A generalization of substitution evolution models of nucleotides to genetic motifs

Emmanuel Benard, Christian J. Michel

www.elsevier.com/locate/yjtbi

A generalization of substitution evolution models of nucleotides to genetic motifs

Emmanuel Benard, Christian J. Michel

PII: S0022-5193(11)00359-6
DOI: doi:10.1016/j.jtbi.2011.07.009
Reference: YJTBI6550

To appear in: *Journal of Theoretical Biology*

Received date: 25 March 2011
Revised date: 15 July 2011
Accepted date: 18 July 2011

Cite this article as: Emmanuel Benard and Christian J. Michel, A generalization of substitution evolution models of nucleotides to genetic motifs, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2011.07.009

A GENERALIZATION OF SUBSTITUTION EVOLUTION MODELS OF NUCLEOTIDES TO GENETIC MOTIFS

EMMANUEL BENARD AND CHRISTIAN J. MICHEL*

Equipe de Bioinformatique Théorique
FDBT, LSIIT (UMR 7005)
Université de Strasbourg
Pôle API, Boulevard Sébastien Brant
67400 Illkirch, France

Email: benard@dpt-info.u-strasbg.fr, michel@dpt-info.u-strasbg.fr

* Corresponding author

Keywords: substitution models, stochastic models, gene evolution, nucleotides, genetic motifs, direct evolution, inverse evolution, analytical solutions, evolution plots, research software.

Abstract

We generalize here the classical stochastic substitution models of nucleotides to genetic motifs of any size. This generalized model gives the analytical occurrence probabilities of genetic motifs as a function of a substitution matrix containing up to three formal parameters (substitution rates) per motif site and of an initial occurrence probability vector of genetic motifs. The evolution direction can be direct (past-present) or inverse (present-past). This extension has been made due to the identification of a Kronecker relation between the nucleotide substitution matrices and the motif substitution matrices. The evolution models for motifs of size 4 (tetranucleotides) and 5 (pentanucleotides) are now included in the SEGM (Stochastic Evolution of Genetic Motifs) web server.

1. Introduction

We present here a generalization of the classical stochastic substitution models of nucleotides to genetic motifs of any size. The first gene evolution model was proposed by Jukes and Cantor [5] with 1-parameter substitution (probability α for all nucleotide substitution types). It was generalized to a 2-parameter substitution model [6] (probability γ for the nucleotide transitions A$\longleftrightarrow$G and C$\longleftrightarrow$T, and probability β for the nucleotide transversions A$\longleftrightarrow$C, A$\longleftrightarrow$T, C$\longleftrightarrow$G and G$\longleftrightarrow$T) and then, to a 3-parameter substitution model [7] (probability a for transitions, probability b for the transversion type A$\longleftrightarrow$T and C$\longleftrightarrow$G, and probability c for the transversion type A$\longleftrightarrow$C and G$\longleftrightarrow$T). Later, these substitution models were generalized to a greater number of substitution parameters, e.g. a 6-parameter substitution model with equal base frequencies [15].

Nucleotide substitution models were extended to genetic motif substitution models, e.g. [1, 2] for the pioneer work. The most recent motif substitution models [11, 12, 13], i.e. trinucleotide models with three substitution rates per motif site, are based on a block matrix factorization [14]. However, this approach cannot be used to generalize the substitution models to genetic motifs of any size. Indeed, the construction of large substitution matrices and their eigenvalues and eigenvectors determination are impossible by applying classical methods of formal calculus with the current software (e.g. Mathematica 8.1 in 2011).

By exploring new mathematical strategies, we identify an unexpected Kronecker relation between the nucleotide substitution matrices and the motif substitution matrices. This new property in this evolution research field allows to construct large motif substitution matrices from nucleotide ones as well as to determine the eigenvalues and eigenvectors associated with genetic motifs from the elementary eigenvalues and eigenvectors associated with nucleotides.

The SEGM (Stochastic Evolution of Genetic Motifs) web server is a web application which was built in 2009 to study evolution of nucleotides, dinucleotides and trinucleotides [3]. Based on this Kronecker property, a new version of this SEGM web server is developed which includes several

improved functionalities (in particular a faster computation) and an extension to motifs of size 4 (tetranucleotides) and 5 (pentanucleotides). This research software extension allows biologists to study evolution of large motifs, for example promoter sites (CAAT box, TATA box, etc.). The theory proposed here is valid for genetic motifs of any size. Thus, application to hexanucleotides is obviously possible, for example, to study evolution of recoding signals stimulating read-through stop codons [4]. The current limit is only the PC power (CPU and memory). Indeed, the computation of the evolution probability of one pentanucleotide leads to an analytical solution with 4^{10} terms, i.e. more than one million of terms. The SEGM web server is freely available at http://lsiit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/SEGM.html.

2. Mathematical model

The generalized model will give here the analytical occurrence probabilities of genetic motifs as a function of a substitution matrix containing up to three formal parameters (substitution rates) per motif site and of an initial occurrence probability vector of genetic motifs. Let us consider a motif of size n on the genetic alphabet {A,C,G,T}. By convention, a genetic motif is represented by its index i, $1 \leq i \leq 4^n$, according to the lexicographical order, e.g. if $n = 3$ (trinucleotides), the index $i = 1$ refers to the first motif AAA and $i = 64$ to the last motif TTT. There are 4^n motifs of size n. For all $1 \leq i \leq 4^n$, we denote by $P_i(t)$, the occurrence probability of motif i of size n at time $t \geq 0$.

2.1. **Stochastic substitution model of genetic motifs of size n.** The substitution process is handled by a differential equation which determines the occurrence probabilities of the 4^n genetic motifs at time $t \geq 0$. The motifs mutate according to constant substitution probabilities. Let us consider two motifs i, j of size n, $1 \leq i, j \leq 4^n$. We denote by $P_T(j \to i)$, the substitution probability of motif j into motif i during time T, $T > 0$. The occurrence probability $P_i(t+T)$ of motif i at time $t+T$ is equal to the sum of probabilities $P_j(t)$ of the 4^n motifs j at previous time t times their substitution probabilities $P_T(j \to i)$ into motif i during T, i.e.

$$P_i(t+T) = \underbrace{\sum_j P_j(t) P_T(j \to i)}_{\text{Probability of motif } i \text{ to appear}} . \tag{2.1}$$

From Equation (2.1), the derivative with respect to time $P_i'(t) = \frac{\partial P_i(t)}{\partial t}$ of the occurrence probability of motif i at time t is

$$\begin{aligned} P_i'(t) &= \lim_{T\to 0}\left[\frac{P_i(t+T)-P_i(t)}{T}\right] \\ &= \lim_{T\to 0}\left[\frac{\sum_j P_j(t)P_T(j\to i)-P_i(t)}{T}\right] \\ &= \lim_{T\to 0}\left[\frac{\sum_{j\neq i} P_j(t)P_T(j\to i)+P_i(t)P_T(i\to i)-P_i(t)}{T}\right] \end{aligned}$$

where $P_T(i \to i)$ represents the probability that motif i does not mutate into a different motif $j \neq i$ during T. Then,

$$\begin{aligned} P_i'(t) &= \lim_{T\to 0}\left[\frac{\sum_{j\neq i} P_j(t)P_T(j\to i)-P_i(t)\left(1-P_T(i\to i)\right)}{T}\right] \\ &= \lim_{T\to 0}\left[\frac{\sum_{j\neq i} P_j(t)P_T(j\to i)-P_i(t)\sum_{j\neq i}P_T(i\to j)}{T}\right] \\ &= \sum_{j\neq i} P_j(t)\lim_{T\to 0}\left[\frac{P_T(j\to i)}{T}\right]-P_i(t)\sum_{j\neq i}\lim_{T\to 0}\left[\frac{P_T(i\to j)}{T}\right]. \end{aligned}$$

For all motifs i, j, the instantaneous substitution probability $P(j \to i)$ of motif j into motif i is assumed to be constant along time. When T is small enough, there is no more than one motif substitution per motif site. Then, the following approximation applies

$$P_T(j\to i) \underset{T\to 0}{=} P(j\to i)T$$

and consequently

$$\lim_{T\to 0}\left(\frac{P_T(j\to i)}{T}\right) = P(j\to i).$$

Finally, for any motif i, the derivative $P_i'(t)$ is

$$\begin{aligned} P_i'(t) &= \sum_{j\neq i} P_j(t)P(j\to i)-P_i(t)\sum_{j\neq i}P(i\to j) \\ &= \sum_{j\neq i} P_j(t)P(j\to i)-P_i(t)\left(1-P(i\to i)\right) \\ &= \sum_{j} P_j(t)P(j\to i)-P_i(t). \end{aligned} \tag{2.2}$$

Let $P_n(t) = [P_i(t)]_{1\leq i\leq 4^n}$ be the column vector of size 4^n made of the probabilities $P_i(t)$ for all $1 \leq i \leq 4^n$. From Equation (2.2), we derive a matrix differential equation which describes the substitution process for genetic motifs

$$\begin{aligned} P_n'(t) &= M_n\cdot P_n(t)-P_n(t) \\ &= (M_n-I_n)\cdot P_n(t) \end{aligned} \tag{2.3}$$

where the symbol $\cdot$ is the matrix product, I_n is the identity matrix $(4^n, 4^n)$ and $M_n = [m_{i,j}]_{1 \leq i,j \leq 4^n}$ is the instantaneous substitution probability matrix whose element $m_{i,j}$ in row i and column j refers to the substitution probability of motif j into motif i

$$m_{i,j} = P(j \to i).$$

The instantaneous substitution probability matrix M_n is stochastic in column. Indeed, for all $1 \leq j \leq 4^n$, the elements of matrix M_n satisfy $\sum_{1 \leq i \leq 4^n} m_{i,j} = \sum_{1 \leq i \leq 4^n} P(j \to i) = 1$. For all $1 \leq j \leq 4^n$, the diagonal elements $m_{j,j}$ of M_n satisfy

$$m_{j,j} = 1 - \sum_{1 \leq i \leq 4^n,\ i \neq j} m_{i,j}.$$

Equation (2.3) is equal to Equation (2) in Michel [11] obtained by a similar approach.

Let $A_n = M_n - I_n$. Then, Equation (2.3) becomes

$$P_n'(t) = A_n \cdot P_n(t). \tag{2.4}$$

If A_n is diagonalizable, i.e. $A_n = Q_n \cdot D_n \cdot Q_n^{-1}$ where D_n is the spectral matrix $(4^n, 4^n)$ and Q_n is its associated eigenvector matrix $(4^n, 4^n)$, then Equation (2.4) becomes

$$P_n'(t) = Q_n \cdot D_n \cdot Q_n^{-1} \cdot P_n(t). \tag{2.5}$$

This differential Equation (2.5) has the classical solution [8]

$$P_n(t) = Q_n \cdot e^{D_n t} \cdot Q_n^{-1} \cdot P_n(0) \tag{2.6}$$

where $e^{D_n t}$ is the exponential spectral matrix $(4^n, 4^n)$ of matrix A_n, Q_n is its associated eigenvector matrix $(4^n, 4^n)$ and $P_n(0)$ is the vector of the 4^n initial occurrence probabilities of motifs at $t = 0$.

2.2. **Substitution matrices of genetic motifs of size n.** For substitution matrices of genetic motifs of size n containing up to three substitution parameters per motif site (extension of the 3-parameter substitution model [7] of nucleotides to any motifs of size n), an unexpected Kronecker property is identified for constructing these classes of substitution matrices. This property was found after a detailed analysis of the dinucleotide matrix δ (Figure 1 in [13]) and the trinucleotide matrix δ (Figure B.1 in [12]).

Let k be the nucleotide site of a genetic motif of size n, $1 \leq k \leq n$. For a given site k, let a_k, b_k and c_k be the parameter of transitions A$\longleftrightarrow$G and C$\longleftrightarrow$T, transversions A$\longleftrightarrow$T and C$\longleftrightarrow$G and transversions A$\longleftrightarrow$C and G$\longleftrightarrow$T, respectively. Thus, a motif of size n has $3n$ substitution parameters.

The substitution matrix A_n $(4^n, 4^n)$ associated with motifs of size n is a block matrix which is classically constructed recursively by varying $k = n$ to $k = 1$ as follows [11, 12, 13]

$$A_k = \begin{pmatrix} A_{k-1} & c_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} \\ c_{n-k+1}I_{k-1} & A_{k-1} & b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} \\ a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} & A_{k-1} & c_{n-k+1}I_{k-1} \\ b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & c_{n-k+1}I_{k-1} & A_{k-1} \end{pmatrix}$$

where I_{k-1} is the identity matrix $\left(4^{k-1}, 4^{k-1}\right)$ with $I_0 = 1$, A_{k-1} is the recursive matrix $\left(4^{k-1}, 4^{k-1}\right)$ with $A_0 = -\sum_{k=1}^{n}(a_k + b_k + c_k)$ and a_k, b_k, c_k, $1 \leq k \leq n$, are the substitution parameters for the kth motif site. As the matrix A_n is real and symmetric, precisely a symmetric circulant matrix, A_n is diagonalizable, i.e. $A_n = Q_n \cdot D_n \cdot Q_n^{-1}$ where D_n is the spectral matrix of A_n and Q_n is its associated eigenvector matrix. This property allows the occurrence probabilities $P_i(t)$ of motifs i to be determined, i.e. Equation 2.6.

Let N_k, $1 \leq k \leq n$, be the nucleotide substitution matrix $(4, 4)$ of a site k of a motif of size n

$$N_k = \begin{pmatrix} d_k & c_k & a_k & b_k \\ c_k & d_k & b_k & a_k \\ a_k & b_k & d_k & c_k \\ b_k & a_k & c_k & d_k \end{pmatrix}$$

with $d_k = -(a_k + b_k + c_k)$. As the matrix N_k is real and symmetric (symmetric circulant matrix), N_k is diagonalizable for all $1 \leq k \leq n$

$$N_k = R \cdot S_k \cdot R^{-1}$$

where the nucleotide spectral matrix S_k of N_k is

$$S_k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(a_k + b_k) & 0 & 0 \\ 0 & 0 & -2(a_k + c_k) & 0 \\ 0 & 0 & 0 & -2(b_k + c_k) \end{pmatrix} \tag{2.7}$$

and its associated nucleotide eigenvectors matrix R is

$$R = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}. \tag{2.8}$$

Remark 1. *For the substitution matrix of nucleotides ($n = 1$), $A_1 = N_1 = Q_1 \cdot D_1 \cdot Q_1^{-1} = R \cdot S_1 \cdot R^{-1}$ leading to $D_1 = S_1$ and $Q_1 = R$.*

We identify an interesting relation between the matrices A_n and N_k.

Indeed, the recursive construction of the substitution matrix A_n is similar to a Kronecker sum of nucleotide substitution matrices N_k associated to each site k of motifs of size n. For a motif size $n > 0$ and $k > 0$,

$$
\begin{aligned}
A_k &= \begin{pmatrix}
A_{k-1} & c_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} \\
c_{n-k+1}I_{k-1} & A_{k-1} & b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} \\
a_{n-k+1}I_{k-1} & b_{n-k+1}I_{k-1} & A_{k-1} & c_{n-k+1}I_{k-1} \\
b_{n-k+1}I_{k-1} & a_{n-k+1}I_{k-1} & c_{n-k+1}I_{k-1} & A_{k-1}
\end{pmatrix} \\
&= \begin{pmatrix}
0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\
c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\
a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\
b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0
\end{pmatrix} \bigotimes I_{k-1} + \begin{pmatrix}
A_{k-1} & & & \\
& A_{k-1} & & \\
& & A_{k-1} & \\
& & & A_{k-1}
\end{pmatrix}
\end{aligned}
$$

where the diagonal block matrix with A_{k-1} on the main diagonal is a matrix $(4^{k-1}, 4^{k-1})$. Then,

$$
A_k = \begin{pmatrix}
0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\
c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\
a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\
b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0
\end{pmatrix} \bigotimes I_{k-1} + I_1 \bigotimes A_{k-1}
$$

with I_1 the identity matrix $(4, 4)$. Therefore, by definition of the Kronecker sum,

$$
A_k = \begin{pmatrix}
0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\
c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\
a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\
b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0
\end{pmatrix} \bigoplus A_{k-1}. \tag{2.9}
$$

Moreover, by noticing that

$$
\begin{aligned}
N_{n-k+1} &= \begin{pmatrix}
0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\
c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\
a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\
b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0
\end{pmatrix} + \begin{pmatrix}
d_{n-k+1} & 0 & 0 & 0 \\
0 & d_{n-k+1} & 0 & 0 \\
0 & 0 & d_{n-k+1} & 0 \\
0 & 0 & 0 & d_{n-k+1}
\end{pmatrix} \\
&= \begin{pmatrix}
0 & c_{n-k+1} & a_{n-k+1} & b_{n-k+1} \\
c_{n-k+1} & 0 & b_{n-k+1} & a_{n-k+1} \\
a_{n-k+1} & b_{n-k+1} & 0 & c_{n-k+1} \\
b_{n-k+1} & a_{n-k+1} & c_{n-k+1} & 0
\end{pmatrix} + d_{n-k+1} \times I_1
\end{aligned}
$$

with $d_{n-k+1} = -(a_{n-k+1} + b_{n-k+1} + c_{n-k+1})$, we can rewrite the recursive Equation (2.9) with the recursive Kronecker sum equation

$$A_k = (N_{n-k+1} - d_{n-k+1} \times I_1) \bigoplus A_{k-1}.$$

Thus,

$$\begin{aligned} A_n &= \bigoplus_{k=1}^{n} (N_k - d_k \times I_1) + A_0 \times I_n \\ &= \bigoplus_{k=1}^{n} N_k - \bigoplus_{k=1}^{n} (d_k \times I_1) + A_0 \times I_n \\ &= \bigoplus_{k=1}^{n} N_k - \sum_{k=1}^{n} d_k \times I_n + A_0 \times I_n \end{aligned}$$

with $A_0 = -\sum_{k=1}^{n} (a_k + b_k + c_k) = -\sum_{k=1}^{n} d_k$, and finally,

$$A_n = \bigoplus_{k=1}^{n} N_k. \tag{2.10}$$

Appendix A illustrates this recurrence relation with a substitution matrix for dinucleotides.

Classical mathematical results [9] allows the spectral matrix D_n and the eigenvectors matrix Q_n to be deduced from S_k and R, respectively

$$\begin{cases} D_n = \bigoplus_{k=1}^{n} S_k \\ Q_n = \bigotimes_{k=1}^{n} R \\ Q_n^{-1} = \left(\bigotimes_{k=1}^{n} R \right)^{-1} = \bigotimes_{k=1}^{n} R^{-1} \end{cases}.$$

Thus, the substitution matrix A_n can be directly determined from the Kronecker sum of the n nucleotide spectral matrices S_k and the Kronecker product of the n nucleotide eigenvectors matrix R as follows

$$A_n = \bigotimes_{k=1}^{n} R \cdot \bigoplus_{k=1}^{n} S_k \cdot \bigotimes_{k=1}^{n} R^{-1}.$$

2.3. **Analytical solutions giving the occurrence probabilities of genetic motifs of size n at time t.** By rewriting Equation (2.6), the occurrence probability $P_n(t)$ of motifs of size n at time t can be expressed as a function of elementary eigenvalues and eigenvectors associated with

nucleotides of each site k

$$\begin{aligned} P_n(t) &= \bigotimes_{k=1}^{n} R \cdot e^{\bigoplus_{k=1}^{n} S_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0) \\ &= \bigotimes_{k=1}^{n} R \cdot \bigotimes_{k=1}^{n} e^{S_k t} \cdot \bigotimes_{k=1}^{n} R^{-1} \cdot P_n(0) \\ &= \bigotimes_{k=1}^{n} \left(R \cdot e^{S_k t} \cdot R^{-1}\right) \cdot P_n(0) \end{aligned} \tag{2.11}$$

where $e^{S_k t}$ is the exponential spectral matrix of matrices N_k.

Proposition 1. *Equation (2.11) gives the occurrence probability $P_n(t)$ of motifs of size n at time t from its past one $P_n(0)$. If we express $P_n(0)$ as a function of $P_n(t)$ in Equation (2.11) then Equation*

$$\tilde{P}_n(t) = \bigotimes_{k=1}^{n} \left(R \cdot e^{-S_k t} \cdot R^{-1}\right) \cdot \tilde{P}_n(0) \tag{2.12}$$

by replacing t by $-t$ gives the past probability $\tilde{P}_n(t)$ of motifs of size n from its current probability $\tilde{P}_n(0)$, i.e. by inverting the direction of the evolution time t.

2.4. **Analytical solution giving the occurrence probability of a genetic motif of size n at time t.** Equation (2.11) determines the analytical solutions for all the 4^n motifs of size n. For $n > 3$ (tetranucleotides, pentanucleotides, etc.), the resolution of this equation system on a current standard PC needs much time and memory. Therefore, for larger motifs, we have also derived an equation allowing to compute directly the occurrence probability $P_{i_1}(t)$ of a genetic motif i_1 of size n at time t. After some algebraic manipulation, we obtain

$$P_{i_1}(t) = \frac{1}{4^n} \sum_{i_2=1}^{4^n} \left(e^{t \times \sum_{k=1}^{n} L_k[\delta(i_2,k)]} \times \sum_{i_3=1}^{4^n} \left(P_{i_3}(0) \times \prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right]\right) \right) \right) \tag{2.13}$$

where $P_{i_3}(0)$ is the initial occurrence probability of motif i_3 of size n at $t = 0$, R the nucleotide eigenvectors matrix (2.8), $\delta(i_x, k) = \left\lfloor \frac{i_x - 1}{4^{n-k}} \right\rfloor [4] + 1$, $1 \leq \delta(i_x, k) \leq 4$, a function associated with the motif i_x and the site k, $\lfloor x \rfloor$ is the integer part of x, and $L_k = [0, -2(a_k + b_k), -2(a_k + c_k), -2(b_k + c_k)]$ is the vector of the 4 eigenvalues of the nucleotide substitution rates matrix N_k (see matrix (2.7)) associated with the site k. The detail of algebraic manipulation is given in Appendix B.

Remark 2. *The 4^n coefficients of initial occurrence probability $P_{i_3}(0)$ in Equation (2.13) are obtained by multiplying the coefficients $R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right]$ of each nucleotide site k. Moreover, the 4^n eigenvalues in Equation (2.13) are obtained from the sum of the k eigenvalues of index $\delta(i_2, k)$ associated with each nucleotide site k.*

Remark 3. *As* $R^{-1} = \frac{1}{4}R$ *here, the coefficients* $R\left[\delta(i_1,k),\delta(i_2,k)\right] \times R\left[\delta(i_2,k),\delta(i_3,k)\right]$ *in Equation (2.13) can be seen as coefficients of matrices* O_k *defined in* [10] *by* $O_k[i,j] = R[i,k] \times R^{-1}[k,j]$.

Remark 4. *Equation (2.13) gives a full simplified analytical solution of the occurrence probability* $P_{i_1}(t)$ *of a motif* i_1 *of size* n *at time* t *composed of* 4^n *exponents, each exponent being multiplied by a sum of the* 4^n *initial occurrence probabilities* $P_{i_3}(0)$, *i.e. a total of* 4^{2n} *terms. Thus, an analytical solution of a pentanucleotide, for example, has more than one million of terms.*

Example 1. *An example of application of Equation (2.13) is given for determining the analytical solution of the occurrence probability of the dinucleotide* AG *at time* t. *The index of* AG *is* $i_1 = 3$ *among* $4^2 = 16$ *dinucleotides* ($n = 2$). *The closed formula of* $P_3(t)$ *obtained is given in Appendix C.*

$$
\begin{aligned}
P_{i_1=3}(t) &= \frac{1}{16}\sum_{i_2=1}^{16}\left(e^{t\times\sum_{k=1}^{2}L_k[\delta(i_2,k)]}\times\sum_{i_3=1}^{16}\left(P_{i_3}(0)\times\prod_{k=1}^{2}\left(R\left[\delta(3,k),\delta(i_2,k)\right]\times R\left[\delta(i_2,k),\delta(i_3,k)\right]\right)\right)\right)\\
&= \frac{1}{16}\sum_{i_2=1}^{16}\Big(e^{t\times(L_1[\delta(i_2,1)]+L_2[\delta(i_2,2)])}\\
&\quad\times\sum_{i_3=1}^{16}\left(P_{i_3}(0)\times\left(R\left[1,\delta(i_2,1)\right]\times R\left[\delta(i_2,1),\delta(i_3,1)\right]\right)\times\left(R\left[3,\delta(i_2,2)\right]\times R\left[\delta(i_2,2),\delta(i_3,2)\right]\right)\right)\Big)\\
&= \frac{1}{16}\sum_{i_2=1}^{16}\Big(e^{t\times(L_1[\delta(i_2,1)]+L_2[\delta(i_2,2)])}\\
&\quad\times\left(P_1(0)\times\left(R\left[1,\delta(i_2,1)\right]\times R\left[\delta(i_2,1),1\right]\right)\times\left(R\left[3,\delta(i_2,2)\right]\times R\left[\delta(i_2,2),1\right]\right)\right.\\
&\quad\left.+\ldots+P_{16}(0)\times\left(R\left[1,\delta(i_2,1)\right]\times R\left[\delta(i_2,1),4\right]\right)\times\left(R\left[3,\delta(i_2,2)\right]\times R\left[\delta(i_2,2),4\right]\right)\right)\Big)\\
&= \frac{1}{16}\Big(e^{t\times(L_1[1]+L_2[1])}\times\left(P_1(0)\times\left(R\left[1,1\right]\times R\left[1,1\right]\right)\times\left(R\left[3,1\right]\times R\left[1,1\right]\right)\right.\\
&\quad\left.+\ldots+P_{16}(0)\times\left(R\left[1,1\right]\times R\left[1,4\right]\right)\times\left(R\left[3,1\right]\times R\left[1,4\right]\right)\right)\\
&\quad+e^{t\times(L_1[1]+L_2[2])}\times\left(P_1(0)\times\left(R\left[1,1\right]\times R\left[1,1\right]\right)\times\left(R\left[3,2\right]\times R\left[2,1\right]\right)\right.\\
&\quad\left.+\ldots+P_{16}(0)\times\left(R\left[1,1\right]\times R\left[1,4\right]\right)\times\left(R\left[3,2\right]\times R\left[2,4\right]\right)\right)\\
&\quad+\ldots+e^{t\times(L_1[4]+L_2[4])}\times\left(P_1(0)\times\left(R\left[1,4\right]\times R\left[4,1\right]\right)\times\left(R\left[3,4\right]\times R\left[4,1\right]\right)\right.\\
&\quad\left.+\ldots+P_{16}(0)\times\left(R\left[1,4\right]\times R\left[4,4\right]\right)\times\left(R\left[3,4\right]\times R\left[4,4\right]\right)\right)\Big)
\end{aligned}
$$

3. Application: extension of the SEGM web server

3.1. **Functionalities.** The biomathematical model developed here allows to extend the SEGM (Stochastic Evolution of Genetic Motifs) web server [3] from trinucleotides to tetranucleotides and pentanucleotides, to improve the computation of analytical solutions with a faster calculus and also to add new functionalities, e.g. the result display. SEGM allows the determination of analytical occurrence probabilities $P(t)$ of genetic motifs of size n (nucleotides to pentanucleotides) at time

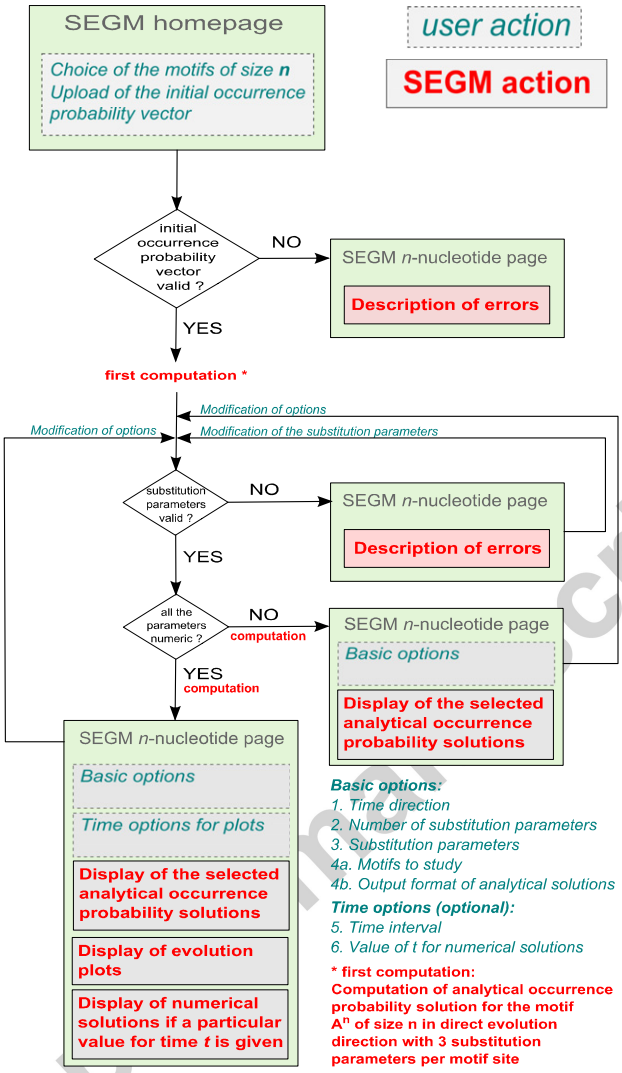

FIGURE 1. Flowchart of the SEGM web server.

t as a function of substitution parameters a_k (A⟷G and C⟷T), b_k (A⟷T and C⟷G) and c_k (A⟷C and G⟷T) per nucleotide site k and an initial occurrence probabilities $P(0)$ of motifs at time $t = 0$. The evolution direction can be direct (past-present) or inverse (present-past). The results are displayed according to several modes defined by the user: general analytical solutions, numerical solutions, evolution plots and analytical solutions converted in C, Fortran or TEX formats in order to facilitate their integration in user-programs. Figure 1 gives the flowchart of the SEGM web server and an overview of its functionalities.

3.2. **Size and initial occurrence probabilities of genetic motifs.** The size n and the initial occurrence probability vector $P(0)$ of studied motifs of size n at time $t = 0$ are chosen on the homepage of the SEGM (Fig. 2). The initial vector $P(0)$ is uploaded thanks to a xls file containing the 4^n initial occurrence probabilities of studied motifs of size n at time $t = 0$. A link gives to

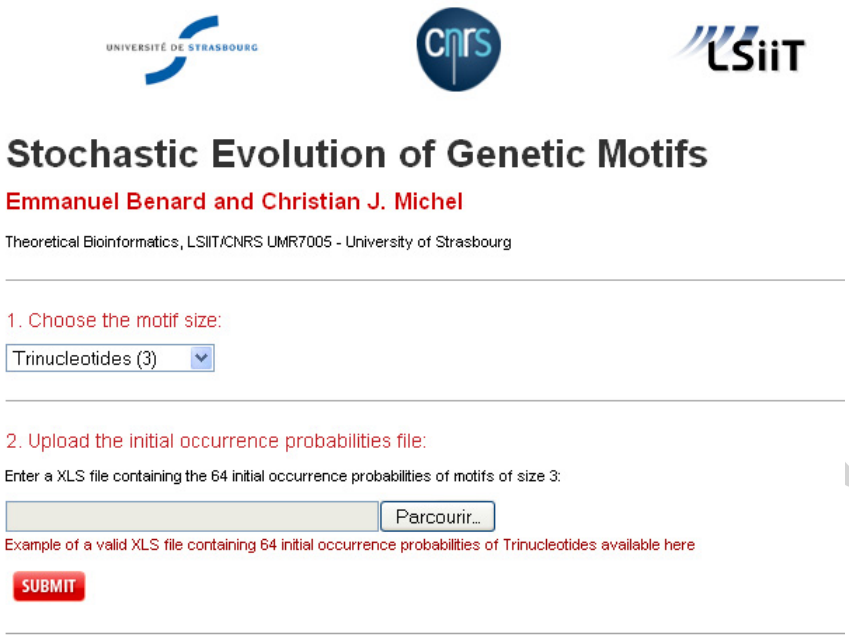

FIGURE 2. Homepage of the SEGM web server: choice of the size of the studied genetic motifs and upload of the initial occurrence probability vector of genetic motifs.

the user the possibility of downloading a template for the xls file. After submission, the vector $P(0)$ is checked by SEGM: its values must be numerical and positive, and its sum must be equal to 1. If errors are detected, a description of these errors is given (Figure 1). Otherwise, the corresponding n-nucleotide page is displayed and the result of a first computation using default options and parameters gives the analytical occurrence probability $P(t)$ of the motif A^n of size n in the direct evolution direction and with three formal substitution parameters a_k , b_k and c_k per motif site. The user can modify these default options and the parameters to get new results with the same initial vector $P(0)$. To study evolution of motifs of different size n or with different initial probabilities $P(0)$ of motifs, he must go back to the homepage of SEGM and upload a new xls file (Figure 2).

Example 2. *The choice of the initial vector* $P(0)$ *depends on the evolutionary problem studied. For example, in order to study "primitive" dinucleotides at donor site, precisely their past occurrence probabilities* $P(t)$ *by inverting the time direction, then the initial vector* $P(0)$ *could be the* 16 *dinucleotides probabilities at donor site at current time, e.g. probabilities obtained from the ICE (Information for the Coordinates of Exons) database from current genes (see Table 2 in* [3]*). Another example, suppose that a DNA sequence is a series of A, e.g. a poly(A) tail to an RNA molecule. Then, the* 256 *occurrence probabilities* $P(t)$ *of tetranucleotides in this sequence subjected*

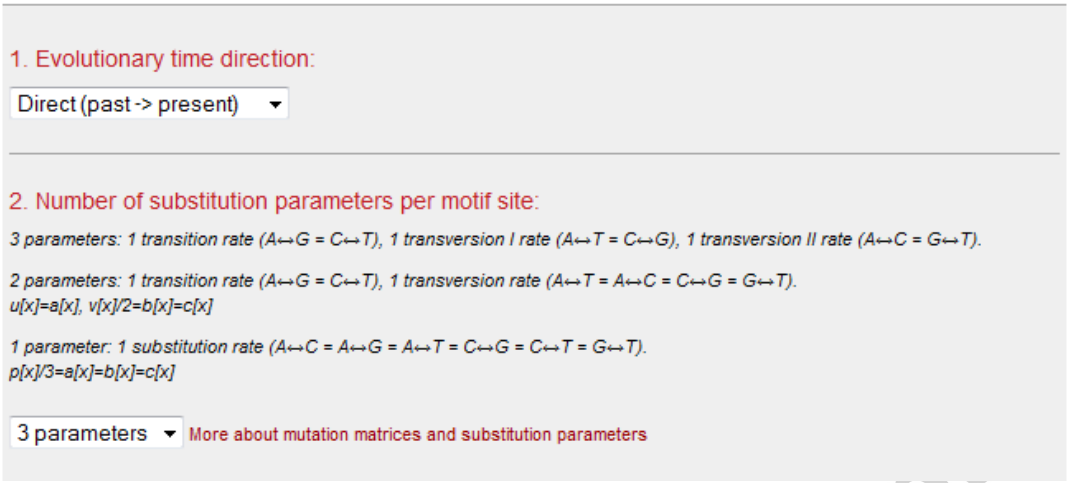

FIGURE 3. Option 1. Choice of the evolutionary time direction. Option 2. Choice of the number of substitution parameters.

to substitutions can be studied with an initial vector $P(0)$ *associated to this sequence, i.e. precisely* $P_1(0) = 1$ *for the motif* $AAAA$ *and* $P_i(0) = 0$ *for the* 255 *other motifs* $i \neq 1$.

3.3. **Basic options.**

3.3.1. *Time direction.* After the submission of the initial occurrence probability vector $P(0)$, the first option is the choice of the evolutionary time direction (Figure 3). The determination of analytical occurrence probabilities $P(t)$ of motifs can be carried out in direct time direction (past-present) using Equation (2.11) or in inverse time direction (present-past) using Equation (2.12). By default, solutions are calculated in direct time direction.

3.3.2. *Number of substitution parameters per motif site.* Option 2 permits to choose the number of substitution parameters per motif site (Figure 3). The biomathematical model of SEGM is an extension of the 3-parameter substitution model [7] of nucleotides to motifs based on three types of substitutions for each motif site k: transitions a_k (A↔G and C↔T), transversions b_k (A↔T and C↔G) and transversions c_k (A↔C and G↔T). It is the model by default. SEGM can also study particular cases extending the 2-parameter nucleotide substitution model [6] and the 1-parameter nucleotide substitution model [5] to motifs. For the model of two substitution parameters per motif site, the parameters are the transitions $u_k = a_k$ (A↔G and C↔T) and the transversions $\frac{v_k}{2} = b_k = c_k$ (A↔C, A↔T, C↔G and G↔T). For the model of one substitution parameter per motif site, the parameters are defined by $\frac{p_k}{3} = a_k = b_k = c_k$. The different models and their associated substitution matrices are explained in a pdf file accessible in Option 2 (Figure 3).

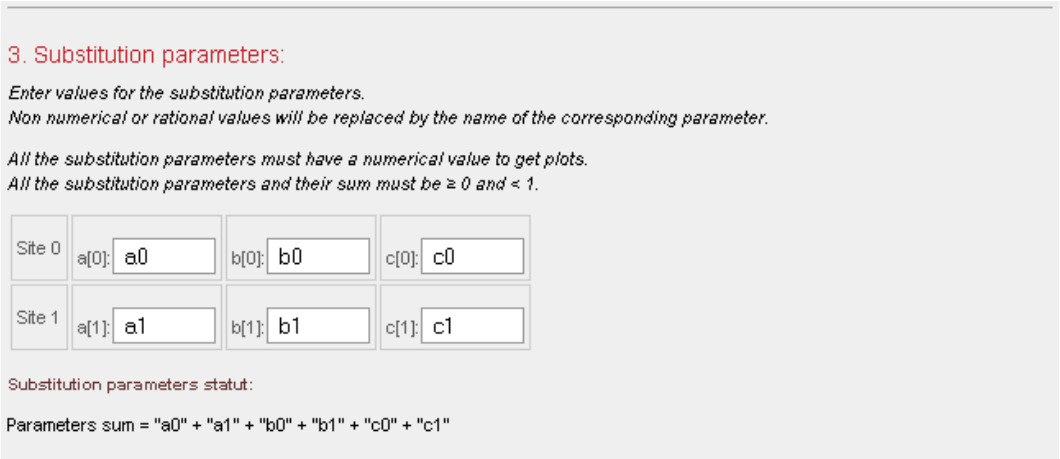

FIGURE 4. Option 3. Substitution parameters which can be formal or numerical.

3.3.3. *Substitution parameters.* The substitution parameters can be left formal or set with numerical values in Option 3 (Figure 4). Numerical values must be positive and their sum must be less or equal to 1. After submission, a description of the encountered errors is displayed if these conditions are not verified (Figure 1). By default, the substitution parameters are left formal to derive formal analytical solutions.

Remark 5. *The motif sites in SEGM are indexed from* 0 *to* $n-1$*,* n *being the size of the studied motifs.*

3.3.4. *Choice of the studied motifs.* Option 4a in this version of SEGM offers the possibility to study evolution up to four large motifs simultaneously. The user selects one motif per list (Figure 5).

3.3.5. *Output format of the analytical solutions.* By default, the analytical occurrence probabilities $P(t)$ of genetic motifs are displayed in a readable text format. In order to facilitate their integration in external user-programs, Option 4b allows others formats: C, Fortran and TEX(Figure 5).

3.4. **Results.** The analytical occurrence probabilities $P(t)$ of genetic motifs are given with formal substitution parameters (Figure 6) in the first computation (Figure 1) or when the substitution parameters are left formal (Figure 4). When all the substitution parameters are set with numerical values (Option 3, Figure 4), evolution plots of the chosen motifs (Option 4a, Figure 5) are displayed: a comparative evolution plot containing the evolution curves of the chosen motifs and a plot drawing their sum curve (Figure 7). In this case, two additional options regarding evolution time t are available.

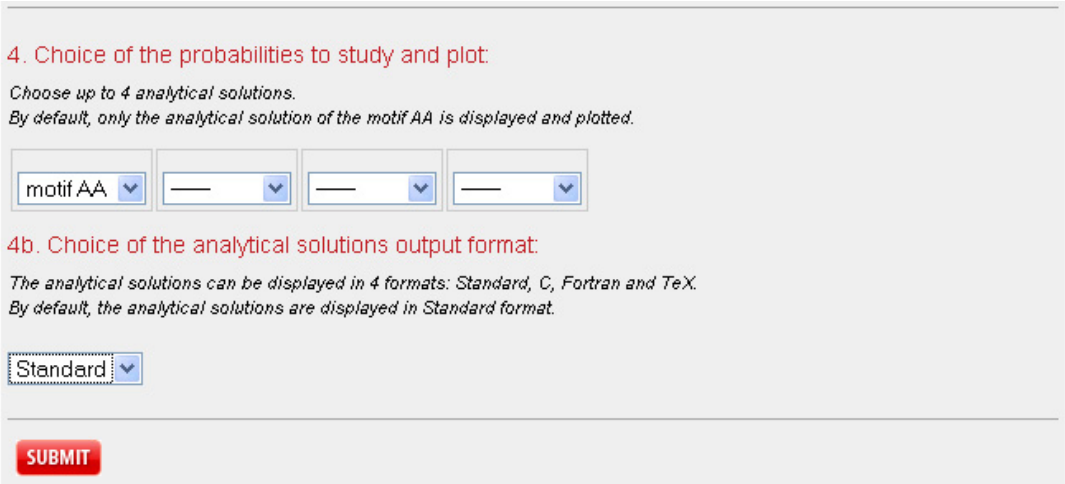

FIGURE 5. Option 4a. Choice of the studied motifs. Option 4b. Choice of the output format for the analytical occurrence probabilities $P(t)$ of genetic motifs at time t.

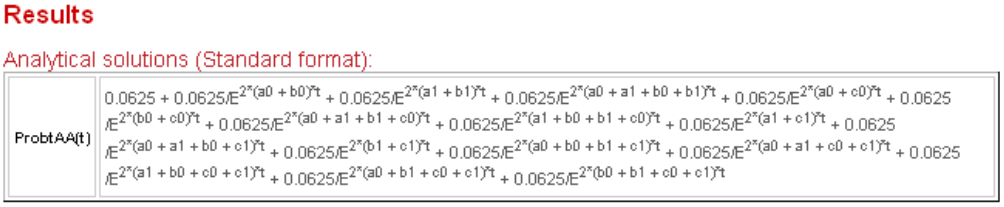

FIGURE 6. Example of analytical occurrence probability $P_1(t)$ of the dinucleotide AA at time t with three formal substitution parameters per site and a particular initial occurrence probability vector of genetic motifs.

The time interval of plots (Figure 7) can be modified thanks to Option 5 (Figure 8). The times $tmin$ and $tmax$ must be always positive even if the inverse time direction is chosen. By default, the time t varies from $tmin = 0$ to $tmax = 5$ whatever the time direction chosen.

A particular value of time t can be set in Option 6 (Figure 8) to have the values of the occurrence probabilities $P(t)$ of studied motifs (Figure 9). As for the evolution plots, a value of the probability sum of studied motifs is also given.

4. DISCUSSION

We proposed here a generalization of stochastic substitution models of nucleotides to genetic motifs of any size. This generalized model is based on the identification of a Kronecker relation between the nucleotide substitution matrices and the motif substitution matrices. It gives the

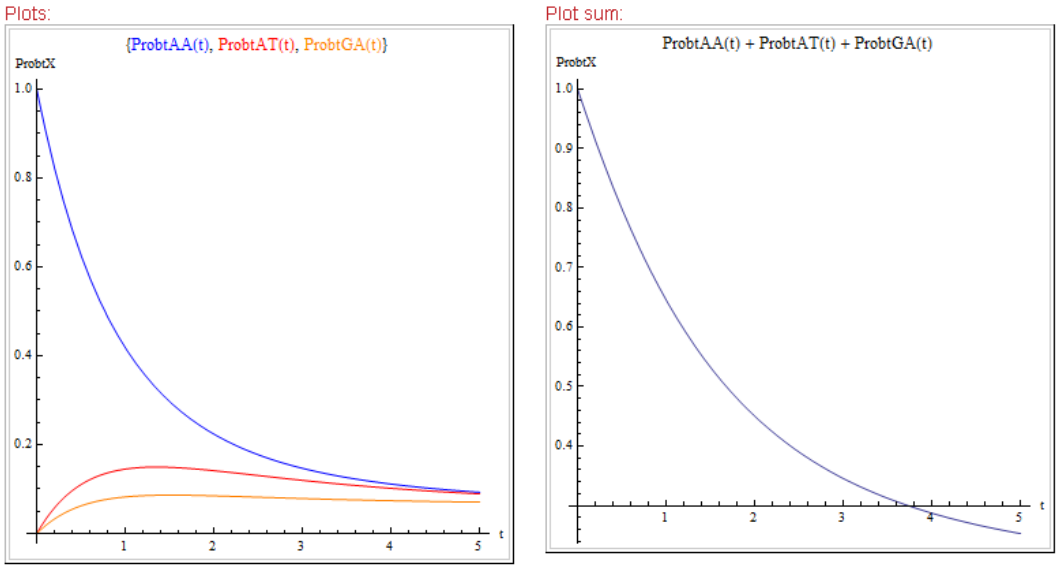

FIGURE 7. Example of evolution curves for the three dinucleotides AA, AT and GA (left) and their sum (right).

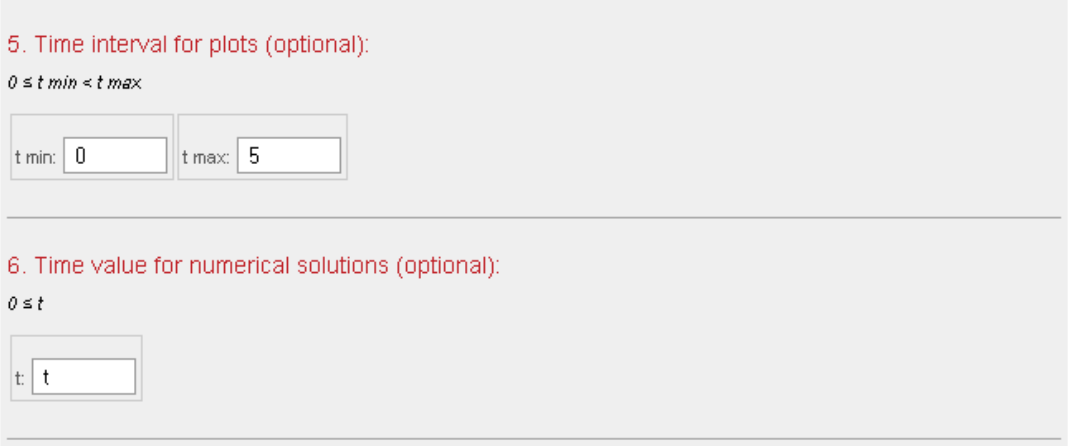

FIGURE 8. Option 5. Choice of the time interval for plots. Option 6. Choice of a time value for numerical solutions.

analytical occurrence probabilities of genetic motifs as a function of a substitution matrix containing up to three formal parameters (substitution rates) per motif site and an initial occurrence probabilities of genetic motifs. This biomathematical model was included in a new version of the SEGM web server offering now several improved functionalities, in particular a faster computation, and an extension to genetic motifs of size 4 (tetranucleotides) and 5 (pentanucleotides). The current limit of the computer implementation of this model is the PC power (CPU and memory). We are currently investigating a parallel approach based on GPU (Mathematica 8.1 integrates GPU programming) in order to allow the evolution analysis of genetic motifs greater than five nucleotides.

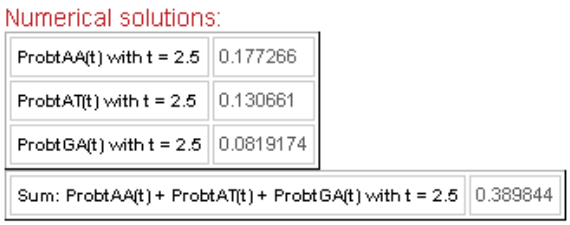

FIGURE 9. Example of numerical solutions for the three dinucleotides AA, AT and GA and their sum at time $t = 2.5$ for particular substitution parameters and a given initial occurrence probability vector of genetic motifs.

Acknowledgement. We thank the reviewers for their advice.

References

[1] Arquès D.G., Michel C.J. (1993). Analytical expression of the purine/pyrimidine codon probability after and before random mutations. Bull. Math. Biol. 55, 1025-1038.

[2] Arquès D.G., Michel C.J. (1995). Analytical solutions of the dinucleotide probability after and before random mutations. J. Theor. Biol. 175, 533-544.

[3] Benard E., Michel C.J. (2009). Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs): An application to splice sites of human genome introns. Comput. Biol. Chem. 33, 245-252.

[4] Harrell A., Melcher U., Atkins J.F. (2002). Predominance of six different hexanucleotide recoding signals 3' of read-through stop codons. Nucleic Acids Res. 30, 2011-2017.

[5] Jukes T.H., Cantor C.R. (1969). Evolution of protein molecules. In H.N. Munro (ed.) Mammalian protein metabolism. Academic Press, New York, 21-132.

[6] Kimura M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111-120.

[7] Kimura M. (1981). Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA 78, 454-458.

[8] Lange K. (2005). Applied Probability. Springer-Verlag, New York.

[9] Laub A.J. (2005). Matrix analysis for scientists and engineers. SIAM Publications, Philadelphia, PA.

[10] Lebre S., Michel C.J. (2010). A stochastic evolution model for residue Insertion-Deletion Independent from substitution. Comp. Biol. Chem. 34, 259-267.

[11] Michel C.J. (2007a). 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, 677-698.

[12] Michel C.J. (2007b). Codon Phylogenetic distance. Comput. Biol. Chem. 31, 36-43.

[13] Michel C.J. (2007c). Evolution probabilities and phylogenetic distance of dinucleotides. J. Theor. Biol. 249, 271-277.

[14] Tian Y., Styan G.P.H. (2001). How to establish block-matrix factorizations. Electron. J. Linear Algebra. 8, 115-127.

[15] Zharkikh A. (1994) Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39, 315-329.

5. Appendix A. Substitution matrix for dinucleotides with the Kronecker method

Construction of the dinucleotide substitution matrix A_2 $(16, 16)$ from the Kronecker sum of the two matrices N_1 and N_2 of size $(4, 4)$ associated to the nucleotide substitution matrices at dinucleotide sites 1 and 2, respectively, using Equation (2.10).

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_1 & d_1 & b_1 & a_1 \\ a_1 & b_1 & d_1 & c_1 \\ b_1 & a_1 & c_1 & d_1 \end{pmatrix} \oplus \begin{pmatrix} d_2 & c_2 & a_2 & b_2 \\ c_2 & d_2 & b_2 & a_2 \\ a_2 & b_2 & d_2 & c_2 \\ b_2 & a_2 & c_2 & d_2 \end{pmatrix}$$

with $d_1 = -(a_1 + b_1 + c_1)$ and $d_2 = -(a_2 + b_2 + c_2)$. Then,

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 & c_1 & a_1 & b_1 \\ c_1 & d_1 & b_1 & a_1 \\ a_1 & b_1 & d_1 & c_1 \\ b_1 & a_1 & c_1 & d_1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes \begin{pmatrix} d_2 & c_2 & a_2 & b_2 \\ c_2 & d_2 & b_2 & a_2 \\ a_2 & b_2 & d_2 & c_2 \\ b_2 & a_2 & c_2 & d_2 \end{pmatrix}$$

$$= \begin{pmatrix} d_1 I_1 & c_1 I_1 & a_1 I_1 & b_1 I_1 \\ c_1 I_1 & d_1 I_1 & b_1 I_1 & a_1 I_1 \\ a_1 I_1 & b_1 I_1 & d_1 I_1 & c_1 I_1 \\ b_1 I_1 & a_1 I_1 & c_1 I_1 & d_1 I_1 \end{pmatrix} + \begin{pmatrix} N_2 & 0 & 0 & 0 \\ 0 & N_2 & 0 & 0 \\ 0 & 0 & N_2 & 0 \\ 0 & 0 & 0 & N_2 \end{pmatrix}$$

where I_1 is the identity matrix $(4, 4)$. Then,

$$N_1 \oplus N_2 = \begin{pmatrix} d_1 I_1 + N_2 & c_1 I_1 & a_1 I_1 & b_1 I_1 \\ c_1 I_1 & d_1 I_1 + N_2 & b_1 I_1 & a_1 I_1 \\ a_1 I_1 & b_1 I_1 & d_1 I_1 + N_2 & c_1 I_1 \\ b_1 I_1 & a_1 I_1 & c_1 I_1 & d_1 I_1 + N_2 \end{pmatrix}$$

$$= \begin{pmatrix}
d & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 \\
c_2 & d & b_2 & a_2 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 \\
a_2 & b_2 & d & c_2 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 \\
b_2 & a_2 & c_2 & d & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 \\
c_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 \\
0 & c_1 & 0 & 0 & c_2 & d & b_2 & a_2 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 \\
0 & 0 & c_1 & 0 & a_2 & b_2 & d & c_2 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 \\
0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & d & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 \\
a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 & c_1 & 0 & 0 & 0 \\
0 & a_1 & 0 & 0 & 0 & b_1 & 0 & 0 & c_2 & d & b_2 & a_2 & 0 & c_1 & 0 & 0 \\
0 & 0 & a_1 & 0 & 0 & 0 & b_1 & 0 & a_2 & b_2 & d & c_2 & 0 & 0 & c_1 & 0 \\
0 & 0 & 0 & a_1 & 0 & 0 & 0 & b_1 & b_2 & a_2 & c_2 & d & 0 & 0 & 0 & c_1 \\
b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & 0 & d & c_2 & a_2 & b_2 \\
0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & 0 & c_2 & d & b_2 & a_2 \\
0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & 0 & a_2 & b_2 & d & c_2 \\
0 & 0 & 0 & b_1 & 0 & 0 & 0 & a_1 & 0 & 0 & 0 & c_1 & b_2 & a_2 & c_2 & d
\end{pmatrix} = A_2.$$

with $d = -(a_1 + b_1 + c_1 + a_2 + b_2 + c_2)$.

6. Appendix B. Detailed derivation of the analytical occurrence probability of a motif i_i of size n at time t

In order to avoid the computation of the occurrence probabilities of the 4^n motifs of size n, we also derived from Equation (2.11) a formula that determines the analytical occurrence probability $P_{i_1}(t)$ at time t of a motif i_1 of size n. Indeed,

$$\begin{aligned} P_{i_1}(t) &= P_n(t)[i_1] \\ &= \left(\bigotimes_{k=1}^{n} \left(R \cdot e^{S_k t} \cdot R^{-1}\right) \cdot P_n(0)\right)[i_1] \\ &= \left(\bigotimes_{k=1}^{n} R\right)[i_1,] \cdot \left(\bigotimes_{k=1}^{n} e^{S_k t}\right) \cdot \left(\bigotimes_{k=1}^{n} R^{-1}\right) \cdot P_n(0). \end{aligned} \tag{6.1}$$

Thus, the computation of the analytical occurrence probability of the motif i_1 does not require the construction of the whole eigenvectors matrix $Q = \bigotimes_{k=1}^{n} R$ but only the i_1th row of $Q[i_1,] = \left(\bigotimes_{k=1}^{n} R\right)[i_1,]$.

Moreover, the matrix Q $(4^n, 4^n)$ is constructed from a Kronecker product of n matrices R $(4, 4)$. Then, the i_1th row of Q can be determined from a Kronecker product of n rows of R of indexes corresponding to the indexes $\delta(i_1, k)$, $1 \leq \delta(i_1, k) \leq 4$, of nucleotides of each site k, $1 \leq k \leq n$, in the motif i_1. For example, the row of Q corresponding to the occurrence probability of the motif GCT of size $n = 3$ and of index $i_1 = 40$ is the result of the Kronecker product of the following three rows of R of indexes $\delta(i_1, k)$: row $\delta(i_1, 1) = 3$ corresponding to the nucleotide G for the site $k = 1$, row $\delta(i_1, 2) = 2$ corresponding to the nucleotide C for the site $k = 2$ and row $\delta(i_1, 3) = 4$ corresponding to the nucleotide T for the site $k = 3$. Thus, the corresponding row of index $i_1 = 40$ in Q is obtained by $Q[40,] = R[\delta(i_1, 1),] \bigotimes R[\delta(i_1, 2),] \bigotimes R[\delta(i_1, 3),] = R[3,] \bigotimes R[2,] \bigotimes R[4,]$. The indexes $\delta(i_1, k)$ of nucleotides of each site k of a motif i_1 of size n are calculated by the formula

$$\delta(i_1, k) = \left\lfloor \frac{i_1 - 1}{4^{n-k}} \right\rfloor [4] + 1$$

where $\left\lfloor \frac{i_1 - 1}{4^{n-k}} \right\rfloor$ is the integer part of $\frac{i_1 - 1}{4^{n-k}}$ and $[\,]$ is the modulo function. As an illustration with the previous example, i.e. the motif GCT of size $n = 3$ and index $i_1 = 40$, $\delta(i_1, 1) = \left\lfloor \frac{40-1}{4^{3-1}} \right\rfloor [4] + 1 = 3$.

Thus, Equation (6.1) can be rewritten

$$P_{i_1}(t) = \left(\bigotimes_{k=1}^{n} R\left[\delta(i_1, k),\right]\right) \cdot \left(\bigotimes_{k=1}^{n} e^{S_k t}\right) \cdot \left(\bigotimes_{k=1}^{n} R^{-1}\right) \cdot P_n(0). \tag{6.2}$$

Equation (6.2) can still be simplified. Indeed, the matrix product $\left(\bigotimes_{k=1}^{n} R\left[\delta(i_1, k),\right]\right) \cdot \left(\bigotimes_{k=1}^{n} e^{S_k t}\right)$ uses a row vector $Q[i_1,]$ and a diagonal matrix e^{St} whose diagonal elements are exponents of eigenvalues of the substitution rates matrix A_n $(4^n, 4^n)$. This matrix e^{St} can also be obtained by

a Kronecker product of n matrices $e^{S_k t}$. The matrix product $Q[i_1,] \cdot e^{St}$ can then be replaced by a scalar product between the row vector $Q[i_1,]$ and a row vector e^{Lt} composed of the diagonal elements of e^{St}. Let L_k be such a row vector associated with the site k of a motif of size n and containing the diagonal elements of S_k, i.e. the eigenvalues of the nucleotide substitution matrix N_k associated with a motif site k, i.e. $L_k = [0, -2(a_k + b_k), -2(a_k + c_k), -2(b_k + c_k)]$ (see matrix (2.7)). By using the row vector $e^{L_k t}$, Equation (6.2) can be rewritten

$$P_{i_1}(t) = \left(\bigotimes_{k=1}^{n} R\left[\delta(i_1,k),\right]\right) \times \left(\bigotimes_{k=1}^{n} e^{L_k t}\right) \cdot \left(\bigotimes_{k=1}^{n} R^{-1}\right) \cdot P_n(0). \tag{6.3}$$

As $\bigotimes_{k=1}^{n} e^{L_k t}$ is also a row vector, Equation (6.3) divides by 4^n the number of operations of Equation (6.2).

Let U_{i_1} be the row vector associated with the motif i_1 and defined by $U_{i_1} = \left(\bigotimes_{k=1}^{n} R\left[\delta(i_1,k),\right]\right) \times \left(\bigotimes_{k=1}^{n} e^{L_k t}\right)$ with its ith element $U_{i_1}[i] = \left(\bigotimes_{k=1}^{n} R\right)[i_1, i] \times e^{\lambda_i t}$ where $e^{\lambda_i t}$ is the ith element of the row vector $\bigotimes_{k=1}^{n} e^{L_k t}$. Let V_{i_1} be the row vector associated with the motif i_1 and defined by $V_{i_1} = U_{i_1} \cdot \left(\bigotimes_{k=1}^{n} R^{-1}\right)$ with its ith element $V_{i_1}[i] = \sum_{j=1}^{4^n} U_{i_1}[j] \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[j, i]$. Then,

$$V_{i_1}[i] = \sum_{j=1}^{4^n} \left(\bigotimes_{k=1}^{n} R\right)[i_1, j] \times e^{\lambda_j t} \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[j, i].$$

From Equation (6.3),

$$\begin{aligned} P_{i_1}(t) &= V_{i_1} \cdot P_n(0) \\ &= \sum_{i_2=1}^{4^n} V_{i_1}[i_2] \times P_{i_2}(0) \\ &= \sum_{i_2=1}^{4^n} \left(\sum_{i_3=1}^{4^n} \left(\left(\bigotimes_{k=1}^{n} R\right)[i_1, i_3] \times e^{\lambda_{i_3} t} \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[i_3, i_2] \right) \times P_{i_2}(0) \right). \end{aligned}$$

By expressing $P_{i_1}(t)$ as a sum of 4^n exponents of eigenvalues $e^{\lambda_{i_2} t}$, each one associated with a sum of 4^n initial occurrence probabilities, we obtain

$$\begin{aligned} P_{i_1}(t) &= \sum_{i_2=1}^{4^n} e^{\lambda_{i_2} t} \times \sum_{i_3=1}^{4^n} \left(P_{i_3}(0) \times \left(\bigotimes_{k=1}^{n} R\right)[i_1, i_2] \times \left(\bigotimes_{k=1}^{n} R^{-1}\right)[i_2, i_3] \right) \\ &= \sum_{i_2=1}^{4^n} e^{t \times \sum_{k=1}^{n} L_k[\delta(i_2,k)]} \times \sum_{i_3=1}^{4^n} \left(P_{i_3}(0) \times \prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right]\right) \times \prod_{k=1}^{n} \left(R^{-1}\left[\delta(i_2,k), \delta(i_3,k)\right]\right) \right) \\ &\sum_{i_2=1}^{4^n} e^{t \times \sum_{k=1}^{n} L_k[\delta(i_2,k)]} \times \sum_{i_3=1}^{4^n} \left(P_{i_3}(0) \times \prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R^{-1}\left[\delta(i_2,k), \delta(i_3,k)\right]\right) \right) \end{aligned} \tag{6.4}$$

with $\delta(i_x,k) = \left\lfloor \frac{i_x-1}{4^{n-k}} \right\rfloor [4] + 1$. As $R^{-1} = \frac{1}{4}R$ (Remark 3), Equation (6.4) simplifies

$$P_{i_1}(t) = \frac{1}{4^n} \sum_{i_2=1}^{4^n} e^{t \times \sum_{k=1}^{n} L_k[\delta(i_2,k)]} \times \sum_{i_3=1}^{4^n} \left(P_{i_3}(0) \times \prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right]\right)\right).$$

7. Appendix C. Analytical solution of the occurrence probability of the dinucleotide AG at time t

$$
\begin{aligned}
P_3(t) \;=\; & \frac{1}{16}\big[e^0(P_1(0)+P_2(0)+P_3(0)+P_4(0)+P_5(0)+P_6(0)+P_7(0)+P_8(0)+\\
& P_9(0)+P_{10}(0)+P_{11}(0)+P_{12}(0)+P_{13}(0)+P_{14}(0)+P_{15}(0)+P_{16}(0))\\
& +e^{-2(a_2+b_2)t}(-P_1(0)-P_2(0)+P_3(0)+P_4(0)-P_5(0)-P_6(0)+P_7(0)+P_8(0)\\
& -P_9(0)-P_{10}(0)+P_{11}(0)+P_{12}(0)-P_{13}(0)-P_{14}(0)+P_{15}(0)+P_{16}(0))\\
& +e^{-2(a_2+c_2)t}(-P_1(0)+P_2(0)+P_3(0)-P_4(0)-P_5(0)+P_6(0)+P_7(0)-P_8(0)\\
& -P_9(0)+P_{10}(0)+P_{11}(0)-P_{12}(0)-P_{13}(0)+P_{14}(0)+P_{15}(0)-P_{16}(0))\\
& +e^{-2(b_2+c_2)t}(P_1(0)-P_2(0)+P_3(0)-P_4(0)+P_5(0)-P_6(0)+P_7(0)-P_8(0)\\
& +P_9(0)-P_{10}(0)+P_{11}(0)-P_{12}(0)+P_{13}(0)-P_{14}(0)+P_{15}(0)-P_{16}(0))\\
& +e^{-2(a_1+b_1)t}(P_1(0)+P_2(0)+P_3(0)+P_4(0)+P_5(0)+P_6(0)+P_7(0)+P_8(0)\\
& -P_9(0)-P_{10}(0)-P_{11}(0)-P_{12}(0)-P_{13}(0)-P_{14}(0)-P_{15}(0)-P_{16}(0))\\
& +e^{-2(a_1+b_1+a_2+b_2)t}(-P_1(0)-P_2(0)+P_3(0)+P_4(0)-P_5(0)-P_6(0)+P_7(0)+P_8(0)\\
& +P_9(0)+P_{10}(0)-P_{11}(0)-P_{12}(0)+P_{13}(0)+P_{14}(0)-P_{15}(0)-P_{16}(0))\\
& +e^{-2(a_1+b_1+a_2+c_2)t}(-P_1(0)+P_2(0)+P_3(0)-P_4(0)-P_5(0)+P_6(0)+P_7(0)-P_8(0)\\
& +P_9(0)-P_{10}(0)-P_{11}(0)+P_{12}(0)+P_{13}(0)-P_{14}(0)-P_{15}(0)+P_{16}(0))\\
& +e^{-2(a_1+b_1+b_2+c_2)t}(P_1(0)-P_2(0)+P_3(0)-P_4(0)+P_5(0)-P_6(0)+P_7(0)-P_8(0)\\
& -P_9(0)+P_{10}(0)-P_{11}(0)+P_{12}(0)-P_{13}(0)+P_{14}(0)-P_{15}(0)+P_{16}(0))\\
& +e^{-2(a_1+c_1)t}(P_1(0)+P_2(0)+P_3(0)+P_4(0)-P_5(0)-P_6(0)-P_7(0)-P_8(0)\\
& -P_9(0)-P_{10}(0)-P_{11}(0)-P_{12}(0)+P_{13}(0)+P_{14}(0)+P_{15}(0)+P_{16}(0))\\
& +e^{-2(a_1+c_1+a_2+b_2)t}(-P_1(0)-P_2(0)+P_3(0)+P_4(0)+P_5(0)+P_6(0)-P_7(0)-P_8(0)\\
& +P_9(0)+P_{10}(0)-P_{11}(0)-P_{12}(0)-P_{13}(0)-P_{14}(0)+P_{15}(0)+P_{16}(0))\\
& +e^{-2(a_1+c_1+a_2+c_2)t}(-P_1(0)+P_2(0)+P_3(0)-P_4(0)+P_5(0)-P_6(0)-P_7(0)+P_8(0)\\
& +P_9(0)-P_{10}(0)-P_{11}(0)+P_{12}(0)-P_{13}(0)+P_{14}(0)+P_{15}(0)-P_{16}(0))\\
& +e^{-2(a_1+c_1+b_2+c_2)t}(P_1(0)-P_2(0)+P_3(0)-P_4(0)-P_5(0)+P_6(0)-P_7(0)+P_8(0)\\
& -P_9(0)+P_{10}(0)-P_{11}(0)+P_{12}(0)+P_{13}(0)-P_{14}(0)+P_{15}(0)-P_{16}(0))\\
& +e^{-2(b_1+c_1)t}(P_1(0)+P_2(0)+P_3(0)+P_4(0)-P_5(0)-P_6(0)-P_7(0)-P_8(0)\\
& +P_9(0)+P_{10}(0)+P_{11}(0)+P_{12}(0)-P_{13}(0)-P_{14}(0)-P_{15}(0)-P_{16}(0))
\end{aligned}
$$

$$+e^{-2(b_1+c_1+a_2+b_2)t}(-P_1(0)-P_2(0)+P_3(0)+P_4(0)+P_5(0)+P_6(0)-P_7(0)-P_8(0)$$

$$-P_9(0)-P_{10}(0)+P_{11}(0)+P_{12}(0)+P_{13}(0)+P_{14}(0)-P_{15}(0)-P_{16}(0))$$

$$+e^{-2(b_1+c_1+a_2+c_2)t}(-P_1(0)+P_2(0)+P_3(0)-P_4(0)+P_5(0)-P_6(0)-P_7(0)+P_8(0)$$

$$-P_9(0)+P_{10}(0)+P_{11}(0)-P_{12}(0)+P_{13}(0)-P_{14}(0)-P_{15}(0)+P_{16}(0))$$

$$+e^{-2(b_1+c_1+b_2+c_2)t}(P_1(0)-P_2(0)+P_3(0)-P_4(0)-P_5(0)+P_6(0)-P_7(0)+P_8(0)$$

$$+P_9(0)-P_{10}(0)+P_{11}(0)-P_{12}(0)-P_{13}(0)+P_{14}(0)-P_{15}(0)+P_{16}(0))]$$

- Generalization of stochastic substitution models of nucleotides

  Determination of analytical occurrence probabilities of genetic motifs at time *t*

  Evolution direction either direct (past-present) or inverse (present-past)

  Kronecker relation between substitution matrices of nucleotides and motifs

  Development of the SEGM (Stochastic Evolution of Genetic Motifs) web server