Available online at www.sciencedirect.com
ScienceDirect
Procedia Computer Science 00 (2012) 1–12
Procedia Computer Science

Predicting the Evolution of Gene *ura*3 in the Yeast *Saccharomyces Cerevisiae*

Jacques M. Bahi[a,c], Christophe Guyeux[a,c], Antoine Perasso[b,c,d]

[a]*Institut femto-st*
[b]*Laboratoire Chrono-Environnement*
[c]*Université de Franche Comté, France*
{jacques.bahi,christophe.guyeux,antoine.perasso}@univ-fcomte.fr
[d]*Authors in alphabetic order*

Abstract

Since the late '60s, various genome evolutionary models have been proposed to predict the evolution of a DNA sequence as the generations pass. Most of these models are based on nucleotides evolution, so they use a mutation matrix of size 4×4. They encompass for instance the well-known models of Jukes and Cantor, Kimura, and Tamura. By essence, all of these models relate the evolution of DNA sequences to the computation of the successive powers of a mutation matrix. To make this computation possible, particular forms for the mutation matrix are assumed, which are not compatible with mutation rates that have been recently obtained experimentally on gene *ura*3 of the Yeast *Saccharomyces cerevisiae*. Using this experimental study, authors of this paper have deduced a simple mutation matrice, compute the future evolution of the rate purine/pyrimidine for *ura*3, investigate the particular behavior of cytosines and thymines compared to purines, and simulate the evolution of each nucleotide.

Keywords:
genome evolutionary models, stochastic processes, nucleotides mutations

1. Introduction

Codons are not uniformly distributed into the genome. Over time mutations have introduced some variations in their frequency of apparition. Mathematical models allow the prediction of such an evolution, in such a way that statistical values observed into current genomes can be recovered from hypotheses on past DNA sequences.

A first model for genomes evolution has been proposed in 1969 by Thomas Jukes and Charles Cantor [1]. This first model is very simple, as it supposes that each nucleotide A, C, G, T has the probability m to mutate to any other nucleotide, as described in the following mutation matrix,

$$\begin{pmatrix} * & m & m & m \\ m & * & m & m \\ m & m & * & m \\ m & m & m & * \end{pmatrix}.$$

In that matrix, the coefficient in row 3, column 2 represents the probability that the nucleotide G mutates in C during the next time interval, *i.e.*, $P(G \to C)$. As diagonal elements can be deduced by the fact that the sum of each row must be equal to 1, they are omitted here.

This first attempt has been followed up by Motoo Kimura [2], who has reasonably considered that transitions ($A \longleftrightarrow G$ and $T \longleftrightarrow C$) should not have the same mutation rate than transversions ($A \longleftrightarrow T$, $A \longleftrightarrow C$, $T \longleftrightarrow G$, and $C \longleftrightarrow G$), leading to the following mutation matrix:

$$\begin{pmatrix} * & b & a & b \\ b & * & b & a \\ a & b & * & b \\ b & a & b & * \end{pmatrix}.$$

This model has been refined by Kimura in 1981 (three constant parameters, to make a distinction between natural $A \longleftrightarrow T$, $C \longleftrightarrow G$ and unnatural transversions), leading to:

$$\begin{pmatrix} * & c & a & b \\ c & * & b & a \\ a & b & * & c \\ b & a & c & * \end{pmatrix}.$$

Joseph Felsenstein [3] has then supposed that the nucleotides frequency depends on the kind of nucleotide A,C,T,G. Such a supposition leads to a mutation matrix of the form:

$$\begin{pmatrix} * & \pi_C & \pi_G & \pi_T \\ \pi_A & * & \pi_G & \pi_T \\ \pi_A & \pi_C & * & \pi_T \\ \pi_A & \pi_C & \pi_G & * \end{pmatrix}$$

with $3\pi_A, 3\pi_C, 3\pi_G$, and $3\pi_T$ denoting respectively the frequency of occurrence of each nucleotide. Masami Hasegawa, Hirohisa Kishino, and Taka-Aki Yano [4] have generalized the models of [2] and [3], introducing in 1985 the following mutation matrix:

$$\begin{pmatrix} * & \alpha\pi_C & \beta\pi_G & \alpha\pi_T \\ \alpha\pi_A & * & \alpha\pi_G & \beta\pi_T \\ \beta\pi_A & \alpha\pi_C & * & \alpha\pi_T \\ \alpha\pi_A & \beta\pi_C & \alpha\pi_G & * \end{pmatrix}.$$

These efforts have been continued by Tamura, who proposed in [5, 6] a simple method to estimate the number of nucleotide substitutions per site between two DNA sequences, by extending the model of Kimura (1980). The idea is to consider a two-parameter method, for the case where a GC bias exists. Let us denote by π_{GC} the frequency of this dinucleotide motif. Tamura supposes that $\pi_G = \pi_C = \dfrac{\pi_{GC}}{2}$ and $\pi_A = \pi_T = \dfrac{1 - \pi_{GC}}{2}$, which leads to the following rate matrix:

$$\begin{pmatrix} * & \kappa(1-\pi_{GC})/2 & (1-\pi_{GC})/2 & (1-\pi_{GC})/2 \\ \kappa\pi_{GC}/2 & * & \pi_{GC}/2 & \pi_{GC}/2 \\ (1-\pi_{GC})/2 & (1-\pi_{GC})/2 & * & \kappa(1-\pi_{GC})/2 \\ \pi_{GC}/2 & \pi_{GC}/2 & \kappa\pi_{GC}/2 & * \end{pmatrix}.$$

All these models lead to the so-called GTR model [7], in which the mutation matrix has the form (using obvious notations):

$$\begin{pmatrix} * & f_{AC}\pi_C & f_{AG}\pi_G & f_{AT}\pi_T \\ f_{AC}\pi_A & * & f_{CG}\pi_G & f_{CT}\pi_T \\ f_{AG}\pi_A & f_{CG}\pi_C & * & \pi_T \\ f_{AT}\pi_A & f_{CT}\pi_C & \pi_G & * \end{pmatrix}.$$

A second category of models focus on di or trinucleotides evolution. From 1990 to 1994, Didier Arquès and Christian Michel have proposed models based on the RY purine/pyrimidine alphabet [8, 9, 10, 11, 12, 13]. These models have been abandoned by their own authors in favor of models over the $\{A, C, G, T\}$ alphabet.

In 2004, Jacques M. Bahi and Christian Michel have published a novel research work in which the model of 1998 has been improved by replacing constants parameters by new time dependent parameters [14]. By this way, it has been possible to simulate a genes evolution that is non-linear. However, the following years, these researchers have been returned to models embedding constant parameters, probably due to the fact that the model of 2004 leads to poor results: only one of the twelve studied cases allows to recover values that are close to reality. For instance, in 2006, Gabriel Frey and Christian Michel have proposed a model that uses 6 constant parameters [15], whereas in 2007, Christian Michel has constructed a model with 9 constants parameters that generalizes those of 1998 and 2006 [16]. Finally, Jacques M. Bahi and Christian Michel have recently introduced in [17], a last model with 3 constant parameters, but *whose evolution matrix evolves over time*.

Due to mathematical complexity, matrices investigated to model evolution of DNA sequences are thus limited either by the hypothesis of symmetry or by the desire to reduce the number of parameters under consideration. These hypotheses allow their authors to solve theoretically the DNA evolution problem by computing directly the successive powers of their mutation matrix. However, one can wonder whether such restrictions on the mutation rates are realistic. Focusing on this question, authors of the present paper have used a recent research work in which the per-base-pair mutation rates of the Yeast *Saccharomyces cerevisiae* have been experimentally measured [18]. Their results are summarized in Table 1.

Mutation	*ura*3
$T \to C$	4
$T \to A$	14
$T \to G$	5
$C \to T$	16
$C \to A$	40
$C \to G$	11
$A \to T$	8
$A \to C$	6
$A \to G$	0
$G \to T$	28
$G \to C$	9
$G \to A$	26
Transitions	46
Transversions	121

Table 1: Summary of sequenced *ura*3 mutations [18]

The mutation matrix of gene *ura*3 can be deduced from this table. It is equal to:

$$\begin{pmatrix} 1-m & \frac{6m}{14} & 0 & \frac{8m}{14} \\ \frac{40m}{67} & 1-m & \frac{11m}{67} & \frac{16m}{67} \\ \frac{26m}{63} & \frac{9m}{63} & 1-m & \frac{28m}{63} \\ \frac{14m}{23} & \frac{4m}{23} & \frac{5m}{23} & 1-m \end{pmatrix},$$

where m is the mutation rate per generation in *ura*3 gene, which is equal to 3.80×10^{-10}/bp/generation, or to 3.0552×10^{-7}/generation for the whole gene [18]. Obviously, none of the existing genomes evolution models can fit such a mutation matrix, leading to the fact that hypotheses must be relaxed, even if this relaxation leads to less ambitious models: current models do not match with what really occurs in concrete genomes, at least in the case of this yeast.

Having these considerations in mind, authors of the present article propose to use the data obtained by Lang and Murray, in order to predict the evolution of the rates or purines and pyrimidines in the two genes studied in [18]. A mathematical proof giving the intended limit for these rates when the generations pass, is reinforced by numerical simulations. The obtained simulations are thus compared with the historical model of Jukes and Cantor, which is still used by current prediction software. A model of size 3×3 with six independent parameters is then proposed and studied in a case that matches with data recorded in [18]. Mathematical investigations and numerical simulations focusing on *ura*3 gene are both given in the case of the yeast *Saccharomyces cerevisiae*.

The remainder of this research work is organized as follows. In Sections 2 and 3, we focus on the evolution of the gene *ura*3. Section 2 is dedicated to the formulation of a non symmetric discrete model of size 2×2. This model translates a genome evolution taking into account purines and pyrimidines mutations. A simulation is then performed to compare this non symmetric model to the classical symmetric Cantor model. Section 3 deals with a 6-parameters non symmetric model of size 3×3, focusing on the one hand on the evolution of purines and on the other hand of cytosines and thymines. This mathematical model is illustrated throughout simulations of the evolution of the purines, cytosines and thymines of gene *ura*3. We finally conclude this work in Section 4.

2. Non-symmetric Model of size 2 × 2

In this section, a first general genome evolution model focusing on purines versus pyrimidines is proposed, to illustrate the method and as a pattern for further investigations. This model is applied to the case of the yeast *Saccharomyces cerevisiae*.

2.1. Theoretical Study

Let R and Y denote respectively the occurrence frequency of purines and pyrimidines in a sequence of nucleotides, and $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the associated mutation matrix, with $a = P(R \to R)$, $b = P(R \to Y)$, $c = P(Y \to R)$, and $d = P(Y \to Y)$ satisfying

$$\begin{cases} a + b = 1, \\ c + d = 1, \end{cases} \tag{1}$$

and thus $M = \begin{pmatrix} a & 1-a \\ c & 1-c \end{pmatrix}$.

The initial probability is denoted by $P_0 = (R_0 \;\; Y_0)$, where R_0 and Y_0 denote respectively the initial frequency of purines and pyrimidines. So the occurrence probability at generation n is $P_n = P_0 M^n$, where $P_n = (R(n) \;\; Y(n))$ is a probability vector such that $R(n)$ (resp. $Y(n)$) is the rate of purines (resp. pyrimidines) after n generations.

Determination of M^n

A division algorithm leads to the existence of a polynomial of degree $n-2$, denoted by $Q_M \in \mathbb{R}_{n-2}[X]$, and to $a_n, b_n \in \mathbb{R}$ such that

$$X^n = Q_M(X)\chi_M(X) + a_n X + b_n, \tag{2}$$

when χ_M is the characteristic polynomial of M. Using both the Cayley-Hamilton theorem and the equality given above, we thus have

$$M^n = a_n M + b_n I_2.$$

In order to determine a_n and b_n, we must find the roots of χ_M. As $\chi_M(X) = X^2 - Tr(M)X + \det(M)$ and due to (1), we can conclude that 1 is a root of χ_M, which thus has two real roots: 1 and x_2. As the roots sum is equal to -tr(A), we conclude that $x_2 = a - c$.

If $x_2 = a - c = 1$, then $a = 1$ and $c = 0$ (as these parameters are in $[0, 1]$), so the mutation matrix is the identity and the frequencies of purines and pyrimidines into the DNA sequence does not evolve. If not, evaluating (2) in both $X = 1$ and $X = x_2$, we thus obtain

$$\begin{cases} 1 = a_n + b_n, \\ (a-c)^n = a_n(a-c) + b_n. \end{cases}$$

Considering that $a - c \neq 1$, we obtain

$$a_n = \frac{(a-c)^n - 1}{a-c-1}, \qquad b_n = \frac{a - c - (a-c)^n}{a-c-1}.$$

Using these last expressions into the equality linking M, a_n, and b_n, we thus deduce the value of $P_n = P_0 M^n$, where

$$M^n = \frac{1}{a-c-1} \begin{pmatrix} (a-1)(a-c)^n - c & (1-a)((a-c)^n - 1)) \\ c((a-c)^n - 1) & -c(a-c)^n + a - 1 \end{pmatrix}. \tag{3}$$

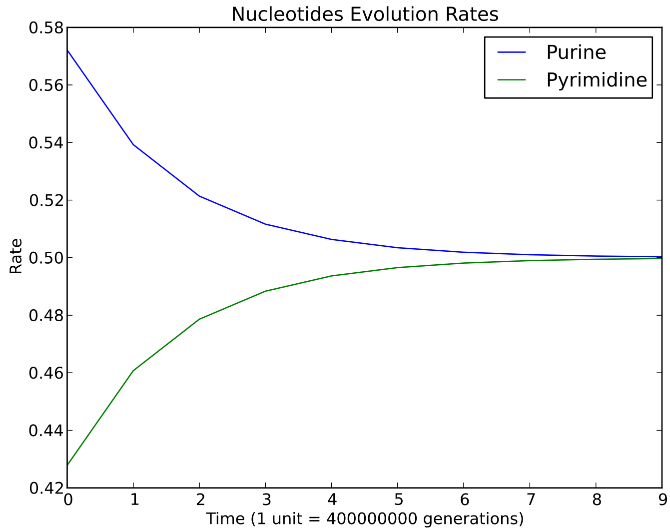

Figure 1: Prediction of purine/pyrimidine evolution of *ura3* gene in symmetric Cantor model.

If $a = 0$ and $c = 1$, then $M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, so M^{2n} is the identity I_2 whereas M^{2n+1} is M. Contrarily, if $(a, c) \notin \{(0, 1); (1, 0)\}$, then the limit of M^n can be easily found using (3), leading to the following result.

Theorem 2.1. *Consider a DNA sequence under evolution, whose mutation matrix is* $M = \begin{pmatrix} a & 1-a \\ c & 1-c \end{pmatrix}$ *with* $a = P(R \to R)$ *and* $c = P(Y \to R)$*.*

- *If* $a = 1, c = 0$*, then the frequencies of purines and pyrimidines do not change as the generation pass.*
- *If* $a = 0, c = 1$*, then these frequencies oscillate at each generation between* $(R_0 \quad Y_0)$ *(even generations) and* $(Y_0 \quad R_0)$ *(odd generations).*
- *Else the value* $P_n = (R(n) \quad Y(n))$ *of purines and pyrimidines frequencies at generation* n *is convergent to the following limit:*

$$\lim_{n\to\infty} P_n = \frac{1}{c+1-a} \begin{pmatrix} c & 1-a \end{pmatrix}.$$

2.2. Numerical Application

For numerical application, we will consider mutations rates in the *ura3* gene of the Yeast *Saccharomyces cerevisiae*, as obtained by Gregory I. Lang and Andrew W. Murray [18]. As stated before, they have measured phenotypic mutation rates, indicating that the per-base pair mutation rate at *ura3* is equal to $m = 3.0552 \times 10^{-7}$/generation. For the majority of Yeasts they studied, *ura3* is constituted by 804 bp: 133 cytosines, 211 thymines, 246 adenines, and 214 guanines. So $R_0 = \frac{246+214}{804} \approx 0.572$, and $Y_0 = \frac{133+211}{804} \approx 0.428$. Using these values in the historical model of Jukes and Cantor [1], we obtain the evolution depicted in Figure 1.

Theorem 2.1 allow us to compute the limit of the rates of purines and pyrimidines:

Computation of probability a. $a = P(R \to R) = (1-m) + P(A \to G) + P(G \to A)$. The use of Table 1 implies that $a = (1-m) + m\left(\frac{0+26}{x}\right)$, where x is such that $1 - a = P(R \to Y) = m\left(\frac{6+8+28+9}{x}\right)$, *i.e.*, $x = 77$, and so $a = 1 - \frac{51m}{77}$.

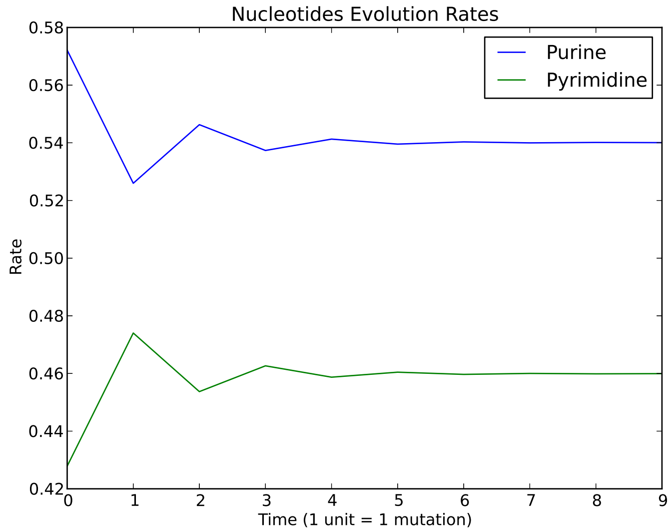

Figure 2: Prediction of purine/pyrimidine evolution of *ura3* gene in non-symmetric Model of size 2×2.

Computation of probability c. Similarly, $c = P(Y \to R) = P(C \to A) + P(C \to G) + P(T \to A) + P(T \to G) = \frac{70m}{y}$, whereas $1 - c = 1 - m + \frac{20m}{y}$. So $c = \frac{7m}{9}$.

The purine/pyrimidine mutation matrix that corresponds to the data of [18] is thus equal to:

$$M = \begin{pmatrix} 1 - \frac{51m}{77} & \frac{51m}{77} \\ \frac{7m}{9} & 1 - \frac{7m}{9} \end{pmatrix}.$$

Using the value of m for the *ura*3 gene leads to $1 - a = 2.02357 \times 10^{-7}$ and $c = 2.37627 \times 10^{-7}$, which can be used in Theorem 2.1 to conclude that the rate of pyrimidines is convergent to 45.992% whereas the rate of purines converge to 54.008%. Numerical simulations using data published in [18] are given in Figure 2, leading to a similar conclusion.

3. A First Non-Symmetric Genomes Evolution Model of size 3 × 3 having 6 Parameters

In order to investigate the evolution of the frequencies of cytosines and thymines in the gene *ura*3, a model of size 3×3 compatible with real mutation rates of the yeast *Saccharomyces cerevisiae* is now presented.

3.1. Formalization

Let us consider a line of yeasts where a given gene is sequenced at each generation, in order to clarify explanations. The $n-$th generation is obtained at time n, and the rates of purines, cytosines, and tymines at time n are respectively denoted by $P_R(n)$, $P_C(n)$, and $P_T(n)$.

Let a be the probability that a purine is changed into a cytosine between two generations, that is: $a = P(R \to C)$. Similarly, denote by b, c, d, e, f the respective probabilities: $P(R \to T)$, $P(C \to R)$, $P(C \to T)$, $P(T \to R)$, and $P(T \to C)$. Contrary to existing approaches, $P(R \to C)$ is not supposed to be equal to $P(C \to R)$, and the same statement holds for the other probabilities. For the sake of simplicity, we will consider in this first research work that a, b, c, d, e, f are not time dependent.

Let

$$M = \begin{pmatrix} 1-a-b & a & b \\ c & 1-c-d & d \\ e & f & 1-e-f \end{pmatrix}$$

be the mutation matrix associated to the probabilities mentioned above, and P_n the vector of occurrence, at time n, of each of the three kind of nucleotides. In other words, $P_n = (P_R(n) \;\; P_C(n) \;\; P_T(n))$. Under that hypothesis, P_n is a probability vector: $\forall n \in \mathbb{N}$,

- $P_R(n), P_C(n), P_T(n) \in [0,1]$,
- $P_R(n) + P_C(n) + P_T(n) = 1$,

Let $P_0 = (P_R(0) \;\; P_C(0) \;\; P_T(0)) \in [0,1]^3$ be the initial probability vector. We have obviously:

$$P_R(n+1) = P_R(n)P(R \to R) + P_C(n)P(C \to R) + P_T(n)P(T \to R).$$

Similarly, $P_C(n+1) = P_R(n)P(R \to C) + P_C(n)P(C \to C) + P_T(n)P(T \to C)$ and $P_T(n+1) = P_R(n)P(R \to T) + P_C(n)P(C \to T) + P_T(n)P(T \to T)$. This equality yields the following one,

$$P_n = P_{n-1}M = P_0 M^n. \tag{4}$$

In all that follows we wonder if, given the parameters a, b, c, d, e, f as in [18], one can determine the frequency of occurrence of any of the three kind of nucleotides when n is sufficiently large, in other words if the limit of P_n is accessible by computations.

3.2. Resolution

The characteristic polynomial of M is equal to

$$\begin{aligned} \chi_M(x) &= x^3 + (s-3)x^2 + (p-2s+3)x - 1 + s - p \\ &= (x-1)\left(x^2 + (s-2)x + (1-s+p)\right), \end{aligned}$$

where

$$\begin{gathered} s = a+b+c+d+e+f, \\ p = ad + ae + af + bc + bd + bf + ce + cf + de, \\ det(M) = 1 - s + p. \end{gathered}$$

The discriminant of the polynomial of degree 2 in the factorization of χ_M is equal to $\Delta = (s-2)^2 - 4(1-s-p) = s^2 - 4p$. Let x_1 and x_2 the two roots (potentially complex or equal) of χ_M, given by

$$x_1 = \frac{-s+2-\sqrt{s^2-4p}}{2} \text{ and } x_2 = \frac{-s+2+\sqrt{s^2-4p}}{2}. \tag{5}$$

Let $n \in \mathbb{N}, n \geqslant 2$. As χ_M is a polynomial of degree 3, a division algorithm of X^n by $\chi_M(X)$ leads to the existence and uniqueness of two polynomials Q_n and R_n, such that

$$X^n = Q_n(X)\chi_2(X) + R_n(X), \tag{6}$$

where the degree of R_n is lower than or equal to the degree of χ_M, *i.e.*, $R_n(X) = a_n X^2 + b_n X + c_n$ with $a_n, b_n, c_n \in \mathbb{R}$ for every $n \in \mathbb{N}$. By evaluating (6) in the three roots of χ_M, we find the system

$$\left\{ \begin{array}{ll} 1 & = a_n + b_n + c_n \\ x_1^n & = a_n x_1^2 + b_n x_1 + c_n \\ x_2^n & = a_n x_2^2 + b_n x_2 + c_n \end{array} \right.$$

This system is equivalent to

$$\begin{cases} c_n \;+\; b_n \;+\; a_n = 1 \\ b_n(x_1-1) + a_n(x_1^2-1) = x_1^n - 1 \\ b_n(x_2-1) + a_n(x_2^2-1) = x_2^n - 1 \end{cases}$$

For the *ura*3 gene, it is easy to check that $x_1 \neq 1$, $x_2 \neq 1$, and $x_1 \neq x_2$ (see numerical applications of Section 3.4). Then standard algebraic computations give

$$\begin{cases} a_n = \dfrac{1}{x_2 - x_1}\left[\dfrac{x_2^n - 1}{x_2 - 1} - \dfrac{x_1^n - 1}{x_1 - 1}\right], \\[2ex] b_n = \dfrac{x_1+1}{x_1-x_2}\dfrac{x_2^n-1}{x_2-1} + \dfrac{x_2+1}{x_2-x_1}\dfrac{x_1^n-1}{x_1-1}, \\[2ex] c_n = 1 - a_n - b_n. \end{cases}$$

Using for $i = 1, 2$ and $n \in \mathbb{N}$ the following notation,

$$X_i(n) = \frac{x_i^n - 1}{x_i - 1}, \tag{7}$$

and since $x_2 - x_1 = \sqrt{\Delta}$, the system above can be rewritten as

$$\begin{cases} a_n = \dfrac{X_2(n) - X_1(n)}{\sqrt{\Delta}}, \\[2ex] b_n = \dfrac{(x_2+1)X_1(n) - (x_1+1)X_2(n)}{\sqrt{\Delta}}, \\[2ex] c_n = 1 + \dfrac{x_1 X_2(n) - x_2 X_1(n)}{\sqrt{\Delta}}. \end{cases} \tag{8}$$

By evaluating (6) in M and due to the theorem of Cayley-Hamilton, we finally have for every integer $n \geqslant 1$,

$$M^n = a_n M^2 + b_n M + c_n I_3, \tag{9}$$

where I_3 is the identity matrix of size 3, a_n, b_n, and c_n are given by (8), and M^2 is given by

$$M^2 = \left(\begin{array}{c|c|c} a^2 + 2ab + ac - 2a & -a^2 - ab - ac & -ab + ad - b^2 \\ +b^2 + be - 2b + 1 & -ad + 2a + bf & -be - bf + 2b \\ \hline -ac - bc - c^2 & ac + c^2 + 2cd - 2c & bc - cd - d^2 \\ -cd + 2c + de & +d^2 + df - 2d + 1 & -de - df + 2d \\ \hline -ae - be + cf & ae - cf - df & be + df + e^2 + 2ef \\ -e^2 - ef + 2e & -ef - f^2 + 2f & -2e + f^2 - 2f + 1 \end{array}\right).$$

3.3. Convergence study

In the case of *ura*3, $|x_1| < 1$ and $|x_2| < 1$ (see the next section). Then $X_i(n) \longrightarrow \dfrac{1}{1 - x_i}$ for $i = 1, 2$ and so

$$a_n \longrightarrow \frac{1}{\sqrt{\Delta}}\left(\frac{1}{1-x_2} - \frac{1}{1-x_1}\right).$$

Denote by a_∞ this limit. We have

$$a_\infty = \frac{x_2 - x_1}{\sqrt{\Delta}(1-x_2)(1-x_1)} = \frac{1}{(1-x_2)(1-x_1)} = \frac{1}{\dfrac{s+\sqrt{\Delta}}{2}\dfrac{s-\sqrt{\Delta}}{2}},$$

and finally

$$a_\infty = \frac{4}{s^2 - \Delta} = \frac{1}{p}.$$

Similarly, $b_n = X_1(n) - a_n(x_1 + 1)$ satisfies

$$b_n \longrightarrow \frac{1}{1 - x_1} - \frac{x_1 + 1}{p}.$$

The following computations

$$\frac{1}{1 - x_1} = \frac{2}{s + \sqrt{\Delta}} = \frac{2(s - \sqrt{\Delta})}{s^2 - \Delta} = \frac{s - \sqrt{\Delta}}{2p},$$

$$\frac{x_1 + 1}{p} = \frac{-s + 4 - \sqrt{\Delta}}{2p},$$

finally yield

$$b_\infty = \frac{s - 2}{p}.$$

So

$$c_n \longrightarrow 1 - a_\infty - b_\infty = \frac{p - s + 1}{p},$$

and to sum up, the distribution limit is given by

$$\begin{cases} a_\infty = \dfrac{1}{p} \\[2ex] b_\infty = \dfrac{s-2}{p} \\[2ex] c_\infty = \dfrac{p - s + 1}{p} \end{cases} \tag{10}$$

Using the latter values in (9), we can determine the limit of M^n, which is $a_\infty M^2 + b_\infty M + c_\infty I_3$. All computations done, we find the following limit for M^n,

$$\frac{1}{p - bf + df} \begin{pmatrix} ce + cf + de - bf + df & ae + af + bf & ad + bc + bd \\ ce + cf + de & ae + af + df & ad + bc + bd \\ ce + cf + de & ae + af + bf & ad + bc + bd - bf + df \end{pmatrix}.$$

Using (4), we can thus finally determine the limit of $P_n = P_0 M^n = (P_R(0) \;\; P_C(0) \;\; P_T(0))M^n$, which leads to the following result.

Theorem 3.1. *The frequencies $P_R(n), P_C(n)$, and $P_T(n)$ of occurrence at time n of purines, cytosines, and thymines in the considered gene ura3 of the yeast Saccharomyces cerevisiae converge to the following values:*

- $P_R(n) \longrightarrow \dfrac{ce + cf + de + (df - bf)P_R(0)}{p - bf + df}$
- $P_C(n) \longrightarrow \dfrac{ae + af + df + (df - bf)P_C(0)}{p - bf + df}$
- $P_T(n) \longrightarrow \dfrac{ad + bc + bd + (df - bf)P_T(0)}{p - bf + df}$

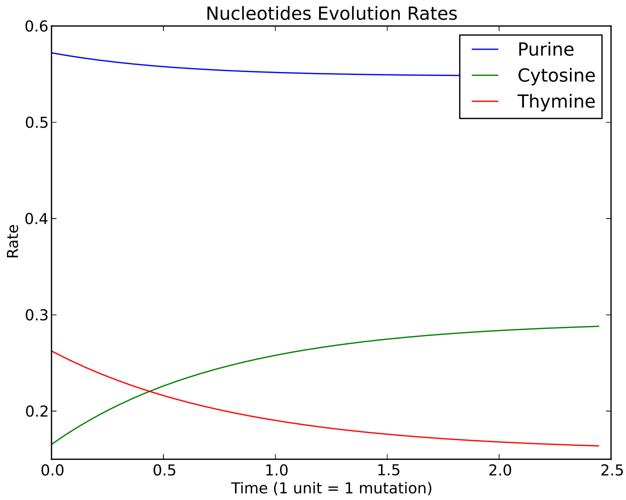

Figure 3: Prediction of evolution concerning the purine, thymine, and cytosine rates in *ura3*. Non-symmetric Model of size 3×3.

3.4. Numerical Application and Simulations

We consider another time the numerical values for mutations published in [18]. Gene *ura3* of the Yeast *Saccharomyces cerevisiae* has a mutation rate of 3.80×10^{-10}/bp/generation [18]. As this gene is constituted by 804 nucleotides, we can deduce that its global mutation rate per generation is equal to $m = 3.80 \times 10^{-10} \times 804 = 3.0552 \times 10^{-7}$. Let us compute the values of a, b, c, d, e, and f. The first line of the mutation matrix is constituted by $1 - a - b = P(R \to R)$, $a = P(R \to T)$, and $b = P(R \to C)$. $P(R \to R)$ takes into account the fact that a purine can either be preserved (no mutation, probability $1 - m$), or mutate into another purine ($A \to G$, $G \to A$). As the generations pass, authors of [18] have counted 0 mutations of kind $A \to G$, and 26 mutations of kind $G \to A$. Similarly, there were 28 mutations $G \to T$ and 8: $A \to T$, so 36: $R \to T$. Finally, 6: $A \to C$ and 9: $G \to C$ lead to 15: $R \to C$ mutations. The total of mutations to consider when evaluating the first line is so equal to 77. All these considerations lead to the fact that $1 - a - b = (1 - m) + m\frac{26}{77}$, $a = \frac{36m}{77}$, and $b = \frac{15m}{77}$. A similar reasoning leads to $c = \frac{19m}{23}$, $d = \frac{4m}{23}$, $e = \frac{51m}{67}$, and $f = \frac{16m}{67}$.

In that situation, $s = a + b + c + d + e + f = \frac{205m}{77} \approx 8.134 \times 10^{-7}$, and $p = \frac{207488m^2}{118657} \approx 1.632 \times 10^{-13}$. So $\Delta = s^2 - 4p = \frac{854221m^2}{9136589} > 0$, $x_1 = 1 - \frac{m}{2}\left(\frac{205}{77} + \sqrt{\frac{854221}{9136589}}\right)$, and $x_2 = 1 - \frac{m}{2}\left(\frac{205}{77} - \sqrt{\frac{854221}{9136589}}\right)$. As $x_1 \approx 0.9999685 \in [0, 1]$ and $x_2 \approx 0.9999686 \in [0, 1]$, we have, due to Theorem 3.1:

- $P_R(n) \longrightarrow \frac{ce + cf + de + (df - bf)P_R(0)}{p - bf + df}$
- $P_C(n) \longrightarrow \frac{ae + af + df + (df - bf)P_C(0)}{p - bf + df}$
- $P_T(n) \longrightarrow \frac{ad + bc + bd + (df - bf)P_T(0)}{p - bf + df}$

Using the data of [18], we find that $P_R(0) = \frac{460}{804} \approx 0.572$, $P_C(0) = \frac{133}{804} \approx 0.165$, and $P_T(0) = \frac{211}{804} \approx 0.263$. So $P_R(n) \longrightarrow 0.549$, $P_C(n) \longrightarrow 0.292$, and $P_T(n) \longrightarrow 0.159$. Simulations corresponding to this example are given in Fig. 3.

4. Conclusion

In this document, the possible evolution of gene *ura*3 of the yeast *Saccharomyces cerevisiae* has been studied. As current models of nucleotides cannot fit the mutations obtained experimentally by Lang and Murray [18], authors of this paper have introduced two new simple models to predict the evolution of this gene. On the one hand, a formulation of a non symmetric discrete model of size 2×2 has been proposed, which studies a DNA evolution taking into account purines and pyrimidines mutation rates. A simulation has been performed, to compare the proposal to the well known Jukes and Cantor model. On the other hand, a 6-parameters non symmetric model of size 3×3 has been introduced and tested with numerical simulations, to make a distinction between cytosines and thymines in the former proposal. These two models still remain generic, and can be adapted to a large panel of applications, replacing either the couple (purines, pyrimidines) or the tuple (purines, cytosines, thymines) by any categories of interest.

The *ura*3 gene is not the unique example of a DNA sequence of interest such that none of the existing nucleotides evolution models cannot be applied due to a complex mutation matrix. For instance, a second gene called *can*1 has been studied too by the authors of [18]. Similarly to gene *ura*3, usual models cannot be used to predict the evolution of *can*1, whereas a study following a same canvas than what has been proposed in this research work can be realized. In future work, the authors' intention is to make a complete mathematical study of the 6-parameters non symmetric model of size 3×3 proposed in this document, and to apply it to various case studies. Biological consequences of the results produces by this model will be systematically investigated. Then, the most general non symmetric model of size 4 will be regarded in some particular cases taken from biological case studies, and the possibility of mutations non uniformly distributed will then be investigated.

[1] T. H. Jukes, C. R. Cantor, Evolution of Protein Molecules, Academy Press, 1969.

[2] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, Journal of Molecular Evolution 16 (1980) 111–120, 10.1007/BF01731581.
URL `http://dx.doi.org/10.1007/BF01731581`

[3] J. Felsenstein, A view of population genetics., Science 208 (4449) (1980) 1253. doi:10.1126/science.208.4449.1253.
URL `http://dx.doi.org/10.1126/science.208.4449.1253`

[4] M. Hasegawa, H. Kishino, T. Yano, Dating of the human-ape splitting by a molecular clock of mitochondrial dna., J Mol Evol 22 (2) (1985) 160–174.

[5] K. Tamura, Estimation of the number of nucleotide substitutions when there are strong transition-transversion and g+c-content biases., Molecular Biology and Evolution 9 (4) (1992) 678–687. arXiv:http://mbe.oxfordjournals.org/content/9/4/678.full.pdf+html.
URL `http://mbe.oxfordjournals.org/content/9/4/678.abstract`

[6] K. Tamura, M. Nei, Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees., Molecular Biology and Evolution 10 (3) (1993) 512–526. arXiv:http://mbe.oxfordjournals.org/content/10/3/512.full.pdf+html.
URL `http://mbe.oxfordjournals.org/content/10/3/512.abstract`

[7] Z. Yang, Estimating the pattern of nucleotide substitution, Journal of Molecular Evolution 10 (1994) 105–111.

[8] D. G. Arquès, C. J. Michel, Periodicities in coding and noncoding regions of the genes., J Theor Biol 143 (3) (1990) 307–318.

[9] D. G. Arquès, C. J. Michel, A model of dna sequence evolution., Bull Math Biol 52 (6) (1990) 741–772.

[10] D. G. Arquès, C. J. Michel, A simulation of the genetic periodicities modulo 2 and 3 with processes of nucleotide insertions and deletions., J Theor Biol 156 (1) (1992) 113–127.

[11] D. G. Arquès, C. J. Michel, K. Orieux, Identification and simulation of new non-random statistical properties common to different populations of eukaryotic non-coding genes., J Theor Biol 161 (3) (1993) 329–342. doi:10.1006/jtbi.1993.1059.
URL `http://dx.doi.org/10.1006/jtbi.1993.1059`

[12] D. G. Arquès, C. J. Michel, Identification and simulation of new non-random statistical properties common to different eukaryotic gene subpopulations., Biochimie 75 (5) (1993) 399–407.

[13] D. G. Arquès, C. J. Michel, Analytical expression of the purine/pyrimidine autocorrelation function after and before random mutations., Math Biosci 123 (1) (1994) 103–125.

[14] J. M. Bahi, C. J. Michel, A stochastic gene evolution model with time dependent mutations., Bull Math Biol 66 (4) (2004) 763–778. doi:10.1016/j.bulm.2003.10.004.
URL `http://dx.doi.org/10.1016/j.bulm.2003.10.004`

[15] G. Frey, C. J. Michel, An analytical model of gene evolution with six mutation parameters: an application to archaeal circular codes., Comput Biol Chem 30 (1) (2006) 1–11. doi:10.1016/j.compbiolchem.2005.09.001.
URL `http://dx.doi.org/10.1016/j.compbiolchem.2005.09.001`

[16] C. J. Michel, An analytical model of gene evolution with 9 mutation parameters: an application to the amino acids coded by the common circular code., Bull Math Biol 69 (2) (2007) 677–698. doi:10.1007/s11538-006-9147-z.
URL `http://dx.doi.org/10.1007/s11538-006-9147-z`

[17] J. M. Bahi, C. Guyeux, Chaotic iterations and topological chaosarXiv:arXiv:0810.3154.
URL `arXiv:0810.3154`

[18] G. I. Lang, A. W. Murray, Estimating the per-base-pair mutation rate in the yeast saccharomyces cerevisiae, Genetics 178 (1) (January 2008) 67–82. arXiv:http://www.genetics.org/content/178/1/67.full.pdf+html, doi:10.1534/genetics.107.071506.
URL `http://www.genetics.org/content/178/1/67.abstract`