



Research Article

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

Emmanuel Benard, Christian J. Michel*

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

ARTICLE INFO

Article history:

Received 11 January 2009

Accepted 23 April 2009

Keywords:

Stochastic evolution

Genetic motifs

Direct and inverse mutations

Evolutionary computation

Web server

Splice sites

Human genome introns

ABSTRACT

We present here the SEGM web server (Stochastic Evolution of Genetic Motifs) in order to study the evolution of genetic motifs both in the direct evolutionary sense (past–present) and in the inverse evolutionary sense (present–past). The genetic motifs studied can be nucleotides, dinucleotides and trinucleotides. As an example of an application of SEGM and to understand its functionalities, we give an analysis of inverse mutations of splice sites of human genome introns. SEGM is freely accessible at <http://lsit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/SEGM.html> directly or by the web site <http://dpt-info.u-strasbg.fr/~michel/>. To our knowledge, this SEGM web server is to date the only computational biology software in this evolutionary approach.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

The SEGM web server (Stochastic Evolution of Genetic Motifs) is a new research software allowing evolution of genetic motifs to be studied both in the direct evolutionary time sense (past–present) and in the inverse evolutionary time sense (present–past). So far, three types of genetic motifs can be analyzed: nucleotides (N), dinucleotides (D) and trinucleotides (T). Thus, SEGM is divided into three functionalities: SEN (Stochastic Evolution of Nucleotides), SED (Stochastic Evolution of Dinucleotides) and SET (Stochastic Evolution of Trinucleotides). The subjacent biomathematical model computes the analytical solutions giving the probability vector at time t of genetic motifs (of a given type N , D or T) from an initial probability vector at time $t_0 = 0$ and as a function of several types of substitution parameters. SEGM is freely accessible at <http://lsit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/SEGM.html> directly or by the web site <http://dpt-info.u-strasbg.fr/~michel/>.

As an example of an application of SEGM, we give an analysis of inverse mutations with dinucleotides. The dinucleotides studied here are the splice sites of human genome introns. In eukaryotic genomes, introns are non-coding sequences inserted between

exons (sequences coding for proteins). These sequences are spliced during the maturation of the mRNA. The boundaries of introns, called splice sites, are involved in the splicing process. The canonical splice pair is GT for the donor site (5') and AT for the acceptor site (3') (Burset et al., 2000; Mount, 1982). The data published in Chong et al. (2004) show that nearly 98% of donor and acceptor sites in human genome are respectively GT and AG. Under the hypothesis that the 2% leaving are mainly due to random substitutions, we can assume that all intron splice sites were GT and AG in past. This evolution model can be analyzed with the SEGM research software which can compute inverse mutations in order to retrieve the dinucleotides GT and AG with the highest possible probabilities (close to 1).

We present here the biomathematical model, the functionalities of the SEGM web server and an application with inverse mutations in splice sites of human genome introns.

2. Methods

2.1. Biomathematical model

The SEGM web server is based on a biomathematical model which unifies the previous models (Michel, 2007a,b).

2.1.1. Substitution parameters

In the three applications SEN, SED and SET, evolution of probabilities of genetic motifs depend on three types of substitutions:

* Corresponding author.

E-mail addresses: Emmanuel.Benard@dpt-info.u-strasbg.fr (E. Benard), michel@dpt-info.u-strasbg.fr (C.J. Michel).

Table 1
Mutation matrix M_m .

	$M_m(x_m, y_m, z_m) =$			
	$1 \dots 4^{m-1}$	$4^{m-1} + 1 \dots 2 \times 4^{m-1}$	$2 \times 4^{m-1} + 1 \dots 3 \times 4^{m-1}$	$3 \times 4^{m-1} + 1 \dots 4^m$
$1 \dots 4^{m-1}$	$M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1})$	$z_m I_{m-1}$	$x_m I_{m-1}$	$y_m I_{m-1}$
$4^{m-1} + 1 \dots 2 \times 4^{m-1}$	$z_m I_{m-1}$	$M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1})$	$y_m I_{m-1}$	$x_m I_{m-1}$
$2 \times 4^{m-1} + 1 \dots 3 \times 4^{m-1}$	$x_m I_{m-1}$	$y_m I_{m-1}$	$M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1})$	$z_m I_{m-1}$
$3 \times 4^{m-1} + 1 \dots 4^m$	$y_m I_{m-1}$	$x_m I_{m-1}$	$z_m I_{m-1}$	$M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1})$

- *transitions*: substitutions between two purines or two pyrimidines $A \leftrightarrow G$ and $C \leftrightarrow T$,
- *transversions I*: substitutions between one purine and one pyrimidine (and reciprocally) $A \leftrightarrow T$ and $C \leftrightarrow G$,
- *transversions II*: substitutions between one purine and one pyrimidine (and reciprocally) $A \leftrightarrow C$ and $G \leftrightarrow T$.

The substitution parameters a, b, c, d, e, f, g, h and k are defined as follows: a, d and g are the transition rates in the first, second and third motif sites; b, e and h are the transversion I rates in the first, second and third motif sites; and c, f and k are the transversion II rates in the first, second and third motif sites. Thus, SET is based on the nine substitution parameters a, b, c, d, e, f, g, h and k (three substitution types for the three trinucleotide sites), SED, on six substitution parameters a, b, c, d, e and f (three substitution types for the two dinucleotide sites), and SEN, on three substitution parameters a, b and c (three substitution types for one nucleotide site).

2.1.2. Matrix equation

Let m be the size of a genetic motif, $1 \leq m \leq 3$, i.e. $m = 1$ for the nucleotides N (SEN), $m = 2$ for the dinucleotides D (SED), $m = 3$ for the trinucleotides T (SET). On the 4-letter genetic alphabet, there are 4^m motifs of size m (4 nucleotides N , 16 dinucleotides D , 64 trinucleotides T). By convention, the indexes $i, j \in \{1, \dots, 4^m\}$ represent the 4^m motifs of size m in alphabetical order. Let $P(j \rightarrow i)$ be the substitution probability of a motif j of size m into a motif i of size m . The probability $P(j \rightarrow i)$ is equal to 0 if the substitution is impossible, i.e. if j and i differ more than one nucleotide as the time interval T is assumed to be enough small that a motif cannot mutate successively two times during T . Otherwise, it is given as a function of the substitution rates in a mutation matrix.

Let $P_i(t)$ be the probability at time t of a motif i of size m . The probability at time $(t + T)$ of the motif i is $P_i(t + T)$ so that $P_i(t + T) - P_i(t)$ represents the probabilities of motifs i which appear and disappear during the interval T

$$P_i(t + T) - P_i(t) = T \sum_{j=1}^{4^m} P(j \rightarrow i) P_j(t) - T P_i(t) \quad (1)$$

Then,

$$\lim_{T \rightarrow 0} \frac{P_i(t + T) - P_i(t)}{T} = P'_i(t) = \sum_{j=1}^{4^m} P(j \rightarrow i) P_j(t) - P_i(t) \quad (2)$$

By considering the column vector $P(t) = [P_i(t)]_{1 \leq i \leq 4^m}$ made of the $4^m P_i(t)$, the differential Eq. (2) can be represented by the following matrix equation

$$P'(t) = M_m P(t) - P(t) = (M_m - I_m) P(t) \quad (3)$$

where $M_m(4^m, 4^m)$ is the mutation matrix containing the 4^{2m} motif substitution probabilities $P(j \rightarrow i)$ and I_m , the identity matrix ($4^m, 4^m$).

The matrix differential Eq. (3) can be written in the following form

$$P'(t) = N_m P(t) \quad (4)$$

with $N_m = M_m - I_m$.

As the matrix M_m is real and also symmetrical by construction, the matrix N_m is also real and symmetrical. Therefore, there exist an eigenvector matrix $Q_m(4^m, 4^m)$ and a diagonal matrix $D_m(4^m, 4^m)$ of eigenvalues λ_k of N_m ordered in the same way as the eigenvector columns in Q_m such that $N_m = Q_m D_m Q_m^{-1}$. Then,

$$P'(t) = Q_m D_m Q_m^{-1} P(t) \quad (5)$$

The matrix differential Eq. (5) has the classical solution (Lange, 2005)

$$P(t) = Q_m e^{D_m t} Q_m^{-1} P(0) \quad (6)$$

where $e^{D_m t}$ is the diagonal matrix of exponential eigenvalues $e^{\lambda_k t}$. The 4^m eigenvalues λ_k of N_m are deduced from the 4^m eigenvalues μ_k of M_m such that $\lambda_k = \mu_k - 1$ (formal eigenvalues detailed in Michel, 2007a,b). The determination of the eigenvalues μ_k is given in Appendix A. The 4^m eigenvectors of N_m associated with the 4^m eigenvalues λ_k computed by formal calculus can be put in a form independent of substitution parameters (non-formal eigenvectors detailed in Michel (2007a,b)).

The formula (6) with the initial probability vector $P(0)$ before the substitution process ($t_0 = 0$), the diagonal matrix $e^{D_m t}$ of exponential eigenvalues $e^{\lambda_k t}$ of N_m , its eigenvector matrix Q_m and its inverse Q_m^{-1} , determine the 4^m probabilities $P_i(t)$ at time t of motifs of size m as a function of substitution parameters.

2.1.3. Mutation matrices

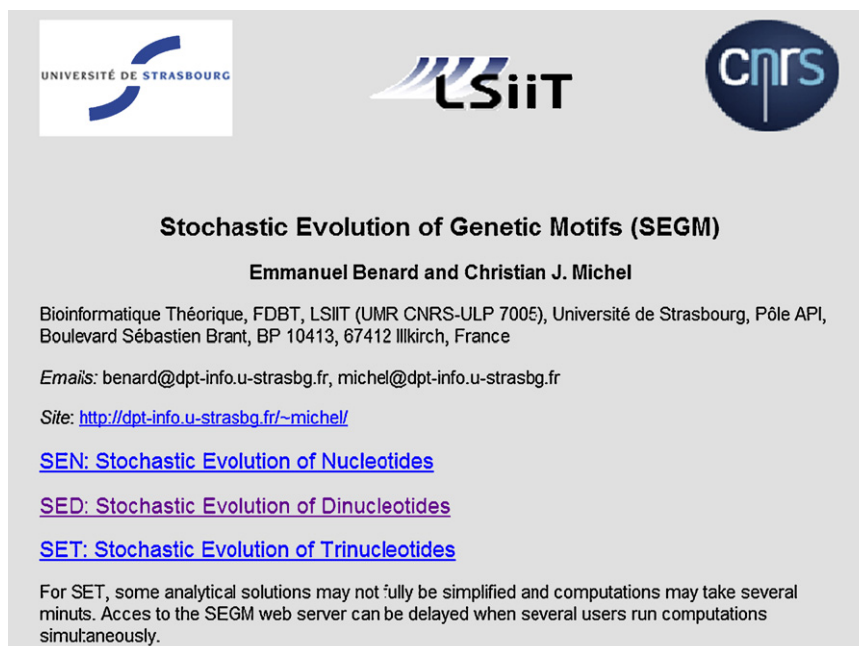
The mutation matrix $M_m(4^m, 4^m)$, $1 \leq m \leq 3$, is associated with the genetic motifs of size m . SEN uses a mutation matrix $M_1(4, 4)$, SED, $M_2(16, 16)$ and SET, $M_3(64, 64)$. A mutation matrix M_m contains 4^{2m} motif substitution probabilities $P(j \rightarrow i)$. As the matrix M_m has a square block structure, it can be defined recursively. The matrix $M_m(x_m, y_m, z_m)(4^m, 4^m)$ with the three substitution parameters noted x_m, y_m and z_m , $1 \leq m \leq 3$, is a square block matrix whose four diagonal elements are square submatrices $M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1})(4^{m-1}, 4^{m-1})$ and whose 12 non-diagonal elements are formed by four square submatrices $x_m I_{m-1}(4^{m-1}, 4^{m-1})$, four square submatrices $y_m I_{m-1}(4^{m-1}, 4^{m-1})$ and four square submatrices $z_m I_{m-1}(4^{m-1}, 4^{m-1})$ with $I_0 = 1$ and $M_0 = 1 - \sum_{i=0}^{m-1} (x_{m-i} + y_{m-i} + z_{m-i})$ (Table 1).

The index ranges $\{1, \dots, 4^{m-1}\}$, $\{4^{m-1} + 1, \dots, 2 \times 4^{m-1}\}$, $\{2 \times 4^{m-1} + 1, \dots, 3 \times 4^{m-1}\}$ and $\{3 \times 4^{m-1} + 1, \dots, 4^m\}$ are associated with the motifs $\{A^m, \dots, AT^{m-1}\}$, $\{CA^{m-1}, \dots, CT^{m-1}\}$, $\{GA^{m-1}, \dots, GT^{m-1}\}$ and $\{TA^{m-1}, \dots, T^m\}$ in alphabetical order, respectively (the empty word being $\varepsilon = I^0$ with $l = \{A, C, G, T\}$).

2.1.4. Time inversion

The formula (6) gives the probabilities at time t of motifs of size m from their past ones $P(0)$. If we express $P(0)$ as a function of $P(t)$ in (6), then the formula

$$\tilde{P}(t) = Q_m e^{-D_m t} Q_m^{-1} \tilde{P}(0) \quad (7)$$



Stochastic Evolution of Genetic Motifs (SEGM)

Emmanuel Benard and Christian J. Michel

Bioinformatique Théorique, FDBT, LSiiT (UMR CNRS-ULP 7005), Université de Strasbourg, Pôle API, Boulevard Sébastien Brant, BP 10413, 67412 Illkirch, France

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

Site: <http://dpt-info.u-strasbg.fr/~michel/>

[SEN: Stochastic Evolution of Nucleotides](#)

[SED: Stochastic Evolution of Dinucleotides](#)

[SET: Stochastic Evolution of Trinucleotides](#)

For SET, some analytical solutions may not fully be simplified and computations may take several minutes. Access to the SEGM web server can be delayed when several users run computations simultaneously.

Fig. 1. Screenshot of the start page of the SEGM web server. The three applications SEN (Stochastic Evolution of Nucleotides), SED (Stochastic Evolution of Dinucleotides) and SET (Stochastic Evolution of Trinucleotides) are accessible from this page.

by replacing t by $-t$ gives the past probabilities of motifs of size m from their current probabilities $\tilde{P}(0)$, i.e. by inverting the direction of the evolutionary time.

3. Results

3.1. Implementation

The SEGM web server is divided into three applications (SEN, SED and SET) hosted on a Tomcat server. It is developed with the calculus software Mathematica 5.2 and webMathematica. In addition, the layout and some functionalities use JavaScript (e.g. tables and updates of substitution rates tables).

On our standard PC, the computations with the SET functionality may take several minutes and the analytical solutions may not be fully simplified.

3.2. Functionalities

The aim of the SEGM web server is to analyze evolution of genetic motifs (nucleotides, dinucleotides, trinucleotides) from a vector of initial probabilities according to three types of substitutions (one transition and two transversions) per site, both in the direct and inverse evolutionary time senses (past–present and present–past). SEGM presents several functionalities:

- **General analytical solutions.** The analytical solutions giving the probabilities $P(t)$ (6) or $\tilde{P}(t)$ (7) at time t are initially computed and determined for all motifs of size m in a formal way as a function of the three substitution parameters per site (parts 1, 2 and 3 of SEGM). The constants in these solutions are in rational or real formats (see Section 3.2.2).
- **Particular analytical solutions.** The analytical solutions $P(t)$ or $\tilde{P}(t)$ can be derived in a formal way as a function of two or one substitution parameters per site (part 4 of SEGM).
- **Partial analytical solutions.** If some numerical values (in rational or real formats) are given for some substitution parameters, partial analytical solutions $P(t)$ or $\tilde{P}(t)$ are derived (part 5 of SEGM).

- **Numerical solutions.** If numerical values (in rational or real formats) are given for all substitution parameters and time t , $P(t)$ or $\tilde{P}(t)$ return a probability (in rational or real formats) (parts 5 and 6 of SEGM).
- **Plots.** If numerical values (in rational or real formats) are given for all substitution parameters, plots of $P(t)$ or $\tilde{P}(t)$ are available as a function of t (part 5 of SEGM). The time interval for the plots, i.e. the lower and upper limits, can be selected (part 7 of SEGM).

The three applications SEN (Stochastic Evolution of Nucleotides), SED (Stochastic Evolution of Dinucleotides) and SET (Stochastic Evolution of Trinucleotides) of SEGM are accessible from the main page (Fig. 1).

Each SEN, SED and SET functionality is divided into eight parts.

3.2.1. Evolutionary time sense

The 1st option is the choice of the evolutionary time sense (Fig. 2). It can be direct (past–present) to compute the probabilities $P(t)$ (6) or inverse (present–past) for the probabilities $\tilde{P}(t)$ (7). By default, the selected sense is direct.

3.2.2. Initial probabilities

The 2nd option is the choice of the initial probability vector $P(0)$ or $\tilde{P}(0)$ at time $t_0 = 0$ of genetic motifs (Fig. 2). Initial probabilities can be entered either in rational format, e.g. $1/3$, or in real format, e.g. 0.3333. If all initial probabilities are entered in rational format, all the analytical solutions obtained for all motifs are in rational format. If one or several initial probabilities are entered in real format, all the analytical solutions obtained for all motifs are in real format. The values of the initial probabilities entered must be numbers in the interval $[0, 1]$ and their sum equal to 1. Error messages appear in red if these two tests are not verified. The sum of the initial probabilities entered is displayed to facilitate the corrections of these values. By default, some initial probabilities are given.

Initial probabilities in rational format allow exact analytical solutions to be determined but computations are faster in real format.

Stochastic Evolution of Dinucleotides (SED)

by Emmanuel Benard & Christian J. Michel, Bioinformatique Théorique, FDBT,
LSIIT (UMR CNRS-ULP 7005), Université de Strasbourg

1. Evolutionary time sense

Direct evolutionary sense (past-present) ▾

2. Initial probabilities Prob(i,t₀) of the 16 dinucleotides i at time t₀

Rational format, e.g. 1/3, or real format, e.g. 0.3333,
such that the initial probability sum is equal to 1
Probabilities must be ≥ 0

ProbAA _{t0} 12/10643	ProbAC _{t0} 14/10643	ProbAG _{t0} 10/10643	ProbAT _{t0} 19/10643
ProbCA _{t0} 11/10643	ProbCC _{t0} 10/10643	ProbCG _{t0} 2/10643	ProbCT _{t0} 18/10643
ProbGA _{t0} 17/10643	ProbGC _{t0} 134/10643	ProbGG _{t0} 17/10643	ProbGT _{t0} 10348/10643
ProbTA _{t0} 5/10643	ProbTC _{t0} 8/10643	ProbTG _{t0} 7/10643	ProbTT _{t0} 11/10643

The initial probabilities sum must be equal to 1 and is **1**

3. Evaluate analytical probabilities of dinucleotides

Fig. 2. Example with the SED application (parts 1–3 of SEGM): (1) evolutionary time sense; (2) initial probabilities of the 16 dinucleotides; (3) determination of the general analytical solutions.

3.2.3. General analytical solutions

If the two previous tests are verified, the probabilities $P(t)$ (6) or $\tilde{P}(t)$ (7) at time t of all motifs of size m are computed as a function of the maximum number of substitution parameters allowed by the functionality: three parameters (a , b and c) for SEN, six parameters (a , b , c , d , e and f) for SED and nine parameters (a , b , c , d , e , f , g , h and k) for SET.

An example of analytical solutions obtained is given in Fig. 3.

3.2.4. Number of substitution parameters

The 3rd option is the selection of the number of substitution parameters to compute the probabilities of motifs of size m (Fig. 4). For each functionality, three types of substitutions are available:

3. Evaluate analytical probabilities of dinucleotides

Results

Probability	Expression
Prob(AA,t)	$\frac{1}{170288} (10643 - 10451 e^{-2(a+b)t} - 10471 e^{-2(a+c)t} + 10499 e^{-2(b+c)t} - 10221 e^{-2(d+e)t} + 10217 e^{-2(a+b+d+e)t} + 10205 e^{-2(a+c+d+e)t} - 10213 e^{-2(b+c+d+e)t} + 10239 e^{-2(d+f)t} - 10191 e^{-2(a+b+d+f)t} - 10223 e^{-2(a+c+d+f)t} + 10203 e^{-2(b+c+d+f)t} - 10481 e^{-2(e+f)t} + 10429 e^{-2(a+b+e+f)t} + 10445 e^{-2(a+c+e+f)t} - 10437 e^{-2(b+c+e+f)t})$
Prob(AC,t)	$\frac{1}{170288} (10643 - 10451 e^{-2(a+b)t} - 10471 e^{-2(a+c)t} + 10499 e^{-2(b+c)t} - 10221 e^{-2(d+e)t} + 10217 e^{-2(a+b+d+e)t} + 10205 e^{-2(a+c+d+e)t} - 10213 e^{-2(b+c+d+e)t} - 10239 e^{-2(d+f)t} + 10191 e^{-2(a+b+d+f)t} + 10223 e^{-2(a+c+d+f)t} - 10203 e^{-2(b+c+d+f)t} + 10481 e^{-2(e+f)t} - 10429 e^{-2(a+b+e+f)t} - 10445 e^{-2(a+c+e+f)t} + 10437 e^{-2(b+c+e+f)t})$
Prob(AG,t)	$\frac{1}{170288} (10643 - 10451 e^{-2(a+b)t} - 10471 e^{-2(a+c)t} + 10499 e^{-2(b+c)t} + 10221 e^{-2(d+e)t} - 10217 e^{-2(a+b+d+e)t} - 10205 e^{-2(a+c+d+e)t} + 10213 e^{-2(b+c+d+e)t} - 10239 e^{-2(d+f)t} + 10191 e^{-2(a+b+d+f)t} + 10223 e^{-2(a+c+d+f)t} - 10203 e^{-2(b+c+d+f)t} - 10481 e^{-2(e+f)t} + 10429 e^{-2(a+b+e+f)t} + 10445 e^{-2(a+c+e+f)t} - 10437 e^{-2(b+c+e+f)t})$
Prob(AT,t)	$\frac{1}{170288} (10643 - 10451 e^{-2(a+b)t} - 10471 e^{-2(a+c)t} + 10499 e^{-2(b+c)t} + 10221 e^{-2(d+e)t} - 10217 e^{-2(a+b+d+e)t} - 10205 e^{-2(a+c+d+e)t} + 10213 e^{-2(b+c+d+e)t} + 10239 e^{-2(d+f)t} - 10191 e^{-2(a+b+d+f)t} - 10223 e^{-2(a+c+d+f)t} + 10203 e^{-2(b+c+d+f)t} + 10481 e^{-2(e+f)t} - 10429 e^{-2(a+b+e+f)t} - 10445 e^{-2(a+c+e+f)t} + 10437 e^{-2(b+c+e+f)t})$

Fig. 3. Example of analytical solution with the SED application: results with the first four dinucleotides AA, AC, AG and AT.

4. Number of substitution parameters

6 parameters :
a is the transition rate A↔G and C↔T in the 1st dinucleotide site
d is the transition rate A↔G and C↔T in the 2nd dinucleotide site

b is the tranversion I rate A↔T and C↔G in the 1st dinucleotide site
e is the tranversion I rate A↔T and C↔G in the 2nd dinucleotide site

c is the tranversion II rate A↔C and G↔T in the 1st dinucleotide site
f is the tranversion II rate A↔C and G↔T in the 2nd dinucleotide site

3 types of substitution per site : transition rates != transversion I rates != transversion II rates

4 parameters : a=u, b=c=v/2, d=w, e=f=x/2
2 types of substitution per site : transition rates and transversion rates, transversion I rates = transversion II rates

2 parameters : a=b=c=p/3, d=e=f=q/3
1 type of substitution per site : transition rates = transversion I rates = transversion II rates

6 parameters ▾

Fig. 4. Example of number of substitution rates with the SED application (part 4 of SEGM): six parameters (three substitution rates per site), four parameters (two substitution rates per site), two parameters (one substitution rate per site).

- m parameters: one substitution rate per motif site,
- $2m$ parameters: two substitution rates per motif site: transitions ($A \leftrightarrow G, C \leftrightarrow T$) and transversions ($A \leftrightarrow T, C \leftrightarrow G, A \leftrightarrow C, G \leftrightarrow T$),
- $3m$ parameters: three substitution rates per motif site: transitions ($A \leftrightarrow G, C \leftrightarrow T$), transversions I ($A \leftrightarrow T, C \leftrightarrow G$) and transversions II ($A \leftrightarrow C, G \leftrightarrow T$).

By default, the maximum number of substitution parameters is selected.

3.2.5. Substitution parameters

The 4th option allows the substitution parameters to leave them formal or to set one, several or all of them with numerical values (Fig. 5). As with the initial probabilities, these rates can be entered either in rational or real formats. Some conditions must be respected to compute the analytical solutions. If all parameters are numerical, their values must be in the interval $[0, 1]$ and their sum equal to 1. If one or several parameters are not numerical, their sum must be equal or less than 1. Error messages appear in red if these two tests are not verified. The sum of the substitution rates entered is displayed to facilitate the corrections of these values. By default, the substitution parameters are formal.

3.2.6. Value of t for numerical results

The 5th option allows the time t to leave it formal or to set it with a numerical value which can be entered in rational or real formats (Fig. 5). This value must be positive whatever the evolutionary time sense chosen. If both the time t and the substitution parameters are numerical, the result is a probability value which can be in rational or real formats. By default, the time t is formal.

3.2.7. Time interval for plots

The 6th option allows the time interval for the plots to be changed by modifying the lower and upper limits t_{min} and t_{max} (Fig. 5). All substitution parameters must be numerical (rational or real formats) in order to obtain plots of analytical solutions as a function of time t . By default, the lower and upper limits of the time interval are $t_{min} = 0$ and $t_{max} = 50$.

3.2.8. Final analytical solutions and plots

If the previous options are entered, the probabilities $P(t)$ (6) or $\tilde{P}(t)$ (7) of genetic motifs are computed and listed in a table divided into three columns: the first column gives the type of motifs, the second column, their analytical solutions either as a function of substitution parameters and the time t or as numerical values, and

the third column, their graphical results (see for example the plots given for the application in Fig. 7).

3.2.9. Remark

By giving particular values for the initial probability vector $P(0)$ with the SEN functionality, i.e. an initial vector $P(0)$ containing only one nucleotide, e.g. A with a probability equal to 1:

$$P(0) = \begin{cases} P_1(0) = 1 \\ P_i(0) = 0 \quad \forall i \in \{2, 3, 4\} \end{cases}$$

then the user can retrieve the classical evolutionary analytical formulae with one, two and three substitution parameters of Jukes and Cantor (1969) and Kimura (1980, 1981) easily.

3.3. An application of the SEGM web server

The SEGM web server is a general research software to study evolution of nucleotides, dinucleotides and trinucleotides, both in the direct evolutionary time sense (past–present) and in the inverse evolutionary time sense (present–past). As an example of an application of SEGM, we use here its SED functionality to analyze inverse mutations of splice sites of human genome introns. The most frequent boundary pairs of introns are GT-AG (Bursat et al., 2000; Mount, 1982) where GT is the dinucleotide at the donor site (5') and AG, the dinucleotide at the acceptor site (3') (Fig. 6).

As GT and AG are the most frequent dinucleotides in donor and acceptor sites of current introns, we make the hypothesis that all splice sites were GT and AG in past and that their evolution is (mainly) based on random substitutions. Such an evolution model can be studied with SEGM by computing inverse mutations on donor and acceptor sites. Therefore, two inverse evolution models, called \mathcal{A} for the acceptor site and \mathcal{D} for the donor site, based on the formula $\tilde{P}(t)$ (7) are studied with an aim of trying to retrieve a past probability close to 1 for each splice site.

The initial probability vector $\tilde{P}(0)$ is computed from the ICE (Information for the Coordinates of Exons) database (Chong et al., 2004) which contains the distribution of donor–acceptor pairs of introns of human genes. These data represent a total of 10,643 donor–acceptor pairs.

The inverse evolution models \mathcal{A} and \mathcal{D} are computed by scanning the six substitution parameters a, b, c, d, e and f (three substitution types for the two dinucleotide sites) and the evolutionary time t . Each substitution parameter varies in the range $[0, 1]$ with a step of

5. Substitution parameters (rates)

Rational format, e.g. 1/3, or real format, e.g. 0.3333,
such that the sum of parameters is :
- equal to 1 if all parameters are numerical
- ≤ 1 if at least one parameter is not numerical

All parameters must be numerical to get graphic results

6 parameters	4 parameters	2 parameters
a : 0.0735	u :	p :
b : 0.0890	v :	q :
c : 0	w :	
d : 0.6705	x :	
e : 0.0835		
f : 0.0835		

The sum of parameters must be ≤ 1 and is **1**.

6. Value of t for numerical results (optional)

$0 \leq t$

t:

7. Time interval for plots

$0 \leq t_{\min} < t_{\max}$

tmin: , tmax:

8. Evaluate

Fig. 5. Example with the SED application (parts 5–8 of SEGM): (5) substitution rates which can be formal or numerical; (6) evolutionary time t which can be formal or numerical; (7) time interval which can be modified for the plots of analytical solutions; (8) determination of the particular analytical solutions.

0.5% such that their probability sum is equal to 1, and t , in the range $[0, 0.5]$.

3.3.1. Inverse evolution model \mathcal{D} for the donor site GT

The dinucleotide GT occurs 10,348 times as donor splice site in ICE, i.e. 97.23% of the 10,643 dinucleotides at donor site (Table 2). The dinucleotide GC occurs with a higher probability, i.e. 1.26%, among the 15 dinucleotides of lower probabilities. The 14 other dinucleotides share the 1.51% leaving occurrences. The probabilities $\tilde{P}(0)$ of the 16 dinucleotides at the donor site in current human introns are given in Table 2.

Table 3 gives the solution space (minimal and maximal values) and the barycenter values such that the probability of GT becomes close to 1 ($\tilde{P}_{GT}(t) \geq 99\%$).

In the solution space of the model \mathcal{D} , one substitution rate d is significantly higher than the others ($d \geq 60\%$). This parameter d corresponds to a transition $A \leftrightarrow G$ and $C \leftrightarrow T$ in the second dinucleotide site (see Section 2.1.1). The reason of this high rate is related

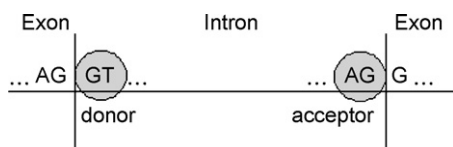


Fig. 6. Most frequent splice junctions between exons and introns.

to the initial frequencies, particularly to the second highest rate of the dinucleotide GC (Table 2). Therefore, the probability increase of GT is mainly provided by the probability decrease of GC which corresponds indeed to a substitution between C and T in the second dinucleotide site modelled by this parameter d .

One substitution rate c in this solution space can be equal to 0, i.e. there is no transversion $A \leftrightarrow C$ and $G \leftrightarrow T$ in the first dinucleotide

Table 2

Probabilities of the 16 dinucleotides at donor site at current time $\tilde{P}(0)$ obtained with the ICE database and past time $\tilde{P}(t)$ determined with the inverse evolution model \mathcal{D} for the donor site GT (Table 3).

Motif	$\tilde{P}(0)$ (%)	$\tilde{P}(t)$ (%)
AA	0.11	0.11
AC	0.13	0.13
AG	0.09	0.09
AT	0.18	0.04
CA	0.10	0.11
CC	0.09	0.09
CG	0.02	0.02
CT	0.17	0.00
GA	0.16	0.00
GC	1.26	0.00
GG	0.16	0.00
GT	97.23	99.11
TA	0.05	0.05
TC	0.08	0.08
TG	0.07	0.07
TT	0.10	0.10

Table 3

Solution space (minimal and maximal values) and barycenter values of the six substitution rates in the inverse evolution model \mathcal{D} for the donor site GT such that $\tilde{P}_{GT}(t) \geq 99\%$. A particular set of values with the rate c equal to 0 is chosen for the plots (Fig. 7).

(%)	a	b	c	d	e	f
Minimum	1.0	0.5	0.0	60.0	0.0	0.0
Maximum	9.5	9.0	5.0	68.0	8.5	8.5
Barycenter	8.1	7.6	3.7	66.6	7.1	7.1
Model \mathcal{D}	7.35	8.90	0.00	67.05	8.35	8.35

site (see Section 2.1.1). As an example, we choose a particular set of values in this solution space with the parameter $c = 0$, called model \mathcal{D} (Table 3), in order to determine the past probabilities $\tilde{P}(t)$ of dinucleotides (Table 2). With this model \mathcal{D} , the probability of GT reaches 99.11% at $t \approx 0.02$ (with a short evolutionary time as $\tilde{P}_{GT}(0)$ is already close to 1). Fig. 7 gives examples of inverse mutations with the dinucleotides AG, GC and GT with this model \mathcal{D} in the time interval $[0.015, 0.025]$.

As the analytical solutions are probabilities, all of them must be in the interval $[0, 1]$. Although $\tilde{P}_{GT}(t)$ continues to increase for $t > 0.02$ (Fig. 7), the optimal solution for $\tilde{P}_{GT}(t)$ is obtained at $t \approx 0.02$ because some dinucleotide probabilities become negative after this time limit (e.g. $\tilde{P}_{GC}(t)$ in Fig. 7).

3.3.2. Inverse evolution model \mathcal{A} for the acceptor site AG

The dinucleotide AG occurs 10,421 times as acceptor splice site in ICE, i.e. 97.91% of the 10,643 dinucleotides at acceptor site (Table 4). The probabilities $\tilde{P}(0)$ of the 16 dinucleotides at the acceptor site in current human introns are given in Table 4.

Table 5 gives the solution space (minimal and maximal values) and the barycenter values such that the probability of AG becomes close to 1 ($\tilde{P}_{AG}(t) \geq 99\%$). In this solution space, three substitution rates can be equal to 0: c , d and f . The barycenter values of these sub-

Table 4

Probabilities of the 16 dinucleotides at acceptor site at current time $\tilde{P}(0)$ obtained with the ICE database and past time $\tilde{P}(t)$ determined with the inverse evolution model \mathcal{A} for the acceptor site AG (Table 5).

Motif	$\tilde{P}(0)$ (%)	$\tilde{P}(t)$ (%)
AA	0.11	0.00
AC	0.28	0.00
AG	97.91	99.00
AT	0.15	0.00
CA	0.15	0.15
CC	0.10	0.10
CG	0.06	0.06
CT	0.24	0.25
GA	0.09	0.08
GC	0.05	0.05
GG	0.30	0.00
GT	0.09	0.08
TA	0.05	0.05
TC	0.05	0.05
TG	0.24	0.00
TT	0.13	0.13

Table 5

Solution space (minimal and maximal values) and barycenter values of the six substitution rates in the inverse evolution model \mathcal{A} for the acceptor site AG such that $\tilde{P}_{AG}(t) \geq 99\%$.

(%)	a	b	c	d	e	f
Minimum	15.5	9.5	0.0	0.0	13.5	0.0
Maximum	30.5	24.5	5.5	11.0	28.5	15.0
Barycenter	27.9	21.9	3.6	8.4	25.9	12.4

stitution rates show that c has the lowest rate (3.6%) in this model \mathcal{A} , similarly to the model \mathcal{D} .

The least rare dinucleotides at acceptor sites are AC and GG (Table 4). The two maximal values of substitution rates in this solution space are a ($A \leftrightarrow G$ and $C \leftrightarrow T$ in the first site) and e ($A \leftrightarrow T$ and C

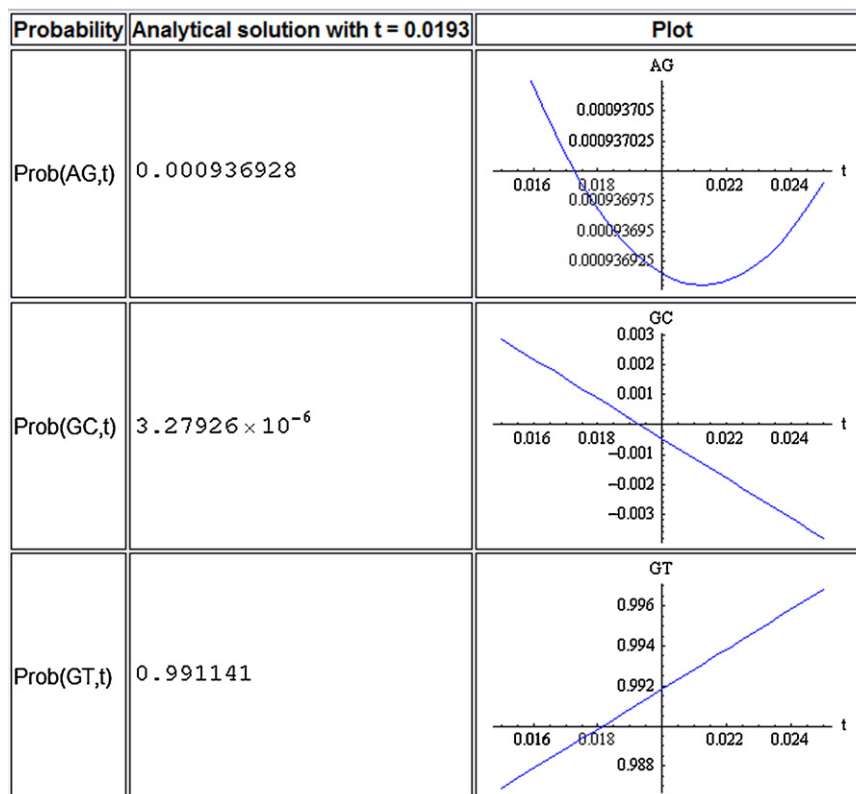


Fig. 7. Plots of the probabilities of the dinucleotides AG, GC and GT in the inverse evolution model \mathcal{D} (Table 3) with a time interval $t \in [0.015, 0.025]$.

↔ G in the second site), about 28% and 26%, respectively (Table 5). Indeed, the increase of AG can be obtained by both the mutation of the first dinucleotide site G of GG into A of AG (parameter a) and the mutation of the second dinucleotide C of AC into G of AG (parameter e).

4. Discussion

We have developed here a new SEGM research software in order to determine analytical solutions giving the probabilities of genetic motifs (nucleotides, dinucleotides, trinucleotides) under mutations, both in the direct evolutionary sense (past–present) and in the inverse evolutionary sense (present–past). This SEGM software is based on a biomathematical model unifying previous models. Therefore, from a theoretical point of view, it can easily

$$M_m(x_m, y_m, z_m) = R_m \cdot \begin{bmatrix} S_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) & 0 & 0 & 0 \\ 0 & T_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) & 0 & 0 \\ 0 & 0 & U_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) & 0 \\ 0 & 0 & 0 & V_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) \end{bmatrix} \cdot R_m$$

with

$$\begin{aligned} S_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) &= M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) + x_m I_{m-1} + y_m I_{m-1} + z_m I_{m-1} \\ T_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) &= M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) - x_m I_{m-1} - y_m I_{m-1} + z_m I_{m-1} \\ U_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) &= M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) + x_m I_{m-1} - y_m I_{m-1} - z_m I_{m-1} \\ V_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) &= M_{m-1}(x_{m-1}, y_{m-1}, z_{m-1}) - x_m I_{m-1} + y_m I_{m-1} - z_m I_{m-1} \end{aligned}$$

and

$$R_m = \frac{1}{2} \begin{bmatrix} I_{m-1} & I_{m-1} & I_{m-1} & I_{m-1} \\ I_{m-1} & I_{m-1} & -I_{m-1} & -I_{m-1} \\ I_{m-1} & -I_{m-1} & I_{m-1} & -I_{m-1} \\ I_{m-1} & -I_{m-1} & -I_{m-1} & I_{m-1} \end{bmatrix}$$

be extended to motifs of higher size (tetranucleotides, pentanucleotides, etc.). However, the current limits are the power of PC and the calculus software as already the computations with the SET functionality may take several minutes and the analytical solutions may not be fully simplified. Indeed, these solutions may have several tens of terms, each term being an exponential function of nine substitution parameters and the evolutionary time t .

The SEGM software is divided into three functionalities: SEN (Stochastic Evolution of Nucleotides), SED (Stochastic Evolution of Dinucleotides) and SET (Stochastic Evolution of Trinucleotides). Each functionality allows to determine:

- general analytical solutions as a function of three substitution parameters per site;
- particular analytical solutions as a function of two or one substitution parameters per site;
- partial analytical solutions when some substitution parameters are numeric (rational or real formats);
- numerical solutions when the substitution parameters and the evolutionary time t are numeric;
- plots as a function of the evolutionary time t when all the substitution parameters are numeric.

As an example of an application of the SEGM software and to understand its functionalities, we proposed a study of inverse mutations of splice sites of human genome introns. Such an analysis allows several biological and evolutionary observations to be deduced: identification of properties with the substitution parameters, e.g. parameters with particular values, interval values, barycenter values, etc. (see for example Tables 3 and 5); analysis of the mutation process for different genetic motifs, e.g.

evolutionary curves with global or local maximum and minimum, increasing or decreasing curves, etc. (see for example the plots in Fig. 7).

In summary, this SEGM software is a general research tool allowing evolution of nucleotides, dinucleotides and trinucleotides in biological applications to be studied. To our knowledge, it is the only research software to compute stochastic evolution of genetic motifs. We are currently adding new functionalities in SEGM.

Appendix A. Determination of eigenvalues

The eigenvalues μ_k of the matrix M_m cannot be determined directly when the matrix size m is greater or equal to 2, i.e. with the dinucleotide matrix M_2 and the trinucleotide matrix M_3 . We apply a block-matrix factorization proposed by Tian and Styan (2001, Corollary 3.3)

The determinant $\det(M_m - \mu I)$ of the factorized characteristic matrix ($M_m - \mu I$) can be determined by using classical computational rules for determinants, particularly the multiplication rule of two matrices A and B ($\det(A \cdot B) = \det(A)\det(B)$) and the rule with a diagonal matrix D ($\det(D) = \prod_i d_{ii}$). Note that $\det(R_m) = 1$. For the dinucleotide matrix M_2 , this factorization is applied directly. For the trinucleotide matrix M_3 , this factorization is applied recursively until $m = 2$. Then, $\det(M_m - \mu I)$ allows the eigenvalues μ_k of the matrix M_m to be deduced.

References

- Burset, M., Seledtsov, I.A., Solovyev, V.V., 2000. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 28, 4364–4375.
- Chong, A., Guanglan, Z., Bajic, V.B., 2004. Information for the coordinates of exons (ice): a human splice sites database. *Genomics* 84, 762–766.
- Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), *Mammalian Protein Metabolism*. Academic Press, pp. 21–132.
- Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16, 111–120.
- Kimura, M., 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *PNAS* 78, 454–458.
- Lange, K., 2005. *Applied Probability*. Springer, New York.
- Michel, C.J., 2007a. Codon phylogenetic distance. *Comput. Biol. Chem.* 31, 36–43.
- Michel, C.J., 2007b. Evolution probabilities of dinucleotides. *J. Theor. Biol.* 249, 271–277.
- Mount, S.M., 1982. A catalogue of splice junction sequences. *Nucleic Acids Res.* 10, 459–472.
- Tian, Y., Styan, G.P.H., 2001. How to establish universal block-matrix factorizations. *The Electronic Journal of Linear Algebra* 8, 115–127.