INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Dynamical Systems in the Analysis of Biological Sequences

Peggy Cénac — Guy Fayolle — Jean-Marc Lasgouttes

N° 5351

October 2004

Thème BIO

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--5351--FR+ENG

Dynamical Systems in the Analysis of Biological Sequences

Peggy Cénac*, Guy Fayolle†, Jean-Marc Lasgouttes†

Thème BIO — Systèmes biologiques
Projet Preval

Rapport de recherche n° 5351 — October 2004 — 47 pages

Abstract: The *Chaos Game Representation* (CGR) maps a sequence of letters taken from a finite alphabet onto the unit square in $\mathbb{R}^2$. While it is a popular tool, few mathematical results have been proved to date. In this report, we show that the CGR gives rise to a limit measure, assuming only the input sequence is stationary ergodic. Some more precise properties are given in the i.i.d. and Markov cases. A new family of statistical tests to characterize the randomness of the inputs is proposed and analyzed. Finally, some basic properties of the CGR are used to generalize the notion of *genomic signature*.

Key-words: Chaos game representation, iterated function system, Hausdorff dimension, d-adic representation, Pearson's χ^2 test, genomic signature.

* INRIA Rocquencourt and Université Paul Sabatier (Toulouse III)
† INRIA Rocquencourt

Systèmes dynamiques pour l'analyse de séquences biologiques

Résumé : La *Chaos Game Representation* (CGR) est une méthode de représentation bidimensionnelle pour des séquences de lettres tirées d'un alphabet fini. Peu de resultats mathématiques existent à ce sujet. Dans ce rapport, nous montrons l'existence d'une mesure limite pour toute séquence d'entrée stationnaire ergodique ; ses caractéristiques sont précisées dans les cas i.i.d. et Markov. Plusieurs tests statistiques d'indépendance ou de dépendance markovienne sont proposés et analysés. Enfin, on utilise les propriétés de la CGR pour généraliser la notion de signature génomique.

Mots-clés : Représentation chaotique, système de fonction itérée, dimension de Hausdorff, écriture d-adique, test du χ^2 de Pearson, signature génomique.

1 Introduction

In the last years, several methods have been used to represent DNA in order to facilitate the visualization of patterns and to detect local or global similarities (see for example Roy *et al.* [29]). The so-called *Chaos Game Representation* (CGR) is both a graphical representation method of sequences and a storage tool. It is an iterative mapping technique that was first applied to genomic sequences by Jeffrey [14]. From a sequence of symbols (such as nucleotides in a DNA sequence or amino acids in a protein, see for example Basu *et al.* [3] or Pleißner *et al.* [27]), one may define trajectories in a continuous space conserving all its statistical properties. There is a one to one correspondence between the subsequences of a gene and the points in the CGR, so that the source sequence can be recovered from its coordinates in the CGR. The representation algorithm is described as follows.

Let $\mathcal{A}$ be a d-letter alphabet. For some bounded Borel set $S \subset \mathbb{R}^q$, where q is a positive integer, let $\{T_u, u \in \mathcal{A}\}$ be a collection of affine functions with contraction factor $0 < \rho < 1$, mapping the set S into itself

$$T_u(x) \stackrel{\text{def}}{=} \rho\big(x + \ell_u\big),\ u \in \mathcal{A},\ x \in S,\ \ell_u \in \mathbb{R}^q,$$

such that

$$T_u(S) \cap T_v(S) = \emptyset,\ \forall\,(u,v) \in \mathcal{A}^2, u \neq v. \tag{1.1}$$

Let $U_N = u_1 \dots u_N$ be a sequence of letters in $\mathcal{A}$. The *Chaos Game Representation* of the sequence U_N on S is a sequence $\{X_0, \dots, X_N\}$ defined by an arbitrary starting position X_0 and the relation

$$X_{i+1} \stackrel{\text{def}}{=} T_{u_{i+1}}(X_i) = \rho\big(X_i + \ell_{u_{i+1}}\big), \tag{1.2}$$

or equivalently

$$X_i = \sum_{j=1}^{i} \rho^{i-j+1}\ell_{u_j} + \rho^i X_0.$$

For example $\mathcal{A}$ can be a 4-letter alphabet for DNA, or a 20-letter amino-acid alphabet necessary to write proteins.

In the context of DNA sequences composed of the 4 nucleotides A (adenine), C (cytosine), G (guanine) and T (thymine), $\mathcal{A}$ will stand for $\{A, C, G, T\}$ and Jeffrey's definition of the CGR is obtained by choosing $S = [0,1[^2$, $\rho = 1/2$ and

$$\ell_A = (0,0), \quad \ell_C = (0,1), \quad \ell_G = (1,1), \quad \ell_T = (1,0),$$

so that (1.2) reads

$$X_{i+1} = \frac{X_i + \ell_{u_{i+1}}}{2},$$

choosing $X_0 = 0$ for instance.

Geometrically, the nucleotides are located at the corners of the unit square in such a way that horizontal sides indicate the base composition, while diagonals display the purine (A,G) and pyrimidine (C,T) composition. The point X_{i+1} is the middle of X_i and the corner corresponding to u_{i+1}.

In order to make a connection between the statistical properties of information theory (mutual information or redundancy), which characterize the dependency structure of sequences, and the properties of the empirical measure given by the points of the CGR, Gutierrez *et al.* [11] define it on the unit segment. In the general case of a d-letter alphabet, this is tantamount to taking $\mathcal{A} = \{0, \dots, d-1\}$, $\rho = 1/d$ and $\ell_u = u$, so that

$$X_i = \sum_{j=1}^{i} \frac{u_j}{d^{i-j+1}} + \frac{X_0}{d^n},$$

the starting point being $X_0 = 0$. We shall return to this point in Section 1.2.

For reasons which will soon become obvious, we associate with a word $w = u_1 \dots u_n$ the set Sw, where

$$Sw \stackrel{\text{def}}{=} \sum_{k=1}^{n} \rho^{n-k+1} \ell_{u_k} + \rho^n S, \tag{1.3}$$

as illustrated in Figure 1.1 for $S = [0,1[^2$.

One of the important properties of the CGR is that any point X_i contains the whole history of the sequence $X_1, \dots, X_i$. Indeed, first notice that (1.2) implies, by construction, $X_i \in Su_i$. Using (1.1), u_i can be recovered from X_i, and one has $X_{i-1} = (X_i - \ell_{u_i})/\rho$. This process can be repeated until $X_i = X_0$, which marks the beginning of the string. It is worth remarking that, for biological sequences, one should be able to decide whether a given point corresponds to a finite sequence, or if it was the last symbol u of a chain Vu, where V is a word consisting of the same letter, say v, repeated an infinite number of times. Then the reverse construction will be well-defined, simply by choosing the initial position X_0 to be one of the fixed points of the linear substitution 1.2. In the square these fixed points are the four vertices.

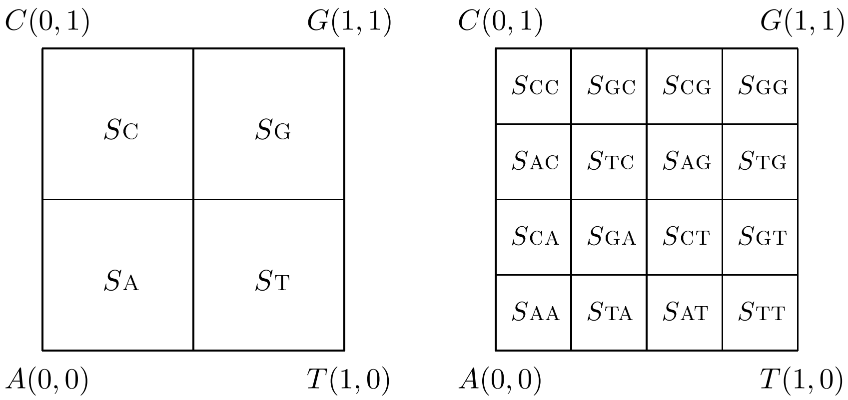


Figure 1.1: Definitions of the squares corresponding to nucleotides (left) and dinucleotides (right) for the CGR on $[0,1[^2$.

This report is an exploratory study of the main mathematical properties of the CGR and of their applications. The main goal pursued here is to try to understand whether the CGR provides more information on the distribution of a sequence than what could be obtained by more classical word-counting methods.

1.1 CGR and word counting

Counting points in some set Sw is tantamount to counting occurrences of the word w (see Figure 1.2 for a trinucleotide frequency matrix). Indeed, Sw consists of all possible sequences (not necessarily finite) having the suffix w. When the CGR is applied to long sequences, one can produce images where intensities of pixels are increasing functions of word frequencies (see for example Figure 1.3). This generalizes the usual tables of word frequencies (see Goldman [10]).

From an arbitrary subdivision of the square into k quadrants, Almeida *et al.* [1] define a distance between two sequences from a correlation coefficient between frequencies of the points in each quadrant. This allows them to make cross table comparisons between genes and to build phylogenetic trees. If the number of quadrants is not of the form 4^n, the frequency table defines frequencies of oligonucleotides with so-called "fractional" length. The authors assert that accessing the frequencies of non integer resolution is relevant for genomic sequences because of the redundancy of the genomic code (Almeida *et al.* [1]). In another study, Oliver *et al.* [25] use CGR of gene structure to determine *entropic profiles* in DNA sequences, by considering

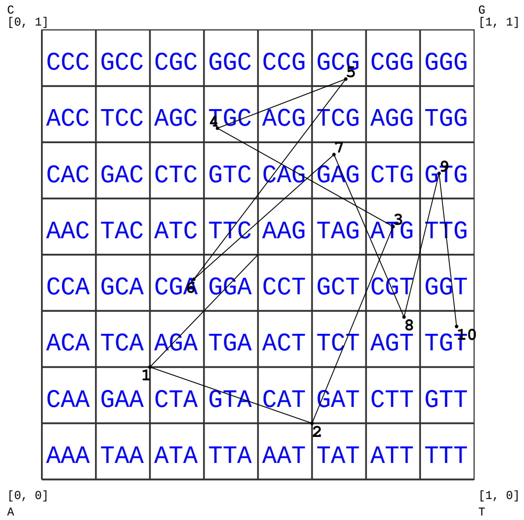


Figure 1.2: Chaos Game Representation of the first 10 nucleotides of the *E. Coli* threonine gene thrA: ATGCGAGTGT. The coordinates for each nucleotide are calculated recursively using $(0.5, 0.5)$ as starting position. The sequence is read from left to right. Point number 3 corresponds to the first 3-letter word *ATG*. It is located in the corresponding quadrant. The second 3-letter word *TGC* corresponds to point 4 and so on.

the entropy at each resolution level. Contrary to previous works (see e.g. Hairiri *et al.* [13]) using entropy related to word counting, the continuous resolution of CGR allows to distinguish between DNA sequences and random sequences.

Ahn *et al.* [2] use subintervals and represent substrings in order to obtain an accurate histogram of the substrings in the complete genome. This histogram is called the *measure representation* of the genome. They provide a characterization of the DNA sequences based on their measure representation, given in the form of the probability density function of the measure. It is determined by the exponent in the multifractal analysis of the cascade.

Most of the time throughout this study, the sequence $U = u_1, u_2, \ldots$ will be random and possibly infinite. Therefore, in the forthcoming section, we analyze some stochastic properties of the CGR needed in the sequel.

1.2 Stochastic properties of the CGR

As remarked in section 1, for arbitrary sequences $U = u_1, u_2 \ldots$ of elements of $\mathcal{A}$, the past of the sequence at time n can be recovered from the value of X_n, so that $(X_i)_{i \geq 0}$ forms a Markov chain of order 1 in the unit square, although in general its generator be not very explicit.

To get some insight into this process, it will be convenient to recast the CGR onto the unit segment $[0, 1[$. Some of the main mathematical features needed later will be stated now, and most of them can be reformulated in the domain formed by the interior of a d-sided convex polygon $\mathcal{R}^2$.

To ensure a somehow self-contained section (at least as for the notation), it will be convenient to consider the recursive scheme

$$Y_n = d^{-1}(Y_{n-1} + Z_n), \quad n \geq 1, \tag{1.4}$$

where $Y_0 \in [0, 1[$ is an arbitrary constant, $d > 1$ is a positive integer and $(Z_m)_{m \in \mathbb{Z}}$ a stationary ergodic sequence taking integer values in the set $\mathcal{A} \equiv \{0, 1, \ldots, d-1\}$, with strictly positive probabilities, and Z will denote an arbitrary Z_i. The evolution equation (1.4) pertains to several deeply explored areas, like stochastic algorithms and linear autoregressive models. In addition, as

$$Y_n = \sum_{i=1}^{n} \frac{Z_i}{d^{n-i+1}} + \frac{Y_0}{d^n}, \tag{1.5}$$

one can view (1.5) as the base-d expansion of $Y_n - d^{-n} Y_0$ (in the reverse order since Z_n appears in first position), which inherently involves the apparition of various beautiful fancy objects like singular distributions and fractals (see for instance Billingsley [4] and Falconer [9]).

Consider the random variable

$$\widetilde{Y}_n = \sum_{k=0}^{\infty} \frac{Z_{n-k}}{d^{k+1}}.$$

Then $\widetilde{Y}_n$ is also stationary ergodic and satisfies (1.5). The variables Z_i being uniformly bounded, we get immediately

$$|Y_n - \widetilde{Y}_n| = cd^{-n},$$

where c is a positive constant, so that

$$\lim_{n\to\infty} |Y_n - \widetilde{Y}_n| = 0, \quad a.s.,$$

and the convergence is geometric. We shall denote by Y_∞ the limiting random variable and by π its corresponding distribution.

As quoted in the introductory section 1, to decide whether a given point in the support of π corresponds to a finite sequence, say $(z_n, z_{n-1}, \ldots, u_1)$, or to an infinite one of the form $(z_n, z, z, \ldots)$ containing the same digit z from a certain position onwards, it suffices to put the starting position at one of the fixed points $a_u, 0 \leq u \leq d-1$, of the linear substitution

$$\sigma_u : x \to \sigma_u(x) = d^{-1}(x+u),$$

which is equal to $a_u = \dfrac{u}{d-1} \in [0,1[$.

Characterizing the limiting distribution π of Y_∞ is a more awkward task. In the classical case when the (Z_i) are independent and identically distributed random variables, with $p_k = \mathbb{P}(Z=k)$, the process (Y_n) enjoys the following properties, see e.g. Billingsley [4] and Falconer [9].

(a) If Z is uniform on $\{0,1,\ldots,d-1\}$, then π is simply the Lebesgue measure over $[0,1[$. Whenever Z is not uniformly distributed, π is continuous, singular with respect to the Lebesgue measure, and we have the strong law of large numbers

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{Z_j=k\}} = p_k, \quad \pi \ a.s.$$

(b) The support S_π of π has a Hausdorff dimension $\alpha \stackrel{\text{def}}{=} \dim_H(S_\pi)$ given by

$$\alpha = -\frac{1}{\log d} \sum_{i=0}^{d-1} p_i \log p_i, \tag{1.6}$$

which in turn shows that the distribution function $F(x) \stackrel{\text{def}}{=} \mathbb{P}(Y_\infty \leq x)$ satisfies a Lipschitz condition of order α on a set of π-measure (see [4, 9]), i.e.

$$|F(x+h) - F(x)| = O(h^\alpha).$$

Our claim is that several of the above properties are still valid under the sole assumption that the input sequence (Z_n) is stationary ergodic *(SE)*.

Proof.

(c) Law of large numbers. It could be obtained as a mere consequence of the geometric ergodicity shown above, see Loeve [23]). A concurrent approch is to remark, from the *(SE)* assumption, that there exists some d-adic interval

$$I_m = \left[\sum_{k=1}^{m} \frac{z_k}{d^{m-k+1}}, \sum_{k=1}^{m} \frac{z_k}{d^{m-k+1}} + \frac{1}{d^m} \right)$$

playing the role of a *recurrent atom*, ensuring in particular the ϕ-irreducibility of the Markov chain (Y_n), according to definitions given in Meyn and Tweedie [24]. So, the law of large number holds, for any Borel set E with $\pi(E) > 0$.

This expected result will be used in section 3 to handle empirical measures associated with U_N

$$\hat{\pi}_N(B) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{X_j \in B\}},$$

for arbitrary Borel set $B \in \mathcal{B}(S)$.

(d) Hausdorff dimension. The argument yielding equation (1.6) relies essentially on the ergodic theorem applied to (Z_n). It is necessary, among other things, to estimate the limiting probability of a sample path (associated with Y_∞) being in a d-adic interval of size d^{-n}. That is, for any sequence $z = (z_n \dots, z_1) \in \mathcal{A}^n$, we compute

$$\mathcal{I} \stackrel{\text{def}}{=} \lim_{n\to\infty} \frac{1}{n} \pi(Z_n = z_n, \dots, Z_1 = z_1). \tag{1.7}$$

For (Z_n) satisfying the *(SE)* condition, without further constraint, there is clearly not much hope of getting a tractable closed formula in (1.7).

On the other hand, in the case (Z_n) form an ergodic Markov chain of first order, we come up with a *relative entropy*, a key quantity in the theory of large deviations. To avoid unnecessary technicalities, the transition matrix of (Z_n) will be taken strictly positive. Consider the consecutive pairs $Z_n Z_{n-1}, \ldots, Z_2 Z_1$: they also form an ergogic Markov chain $(W_n)_{n\geq 0}$, with state space $\mathcal{A}^2$, transition matrix $\mathbf{R} \stackrel{\text{def}}{=} [r(i,j)]$ and invariant measure denoted by ζ. Then (1.7) gives

$$\mathcal{I} = \sum_{(i,j)\in\mathcal{A}^2} \zeta(i,j) \log r(i,j), \tag{1.8}$$

whence the following Hausdorff dimension

$$\beta = -\frac{1}{\log d} \sum_{(i,j)\in\mathcal{A}^2} \zeta(i,j) \log r(i,j).$$

By extending the argument in [9], one can also achieve a finer multifractal analysis. Details are omitted. □

1.3 Some applications

It is possible to give a simple criterion telling when a CGR has been obtained from an i.i.d. sequence. While simple, this result gives rise to interesting applications, briefly presented here.

Let $\mathcal{B}(S)$ denote the σ-field of Borel sets in S. The one-to-one mapping between the set of possible sequences and points of S suggests to abuse slightly the notation and write, for any finite word $w = u_1 \ldots u_N$ and for any Borel set $B \in \mathcal{B}(S)$,

$$Bw \stackrel{\text{def}}{=} T_{u_N} \circ \cdots \circ T_{u_1}(B). \tag{1.9}$$

This definition clearly coincides with (1.3) when $B = S$.

Assume that U is a stationary ergodic sequence, and let π be the stationary distribution of its CGR on S. The existence of π is a direct application of Section 1.2 in the unit segment case and the following property of π will be useful in the sequel.

Proposition 1.1. *The stationary random sequence U is i.i.d. if and only if*

$$\pi(Bu) = \pi(B)\pi(Su), \quad \forall u \in \mathcal{A},\ \forall B \subset \mathcal{B}(S). \tag{1.10}$$

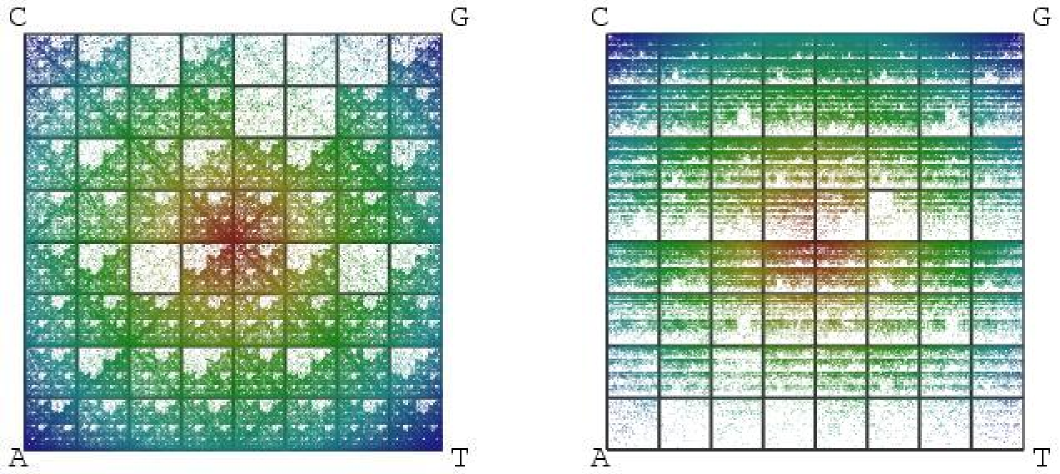


Figure 1.3: Chaos Game Representation of the 400000 first nucleotides of Chromosome 2 of *Homo Sapiens* (homsa1 in Table 4.1) on the left, and of *Streptomyces Coelicolor* (scoe) on the right.

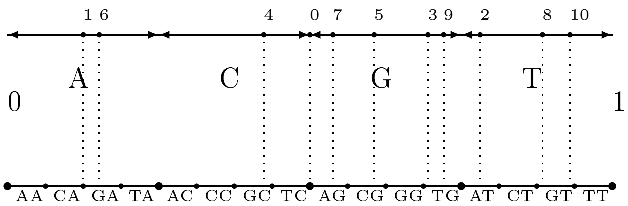


Figure 1.4: Chaos Game Representation of the 10 first nucleotides of *E. Coli* gene thrA (ATGCGAGTGT) on the unit segment

Proof. When U has i.i.d. components, the event $\{X_{n+1} \in Su\} = \{u_{n+1} = u\}$ is independent of the event $\{X_n \in B\}$. Hence, (1.2) implies

$$\mathbb{P}(X_{n+1} \in Su)\mathbb{P}(X_n \in B) = \mathbb{P}(X_{n+1} \in Su,\, X_n \in B) = \mathbb{P}(X_{n+1} \in Bu).$$

Conversely, assume that (1.10) holds and pick an arbitrary finite sequence $v_1 \dots v_N$. The choice $u = v_N$ and $B = Sv_1 \dots v_{N-1}$ yields

$$\pi(Sv_1 \dots v_N) = \pi\big((Sv_1 \dots v_{N-1})v_N\big) = \pi(Sv_1 \dots v_{N-1})\pi(Sv_N),$$

and, through a straightforward recurrence scheme,

$$\pi(Sv_1 \dots v_N) = \pi(Sv_1)\pi(Sv_2)\cdots\pi(Sv_N).$$

Since $v_1 \dots v_N$ is arbitrary, the random sequence U is i.i.d. □

The first idea coming to mind to make Proposition 1.1 profitable is to construct a test of independence for a stationary sequence of letters in $\mathcal{A}$. This is the subject of Section 3, where Markov chains of arbitrary order will also be considered.

Deschavanne *et al.* [7] use the CGR with a view to characterizing and classifing species. They argue that this specificity of genomic structure is the consequence of environment action on the one hand, and of structure constraints (replication, repairing, recombination) on the other hand.

Analysis of word frequency along a gene also highlights similarities and differences between each species. Karlin and Burge [17] and Karlin and Mràzek [20] use profile of *dinucleotide relative abundance* values as a genomic signature. They show that the variation of word frequency between genes of a given species is smaller than for genes of different species, and build phylogenetic trees based on these profiles.

In the terms of Proposition 1.1, dinucleotide relative abundance for nucleotide uv can be written

$$\rho_{uv} \stackrel{\text{def}}{=} \frac{\pi(Suv)}{\pi(Su)\pi(Sv)}. \tag{1.11}$$

It is therefore tempting to define a more general *CGR-based relative abundance* as

$$\rho(B, v) \stackrel{\text{def}}{=} \frac{\pi(Bv)}{\pi(B)\pi(Sv)}, \tag{1.12}$$

which trivially satisfies $\rho(Su, v) = \rho_{uv}$. Section 4 is devoted to testing the performance of this new ratio.

2 Asymptotic behavior of the Fourier transform

We consider the case where $U = u_1, u_2, \ldots$ is a sequence of independent realizations of a random variable u, taking its values in $\mathcal{A} = \{0, \ldots, d-1\}$. This situation corresponds to the evolution equation (1.2) for the Markov chain $(X_i)_{i\geq 0}$. Let $\mathcal{F} = \{\mathcal{F}_i, i \geq 0\}$ where $\mathcal{F}_i$ is the σ-algebra of the events occurring up to time i.

It was shown in Section 1.2 that X_n converges almost surely to a random variable X_∞. Some preliminary results about the distribution π of X_∞ are obtained in the next proposition, giving in particular the exact Hölder exponent of π at $x = 0$, in agreement with the assertions presented earlier.

Proposition 2.1. *Let $\Phi(t)$ denote the characteristic function of the distribution of X. Then*

$$\Phi(t) = \prod_{j=0}^{+\infty} g(\rho^j t), \quad \text{with} \quad g(t) \stackrel{\text{def}}{=} \sum_{v\in\mathcal{A}} p_v e^{i\rho t \ell_v} \quad \text{and} \quad p_v \stackrel{\text{def}}{=} \mathbb{P}(u_1 = v). \tag{2.1}$$

Proof. It immediately follows from (1.2) that

$$\begin{aligned} \mathbb{E}\Big[e^{itX_{j+1}} \,\Big|\, \mathcal{F}_j\Big] &= \sum_{v\in\mathcal{A}} e^{it\rho\left(X_j+\ell_v\right)}\, \mathbb{E}\Big[\mathbb{1}_{\{u_{j+1}=v\}} \,\Big|\, \mathcal{F}_j\Big] \\ &= e^{i\rho t X_j} \sum_{v\in\mathcal{A}} e^{it\rho\ell_v}\, \mathbb{P}(u_{j+1} = v|\mathcal{F}_j). \end{aligned}$$

Let $\Phi_j(t) \stackrel{\text{def}}{=} \mathbb{E}[e^{itX_j}]$. Since the variables $\{u_i, i \geq 1\}$ are independent, one has

$$\Phi_{j+1}(t) = g(t)\ \Phi_j(\rho\ t) \quad \text{i.e.} \quad \Phi_{j+1}(t) = \Phi_0(\rho^{j+1}t) \prod_{k=0}^{j} g(\rho^k t). \tag{2.2}$$

The sequence of analytic functions $\Phi_j(t)$ converges to the function $\Phi(t)$ defined by (2.1). Indeed, the limit of the infinite product in (2.2) exists because the series

$$\sum_{j=0}^{+\infty} \Big|\Big(\sum_{v\in\mathcal{A}} p_v e^{i\rho^j t\ell_v}\Big) - 1\Big|$$

converges. $\square$

The function $g(t)$ is a trigonometric polynomial of degree $d-1$. Let us define the polynomial

$$p(x) \stackrel{\text{def}}{=} \sum_{1\le i\le d} p_i x^{i-1}.$$

Then $g(t) = p(e^{i\rho t})$, where p factorizes as the product of $d-1$ polynomials of degree 1. In order to study the asymptotic behavior of $\Phi(t)$, it is therefore enough to consider a basic infinite product of the form

$$P(t) = \prod_{j=0}^{+\infty}\Big(1-a+ae^{i\rho^{j+1}t}\Big), \tag{2.3}$$

where a is a complex number with $\Re(a) \le \frac{1}{2}$. Indeed, up to replacing a by $1-a$, one can assume $\left|\frac{a}{1-a}\right| \le 1$.

Let us suppose $\Re(a) < \frac{1}{2}$. In the particular case $\mathcal{A} = \{A, C, G, T\}$, this assumption holds if

$$p_A - p_C + p_G - p_T \neq 0,$$

$$p_A(p_A - p_G) \neq p_T(p_T - p_C) \quad \text{or} \quad p_A - p_G > p_T,\ p_T - p_C > p_A.$$

However, Chargaff's second rule for DNA sequences states that $p_A = p_T$, $p_C = p_G$, which means that hence we need to go further in the study of the asymptotic behavior of $\Phi(t)$ when $\Re(a) = \frac{1}{2}$.

Theorem 2.2. *Let us introduce the Lerch function*

$$\hat{\Phi}(z,s,v) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} \frac{z^{k+1}}{(v+k)^s},$$

where $-v \notin \mathbb{N}$ *and* $|z| < 1$ *or* $|z| = 1$ *and* $\Re(s) > 1$. *Then for all* $\sigma_1 > 0$, *we have*

$$\log P(t) = Res(F_t, 0) + \sum_{k\in\mathbb{Z}^*} Res(F_t, \theta_k) + \frac{1}{2i\pi}\int_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} F_t(\theta)\, d\theta \tag{2.4}$$

where

$$F_t(\theta) \stackrel{\text{def}}{=} \frac{1}{\rho^\theta - 1}\Gamma(\theta)\hat{\Phi}\Big(\frac{-a}{1-a}, \theta+1, 1\Big)e^{\frac{i\theta\pi}{2}}t^{-\theta}.$$

Corollary 2.3.

$$\log P(t) = \frac{\log(1-a)}{\log\rho}\log t + \mathcal{O}(1).$$

Proofs of Theorem 2.2 and Corollary 2.3 are given in Appendix A.

Corollary 2.4.

$$\log \Phi(t) = \frac{\log p_0}{\log \rho} \log(t) + \mathcal{O}(1).$$

Proof. Corollary 2.4 is a direct consequence of Corollary 2.3, using the decomposition of the polynomial $p(t)$. □

3 Testing the structure of a sequence

Let H_0, H_m and H respectively denote the hypothesis "$U_N = u_1 \dots u_N$ *is an i.i.d. sequence*", "U_N *is a Markov chain of order* m" and "U_N *is a stationary ergodic sequence*". We consider the problem of testing H_0 against $H \setminus H_0$, then H_m against $H \setminus H_m$.

3.1 Construction of the test H_0 against $H \setminus H_0$

Proposition 1.1 suggests using a reject region of the form

$$\left| \hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu) \right| > \varepsilon.$$

To build an asymptotic test of level α, we shall adjust ε by use of the central limit theorem. This is done in the following theorem.

Theorem 3.1. *Let u_α be the $(1-\frac{\alpha}{2})$-quantile of the normal law, that is to say the real u_α such that $P(|Y| > u_\alpha) \leq \alpha$ where Y has a normal distribution. Define also*

$$\hat{\sigma}_N(B,u) \stackrel{\text{def}}{=} \sqrt{\Big(\hat{\pi}_N(Su)\big(1-\hat{\pi}_N(Su)\big)\Big)\Big(\hat{\pi}_N(B)\big(1-\hat{\pi}_N(B)\big)\Big)}.$$

Then the set

$$\left\{ \left| \hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu) \right| > u_\alpha \frac{\hat{\sigma}_N(B,u)}{\sqrt{N}} \right\} \tag{3.1}$$

is a reject region of a test of asymptotic level α, of the null hypothesis H_0 against the hypothesis $H \setminus H_0$, that is to say

$$\lim_{N\to+\infty} \mathbb{P}_{H_0}\left(\left| \hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu) \right| > u_\alpha \frac{\hat{\sigma}_N(B,u)}{\sqrt{N}} \right) \leq \alpha,$$

The proof of Theorem 3.1 is given in Appendix B.

In Theorem 3.1, one has to choose a nucleotide u and a set $B \subset \mathcal{B}(S)$. The choice of the most suitable B and u depends on the law $(p_u)_{u\in\mathcal{A}}$ which is unknown in practice. It is however possible to build a test of H_0 against $H \setminus H_0$, with a reject region of the form

$$\left\{ \sum_{\substack{1\le i\le k\\ u\in\mathcal{A}}} R_N(B_i,u)^2 > q_\alpha \right\}, \tag{3.2}$$

for any partition $\{B_1,\dots,B_k\}$ of $[0,1[^2$, where

$$R_N(B,u) \stackrel{\text{def}}{=} \frac{\sqrt{N}\big(\hat\pi_N(Su)\hat\pi_N(B) - \hat\pi_N(Bu)\big)}{\sqrt{\hat\pi_N(Su)\hat\pi_N(B)}}. \tag{3.3}$$

This statistics generalizes of the Pearson statistics used by Reinert *et al.* [28]

$$X^2 \stackrel{\text{def}}{=} \sum_{u,v\in\mathcal{A}} \frac{\big(N(uv) - N(u\cdot)N(\cdot v)/(N-1)\big)^2}{N(u\cdot)N(\cdot v)/(N-1)}, \tag{3.4}$$

to test H_0 against H_1, where $N(uv)$ counts the occurences of uv in the sequence, $N(u\cdot)$ (resp. $N(\cdot v)$) is the number of dinucleotides beginning with u (resp. ending with v). Under H_0, X^2 follows asymptotically a chi-square distribution with 9 degrees of freedom. This test can be viewed as a generalized likelihood ratio test. In the special case when the partition is $\{B_u, u \in \mathcal{A}\}$, the statistics (3.2) asymptotically yields to X^2.

Theorem 3.2. *For all $B \subset \mathcal{B}(S)$,*

$$\frac{1}{\sqrt{1-\hat\pi_N(B)}} \sum_{u\in\mathcal{A}} R_N(B,u)^2 \stackrel{\mathcal{L}}{\rightsquigarrow} \chi^2(d-1), \tag{3.5}$$

Let us denote $q_\alpha(r)$ the $(1-\alpha)$-quantile of the chi-square law $\chi^2(r)$. Then the set

$$\left\{ \frac{1}{\sqrt{1-\hat\pi_N(B)}} \sum_{u\in\mathcal{A}} R_N(B,u)^2 > q_\alpha(d-1) \right\} \tag{3.6}$$

is a reject region of a test of asymptotic level α, of H_0 against $H \setminus H_0$.

More generally, for any partition $\{B_1, \ldots, B_k\}$ *of* S*, with* $k > 1$*,*

$$\sum_{\substack{1 \le i \le k \\ u \in \mathcal{A}}} R_N(B_i, u)^2 \overset{\mathcal{L}}{\rightsquigarrow} \chi^2((d-1)(k-1)), \tag{3.7}$$

and the set

$$\left\{ \sum_{\substack{1 \le i \le k \\ u \in \mathcal{A}}} R_N(B_i, u)^2 > q_\alpha((d-1)(k-1)) \right\} \tag{3.8}$$

is a reject region of a test of asymptotic level α*, of* H_0 *against* $H \setminus H_0$*.*

The proof of Theorem 3.2 is given in Appendix C.

In order to appreciate the quality of the tests defined in Theorem 3.2, we need to calculate the asymptotic power of the test, that is to say, the asymptotic probability under $H \setminus H_0$ of the sets (3.6) and (3.8).

Theorem 3.3. *Assume that* $H \setminus H_0$ *holds, and let* $B \in \mathcal{B}(S)$ *and* $u \in \mathcal{A}$ *be such that*

$$\pi(Bu) \neq \pi(Su)\pi(B). \tag{3.9}$$

Then the asymptotic power of the test built from the rejection set (3.6) is 1*, that is to say, the test is asymptotically consistent. Likewise, if* $\{B_1 = B, B_2 \ldots, B_k\}$ *is a partition of the unit square, the test built from the rejection test (3.8) is asymptotically consistent.*

Proof. According to the convergence of the empirical estimate, almost surely

$$\hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu) \longrightarrow p_u\pi(B) - \pi(Bu).$$

Assuming (3.9),

$$\sqrt{N}\left|\hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu)\right| \xrightarrow[\mathrm{N}\to\infty]{\text{a.s.}} +\infty,$$

and therefore,

$$\sum_{v \in \mathcal{A}} R_N(B, v)^2 \xrightarrow[\mathrm{N}\to\infty]{\text{a.s.}} +\infty.$$

The assertions of Theorem 3.3 follow easily. $\square$

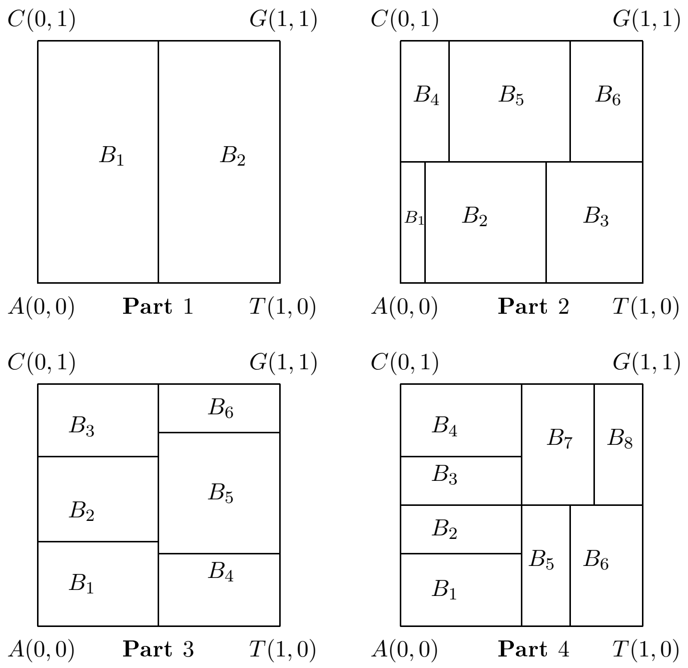


Figure 3.1: The 4 different partitions of the square $[0,1[^2$ chosen for the test.

The statistics defined in Theorem 3.2 is more general than Pearson's since it does not limit itself to some particular markovian structure. To illustrate this point, let us consider the following family of Markov chains of order $m > 1$: first, m independent Markov chains $U^{(1)}, \ldots, U^{(m)}$ of order 1 are generated as above; then the final sequence U is obtained by the following aggregation

$$u_{km+i} = u_k^{(i)}, \text{ for all } k \geq 0, 1 \leq i \leq m. \tag{3.10}$$

U is a Markov chain of order m, where each nucleotide u_i depends only of the nucleotide u_{i-m}, and is independent of nucleotides u_{i-k} for $1 \leq k < m$. However, from the point of view of the X^2 statistics of (3.4), it behaves like an i.i.d. process and the probability of rejection of H_0 for this particular process will not tend to 1.

3.2 Numerical experiments

This section shows numerical results comparing the test described in Theorem 3.2 to Pearson's test on several types of sequences, when the CGR is computed on the unit square ($S = [0, 1[^2$), on the 4-letter alphabet of nucleotides. First, for various values of N, 1000 sequences of length N of i.i.d. random variables (which distribution is itself chosen uniformly) are generated. In order to see how the choice of the sets B_i affects the accuracy of the test, several arbitrary partitions have been selected, as described in Figure 3.1.

Table 3.1 shows the fraction of cases where hypothesis H_0 is accepted, for a level $\alpha = 0.05$. For convenience, the results of Pearson's test have been added to the list. The level of the different tests are comparable and already close to 0.05 for a sequence length $N = 500$. In a second set of computations, 1000 sequences of a Markov chain of order 1 with random transition matrix are generated and tested (see the "order 1" part of Table 3.2). Here the choice of the partition $\{B_1, \ldots, B_k\}$ is crucial for the power of the test and, more importantly, Pearson statistics gives the best results in order to distinguish an i.i.d. sequence from a Markov chain of order 1. This is not surprising because the statistics (3.4) is based on count of words of two letters (dependency of order 1). Pearson test is asymptotically equivalent to the likelihood ratio statistics (see for example Van der Vaart [32]), which is "asymptotically efficient".

However, when we use the special case of order-m Markov processes defined in (3.10), the situation is very different. Table 3.2 shows the fraction of cases where hypothesis H_0 is rejected for $m > 1$. The results shows that the CGR-based test is very effective on this particular set of Markov chains, whereas Pearson's test fares as expected very poorly. Interestingly, the proportion of rejected sequences is not affected notably by the order of dependency. This illustrates the strength of the CGR for such tests: it does not impose any constraint on the input sequence besides stationarity. The alternative $H \setminus H_0$ is much more general than H_1.

3.3 Test for the appropriate order of a Markov model

We consider the problem of testing H_m against the hypothesis $H \setminus H_m$. The construction of the test relies on the following generalization of Proposition 1.1.

Proposition 3.4. *The stationary random sequence U is a Markov chain of order m if and only if $\forall B \subset \mathcal{B}(S)$, $\forall w \in \mathcal{A}^m$, $\forall u \in \mathcal{A}$,*

$$\pi(Sw)\pi(Bwu) = \pi(Swu)\pi(Bw). \tag{3.11}$$

N	**Pearson**	**Part** 1	**Part** 2	**Part** 3	**Part** 4
50	85.50%	94.70%	65.60%	94.40%	87.30%
100	91.00%	94.70%	86.70%	95.10%	93.40%
500	94.80%	95.20%	95.30%	95.40%	94.20%
1000	94.10%	94.90%	95.10%	95.10%	94.70%
10000	94.70%	95.70%	94.60%	94.60%	95.50%

Table 3.1: Fraction of cases where H_0 is accepted, simulating i.i.d sequences of random law, for a level $\alpha = 0.05$.

order	N	**Pearson**	**Part** 1	**Part** 2	**Part** 3	**Part** 4
1	50	44.6%	9.80%	24.50%	6.80%	17.50%
	100	85.6%	14.40%	11.20%	7.80%	10.50%
	500	100%	63.60%	28.80%	40.30%	36.50%
	1000	100%	92.60%	64.50%	77.10%	75.80%
	10000	100%	100%	100%	100%	100%
2	50	3.2%	8.90%	23.30%	8.60%	15.90%
	100	4.7%	15.50%	9.80%	9.30%	13.10%
	500	8.7%	64.30%	29.90%	39.50%	38.00%
	1000	7.3%	91.40%	64.50%	80.00%	71.90%
	10000	7.2%	100%	100%	100%	100%
10	50	2.5%	9.50%	21.20%	8.60%	16.40%
	100	3.3%	13.80%	10.10%	10.60%	12.60%
	500	6.1%	64.50%	29.70%	42.90%	41.20%
	1000	5.6%	91.50%	65.20%	75.70%	76.40%
	10000	4.8%	100%	100%	100%	100%

Table 3.2: Reject rate of H_0, when the Markov chain U is defined as in (3.10), for an asymptotic level $\alpha = 0.05$.

Proof. When U is a Markov chain of order m,

$$\begin{aligned}\mathbb{P}\big[X_{j+1} \in Bwu \bigm| \mathcal{F}_j\big] &= \mathbb{P}\big[X_{j-m} \in B, X_j \in Sw, X_{j+1} \in Su \bigm| \mathcal{F}_j\big] \\ &= \mathbb{1}_{\{X_{j-m}\in B\}}\mathbb{1}_{\{X_j \in Sw\}}\mathbb{P}\big[X_{j+1} \in Su \bigm| \mathcal{F}_j\big] \\ &= \mathbb{1}_{\{X_{j-m}\in B\}}\mathbb{1}_{\{X_j \in Sw\}}\frac{\pi(Swu)}{\pi(Sw)},\end{aligned}$$

which implies (3.11). Conversely, assume that (3.11) holds and choose an arbitrary finite sequence $v_{-n}\dots v_0$ with $n > m$. The choice $u = v_0$, $w = v_{-m}\dots v_{-1}$ and $B = Sv_{-n}\dots v_{-m-1}$ yields

$$\pi(Sv_{-n}\dots v_0) = \frac{\pi(Sv_{-m}\dots v_0)}{\pi(Sv_{-m}\dots v_{-1})}\pi(Sv_{-n}\dots v_{-1})$$

and therefore

$$\begin{aligned}\mathbb{P}(u_0 = v_0 | u_{-n} = v_{-n}, \dots u_{-1} = v_{-1}) &= \frac{\pi(Sv_{-n}\dots v_0)}{\pi(Sv_{-n}\dots v_{-1})} \\ &= \frac{\pi(Sv_{-m}\dots v_0)}{\pi(Sv_{-m}\dots v_{-1})}.\end{aligned}$$

Since this is true for any $n > m$, the sequence U is markovian of order at most m. □

The corresponding statistics to (3.3) in the markovian case is the ratio

$$R_N(B,w,u) \stackrel{\text{def}}{=} \sqrt{N}\frac{\big(\hat{\pi}_N(Sw)\hat{\pi}_N(Bwu) - \hat{\pi}_N(Swu)\hat{\pi}_N(Bw)\big)}{\sqrt{\hat{\pi}_N(Sw)\hat{\pi}_N(Swu)\hat{\pi}_N(Bw)}}$$

Theorem 3.5. *Let us denote $q_\alpha(r)$ the $(1-\alpha)$-quantile of the chi-square law $\chi^2(r)$. For any partition $\{B_1,\dots,B_k\}$ of S, with $k > 1$,*

$$\sum_{\substack{wu\in\mathcal{A}^m\times\mathcal{A} \\ 1\le i\le k}} R_N(B_i,w,u)^2 \overset{\mathcal{L}}{\rightsquigarrow} \chi^2\Big(4^m(d-1)(k-1)\Big), \tag{3.12}$$

Then the set

$$\left\{\sum_{\substack{wu\in\mathcal{A}^m\times\mathcal{A} \\ 1\le i\le k}} R_N(B_i,w,u)^2 > q_\alpha\Big(4^m(d-1)(k-1)\Big)\right\} \tag{3.13}$$

is a reject region of a test of asymptotic level α, of the null hypothesis H_m against the hypothesis $H \setminus H_m$.

The proof of Theorem 3.5 is given in Appendix D.

N	Order 1		Order 2		Order 3		Order 2 mixed		Order 3 mixed	
	CGR	Pearson	CGR	Pearson	CGR	Pearson	CGR	Pearson	CGR	Pearson
50	1.1%	0.6%	7.6%	97.1%	0%	99.9%	4.7%	9.8%	2.9%	1.7%
100	4.9%	1.3%	90.3%	100%	59%	100%	24.9%	57.1%	8.3%	3.6%
500	5.7%	5.3%	100%	100%	100%	100%	92.1%	99.1%	32.3%	7.1%
1000	6.6%	4.6%	100%	100%	100%	100%	99%	100%	55.5%	8.9%
5000	5.9%	3.9%	100%	100%	100%	100%	100%	100%	95.7%	4.3%
10000	5.7%	4.9%	100%	100%	100%	100%	100%	100%	99.6%	7.6%

Table 3.3: Reject rate of H_1, for an asymptotic level $\alpha = 0.05$.

Remark. For a single set B, a similar proof yields

$$\sum_{wu \in \mathcal{A}^m \times \mathcal{A}} R_N(B,w,u)^2 \overset{\mathcal{L}}{\rightsquigarrow} \chi^2\Big(4^m(k-1)\Big).$$

For various values of N, 1000 sequences of length N of order d Markov chains, taking its values in $\mathcal{A} = \{A, C, G, T\}$, (which distribution is chosen uniformly) are generated. The CGR is computed on the unit square. Table 3.3 shows the fractions of cases where hypothesis H_1 is rejected, for a level $\alpha = 0.05$ and for the partition $\{B_1, \ldots, B_4\}$ depicted in Figure 4.2. The results of Pearson's test have been added to the list. Pearson statistics for the test of H_1 against H_2 is defined (see for instance Dacunha-Castelle and Duflo [6]) in the following way

$$X^2 \stackrel{\text{def}}{=} \sum_{u,v,w \in \mathcal{A}} \frac{\Big(N(uvw) - N(uv)N(vw)/N(v.)\Big)^2}{N(uv)N(vw)/N(v.)}.$$

Unsurprisingly Pearson statistics gives the best results in order to distinguish order 1 Markov chains from order 2, as well as mixed order 2 defined in (3.10). However the CGR-based test is very effective in the particular set of mixed Markov chains of order > 2.

4 Genomic signature

This section is devoted to evaluating the performance of CGR-based relative abundance (1.12) as a way to characterize DNA sequences. Karlin and Mràzek [21], Campbell *et al.* [5] study the profile of dinucleotide relative abundance of different DNA sequences. This profile is the set of all

$$\hat{\rho}_{uv} \stackrel{\text{def}}{=} \frac{\hat{\pi}_N(Suv)}{\hat{\pi}_N(Su)\hat{\pi}_N(Sv)},$$

for all nucleotides u and v, where the empirical measures are computed from the sequence concatenated with its inverted complement. This set constitutes a "genomic signature" (Karlin and Burge [17], Karlin and Cardon [18]) that may reflect the influence of factors such as functions of the replication and repair machinery or context dependent mutation rates. Dinucleotide relative abundance values are equivalent to the "general designs" derived from biochemical nearest-neighbor frequency analysis (Josse *et al.* [16], Russel and Subak-Sharpe [30]).

It seems remarkable that dinucleotide relative abundance profiles exhibit local stability in the sense that, when computed for any 50kb window on a given genome, the profile is about the same as when computed globally from all the DNA of the organism (Karlin and Mràzek [21], Karlin *et al.* [22]). The stability of this profile could result from constraints on dinucleotide stacking energy and DNA helicity, context dependent mutation pressures, and replication and repair mechanisms (Karlin and Mràzek [21], Karlin *et al.* [22], Karlin and Burge [17] and Blaisdell *et al.* [19]). Jernigan and Baran [15] test the hypothesis that patterns of dinucleotide over-and-under representation in a given genome are invariant.

Campbell *et al.* [5] highlight the advantages of using genomic signature in order to reconstruct phylogenetic trees. Conventional methods employ similarity or dissimilarity assesments of aligned homologous genes or regions. Alignments of related long sequences are generally not feasible and different phylogenetic reconstructions may result for the same set of organisms based on analysis of different sequences. Because of the local stability of dinucleotide relative abundance profile, a tree based on matrice differences of profiles is independent of which genome segments of 50kb is used in its construction. Moreover the signature pervades both coding and non coding DNA.

In this section, we aim at comparing the performance of the CGR-based relative abundance vs. dinucleotide relative abundance when it comes to classifying DNA sequences. To this end, several sequences have been selected according to Table 4.1 (see Figure 4.1 for the the corresponding lineage tree). Subsequences of length 100kb are extracted from them and concatenated with their inverted complement.

Abbr	Sequence	GenBank
homsa1	Homo Sapiens	NT_022184.13
homsa2	Homo Sapiens	NT_005403.14
homsa3	Homo Sapiens	NT_025741.13
homsa4	Homo Sapiens	NT_011520.9
homsa5	Homo Sapiens	NT_011757.13
mmus	Musmusculus	NT_078586.1
ratn1	Rattus Norvegicus	NC_005118
ratn2	Rattus Norvegicus	NC_005117
ratn3	Rattus Norvegicus	NC_005107
ratn4	Rattus Norvegicus	NC_005105
gal1	Gallus gallus	NC_006097.1
gal2	Gallus gallus	NC_006096.1
gal3	Gallus gallus	NC_006095.1
gal4	Gallus Gallus	NC_006094.1
gal5	Gallus Gallus	NC_006093.1
gal6	Gallus Gallus	NC_006092.1
gal7	Gallus Gallus	NC_006091.1
agam1	Anopheles gambiae	NW_045719.1
agam2	Anopheles gambiae	NW_045746.1
agam3	Anopheles gambiae	NW_045763.1
agam4	Anopheles gambiae	NW_045815.1
dmela1	Drosophila melanogaster	NC_004354.1
dmela2	Drosophila melanogaster	NT_033779.2
dmela3	Drosophila melanogaster	NT_033778.1
dmela4	Drosophila melanogaster	NT_037436.1
dmela5	Drosophila melanogaster	Arm X
dmela6	Drosophila melanogaster	Arm2R
dmela7	Drosophila melanogaster	Arm 2L
dmela8	Drosophila melanogaster	Arm3L
dmela9	Drosophila melanogaster	Arm4
dmela10	Drosophila melanogaster	Arm3R

Abbr	Sequence	GenBank
celeg1	Caenorhabditis elegans	CHR_I
celeg2	Caenorhabditis elegans	CHR_II
celeg3	Caenorhabditis elegans	CHR_III
celeg4	Caenorhabditis elegans	CHR_IV
celeg5	Caenorhabditis elegans	CHR_V
celeg6	Caenorhabditis elegans	CHR_X
pfal	Plasmodium Falciparum	NC_004317
ylip1	Yarrowia Lipolytica	NC_006072
ylip2	Yarrowia Lipolytica	NC_006071
ylip3	Yarrowia Lipolytica	NC_006070
ylip4	Yarrowia Lipolytica	NC_006069
osat1	Oryza Sativa	NT_036323
osat2	Oryza Sativa	NT_079973
osat3	Oryza Sativa	NT_080060
osat4	Oryza Sativa	NT_080067
osat5	Oryza Sativa	NT_080068
athal1	Arabidopsis thaliana	NC_003070
athal2	Arabidopsis thaliana	NC_003071.3
athal3	Arabidopsis thaliana	NC_003074.4
athal4	Arabidopsis thaliana	NC_003075.3
athal5	Arabidopsis thaliana	NC_003076.4
bacc	Bacillus cereus	NC_004722
braj	Bradhyrhizobium japonicum	NC_004463
ccre	Caulobacter crescentus	NC_002696
mlot	Mesorhizobium loti	NC_002678
mbov	Mycobacterium bovis	AF2122/97
save	Streptomyces Avermitilis	NC_003155
scoe	Streptomyces Coelicolor	NC_003888
mace	Methanosarcina Acetivorans C2A	NC_003552
maze	Methanosarcina Mazei	NC_003901
ssol	Sulfolobus Solfataricus P2	NC_002754.1

Table 4.1: List of sequences used in the simulation for genomic signature. All these sequences are available from `http://www-rocq.inria.fr/~cenac/sequences.html`

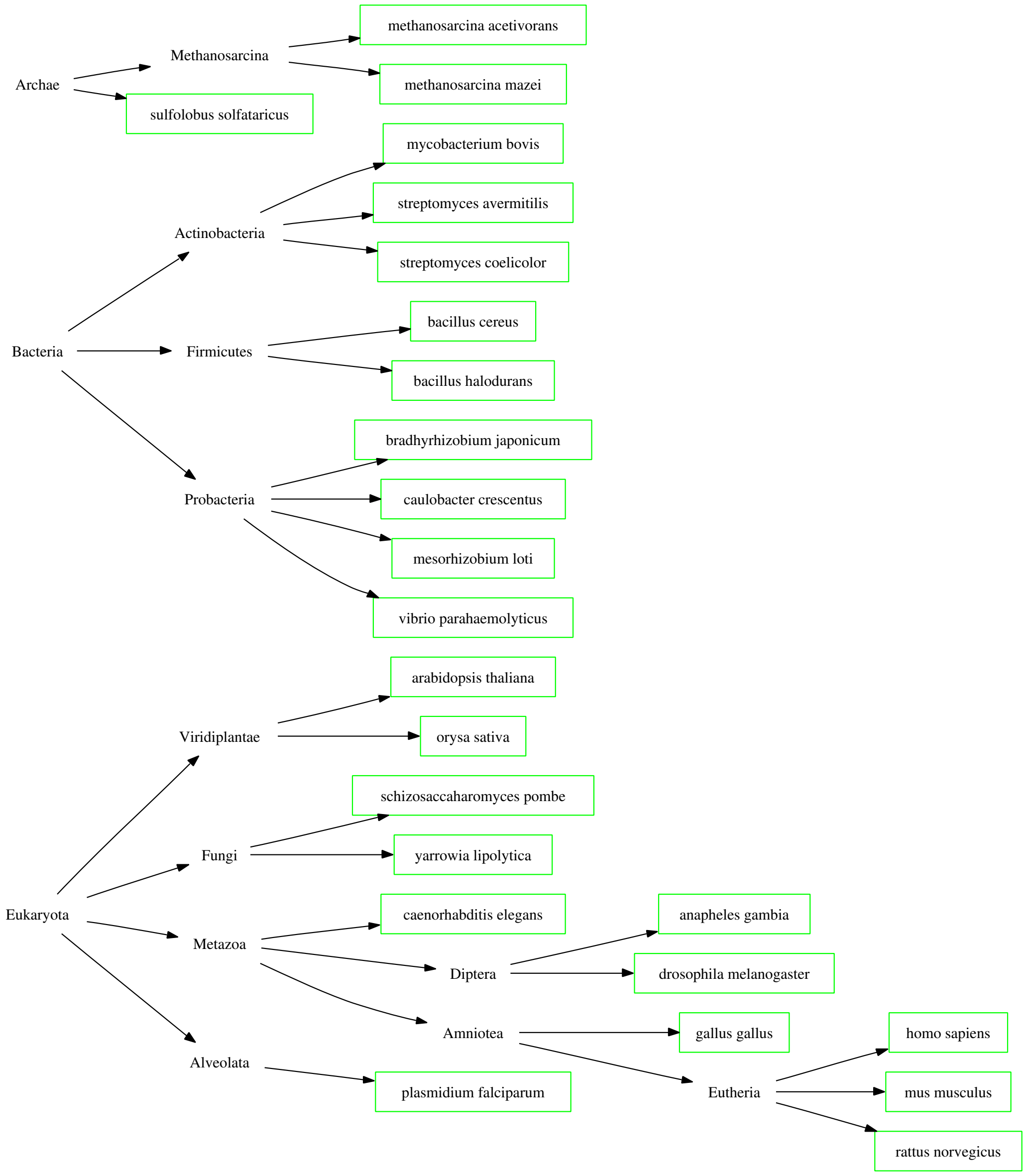


Figure 4.1: Lineage tree of the species used in the experiments.

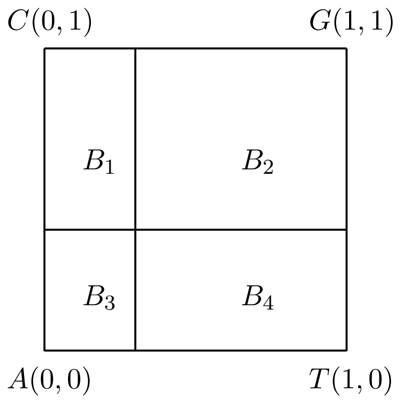


Figure 4.2: The partition of the square $[0,1[^2$ in 4 rectangles used for the computation of the CGR-based relative abundance in the experimentation.

The CGR-based relative abundance (1.12) is approximated by the empirical CGR-based relative abundance

$$\hat{\rho}(B,v) \stackrel{\text{def}}{=} \frac{\hat{\pi}_N(Bv)}{\hat{\pi}_N(B)\hat{\pi}_N(Sv)}.$$

For a species Σ (resp. Σ') containing N (resp. N') sequences, let us denote by $\hat{\rho}_i(B,v)$ (resp. $\hat{\rho}'_i(B,v)$) the empirical CGR-based relative abundance for sequence i. For a given partition $\{B_1,\dots,B_K\}$ of the unit square, the *CGR-based relative abundance difference* is defined for $\Sigma \neq \Sigma'$ as

$$\delta(\Sigma,\Sigma') = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N'}\sum_{j=1}^{N'}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{4}\sum_{v\in\mathcal{A}}\left|\hat{\rho}_i(B_k,v)-\hat{\rho}'_j(B_k,v)\right|,$$

with a special casing when $\Sigma = \Sigma'$:

$$\delta(\Sigma,\Sigma) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{N-1}\sum_{j\neq i}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{4}\sum_{v\in\mathcal{A}}\left|\hat{\rho}_i(B_k,v)-\hat{\rho}_j(B_k,v)\right|.$$

When the partition used is $\{Su, u\in\mathcal{A}\}$, δ coincides with the *dinucleotide relative abundance difference* defined by Karlin and Mràzek [21], which is tabulated in Table 4.3. Alternatively, the partition of $S=[0,1[^2$ depicted in Figure 4.2 gives rise to a CGR-based relative abundance difference shown in Table 4.2. In the two cases, within-species differences are very few and *Eutheria*, *Diptera*, *Plants*, *Firmicutes*, *Probacteria*, *Actinobacteria* and *Archae* form coherent groups. Comparing maximal within-group differences to minimal between-group differences, one can observe that 14 species are "closer" to their group than the others in the simulation based on CGR-based relative abundance, whereas this number is only 10 species for the dinucleotides relative abundance. The demarcation is made more evident for the following species: *Streptomyces Coelicolor*, *Streptomyces Avermitilis*, the *Bacillus*, *Plasmidium falciparum*, *Yarrowia Lipolytica*.

Karlin and Mràzek [21] build phylogenetic trees from the matrices of measures of differences. The trees in Figure 4.3 are generated with NJPLOT method (Neighbor Joining method, Saitou and Nei [31], Perrière and Gouy [26]) from the matrices in Table 4.3 and Table 4.4 respectively. Moreover, a phylogenetic tree of all the Bacteria sequences only can be built (Figure 4.4). The sequences are grouped by species, excepted one sequence of *Streptomyces Coelicolor* and one sequence of *Streptomyces Avermitilis* for the both experiences. In the tree built on CGR-based δ, the three groups *Firmicutes*, *Probacteria* and *Actinobacteria* are clearly separated in three distinct families.

To investigate the influence of the size of the partition, the same computation have been carried on a regular 10×10 grid, which corresponds in Almeida *et al.* [1] terminology to counting polynucleotides with fractional length $\log_2 10 \approx 3.32$. Table 4.4 gives the average difference δ (multiplied by 1000) between all the ratios for the partition of $(B_i)_{1 \leq i \leq 100}$ defined by the grid. The number (17) of species having maximal within-group differences smaller than minimal between-group differences is even larger. *Drosophila Melanogaster* and *Oryza Sativa* are demarcated. The tree built with Table 4.4 is given in Figure 4.5.

	mace	maze	ssol	mbov	save	scoe	bacc	bach	braj	ccre	mlot	pfal	ylip	celeg	gal	homsa	mmus	ratn	agam	dmela	athal	osat
samples	(40)	(40)	(29)	(40)	(41)	(41)	(40)	(40)	(41)	(40)	(41)	(32)	(43)	(22)	(68)	(53)	(43)	(77)	(10)	(45)	(28)	(42)
mace	23	28	*103*	162	173	173	80	*68*	139	111	130	140	115	91	135	114	144	140	116	106	94	87
maze		18	*95*	177	187	187	84	*82*	151	119	140	140	124	103	126	110	138	136	126	111	102	88
ssol	*103*		24	189	185	190	*98*	128	193	143	203	140	127	163	122	109	112	114	164	153	114	111
mbov				19	109	*118*	121	101	107	*88*	106	182	113	126	198	200	202	193	107	143	133	148
save				**109**	**36**	**37**	151	137	130	112	150	240	135	184	224	230	223	215	182	212	161	199
scoe				**118**	**37**	**29**	158	144	144	122	165	246	140	187	228	235	227	219	186	220	165	208
bacc							**27**	**49**	130	104	129	101	89	88	118	106	133	129	83	78	75	61
bach								**17**	104	85	101	129	86	69	135	123	150	146	77	83	76	76
braj									**31**	**70**	**43**	213	155	148	204	201	223	219	168	171	161	157
ccre										**20**	**78**	189	116	134	156	156	168	165	142	152	127	139
mlot											**22**	203	151	135	197	191	219	215	156	155	154	149
pfal												36	142	131	156	127	164	160	131	106	122	93
ylip													28	110	104	110	105	97	106	122	52	119
celeg														54	151	137	164	162	74	80	98	90
gal															**34**	**60**	**47**	**50**	147	145	87	135
homsa																**37**	**61**	**60**	137	125	88	104
mmus																	**27**	**34**	162	166	94	148
ratn																		**34**	156	163	87	146
agam																			**43**	**67**	98	90
dmela																			*67*	36	108	*57*
athal													*52*								34	*96*
osat																				*57*		28

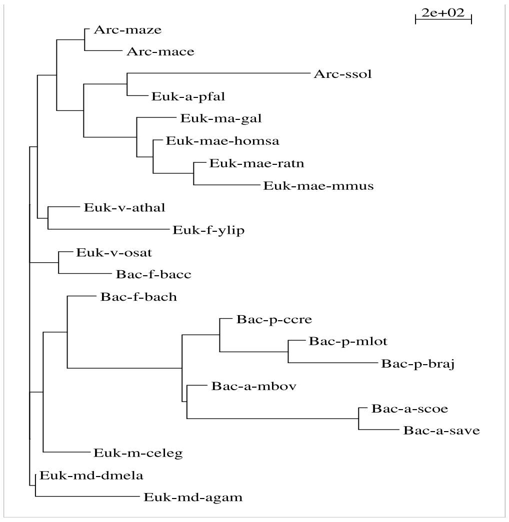
Table 4.2: The CGR-based relative abundance difference δ (multiplied by 1000) between the species described in Table 4.1, using the partition of $[0,1[^2$ given in Figure 4.2. When the maximal within-group difference is smaller than all between-group differences, the corresponding values are in bold. Otherwise, the problematic values are in italics.

	mace	maze	ssol	mbov	save	scoe	bacc	bach	braj	ccre	mlot	pfal	ylip	celeg	gal	homsa	mmus	ratn	agam	dmela	athal	osat
samples	(40)	(40)	(29)	(40)	(41)	(41)	(40)	(40)	(41)	(40)	(41)	(32)	(43)	(22)	(68)	(53)	(43)	(77)	(10)	(45)	(28)	(42)
mace	26	32	*108*	202	208	209	109	89	203	164	184	168	122	102	166	131	156	149	148	127	*83*	90
maze		20	*102*	209	224	225	113	97	213	174	193	172	129	113	152	119	148	142	153	132	*87*	95
ssol	**108**	**102**	**31**	243	246	248	145	171	267	218	272	147	151	183	165	124	136	131	212	188	125	123
mbov				20	118	*134*	143	122	130	*97*	115	195	132	147	237	246	253	241	114	162	157	155
save				*134*	36	36	175	150	154	*112*	168	254	143	183	277	280	279	268	172	217	175	191
scoe				*134*		25	181	157	166	*122*	183	268	147	186	281	285	282	271	177	226	178	201
bacc							32	*71*	171	131	157	155	121	94	146	152	181	168	77	69	88	*48*
bach							*71*	20	135	99	114	184	121	*66*	179	183	206	198	78	87	97	84
braj									**35**	**63**	**52**	270	213	179	279	290	309	302	180	195	204	186
ccre										**19**	**69**	236	162	144	233	242	257	250	148	170	153	140
mlot											**23**	256	207	151	266	278	302	295	150	167	190	175
pfal												38	174	183	224	178	199	193	193	168	161	128
ylip													27	129	148	149	144	131	125	141	64	123
celeg														59	172	173	191	183	90	95	102	102
gal															38	68	68	67	174	155	125	140
homsa																**35**	**57**	**54**	189	163	123	129
mmus																	**31**	**39**	215	196	132	160
ratn																		**38**	201	186	117	150
agam																			**48**	**76**	116	103
dmela							*69*												*76*	43	114	76
athal													*64*								34	*86*
osat							*48*														*86*	31

Table 4.3: The dinucleotides relative abundance difference (multiplied by 1000) between the species described in Table 4.1. When the maximal within-group difference is smaller than all between-group differences, the corresponding values are in bold. Otherwise, the problematic values are in italics.

	mace	maze	ssol	mbov	save	scoe	bacc	bach	braj	ccre	mlot	pfal	ylip	celeg	gal	homsa	mmus	ratn	agam	dmela	athal	osat
samples	(5)	(5)	(3)	(5)	(6)	(6)	(5)	(5)	(6)	(5)	(6)	(4)	(8)	(22)	(68)	(53)	(8)	(77)	(10)	(45)	(28)	(7)
mace	21	40	*171*	259	275	295	*156*	*156*	259	260	246	243	182	161	194	192	195	191	192	171	157	178
maze		18	*178*	280	294	313	*170*	174	274	275	262	251	197	176	185	188	191	189	206	182	170	187
ssol			42	284	306	325	*156*	193	304	302	309	240	190	212	205	204	202	198	224	216	158	186
mbov				19	*165*	191	205	186	*146*	167	148	307	191	209	312	325	324	317	194	213	211	233
save				**165**	**56**	**65**	249	223	198	183	215	332	214	253	334	352	342	335	252	280	245	279
scoe				**191**	**65**	**42**	273	248	221	202	239	352	237	275	354	372	363	356	275	304	267	297
bacc							**28**	**106**	229	239	227	226	157	146	194	220	224	217	129	133	138	150
bach								**19**	187	203	184	251	158	145	222	244	240	233	143	155	137	174
braj									**33**	**120**	**70**	347	246	235	335	353	351	346	235	248	245	261
ccre										**23**	**120**	345	234	245	330	342	332	328	256	261	243	269
mlot											**28**	342	244	226	328	343	341	336	232	235	245	256
pfal												45	221	237	262	242	240	235	243	235	215	213
ylip													28	148	191	197	183	176	157	154	99	146
celeg														**65**	195	200	206	200	**123**	**115**	131	134
gal															**43**	**100**	**95**	**97**	203	183	179	185
homsa																**49**	**81**	**86**	224	192	177	173
mmus																	**38**	**46**	222	201	168	185
ratn																		**43**	215	198	162	182
agam																			**50**	**101**	148	142
dmela																				**35**	146	125
athal													*99*								37	*120*
osat																					**120**	**31**

Table 4.4: The CGR-based relative abundance difference δ (multiplied by 1000) between the species described in Table 4.1, using a regular 10×10 grid as partition of $[0,1[^2$. When the maximal within-group difference is smaller than all between-group differences, the corresponding values are in bold. Otherwise, the problematic values are in italics.



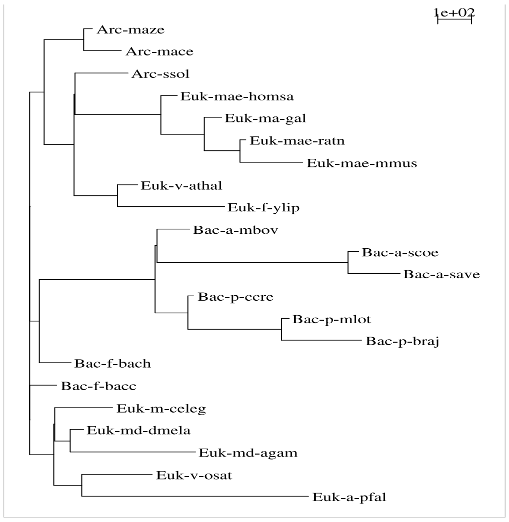


Figure 4.3: Unrooted lineage trees built using the NJPLOT method from the dinucleotide relative abundance differences of Table 4.3 (top) and from the CGR-based relative abundance differences of Table 4.2 (bottom). *Amniotea*, *Probacteria* and *Actinobacteria* form coherent groups. On the contrary, *Firmicutes*, *Metazoa*, *Viridiplantae* are separated and *Arcae* are mixed with some *Eukaryota*.

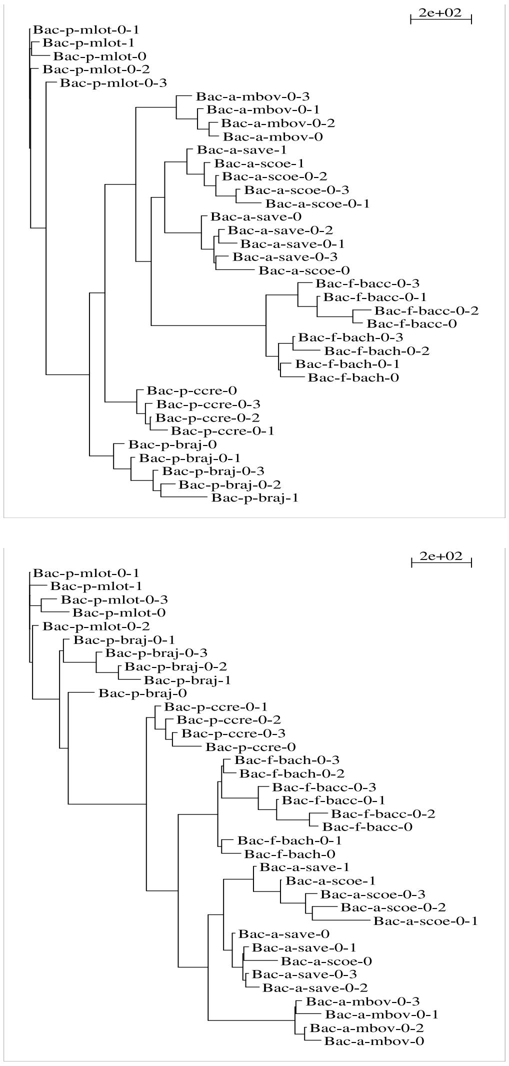


Figure 4.4: Unrooted lineage trees built using the NJPLOT method from the dinucleotide relative abundance differences (top) and the CGR-based relative abundance differences of (bottom) for all the sequences of Bacteria. With the CGR, the 3 groups *Firmicutes*, *Probacteria* and *Actinobacteria* are demarcated, whereas *Firmicutes* appears as a family of *Actinobacteria* in the top tree.

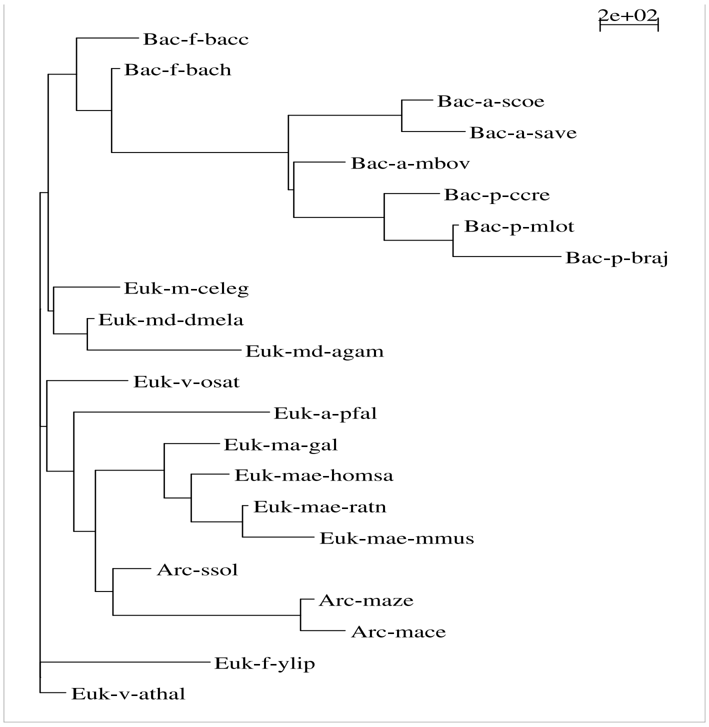


Figure 4.5: Unrooted lineage trees built using the NJPLOT method from the CGR-based relative abundance differences of Table 4.4 with a regular 10×10 partition of the unit square.

A Proof of Theorem 2.2

Lemma A.1. *The infinite product $P(t)$ defined in (2.3) satisfies the following equation for $-1 < \sigma < 0$,*

$$\log P(t) = \frac{1}{2i\pi} \int_{\sigma - i\infty}^{\sigma + i\infty} \frac{1}{\rho^\theta - 1} \Gamma(\theta) \hat{\Phi}\Big(\frac{-a}{1-a}, \theta + 1, 1 \Big) e^{\frac{i\theta\pi}{2}} t^{-\theta} \, d\theta. \tag{A.1}$$

Proof. The proof relies on a Mellin transform argument. Let us just recall the definition. The Mellin Transform $f^*(s)$ of a complex-valued function $f(x)$ defined over the positive real line is

$$\mathcal{M}[f(x); s] \stackrel{\text{def}}{=} f^*(s) \stackrel{\text{def}}{=} \int_0^{+\infty} f(x) x^{s-1} \, dx$$

with s being a complex number. The Mellin transform allows the reduction of certain functional equations to algebraic ones. The usefulness of the Mellin transform stems for its asymptotic properties. There is a direct mapping between asymptotic expansions of a function near zero or infinity and the set of singularities of the transform in the complex plane.

Note that $\log P(t)$ is defined for $t > 0$, since $\log(1 - a + ae^{it})$ keeps the same determination whenever $\left| \frac{a}{1-a} \right| < 1$. For $-1 < \theta < 0$, the Mellin transform of $\log P(t)$ is given by

$$\begin{aligned}
(\log P)^*(\theta) &= \int_0^{+\infty} \sum_{j=0}^{+\infty} \log\bigg(1 - a + a \exp\Big(\frac{it}{4^{j+1}} \Big) \bigg) t^{\theta - 1} \, dt \\
&= \frac{1}{\rho^\theta - 1} \int_0^{+\infty} \log(1 - a + ae^{iu}) u^{\theta - 1} \, du.
\end{aligned}$$

As a consequence, it remains to study the generic quantity

$$\begin{aligned}
k^*(\theta) &\stackrel{\text{def}}{=} \int_0^{+\infty} \log(1 - a + ae^{it}) t^{\theta - 1} \, dt. \\
&= -\frac{ia}{\theta} \int_0^{+\infty} \frac{e^{it} t^\theta}{1 - a + ae^{it}} \, dt.
\end{aligned} \tag{A.2}$$

The second equality is the result of an integration by parts. Indeed, $\log\big(1 - a + ae^{it}\big)$ is bounded in t and $t^\theta \underset{t \to +\infty}{\longrightarrow} 0$, and $\log\big(1 - a + ae^{it}\big) t^\theta \underset{t \to +0}{\longrightarrow} 0$, for $-1 < \theta < 0$. Moreover, the integral in (A.2) is convergent because $t \mapsto t^\theta$ is decreasing and positive for $-1 < \theta < 0$. Decomposing the function $\frac{1}{1-u}$ in its series expansion, we have

$$\begin{aligned} k^*(\theta) &= \frac{ia}{\theta}\int_0^{+\infty}\frac{e^{it}t^\theta}{1-a}\sum_{n=0}^{+\infty}\Big(\frac{-a}{1-a}\Big)^n e^{int}\,dt \\ &= -\frac{i}{\theta}\sum_{n=0}^{+\infty}\frac{1}{(n+1)^{\theta+1}}\Big(\frac{-a}{1-a}\Big)^{n+1}\int_0^{+\infty} e^{iu}u^\theta\,du. \end{aligned}$$

It ensures that

$$k^*(\theta) = -\frac{i}{\theta}\hat\Phi\Big(\frac{-a}{1-a},\theta+1,1\Big)\int_0^{+\infty} e^{iu}u^\theta\,du.$$

On the other hand, it is easy to prove (see for example Dieudonné [8])

$$\int_0^{+\infty} e^{it}t^\theta\,dt = e^{\frac{(\theta+1)i\pi}{2}}\Gamma\big(\theta+1\big).$$

Hence, we obtain for $-1 < \theta < 0$,

$$k^*(\theta) = \hat\Phi\Big(\frac{-a}{1-a},\theta+1,1\Big)e^{\frac{i\theta\pi}{2}}\Gamma(\theta). \tag{A.3}$$

According to the analytic extension theorem, the previous equality can be spread to $-1 < \Re(\theta) < 0$. Then Mellin's inversion formula leads to the following equation, for $-1 < \Re(\theta) < 0$

$$\log P(t) = \frac{1}{2i\pi}\int_{\sigma-i\infty}^{\sigma+i\infty} h^*(\theta)t^{-\theta}\,d\theta, \quad \text{where} \quad h^*(\theta) = \frac{k^*(\theta)}{\rho^\theta - 1}.$$

This equality, together with (A.1), conclude the proof of Lemma A.1. □

Proof of Theorem 2.2. We apply the residues formula to the function F_t along the path Γ depicted in Figure A.1. It follows that

$$\frac{1}{2i\pi}\int_{\sigma-i\infty}^{\sigma+i\infty} F_t(\theta)\,d\theta = -\sum_{k\in\mathbb{Z}} Res(F_t,\theta_k) + \frac{1}{2i\pi}\int_{\sigma_1-i\infty}^{\sigma_1+i\infty} F_t(\theta)\,d\theta,$$

with $\theta_k = -\frac{2ik\pi}{\log\rho}$. Indeed, according to Jordan's lemma,

$$\lim_{R\to+\infty}\int_{\gamma_3} F_t(\gamma_3(\theta))\gamma_3'(\theta)\,d\theta = \lim_{R\to+\infty}\int_{\gamma_4} F_t(\gamma_4(\theta))\gamma_4'(\theta)\,d\theta = 0,$$

concluding the proof of (2.4). □

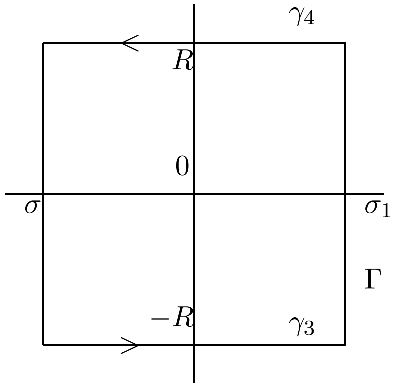


Figure A.1: Integration path Γ

Proof of Corollary 2.3. The computation of the residue at $\theta_k = -\frac{2ik\pi}{\log\rho}$ (simple pole), for $k \in \mathbb{Z}^*$, and the residueat 0 (double pole) gives

$$
\begin{aligned}
Res(F_t, \theta_k) &= \frac{\hat{\Phi}(\frac{-a}{1-a}, \theta_k + 1, 1) e^{\frac{i\theta_k \pi}{2}} \Gamma(\theta_k)(t\rho)^{-\theta_k}}{\log\rho}, \\
Res(F_t, 0) &= \frac{\log(1-a)}{\log\rho} \log t + \log(1-a)\Big(\frac{\gamma}{\log\rho} + 1 - \frac{i\pi}{2\log\rho}\Big) \\
&\quad + \frac{1}{\log\rho} \sum_{k=0}^{+\infty} \Big(\frac{-a}{1-a}\Big)^{k+1} \frac{\log(k+1)}{k+1}.
\end{aligned}
$$

For all $\sigma_1 > 0$,

$$
\frac{1}{2i\pi} \int_{\sigma_1 - i\infty}^{\sigma_1 + i\infty} F_t(\theta)\, d\theta = o\big(t^{-\sigma_1}\big).
$$

Hence, by Theorem 2.2,

$$
\begin{aligned}
\log P(t) = &\quad \frac{\log(1-a)}{\log\rho} \log t + \log(1-a)\Big(\frac{\gamma}{\log\rho} + 1 - \frac{i\pi}{2\log\rho}\Big) \\
+ &\quad \frac{1}{\log\rho} \sum_{k=0}^{+\infty} \Big(\frac{-a}{1-a}\Big)^{k+1} \frac{\log(k+1)}{k+1} \\
+ &\quad \sum_{k \in \mathbb{Z}^*} Res(F_t, \theta_k) + o\big(t^{-\sigma_1}\big).
\end{aligned}
$$

We are left with the study of the oscillating series

$$\sum_{k\in\mathbb{Z}^*} e^{\frac{-k\pi}{2}}\Gamma\Big(\frac{-2ik\pi}{\log\rho}\Big)t^{\frac{2ik\pi}{\log\rho}}\sum_{n=0}^{+\infty}\Big(\frac{-a}{1-a}\Big)^{n+1}\frac{1}{(n+1)^{\frac{-2ik\pi}{\log\rho}+1}}. \tag{A.4}$$

As $k\to\pm\infty$, the following well known equivalent holds

$$\Big|\Gamma\Big(\frac{-2ik\pi}{\log\rho}\Big)\Big| \sim e^{\frac{k\pi^2}{\log\rho}}\Big|\frac{k}{\log\rho}\Big|^{-1/2}.$$

So the series in (A.4) converges uniformly for $k\in\mathbb{N}$. When k is a negative integer, a more precise analysis is necessary. The Lerch function has the integral form

$$\sum_{n=1}^{+\infty}\Big(\frac{a}{a-1}\Big)^n\frac{1}{n^{i\theta_k+1}} = \frac{1}{\Gamma(\theta_k+1)}\int_0^{+\infty}\frac{ae^{-t}}{a-1-ae^{-t}}t^{-i\frac{2k\pi}{\log\rho}\log t}\,dt.$$

The integral in the last formula is the subject of the next lemma.

Lemma A.2.

$$\Big|\int_0^{+\infty}\frac{e^{-i\alpha\log t}}{e^t-u}\,dt\Big| = \mathcal{O}\Big(\frac{e^{-\theta\alpha}}{\mid\alpha\mid}\Big).$$

Proof. An integration by parts yields

$$\int_0^{+\infty}\frac{e^{-i\alpha\log t}}{e^t-u}\,dt = \int_0^{+\infty}\frac{te^{-i\alpha\log t}e^t}{(1-i\alpha)(e^t-u)^2}\,dt.$$

Setting

$$f(z) \stackrel{\text{def}}{=} \frac{ze^{-i\alpha\log z}e^z}{(e^z-u)^2},$$

we have

$$\int_0^{+\infty}\frac{e^{-i\alpha\log t}}{e^t-u}\,dt = \frac{1}{1-i\alpha}\int_0^{+\infty}f(t)\,dt.$$

Consider the integration path drawn in Figure A.2. $\mathcal{C}$ is the union of the paths γ_1, γ_2, and γ_3. $\mathcal{L}$ is the union of the paths γ_4, γ_5, γ_6, γ_7 and γ_8. We have

$$\int_{\mathcal{C}} f(z)\,dz = -\int_{\mathcal{L}} f(z)\,dz.$$

Then

$$\lim_{\substack{\epsilon\to0\\ \delta\to0\\ A\to+\infty}}\int_{\mathcal{C}} f(z)\,dz = (e^{2\pi\alpha}-1)\int_0^{+\infty}\frac{te^{-i\alpha\log t}e^t}{(e^t-u)^2}\,dt,$$

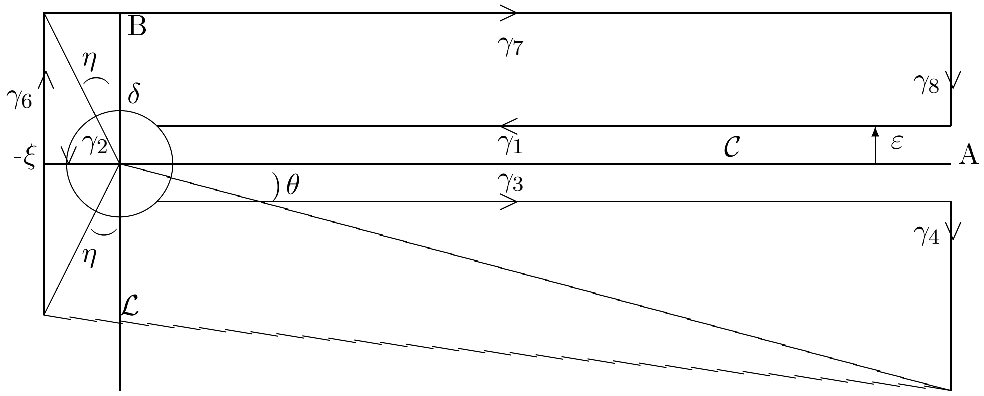


Figure A.2: Integral paths $\mathcal{C}$ and $\mathcal{L}$.

so that

$$\int_0^{+\infty} \frac{e^{-i\alpha \log t}}{e^t - u}\, dt = \frac{1}{1-i\alpha}\, \frac{1}{e^{2\pi\alpha}-1} \lim_{\substack{\epsilon\to 0\\ \delta\to 0\\ A\to+\infty}} - \int_{\mathcal{L}} f(z)\, dz.$$

To study the asymptotic behavior of $f(z)$ on the path Γ, we note on the one hand,

$$\lim_{A\to+\infty} \left| \int_{\gamma_7} f(z)\, dz \right| \quad \leq \quad e^{(\frac{\pi}{2}+\eta)\alpha} \int_{-\xi}^{+\infty} \frac{e^t \mid \gamma_7(t) \mid}{\mid e^t - u \mid^2}\, dt. \tag{A.5}$$

On the other hand, it is easy to see that

$$\lim_{\substack{\epsilon\to 0\\ A\to+\infty}} \left| \int_{\gamma_6} f(z)\, dz \right| = 0 \quad \text{and} \quad \lim_{A\to+\infty} \left| \int_{\gamma_4} f(z)\, dz \right| = 0.$$

Moreover

$$\left| \int_{\gamma_5} f(z)\, dz \right| \leq e^{(2\pi-\theta)\alpha} \int_{-\xi}^{A} \frac{e^t \mid \gamma_5(t) \mid}{(e^t - \mid u \mid)^2}\, dt,$$

the integral on the right-hand side being convergent, whence there exists a constant $C(u,\xi)$ such that

$$\lim_{A\to+\infty} \left| \int_{\gamma_5} f(z)\, dz \right| \leq C(u,\xi) e^{(2\pi-\theta)\alpha}.$$

Thus Lemma A.2 follows. $\square$

Getting through the limit for $\theta \to \frac{\pi}{2}$ Corollary 2.3 follows, since

$$\Big| \sum_{n=1}^{+\infty} \Big(\frac{a}{a-1} \Big)^n \frac{1}{n^{i\theta_k + 1}} \Big| = \mathcal{O}\Big(\frac{1}{\mid k \mid^{3/2}} \Big).$$

□

B Proof of Theorem 3.1

Lemma B.1. *Under the hypothesis H_0, the asymptotic behavior of the difference*

$$D_N(B,u) \stackrel{\text{def}}{=} \sqrt{N}\big(\hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu)\big)$$

is the following.

$$D_N(B,u) \stackrel{a.s.}{\sim} \frac{1}{\sqrt{N}} \sum_{j=1}^{N} \varepsilon_j(u) V_{j-1}(B),$$

where $\varepsilon_j(u) \stackrel{\text{def}}{=} \mathbb{1}_{\{u_j = u\}} - p_u$ and $V_j(B) \stackrel{\text{def}}{=} \pi(B) - \mathbb{1}_{\{B\}}(X_j)$.

Proof. Under the hypothesis H_0, we study the asymptotic behavior of the quantity

$$\begin{aligned} D_N(B,u) \quad &= \quad \sqrt{N}\Big(\big(\hat{\pi}_N(Su) - p_u\big)\big(\hat{\pi}_N(B) - \pi(B)\big) - \hat{\pi}_N(Bu) \\ &\qquad + p_u \hat{\pi}_N(B) + \pi(B)\big(\hat{\pi}_N(Su) - p_u\big)\Big), \end{aligned}$$

where $p_u \stackrel{\text{def}}{=} \pi(Su) = \mathbb{P}(v_1 = u)$, by the strong law of large numbers. Then, the central limit theorem yields

$$\sqrt{N}\big(\hat{\pi}_N(Su) - p_u\big) \stackrel{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, p_u(1-p_u)).$$

Thus, the convergence of the empirical estimate, together with Slutsky lemma, ensure that

$$\sqrt{N}\Big(\hat{\pi}_N(Su) - p_u\Big)\Big(\hat{\pi}_N(B) - \pi(B)\Big) \xrightarrow[\text{N}\to\infty]{\text{a.s.}} 0. \tag{B.1}$$

It remains to study the asymptotic behavior of the quantity

$$\begin{aligned}
&\sqrt{N}\Big[-\hat{\pi}_N(Bu) + p_u\hat{\pi}_N(B) + \pi(B)\big(\hat{\pi}_N(Su) - p_u\big)\Big] \\
&\quad = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\Big[-\mathbb{1}_{\frac{B+\ell_u}{2}}(X_j) + p_u\mathbb{1}_B(X_j) + \pi(B)\varepsilon_j(u)\Big] \\
&\quad = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\Big[-\mathbb{1}_{\{u_j=u\}}\mathbb{1}_B(X_{j-1}) + p_u\mathbb{1}_B(X_{j-1}) + \pi(B)\varepsilon_j(u)\Big] \\
&\qquad - \frac{1}{\sqrt{N}}p_u\big(\mathbb{1}_B(X_0) - \mathbb{1}_B(X_N)\big). \\
&\quad = \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\varepsilon_j(u)V_{j-1}(u) - \frac{1}{\sqrt{N}}p_u\big(\mathbb{1}_B(X_0) - \mathbb{1}_B(X_N)\big).
\end{aligned}$$

The second term tends to 0, and Lemma B.1 follows. □

Proof of Theorem 3.1. According to Lemma B.1,

$$D_N(B,u) \overset{a.s.}{\sim} \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\varepsilon_j(u)V_{j-1}(B) \overset{\text{def}}{=} \frac{1}{\sqrt{N}}M_N(B,u). \tag{B.2}$$

$M_N(B,u)$ is a $\mathcal{F}$-martingale with increasing process

$$\langle M(B,u)\rangle_N = \sum_{j=0}^{N}\mathbb{E}\big[\varepsilon_j^2 \mid \mathcal{F}_{j-1}\big]V_{j-1}^2(B) = p_u(1-p_u)\sum_{j=0}^{N}V_{j-1}^2(B),$$

and, under hypothesis H_0, it is easy to see that

$$\sigma^2 \overset{\text{def}}{=} \lim_{N\to\infty}\frac{1}{N}\langle M(B,u)\rangle_N = p_u(1-p_u)\pi(B)\big(1-\pi(B)\big).$$

Since $M_N(u,B)$ has bounded jumps, Lindeberg's condition is clearly satisfied and the central limit theorem for martingales (see for example Hall and Heyde [12]) yields

$$\frac{1}{\sqrt{N}}M_N(B,u) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0,\sigma^2). \tag{B.3}$$

Hence, under hypothesis H_0, (B.3) together with (B.2) imply

$$\sqrt{N}\Big(\hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu)\Big) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0,\sigma^2).$$

Since $\hat{\sigma}_N(u, B)$ is a consistent estimate of σ, Slutsky's lemma ensures that

$$\frac{\sqrt{N}}{\hat{\sigma}_N(B,u)}\Big(\hat{\pi}_N(Su)\hat{\pi}_N(B) - \hat{\pi}_N(Bu)\Big) \overset{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0,1),$$

and Theorem 3.1 follows. □

C Proof of Theorem 3.2

Lemma C.1.

$$\sum_{i=1}^{k}\sum_{u\in\mathcal{A}} R_N(B_i,u)^2 \quad \overset{a.s.}{\sim} \quad \sum_{i=1}^{k}\sum_{u\in\mathcal{A}} \frac{1}{N}\Big(\sum_{j=1}^{N} \frac{V_{j-1}(B_i)}{\sqrt{\pi(B_i)}}\frac{\varepsilon_j(u)}{\sqrt{p_u}}\Big)^2$$

Proof. We study the asymptotic behavior of

$$\begin{aligned} R_N(B_i,u) &= \frac{D_N(B_i,u)}{\sqrt{\hat{\pi}_N(Su)\hat{\pi}_N(B_i)}} \\ &= \frac{D_N(B_i,u)}{\sqrt{p_u\pi(B_i)}}\ \frac{\sqrt{p_u\pi(B_i)}}{\sqrt{\hat{\pi}_N(Su)\hat{\pi}_N(B_i)}}, \end{aligned}$$

Under H_0 the empirical estimate $\hat{\pi}_N(B_i)$ is consistent, so we have

$$R_N(B_i,u) \overset{a.s.}{\sim} \frac{D_N(B_i,u)}{\sqrt{p_u\pi(B_i)}}.$$

According to Lemma B.1,

$$R_N(B_i,u) \overset{a.s.}{\sim} \frac{1}{\sqrt{N}}\sum_{j=1}^{N}\frac{V_{j-1}(B_i)\varepsilon_j(u)}{\sqrt{\pi(B_i)}\sqrt{p_u}}.$$

Lemma C.1 follows immediately. □

Proof of Theorem 3.2. Let us introduce the two following column vectors

$$\begin{aligned} \xi_j &\overset{\text{def}}{=} \Big(\frac{V_{j-1}(B_i)}{\sqrt{\pi(B_i)}}\Big)_{1\le i\le k} \otimes \Big(\frac{\varepsilon_j(u)}{\sqrt{p_u}}\Big)_{u\in\mathcal{A}} \\ M_N &\overset{\text{def}}{=} \sum_{j=1}^{N}\xi_j. \end{aligned}$$

$(M_N)_N$ is a $\mathcal{F}$-martingale with the associated increasing process

$$\langle M\rangle_N \stackrel{\text{def}}{=} \sum_{j=1}^{N} \mathbb{E}\Big[\xi_j\xi_j^t \mid \mathcal{F}_{j-1}\Big].$$

Under hypothesis H_0, it is easy to see that, for all $u, v \in \mathcal{A}$,

$$\mathbb{E}\left[\frac{\varepsilon_j(u)}{\sqrt{p_u}}\frac{\varepsilon_j(v)}{\sqrt{p_v}} \;\middle|\; \mathcal{F}_{j-1}\right] = 1\!\!1_{\{v=u\}} - \sqrt{p_u p_v},$$

and for all $B, B' \in \{B_1, \dots B_k\}$,

$$\lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} \frac{V_j(B)}{\sqrt{\pi(B)}}\frac{V_j(B')}{\sqrt{\pi(B')}} = 1\!\!1_{\{B=B'\}} - \sqrt{\pi(B)\pi(B')}.$$

Letting the two column vectors

$$\begin{aligned}
\sqrt{p} &\stackrel{\text{def}}{=} \Big(\sqrt{p_u}\Big)_{u\in\mathcal{A}} \\
\sqrt{\pi} &\stackrel{\text{def}}{=} \Big(\sqrt{\pi(B_1)}, \sqrt{\pi(B_2)}, \dots, \sqrt{\pi(B_k)}\Big)^t,
\end{aligned}$$

the following limit holds.

$$\lim_{N\to\infty} \frac{1}{N}\langle M\rangle_N \stackrel{\text{def}}{=} \Gamma = \Delta \otimes S,$$

where

$$\begin{aligned}
S &\stackrel{\text{def}}{=} I_d - \sqrt{p}\sqrt{p}^t, \\
\Delta &\stackrel{\text{def}}{=} I_k - \sqrt{\pi}\sqrt{\pi}^t,
\end{aligned}$$

are the orthogonal projection matrices respectively on $(\sqrt{p})^\perp$ and $(\sqrt{\pi})^\perp$, remarking that $\sum_{i=1}^k \pi(B_i) = 1$.

As in Theorem 3.1, the central limit theorem for martingales yields

$$\frac{1}{\sqrt{N}} M_N \stackrel{\mathcal{L}}{\leadsto} \mathcal{N}_{4k}(0, \Gamma), \tag{C.1}$$

and, by the continuity of the Euclidean norm,

$$\sum_{i=1}^{k}\sum_{u\in\mathcal{A}} R_N(B_i, u)^2 \stackrel{a.s.}{\sim} \frac{1}{N}\|M_N\|^2 \stackrel{\mathcal{L}}{\leadsto} \|Z\|^2, \quad \text{where} \quad Z \stackrel{\mathcal{L}}{\sim} \mathcal{N}_{dk}(0, \Gamma).$$

Since Γ is symmetric and orthogonal (as a tensor product of two orthogonal projections), $Z \stackrel{\mathcal{L}}{=} \Gamma Y$, where $Y \stackrel{\mathcal{L}}{\sim} \mathcal{N}_{dk}(0, I_{dk})$. Therefore, Cochran's theorem applies (see e.g. [6]) and $\|Z\|^2 \stackrel{\mathcal{L}}{=} \chi^2(\text{rank}(\Gamma))$. When $k > 1$, the rank of Γ is $(d-1)(k-1)$ and (3.7) follows.

In the particular case $k = 1$, $\Gamma = (1 - \pi(B))S$ and Cochran's theorem yields (3.5). This concludes the proof of Theorem 3.2. □

D Proof of Theorem 3.5

Lemma D.1. *Under the hypothesis H_m, the asymptotic behavior of the difference*

$$D_N(B, w, u) \stackrel{\text{def}}{=} \sqrt{N}\Big(\hat{\pi}_N(Sw)\hat{\pi}_N(Bwu) - \hat{\pi}_N(Swu)\hat{\pi}_N(Bw)\Big)$$

is the following.

$$D_N(B, w, u) \stackrel{a.s.}{\sim} \frac{1}{\sqrt{N}} \sum_{j=1}^{N-m-1} \varepsilon_{j+m+1}(w, u) V_j(B, w), \tag{D.1}$$

where

$$\begin{aligned} \varepsilon_j(w, u) &\stackrel{\text{def}}{=} \mathbb{1}_{Sw}(X_{j-1})(\mathbb{1}_{\{u_j = u\}} - Q(w, u)), \\ V_j(B, w) &\stackrel{\text{def}}{=} p_w \mathbb{1}_B(X_j) - \pi(Bw). \end{aligned}$$

Proof. The proof is almost the same as the proof of Lemma B.1. First all the terms in $D_N(B, w, u)$ are centered. For all word w of length $\geq m$

$$\sqrt{N}\bigg(\hat{\pi}_N(Sw) - p_w\bigg) \stackrel{\mathcal{L}}{\rightsquigarrow} \mathcal{N}(0, \sigma^2).$$

$$\hat{\pi}_N(Bw) \xrightarrow[\text{N}\to\infty]{\text{a.s}} \pi(Bw).$$

Exchanging now the variables in the sum (the limit is not modified since all terms are bounded and divided by $\sqrt{N}$), we get immediately (D.1). □

Lemma D.2.

$$\sum_{\substack{wu\in\mathcal{A}^m\times\mathcal{A}\\1\le i\le k}} R_N(B_i,w,u)^2$$

$$\overset{a.s.}{\sim} \quad \sum_{\substack{wu\in\mathcal{A}^m\times\mathcal{A}\\1\le i\le k}} \frac{1}{N}\left(\frac{1}{\sqrt{p_w}}\sum_{j=1}^{N-m-1}\frac{V_j(B_i,w)}{\sqrt{\pi(B_iw)}}\frac{\varepsilon_{j+m+1}(w,u)}{\sqrt{p_{wu}}}\right)^2.$$

Proof. The proof follows directly from Lemma D.1, by using the consistency of the empirical estimate and Slutsky's lemma. □

Proof of Theorem 3.5. The argument mimics that of Theorem 3.1, by means of the column vectors

$$\xi_{j+m+1} \stackrel{\text{def}}{=} \left(\frac{1}{\sqrt{p_w}}\Big(\frac{V_j(B_i,w)}{\sqrt{\pi(B_iw)}}\Big)_{1\le i\le k}\otimes\Big(\frac{\varepsilon_{j+m+1}(w,u)}{\sqrt{p_{wu}}}\Big)_{u\in\mathcal{A}}\right)_{w\in\mathcal{A}^d},$$

$$M_N \stackrel{\text{def}}{=} \sum_{j=1}^{N-d-1}\xi_j.$$

Then, under hypothesis H_m, it is easy to see that, for all $u,v\in\mathcal{A}$, $w\in\mathcal{A}^m$,

$$\mathbb{E}\left[\frac{\varepsilon_{j+1}(w,u)}{\sqrt{p_{wu}}}\frac{\varepsilon_{j+1}(w,v)}{\sqrt{p_{wv}}}\middle|\mathcal{F}_j\right] = \frac{\mathbb{1}_{Sw}(X_j)}{p_w}\Big(\mathbb{1}_{\{v=u\}}-\sqrt{Q(w,u)Q(w,v)}\Big),$$

and for all $B,B'\in\{B_1,\dots B_k\}$,

$$\frac{1}{N}\sum_{j=1}^{N-m-1}\frac{\mathbb{1}_{Sw}(X_{j+m})}{p_w^2}\frac{V_j(B,w)V_j(B',w)}{\sqrt{\pi(Bw)\pi(B'w)}} \underset{N\to\infty}{\longrightarrow} \mathbb{1}_{\{B=B'\}}-\frac{\sqrt{\pi(Bw)\pi(B'w)}}{p_w}.$$

Introduce two column vectors

$$\sqrt{Q_w} \stackrel{\text{def}}{=} \Big(\sqrt{Q(w,u)}\Big)_{u\in\mathcal{A}}$$

$$\sqrt{\pi_w} \stackrel{\text{def}}{=} \left(\sqrt{\frac{\pi(B_1w)}{p_w}},\sqrt{\frac{\pi(B_2w)}{p_w}},\dots,\sqrt{\frac{\pi(B_kw)}{p_w}}\right)^t.$$

Then

$$\lim_{N\to\infty}\frac{1}{N}\langle M\rangle_N \stackrel{\text{def}}{=} \Gamma,$$

Γ being block-diagonal, with blocks of the form $\Delta_w \otimes \Sigma_w$, where

$$\begin{aligned}\Sigma_w &\stackrel{\text{def}}{=} I_d - \sqrt{Q_w}\sqrt{Q_w}^t,\\ \Delta_w &\stackrel{\text{def}}{=} I_k - \sqrt{\pi_w}\sqrt{\pi_w}^t,\end{aligned}$$

are the orthogonal projection matrices respectively on $(\sqrt{Q_w})^\perp$ and on $(\sqrt{\pi_w})^\perp$.

□

Acknowledgements We would like to thank B. Bercu for useful remarks on the statistical issues. The first author is also grateful to M. Guesdon for his computer programming assistance.

References

[1] J. Almeida, J. Carriço, A. Maretzek, P. Noble, and F. M., *Analysis of genomic sequences by Chaos Game Representation*, Bioinformatics, 17 (2001), pp. 429–437.

[2] V. Anh, K. Lau, and Z.-G. Yu, *Multifractal characterisation of complete genomes*, Physica A, 301 (2001), pp. 351–361.

[3] S. Basu, A. Pam, C. Dutta, and J. Das, *Chaos Game Representation of proteins*, J. Mol. Graph. Model., 15 (1997), pp. 279–289.

[4] P. Billingsley, *Probability and Measure*, Wiley Series in Probability & Mathematical Statistics: Probability and Mathematical Statistics, 1995.

[5] A. Campbell, J. Mràzek, and S. Karlin, *Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA*, Proc. Natl. Acad. Sci. USA, 96 (1999), pp. 9184–9189.

[6] D. Dacunha-Castelle and M. Duflo, *Probabilités et statistiques*, Masson, 1982.

[7] P. Deschavanne, A. Giron, J. Vilain, G. Fagot, and F. B., *Genomic signature: Characterization and classification of species assessed by Chaos Game Representation of sequences*, Mol. Bio. Evol., 16 (1999), pp. 1391–1399.

[8] J. Dieudonné, *Calcul Infinitésimal*, Hermann, 1997.

[9] K. J. Falconer, *Fractal Geometry: Mathematical Foundations and Applications*, J. Wiley and sons, 1990.

[10] N. Goldman, *Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences*, Nucleic Acids Res., 21 (1993), pp. 2487–2491.

[11] J. Gutiérrez, M. Rodríguez, and G. Abramson, *Multifractal analysis of DNA sequences using a novel Chaos Game Representation*, Physica A, 300 (2001), pp. 271–284.

[12] D. Hall and C. Heyde, *Martingale Limit Theory and its Applications*, Academic press, 1980.

[13] A. Hariri, B. Weber, and J. Olmsted, *On the validity of shannon information calculations for molecular biological sequences*, J. Theor. Biol., 147 (1990), pp. 235–254.

[14] H. Jeffrey, *Chaos Game Representation of gene structure*, Nucleic Acid. Res, 18 (1990), pp. 2163–2170.

[15] R. Jernigan and R. Baran, *Pervasive properties of the genomic signature*, BMC Genomics, 3 (2002).

[16] J. Josse, A. Kaiser, and A. Kornberg, *Enzymatic synthesis of deoxyribonucleic acid. VIII. frequencies of nearest neighbor base sequences in deoxyribonucleic acid*, J. Biol. Chem, 263 (1961), pp. 864–875.

[17] S. Karlin and C. Burge, *Dinucleotide relative abundance extremes: a genomic signature*, Trends Genet., 7 (1995), pp. 283–290.

[18] S. Karlin and L. Cardon, *Computational DNA analysis*, Annu. Rev. Microbiol., 48 (1994), pp. 619–654.

[19] S. Karlin, I. Landunga, and B. Blaisdell, *Heterogeneity of genomes : measures and values*, Proc Natl Acad Sci USA, 91 (1994), pp. 12837–12841.

[20] S. Karlin and J. Mràzek, *Compositional differences within and between eukaryotic genomes*, Proc. Natl. Acad. Sci. USA, 94 (1997), pp. 10227–10232.

[21] ——, *Strand compositional asymmetry in bacterial and large viral genomes*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 3720–3725.

[22] S. Karlin, J. Mràzek, and A. Campbell, *Compositional biases of bacterial genomes and evolutionary implications*, J. Bacteriol., 179 (1997), pp. 3899–3913.

[23] Loève, *Probability Theory*, Springer, 4th ed., 1978.

[24] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer, 1993.

[25] J. Oliver, P. Bernaola-Galván, J. Guerrero-Garcia, and R. Romaán Roldan, *Entropic profiles of DNA sequences through chaos-game derived images*, Journal of Theor. Biology, 160 (1993), pp. 457–470.

[26] G. Perrière and M. Gouy, *www-query: An on-line retrieval system for biological sequence banks*, Biochimie., 78 (1996), pp. 364–369.

[27] K. Pleiss ner, L. Wernisch, H. Osvald, and E. Fleck, *Representation of amino acid sequences as two-dimensional point patterns*, Electrophoresis., 18 (1997), pp. 2709–2713.

[28] G. Reinert, S. Schbath, and M. Waterman, *Probabilistic and statistical properties of words: An overview*, Journal of Computational Biology, 7 (2000), pp. 1–46.

[29] A. Roy, C. Raychaudhury, and A. Nandy, *Novel techniques of graphical representation and analysis of DNA sequences – a review*, J. Biosci., 23 (1998), pp. 55–71.

[30] G. Russel and J. Subak-Sharpe, *Similarity of the general designs of protochordates and invertebrates*, Nature(London), 266 (1977), pp. 533–535.

[31] N. Saitou and M. Nei, *The neighbor-joining method : A new method for reconstructing phylogenetic trees*, Mol. Biol. Evol, 4 (1987), pp. 406–425.

[32] A. Van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399