the null hypothesis, that of time reversibility, but the simplicity of the index allowed us to simulate the distribution with little effort. From each window's inferred rate matrix we constructed an approximated version of the original one with the added property of time reversibility. The construction method uses the fact that any rate matrix $Q$, with equilibrium distribution $\pi$, can be written in the following way:

$$Q = D(\pi)F = \begin{pmatrix} \pi_{\mathtt{A}} & 0 & 0 & 0 \\ 0 & \pi_{\mathtt{C}} & 0 & 0 \\ 0 & 0 & \pi_{\mathtt{G}} & 0 \\ 0 & 0 & 0 & \pi_{\mathtt{T}} \end{pmatrix} \begin{pmatrix} \cdot & F_{12} & F_{13} & F_{14} \\ F_{21} & \cdot & F_{23} & F_{24} \\ F_{31} & F_{32} & \cdot & F_{34} \\ F_{41} & F_{42} & F_{43} & \cdot \end{pmatrix}, \tag{4.21}$$

For a suitably chosen matrix $F$. The dotted elements are again constrained by the fact that the sum of the elements in a column of the rate matrix must be zero, $F_{\alpha\alpha} = -\sum_{\beta\neq\alpha} \pi_\beta F_{\beta\alpha}$.

We now substitute $F$ with its symmetrized version, and obtain a time reversible generator $\hat{Q}$ with the following off diagonal elements:

$$\hat{Q}_{\alpha\beta} = \pi_\alpha \left[ \frac{F + F^t}{2} \right]_{\alpha\beta}, \tag{4.22}$$

while the diagonal elements are defined as $\hat{Q}_{\alpha\alpha} = -\sum_{\alpha\neq\beta} \hat{Q}_{\alpha\beta}$. This generator still has $\pi$ as equilibrium distribution.

We have used the symmetrized rate matrix to evolve the present day *D. simulans* sequences contained in each window. We made this in order to simulate evolution under a time reversible model. We could have used the inferred ancestral *sechellia-simulans* sequence as starting point of the evolution, but since ancestral and present day sequences have about 13 mismatches per 1000 bases this approximation does not affect the following results in any way.

We have then used the RCS model to estimate again the rates, comparing present day sequences and their evolved counterparts. As a result we got a second $\mathrm{IRI}_1$ distribution which we have used as null distribution, calling it $\mathrm{IRI}_{\mathrm{Null}}$. The plot is shown in Fig. 4.4 and as expected it is centered in zero.

We performed a two sample t-test to test the null hypothesis that the distributions of $\mathrm{IRI}_1$ and $\mathrm{IRI}_{\mathrm{Null}}$ have the same mean. The extremely low p-value of $10^{-15}$ shows that there is strong evidence against the null hypothesis. In other words, the process is not reversible even when the equilibrium distribution is reached.

## 4.6 Measurements of IRI in human genome

As a further example we have measured the STI and IRI for the *Homo sapiens* lineage using a triple alignment of *Homo sapiens*, *Pan troglodytes* and *Macaca mulatta* as an outgroup. Whole genome DNA alignments of these species are available from the Ensembl website [24].

Like in the previous case we have removed all coding regions and all gaps using Ensembl as a source of annotations. We have split the genome in 2413 windows of 1 Mbp size. For the analysis of nucleotide substitutions in vertebrates we have to include neighbor dependencies due to the CpG deamination process and have to use the extended model introduced before.

Distributions of the STIs are shown in Fig. 4.5. A $\chi^2$-test like the one used above for the human case does not reject the stationarity hypothesis in only 17 tiles and rejects it in 2396 tiles. Note that we should in principle also check whether the dinucleotide distribution is stationary. However, since the results show that in the vast majority of tiles already single nucleotides are out of equilibrium, we disregard such an analysis here.

For analyzing time reversibility it is necessary to use the two indices $\mathrm{IRI}_1$ and $\mathrm{IRI}_{\mathtt{CpG}}$ introduced in the last part of the Methods section. The resulting plots and statistics are shown in Fig. 4.6. The same t-test discussed in the previous section for the equality of the means of $\mathrm{IRI}_1$ and $\mathrm{IRI}_{\mathrm{Null}}$ also gives a p-value smaller than $10^{-15}$. The $\mathrm{IRI}_{\mathrm{Null}}$ distribution in this case has a smaller variance then the $\mathrm{IRI}_1$ distribution. This is because in addition to the variance introduced by finite sequence length, as discussed for Drosophila, in the Human genome one finds an intrinsic variation in rates due to its structured nature [5]. Time symmetrizing the matrix reduces the dimension of the parameter space and as a consequence reduces heterogeneity in the rates, thus reducing total variance of the $\mathrm{IRI}_1$ in the null model.

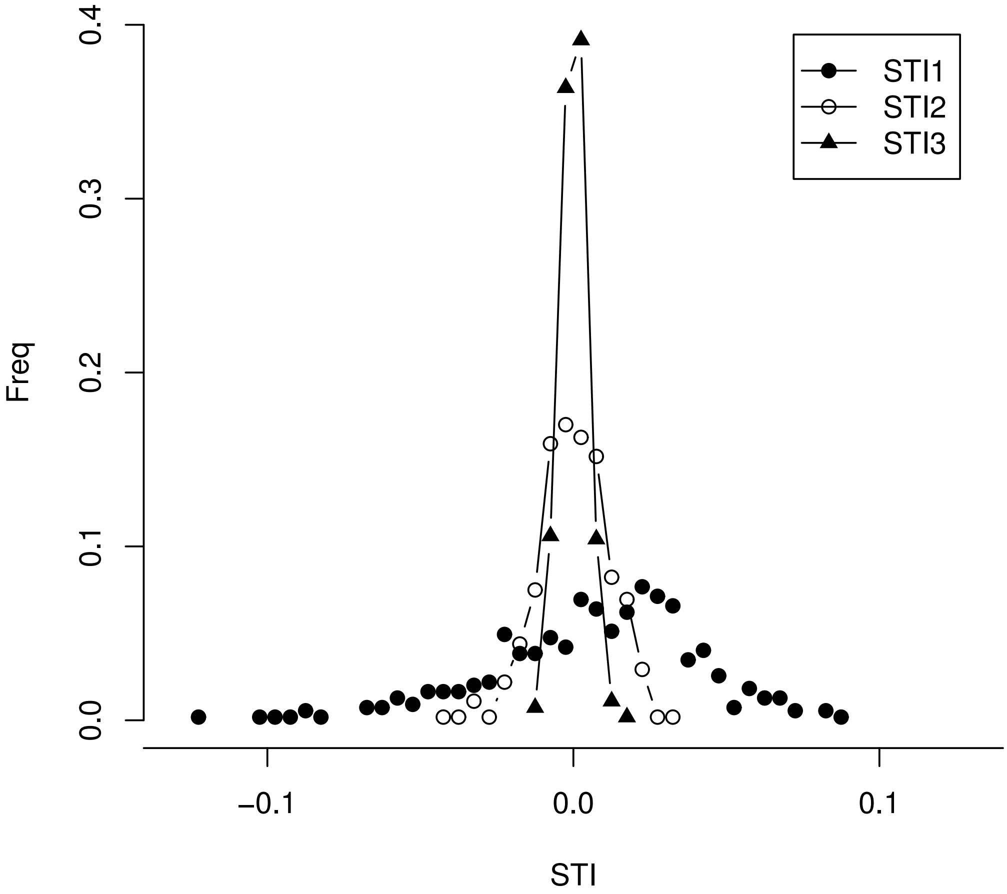

**Figure 4.3:** The distribution of the $\mathsf{STI}_1, \mathsf{STI}_2$, and $\mathsf{STI}_3$ in the *D. Simulans* genome. Means and standard deviations are: $\mathsf{STI}_1 = 0.007 \pm 0.034$, $\mathsf{STI}_2 = 0.000 \pm 0.011$, $\mathsf{STI}_3 = 0.000 \pm 0.004$.

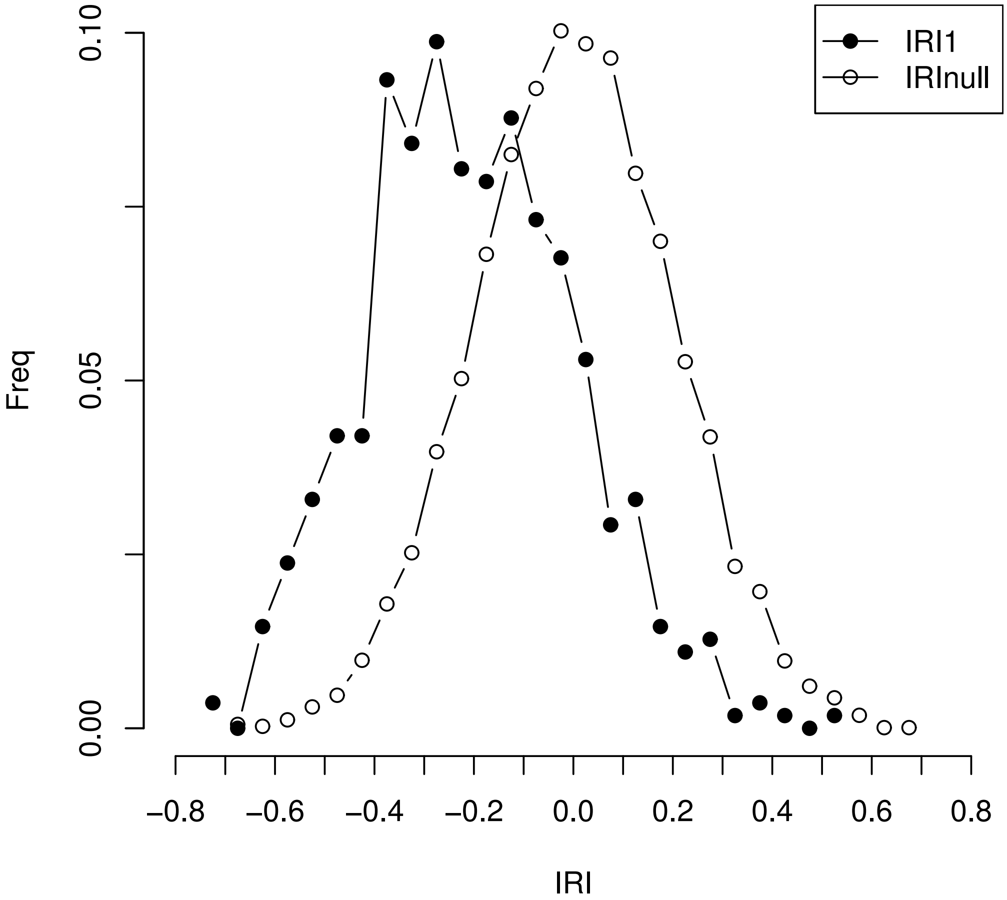

**Figure 4.4:** The distribution of the IRI in the *D. simulans* genome alongside with the distribution of the IRI for the null model. Means and standard deviations are: $\text{IRI}_1 = -0.204 \pm 0.208$ for *D. simulans* and $\text{IRI}_1 = 0.002 \pm 0.197$ for the null model.

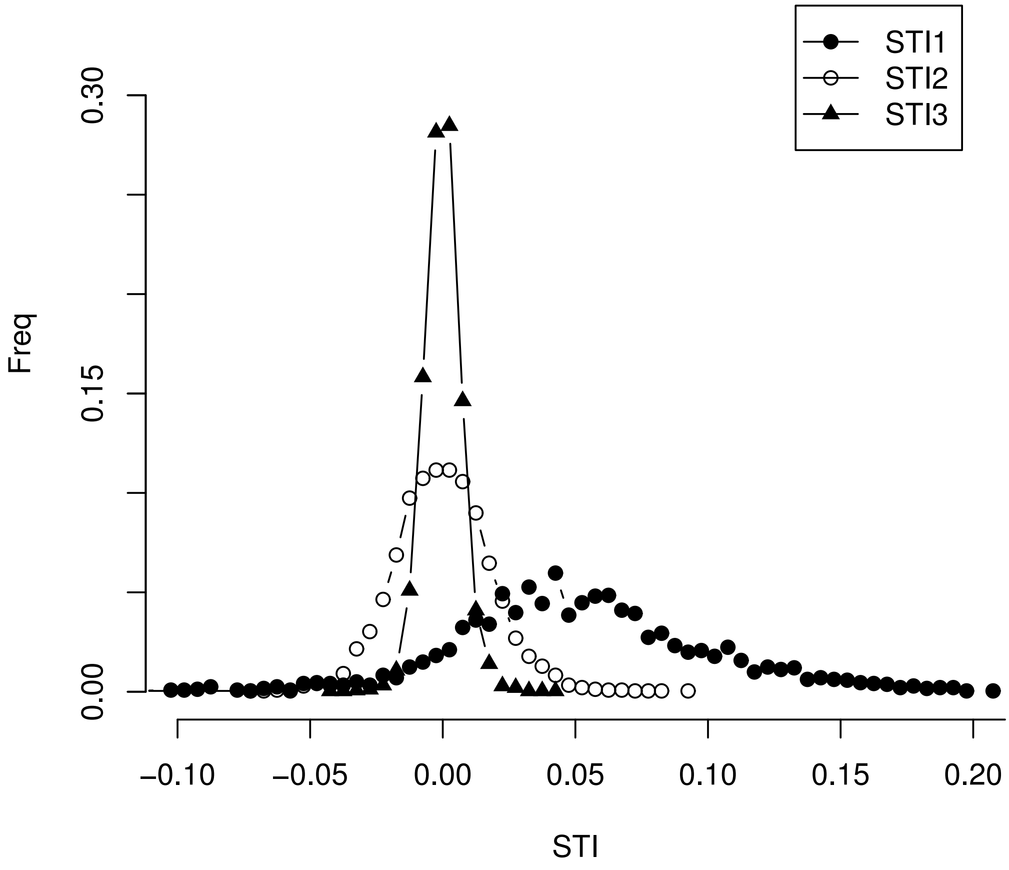

**Figure 4.5:** The distribution of the $\text{STI}_1, \text{STI}_2$, and $\text{STI}_3$ in the Human genome. Means and standard deviations are: $\text{STI}_1 = 0.052 \pm 0.048$, $\text{STI}_2 = 0.000 \pm 0.018$, $\text{STI}_3 = 0.000 \pm 0.007$.

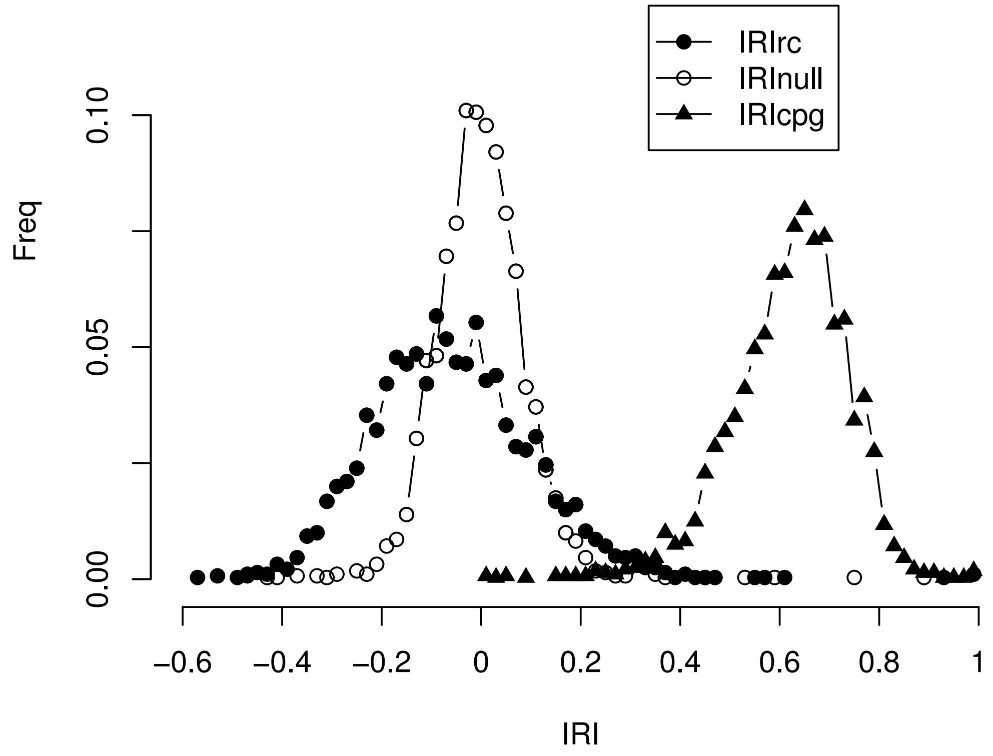

**Figure 4.6:** The distribution of the $\mathsf{IRI}_1$ and $\mathsf{IRI}_{\mathtt{CpG}}$ in the Human genome and $\mathsf{IRI}_1$ for the null case. Means and standard deviations are: $\mathsf{IRI}_1 = -0.060 \pm 0.161$, $\mathsf{IRI}_{\mathtt{CpG}} = 0.620 \pm 0.117$ and $\mathsf{IRI}_1 = -0.002 \pm 0.094$ for the null model

# Chapter 5

# Summary

The aim of this thesis was to present the concepts of stationarity and reversibility in the modeling of the evolution of DNA nucleotide sequences, and to check whether they are valid for evolution of real genomes. To this end we have introduced the Stationarity Indices, STIs, which compare the current nucleotide distribution to the stationary one, and the Irreversibility Indices, IRIs, which are based on the Kolmogorov cycle conditions for the time reversibility of a Markov process. The indices can be easily computed once we have at disposal, using for example a maximum likelihood estimation, the rates of the process.

We derived explicit expressions of the indices for the general 12 parameters model of nucleotide evolution with independent sites. It is interesting to note that assuming time-reversibility, which amounts to setting the IRI indices to zero, defines a 9-dimensional sub-manifold of the 12-dimensional space of all possible models. This manifold is the one spanned by the GTR model and its nested sub-models.

We analyzed the analytical formulation of the indices for the reverse complement symmetric models. This particular parameterization arises in a natural way when describing evolution of neutrally evolving sequences. In this case it turns out that both STI and IRI have a simpler form. In particular one needs only one index, $\mathrm{IRI}_1$, in order to test time reversibility. So imposing the constraint of time reversibility restricts the space of models to a 5-dimensional manifold in the 6-dimensional space of all the possible reverse complement symmetric models. We have successively extended the scope of our study to an evolutionary model which takes into account the `CpG` decay process, the predominant substitution process in vertebrates.

This approach based on a set of indices is complementary to the one using a likelihood ratio test, and it has the advantage that it simultaneously assesses stationarity and time-reversibilty for all branches of a given phylogeny once the rate matrices have been estimated. On the contrary, a likelihood ratio test requires a comparison of different hypotheses on different branches and a new estimation of the parameters for each of them.

When testing for all combinations the number of likelihood ratio tests required grows exponentially with the number of branches in the phylogeny.

As an application of the theory we have measured the STI and IRI in two different species lineages, *D. simulans* and *H. sapiens*. Using a sliding window analysis and the maximum likelihood estimation method we have derived the distributions of STI and $\mathrm{IRI}_1$ for Drosophila, and of STI, $\mathrm{IRI}_1$ and $\mathrm{IRI}_{\mathtt{CpG}}$ for human. In both cases we find statistically significant deviations from equilibrium and time reversibility. In *D. simulans*, the values of STI and $\mathrm{IRI}_1$ are close to zero, suggesting that it is legitimate to use a time reversible Markov model in bioinformatics algorithms, for instance in those used for phylogenetic reconstruction. However, in the human lineage, we find substantial deviations from equilibrium and time-reversibility due to the `CpG` methylation deamination process, in particular $\mathrm{IRI}_{\mathtt{CpG}} \approx 1$. In this case, the lack of equilibrium and time-reversibility is an important feature of the probabilistic model and consequently should not be disregarded.

# Chapter 6

# Zusammenfassung

Ziel dieser Arbeit war es, die Bedingungen für Stationarität und Zeitreversibilität in Bezug auf die Modellierung der Evolution von DNS Sequenzen vorzustellen und zu überprüfen, ob diese Gegebenheiten bei der Evolution von genomischen DNS Sequenzen zutreffen. Zu diesem Zweck wurden Statinaritätsindices (STIs) die die derzeitige Nukleotidverteilung mit stationärer Nukleotidverteilung vergleichen, und Irrevesibilitätsindizes (IRIs), die auf Kolmogorovs Bedingungen für Zyklen zurückgehen, eingeführt. Diese Indizes können einfach errechnet werden, sobald die Raten des evolutionären Prozesses bekannt sind, z.B. durch eine Schätzung mittels Maximum Likelihood Verfahren.

Es wurden explizite Ausdrücke für diese Indizes für das generelle 12 Parameter Modell der Evolution von DNS Sequenzen ohne Nachbarabhängigkeiten hergeleitet. Es ist interessant zu beobachten, dass unter der Annahme von Zeitreversibilität die drei IRIs verschwinden müssen und diese Bedingungen eine 9-dimensionale Untermannigfaltigkeit in dem 12-dimensionalen Raum aller Modelle aufspannen. Diese Untermannigfaltigkeit ist die des GTR Modells und aller seiner Untermodelle.

Des Weiteren wurden diese Indizes für Modelle mit einer zusätzlichen Symmetrie, der reversen Komplementarität, die bei der Beschreibung von neutraler Evolution der doppelsträngigen DNS gegeben ist, hergeleitet. Unter dieser Symmetrie nehmen die Indizes eine einfachere Form an. Insbesondere gibt es nur noch einen Irreversibilitätsindex. Im zeitreversiblen Fall wird dadurch eine 5-dimensionale Untermannigfaltigkeit in dem 6-dimensionalen Raum der reversen komplementen Modelle beschrieben. Darüber hinaus wurden diese Konzepte auch für die Evolution von DNS Sequenzen mit Nachbarabhängigkeiten verallgemeinert, wie sie zum Beispiel durch den CpG Methylierungs- und Deaminationsprozess, der vor allem in Wirbeltieren ein sehr verbreiteter Mutationsprozess ist, entstehen.

Dieser Zugang, die Stationarität und Zeitreversibilität anhand von Indizes zu prüfen, ist insbesondere bei großen phylogenetischen Bäumen einem Likelihood Ratio Test vorzuziehen, da er eine unabhängige überprüfung dieser Annahmen auf jedem Ast der

Phylogenie zulässt. Ein Likelihood Ratio Test müsste demgegenüber alle möglichen Kombinationen berücksichtigen und deshalb exponentiell viel häufiger ausgeführt werden.

Im Rahmen einer Anwendung unserer theoretischen überlegungen, wurden die IRIs und STIs für die Nukleotidevolution in der menschlichen Linie (*Homo Sapiens*) und in der Fruchtfliege (*Drosophila Simulans*) berechnet. Die Indizes wurden in verschiedenen Regionen aus den Mutationsraten berechnet, welche mittels Maximum Likelihood Methode gemessenen worden waren. In beiden Spezies fanden wir statistisch signifikante Abweichungen der Stationarität und Zeitreversibilität. In der Fruchtfliege sind die Abweichungen klein und die Verwendung von bioinformatischen Methoden, die diese Annahmen machen, erscheint legitim. In der menschlichen Linie allerdings sind die Abweichungen substanziell größer, was zuallererst auf die CpG Methylierung und Deamination zurückzuführen ist. In diesem Falle ist das Nichtvorhandensein von Stationarität und Zeitreversibilität eine Tatsache, die bei der statistischen Beschreibung und Modellierung nicht vernachlässigt werden sollte.

# Bibliography

[1] F. Ababneh, L. S. Jermiin, C. Ma, and J. Robinson. Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, 22(10):1225–1231, May 2006.

[2] D. C. Allis, T. Jenuwein, D. Reinberg, and M. L. Caparros. *Epigenetics*. Cold Spring Harbor Laboratory Press, October 2008.

[3] P. F. Arndt, C. B. Burge, and T. Hwa. DNA sequence evolution with neighbor-dependent mutation. *J Comput Biol*, 10(3-4):313–322, 2003.

[4] P. F. Arndt and T. Hwa. Identification and measurement of neighbor-dependent nucleotide substitution processes. *Bioinformatics*, 21(10):2322–2328, May 2005.

[5] P. F. Arndt, T. Hwa, and D. A. Petrov. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. *J Mol Evol*, 60(6):748–763, Jun 2005.

[6] P. F. Arndt, D. A. Petrov, and T. Hwa. Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation. *Mol Biol Evol*, 20(11):1887–1896, Nov 2003.

[7] J. A. Bailey and E. E. Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*, 7(7):552–564, 2006.

[8] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*, 11(6):1005–1017, Jun 2001.

[9] J. P. Bielawski and J. R. Gold. Mutation patterns of mithocondrial h– and l–strand dna in closely related cyprinid fishes. *Genetics*, 161(12):1589–1597, 2002.

[10] E. Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6(6):201–209, Jun 1950.

[11] C. Coulondre, J. H. Miller, P. J. Farabaugh, and W. Gilbert. Molecular basis of base substitution hotspots in Escherichia coli. *Nature*, 274(5673):775–780, Aug 1978.

[12] F. H. Crick. On protein synthesis. *Symp Soc Exp Biol*, 12:138–63, 1958.

[13] F. H. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[14] C. Darwin. *The Origin Of Species.* Signet Classics, September 2003.

[15] L. Duret. The GC content of primates and rodents genomes is not at equilibrium: a reply to Antezana. *J Mol Evol*, 62(6):803–806, Jun 2006.

[16] L. Duret and P. F. Arndt. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*, 4(5):e1000071, May 2008.

[17] A. Eyre-Walker. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics*, 152(2):675–683, Jun 1999.

[18] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[19] R. A. Fisher. *The genetical theory of natural selection.* 1930.

[20] R. E. Green, J. Krause, S. E. Ptak, A. W. Briggs, M. T. Ronan, J. F. Simons, L. Du, M. Egholm, J. M. Rothberg, M. Paunovic, and S. Pääbo. Analysis of one million base pairs of Neanderthal DNA. *Nature*, 444(7117):330–336, Nov 2006.

[21] M. Hasegawa, H. Kishino, and T. Yano. Dating of the Human-Ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, 22(2):160–174, 1985.

[22] B. Haubold and T. Wiehe. *Introduction to computational biology: an evolutionary approach.* Birkhauser, 2006.

[23] J. R. Helliwell. Synchrotron X-radiation protein crystallography: instrumentation, methods and applications. *Reports on Progress in Physics*, 47(11):1403–1497, 1984.

[24] T. J. P. Hubbard, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, and E. Birney. Ensembl 2007. *Nucleic Acids Res*, 35:D610–D617, Dec 2006.

[25] D. G. Hwang and P. Green. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc Natl Acad Sci U S A*, 101(39):13994–14001, Sep 2004.

[26] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[27] E. Jablonka and M. J. Lamb. *Evolution in four dimensions: genetic, epigenetic, behavioural and symbolic variation in the history of life.* MIT Press, May 2005.

[28] T. H. Jukes and C. R. Cantor. Evolution of protein molecules. In H. N. Munro, editor, *Mammalian protein metabolism*, pages 21–123. Academic Press, New York, 1969.

[29] F. P. Kelly. *Reversibility and stochastic networks.* John Wiley & Sons Ltd., Chichester, 1979.

[30] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–719, Jun 1962.

[31] M. Kimura. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16:111–120, 1980.

[32] D. E. Knuth. *The Art of Computer Programming*, volume 1. Addison Wesley, Boston, 1997.

[33] C. Lanave, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *J Mol Evol*, 20(1):86–93, 1984.

[34] J. R. Lobry. Properties of a general model of DNA evolution under no-strand-bias conditions. *J Mol Evol*, 40(3):326–330, Mar 1995.

[35] J. R. Lobry and C. Lobry. Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Mol Biol Evol*, 16(6):719–723, Jun 1999.

[36] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–5, 2000.

[37] M. A. Martí-Renom, A. C. Stuart, A. Fiser, R. Sánchez, F. Melo, and A. Šali. Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29(1):291–325, 2000.

[38] G. McLachlan and T. Krishnan. *The EM algorithm and extensions.* John Wiley & Sons Inc., New York, 1997.

[39] G. Mendel. Versuche über Plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn*, 1885.

[40] P. W. Messer and P. F. Arndt. The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol*, 24(5):1190–7, 2007.

[41] W. Miller, D. I. Drautz, A. Ratan, B. Pusey, J. Qi, A. M. Lesk, L. P. Tomsho, M. D. Packard, F. Zhao, A. Sher, A. Tikhonov, B. Raney, N. Patterson, K. Lindblad-Toh, E. S. Lander, J. R. Knight, G. P. Irzyk, K. M. Fredrikson, T. T. Harkins, S. Sheridan, T. Pringle, and S. C. Schuster. Sequencing the nuclear genome of the extinct woolly mammoth. *Nature*, 456(7220):387–390, Nov 2008.

[42] C. Moler and V. C. Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM review*, 45, 2003.

[43] J. P. Noonan, G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J. K. Pritchard, and E. M. Rubin. Sequencing and analysis of Neanderthal genomic DNA. *Science*, 314(5802):1113–1118, Nov 2006.

[44] S. Ohno. *Evolution by gene duplication*. Springer-Verlag, Berlin, New York,, 1970.

[45] D. G. Reid, L. K. MacLachlan, A. J. Edwards, J. A. Hubbard, and P. J. Sweeney. Introduction to the NMR of proteins. 60, July 1997.

[46] F. Rodríguez, J. L. Oliver, A. Marín, and J. R. Medina. The general stochastic model of nucleotide substitution. *J Theor Biol*, 142(4):485–501, Feb 1990.

[47] R. Rudner, J. D. Karkas, and E. Chargaff. Separation of B. subtilis DNA into complementary strands. 3. direct analysis. *Proc Natl Acad Sci U S A*, 60(3):921–922, Jul 1968.

[48] A. Rzhetsky and M. Nei. Tests of applicability of several substitution models for DNA sequence data. *Mol Biol Evol*, 12(1):131–151, Jan 1995.

[49] C. Saccone, C. Lanave, G. Pesole, and G. Preparata. Influence of base composition on quantitative estimates of gene evolution. *Methods Enzymol*, 183:570–583, 1990.

[50] F. Squartini and P. F. Arndt. Quantifying the equilibrium and irreversibility properties of the nucleotide substitution process. *Molecular Biology and Evolution*, 25(12):2525–35, 2008.

[51] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, H. F. curators, B. D. G. Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S.-W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van

Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart, and M. Kellis. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, Nov 2007.

[52] N. Sueoka. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J Mol Evol*, 40(3):318–325, Mar 1995.

[53] K. Tamura. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*, 9(4):678–687, Jul 1992.

[54] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*, 10(3):512–526, May 1993.

[55] S. Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, 17:57–86, 1986.

[56] M. D. Topal and J. R. Fresco. Complementary base pairing and the origin of substitution mutations. *Nature*, 263(5575):285–289, Sep 1976.

[57] C. Tuffley and M. Steel. Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci*, 147(1):63–91, Jan 1998.

[58] J. D. Watson and F. H. C. Crick. Genetical implications of the structure of deoxyribonucleic acid. *Nature*, 171(4361):964–967, May 1953.

[59] A. Wlodawer, W. Minor, Z. Dauter, and M. Jaskolski. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS Journal*, 275:1–21, 2008.

[60] K. M. Wong, M. A. Suchard, and J. P. Huelsenbeck. Alignment uncertainty and genomic analysis. *Science*, 319(5862):473–6, 2008.

[61] S. Wright. Evolution in Mendelian populations. *Genetics*, 16:97–159, 1931.

[62] C. I. Wu and N. Maeda. Inequality in mutation rates of the two strands of DNA. *Nature*, 327(6118):169–170, 1987.

[63] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol*, 10(6):1396–1401, Nov 1993.

[64] E. Zuckerkandl and L. Pauling. Evolutionary Divergence and Convergence in Proteins. Academic Press, New York, 1965.

[65] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, 1989.

[66] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.