Drifting Markov Models with Polynomial Drift and Applications to DNA Sequences

Nicolas Vergne

An Article Submitted to

Statistical Applications in Genetics and Molecular Biology

Manuscript 1326

Drifting Markov Models with Polynomial Drift and Applications to DNA Sequences

Nicolas Vergne*

*University of Evry, nvergne@genopole.cnrs.fr

Drifting Markov Models with Polynomial Drift and Applications to DNA Sequences*

Nicolas Vergne

Abstract

In this article, we introduce the drifting Markov models (DMMs) which are inhomogeneous Markov models designed for modeling the heterogeneities of sequences (in our case DNA or protein sequences) in a more flexible way than homogeneous Markov chains or even hidden Markov models (HMMs). We focus here on the polynomial drift: the transition matrix varies in a polynomial way. To show the reliability of our models on DNA, we exhibit high similarities between the probability distributions of nucleotides obtained by our models and the frequencies of these nucleotides computed by using a sliding window. In a further step, these DMMs can be used as the states of an HMM: on each of its segments, the observed process can be modeled by a drifting Markov model. Search of rare words in DNA sequences remains possible with DMMs and according to the fits provided, DMMs turn out to be a powerful tool for this purpose. The software is available on request from the author. It will soon be integrated on *seq++* library (http://stat.genopole.cnrs.fr/seqpp/).

KEYWORDS: drifting Markov models, Markov models, DNA sequences, heterogeneity, rare words

---

*We are grateful to Bernard Prum and Catherine Matias for useful and numerous comments, and Vincent Miele for helpful discussions and important help in the software conception.

1 Introduction

Modeling DNA sequences with stochastic models and developing statistical methods to analyze the enormous set of data that results from the multiple projects of DNA sequencing are challenging questions for statisticians and biologists. The most popular model in this domain is the Markov model on the nucleotides that gives a description of the local behaviour of the sequence (see Almagor (1983), Blaisdell (1985), Phillips et al. (1987), Gelfand et al. (1992)).

Thanks to the statistical properties of these Markov models, we can enlighten different biological properties of DNA or protein sequences. Different Markov models may be proposed. First of all, classical homogeneous with some order k Markov chains provide a general description of a sequence (for instance, the different frequencies of the dinucleotides). Simons et al. (2005) provides a good discussion about this global Markov model. See also Almagor (1983) or Blaisdell (1985) on this point. Schbath et al. (1995) identifies exceptional motifs in sequences using Markov models. Reinert & Schbath (1998) gives another way to detect rare words in biological sequences and Nuel (2001) proposes a comparison between the most used methods for discovering relevant patterns in sequences modeled by classical Markov chains.

The main drawback in considering classical Markov models for the analysis of sequences is that it supposes the homogeneity of sequences, whereas it turns out that long biological sequences are inhomogeneous. A way to take into account this heterogeneity is the use of hidden Markov models (HMMs). HMM is largely used for modeling biological sequences. For instance Churchill (1989) analyzes the heterogeneity of DNA sequences using HMMs. See for example Stanke & Waack (2003) or Krogh et al. (1994) for applications to gene prediction. Thanks to HMMs, one can detect coding or non-coding regions, exons or introns, but also homologies between sequences or discover horizontal transfers (Nicolas et al. 2002). HMM corresponds to the biological fact that some signals succeed one another along the sequence. For example, on a DNA strand, a gene may be followed by a non-coding region, then by a promoter, an other gene and so on. Proteins are often composed of various "domains" separated by hinge. It is natural to think that the way the letters succeed differs from one of these regions to the others, and this explains the success of HMM in the search of regions with different biological roles.

Nevertheless, it is common to observe gradual variations along a biological sequence, either at a global level, either within one of the regions we just mentioned. For example, the gc-richness of a sequence varies according to the position. A first model refers to two kinds of behaviours: high percentage of gc (denoted by H) against low percentage of gc (denoted by L). Then a refined model has been developed, introducing H1, H2, H3, and L1, L2 regions. But there is a broad consensus

about the simplifying aspect of this model, in particular its inability of a sharp determination of the limits of the regions belonging to one of the two (or five) categories: a soft transition from a gc-richness to another one is always observed. As an example, we will use our model on *Phage lambda* complete genome (see Wu & Taylor (1971)). Figures 3(a), 3(b), 3(c) and 3(d) show an estimation of the richness of each of the four nucleotides as a function of the position along this genome (this estimation was obtained using a sliding window of width 2000). The figure shows that "at each position" at least one of the 4 curves has a soft variation. Even around positions 22000, where 3 richness curves seem to have a discontinuity (smoothed by the usage of a window), the fourth one (corresponding to the a nucleotide) has a continuous variation. Even inside genes, for example, this type of behaviour is observed (see Nicolas et al. (2002)).

It is then necessary to develop mathematical tools to account for such gradual changes and we propose such a model, the drifting Markov model (DMM, see precise definition below). It can be seen as a competitive model to the HMM one: a DMM can be adjusted to a whole sequence; and it turns out that the classical problem of the search of rare words remains tractable with this model. But it over all can be understood as a complementary tool: the hidden models of an HMM, usually fixed Markov chains, can be replaced by DMM. This second approach will be treated in a further paper, the present one presenting the necessary tool and first results about its ability for the modeling of biological sequences.

Walking Markov models (WMM), introduced by Fickett et al. (1992) were the first models with a continuous change of base composition. They want to model gc and at composition in a DNA sequence as we just discussed above isochores. For example, they cut a sequence in 1000-base windows and estimate a Markov model on all the windows containing between 300 and 400 at, 400 and 500 at, 500 and 600 at, 600 and 700 at, to have four Markov models. Then for any value w (the at-content), a Markov model M_w is defined by linear interpolation of these primary processes. At last, WMM is defined by a random walk on w: they choose an initial value for w between $1/3$ and $2/3$ (that changes according to the studied sequence), and to choose each succeeding base, they add or subtract (with probability 0.5) 0.0015 from w and use M_w to generate the next base. We use a totally different way to define our DMMs. First, we do not use a random walk to choose our transition matrix: our models are based on the sequence. Second, our models are adapted for any size of state space without a lot of preliminary treatments such as the estimation of some Markov models. It would be difficult to adapt WMM to state space of size 20. Of course, WMMs, just as well as DMMs, do not model detailed local structure, such as the local structure of genes. They are intended to model the large-scale background variation of base composition in the genome.

At now, let us explain the principle of a DMM. Instead of fitting a transition matrix on a whole sequence (homogeneous Markov model) or different transition matrices on different homogeneous parts of the sequence (HMM), we allow the transition matrix to vary (to drift) from the beginning to the end of the sequence. At each position, we obtain a different transition matrix. Our models are thus constrained heterogeneous Markov models. In this paper, we focus on a polynomial drift. The use of such models, where the transition matrix on DNA-alphabet or protein-alphabet (state space) may vary along the genome is a completely new approach.

In the second part of this work, the correct adjustment of probability distributions of nucleotides in DMMs to nucleotide frequencies computed on real sequences shows that our new models provide a more flexible, higher-dimensional parameterization of the data that can be hoped to result in better fits than homogeneous Markov models or HMMs (see Figures from 3(a) to 3(d) and from 4(a) to 4(h)). Then, we compute some model selection criteria (AIC and BIC) to compare different models. Two applications of our models are proposed here. Relying on the compositional asymmetries between the leading and the lagging strand of replication, the program ORILOC (Lobry 2000) helps to predict replication origins in bacterial genomes. We propose an alternative method based on our modeling to detect replication origins which present the advantage of being able to compute analytically a maximum. At last, we discuss a new application for the search of rare words in sequences modeled by a DMM. We offer a simple example with the Chi (`gctggtgg` motif) of *Escherichia coli* and we give different classifications of words according to different models. Many papers treat of rare words and patterns in biological sequences modeled by Markov chains (Schbath et al. 1995, Reinert & Schbath 1998), but all of them are based on Markov chains and their homogeneity. We offer the possibility to study rare words with a model which better correspond to the real sequence, so we can assume the reliability of our result in a better way than before.

This paper is organized in the following way. In Section 2, we describe the drifting Markov models with polynomial drift. Different methods of estimation are proposed and explained. In Section 3, we give first results concerning these new models. We establish reliability of DMMs by adjusting probability distributions of nucleotides and nucleotide frequencies (Section 3.1). We compare different models using AIC and BIC. We propose an alternative to the software ORILOC (Lobry 2000) for detecting replication origins (Section 3.3) and another application to the search of rare words in DNA sequences (Section 3.4). At last, in Section 4, we discuss our results and offer perspectives about these models.

2 Drifting Markov models

A sequence modeled by an order k drifting Markov model is a sequence of random variables X_t:

$$X = (X_t)_{t\in\{0,\dots,n\}}$$

where $n+1$ is the length of the sequence and where instead of fitting only one transition matrix on the whole sequence, we fit a possibly different transition matrix at each position in the sequence. Hence, the distribution of X_t is defined in the following way:

$$\mathbb{P}\left(X_t = v | X_{t-k}\dots X_{t-1} = u\right) = \Pi_{\frac{t}{n}}(u,v)$$

with $u = u_1u_2\dots u_k$ a k-word and $(u_1,u_2,\dots,u_k,v) \in \mathcal{A}^{k+1}$ where $\mathcal{A}$ is the state space (the alphabet $\mathcal{A} = \{\mathtt{a},\mathtt{c},\mathtt{g},\mathtt{t}\}$ for example). Drifting Markov models are inhomogeneous Markov models and without constraints they cannot be estimated. Thus we propose a polynomial evolution of the transition matrix, according to the position in the sequence. We begin by using a linear drift (and later we will more generally use polynomial drifts).

2.1 Drifting Markov models: linear drift

We fix a transition matrix Π_0 at the beginning of the sequence and a transition matrix Π_1 at the end of the sequence and we allow the transition matrix to vary linearly from Π_0 to Π_1:

$$\Pi_{\frac{t}{n}} = \left(1 - \frac{t}{n}\right)\Pi_0 + \frac{t}{n}\Pi_1.$$

Polynomials $(1-t/n)$ and t/n are chosen to establish the stochasticity of Π_0 to Π_1. Obviously, role of Π_0 to Π_1 is artificial as any model parameters but stochastic matrices make easier the understanding of the model. We want to estimate these two matrices in order to build the model. In the case of a simple Markov model, the method of maximum likelihood is successfully used but, because of numerical complexity, we cannot use it here. Hence, we propose two different methods to estimate Π_0 and Π_1: a matrix regression method and a point by point method. We describe these two methods in the following subsections.

We just give the example of an attempt of likelihood maximisation for a DMM of order 1 and degree 1, to conclude that numerical complexity precludes the use of the estimation by maximum likelihood. The likelihood ℓ of a DMM of order 1

and degree 1 is the product over t of the transition matrices $\Pi_{\frac{t}{n}}(X_{t-1}, X_t)$. Then the log-likelihood L is:

$$L(X, \Pi_0, \Pi_1) = \log \nu_0(X_0) + \sum_{t=1}^{n} \sum_{u \in \mathcal{A}} \mathbf{1}_{\{X_{t-1}=u\}} \sum_{v \in \mathcal{A}} \mathbf{1}_{\{X_t=v\}} \log \left(\Pi_{\frac{t}{n}}(u, v) \right).$$

In order to obtain the maximum likelihood, we look for the zero of the derivative of L. We obtain a system of $2|\mathcal{A}|(|\mathcal{A}| - 1)$ equations with $2|\mathcal{A}|(|\mathcal{A}| - 1)$ variables. In fact, it reduces to $|\mathcal{A}|$ systems of $2(|\mathcal{A}| - 1)$ equations with $2(|\mathcal{A}| - 1)$ variables. In Appendix A, we give an example of one of these systems with alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, in order to see that all the variables are in the denominator and that it is unthinkable to solve these systems by analytical or numerical methods. Obviously the same problem exists with polynomial DMMs of higher order or higher degree. It is sheer madness to envisage to solve numerically such a system.

2.1.1 Estimation by a matrix regression method.

A first idea to obtain estimators of the matrices Π_0 and Π_1 is to divide the sequence in N segments of size m (size m will be chosen later). The idea of this method is to use an "approximated homogeneity" on each segment. Then, we fit on each segment S_ℓ, for $1 \leq \ell \leq N$, a Markov model $\widehat{\Pi}_{S_\ell}$ classically estimated by maximum likelihood estimation. In order to fit our heterogeneous model on the whole sequence, we choose one point in each segment. We choose the N centers τ_ℓ of the N segments S_ℓ because $\mathbb{E}(\widehat{\Pi}_{S_\ell})$ tends to Π_{τ_ℓ} as m goes to infinity. We could choose more than one point by segment but that induces numerical complexities without improving the estimation. We want our matrix $\Pi_{\frac{t}{n}}$ to be the nearest possible to each $\widehat{\Pi}_{S_\ell}$ at the center of each segment S_ℓ. Thus, for the matrix regression, we minimize the sum of distances between the estimated matrices on each segment S_ℓ and the transition matrices $\Pi_{\frac{t}{n}}$ at the center τ_ℓ of the ℓ^{th}segment:

$$\sum_{\ell=1}^{N} d \left(\widehat{\Pi}_{S_\ell}, (1 - \tau_\ell)\Pi_0 + \tau_\ell \Pi_1 \right).$$

We choose a quadratic distance for d. Hence, we minimize with respect to $\Pi_0(u, v)$ and $\Pi_1(u, v)$ the following function:

$$\sum_{\ell=1}^{N} \sum_{u \in \mathcal{A}^k} \sum_{v \in \mathcal{A}} \left(\widehat{\Pi}_{S_\ell}(u, v) - (1 - \tau_\ell)\Pi_0(u, v) - \tau_\ell \Pi_1(u, v) \right)^2.$$

For all u in $\mathcal{A}^k$ and v in $\mathcal{A}$, we obtain the following estimators by Lagrange minimization:

$$\widehat{\Pi}_0(u,v) = \frac{b_1 c_2(u,v) - b_2 c_1(u,v)}{a_2 b_1 - a_1 b_2}$$

$$\widehat{\Pi}_1(u,v) = \frac{a_2 c_1(u,v) - a_1 c_2(u,v)}{a_2 b_1 - a_1 b_2}$$

where

$$a_1 = \sum_{\ell=1}^{N}(1-\tau_\ell), \qquad a_2 = \sum_{\ell=1}^{N}\tau_\ell(1-\tau_\ell),$$
$$b_1 = \sum_{\ell=1}^{N}\tau_\ell, \qquad b_2 = \sum_{\ell=1}^{N}{\tau_\ell}^2,$$
$$c_1(u,v) = \sum_{\ell=1}^{N}\widehat{\Pi}_{S_\ell}(u,v), \quad c_2(u,v) = \sum_{\ell=1}^{N}\tau_\ell\widehat{\Pi}_{S_\ell}(u,v).$$

Matrices $\widehat{\Pi}_0$ and $\widehat{\Pi}_1$ are stochastic. In some cases, for small values of N, it is possible to obtain negative terms in the estimated matrices. This problem is solved by a proportional rescaling of the values. Note that we do not obtain a homogeneous model on the segments and that this assumption is only used to get preliminary estimators $\widehat{\Pi}_{S_\ell}$. Size m of the segments is chosen in order to minimize the variance of the estimators. Simulations led us to conclude that the value of m minimizing the variance is $\sqrt{n}$, where n is the length of the sequence. Variance of estimators is analytically obtained using expectation and variance of estimators on each segment.

2.1.2 Estimation by a point by point method.

Another way to estimate Π_0 and Π_1 is a least squares method. We minimize a quadratic form of the different parameters which is the sum of "prediction errors". At each position t in the sequence, knowing the k-word $u = X_{t-k}\dots X_{t-1}$ preceding X_t, we want $\Pi_{\frac{t}{n}}(u,v)$ to be the nearest possible to 1 if $X_t = v$ or the nearest possible to 0 if $X_t \neq v$. We minimize the sum of error squares:

$$\mathbf{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\Pi_{\frac{t}{n}}(u,v) - \mathbf{1}_{\{X_{t-k}\dots X_t=uv\}}.$$

Let us note $\mathbf{1}_u$ for $\mathbf{1}_{\{X_{t-k}\dots X_{t-1}=u\}}$ and $\mathbf{1}_{uv}$ for $\mathbf{1}_{\{X_{t-k}\dots X_{t-1}=u, X_t=v\}}$. We choose a quadratic distance and then we minimize the following function:

$$\sum_{t=1}^{n}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\mathbf{1}_{\{X_{t-k}\dots X_{t-1}=u\}}\left(\Pi_{\frac{t}{n}}(u,v) - \mathbf{1}_{\{X_t=v\}}\right)^2.$$

For all u in $\mathcal{A}^k$ and v in $\mathcal{A}$, we obtain the following estimators by Lagrange minimization

$$\widehat{\Pi}_0(u,v) = \frac{B_2(u)C_1(u,v) - B_1(u)C_2(u,v)}{A_1(u)B_2(u) - A_2(u)B_1(u)}$$

$$\widehat{\Pi}_1(u,v) = \frac{A_1(u)C_2(u,v) - A_2(u)C_1(u,v)}{A_1(u)B_2(u) - A_2(u)B_1(u)}$$

with

$$A_1(u) = 2\sum_{t=1}^{n} \mathbf{1}_u \left(1 - \frac{t}{n}\right)^2, \qquad A_2(u) = 2\sum_{t=1}^{n} \mathbf{1}_u \left(1 - \frac{t}{n}\right)\left(\frac{t}{n}\right),$$
$$B_1(u) = 2\sum_{t=1}^{n} \mathbf{1}_u \left(1 - \frac{t}{n}\right)\left(\frac{t}{n}\right), \qquad B_2(u) = 2\sum_{t=1}^{n} \mathbf{1}_u \left(\frac{t}{n}\right)^2,$$
$$C_1(u,v) = 2\sum_{t=1}^{n} \mathbf{1}_{uv} \left(1 - \frac{t}{n}\right), \qquad C_2(u,v) = 2\sum_{t=1}^{n} \mathbf{1}_{uv} \left(\frac{t}{n}\right).$$

Once again, matrices $\widehat{\Pi}_0$ and $\widehat{\Pi}_1$ are stochastic except in rare cases where negative terms appear. But they are then modified by a proportional rescalement.

2.2 Drifting Markov models: polynomial drift

Up to now, we have only considered a linear variation of the transition matrix (DMM of degree 1), but we can generalize to DMMs of higher degree. Thus DMMs have two order parameters: the order k of the Markov model and the degree d of the polynomial drift. To describe such a polynomial model of degree d, we need $d+1$ points of support. For linear drift ($d = 1$), the model was based on the only two matrices of parameters Π_0 and Π_1. Now, we base our model on $d+1$ matrices $\Pi_{\frac{i}{d}}$, for $0 \leq i \leq d$. We choose $\Pi_{\frac{i}{d}}$ uniformly spaced along the sequence. Any other choice would not be penalizing. Indeed, simulations show that the obtained transition matrices $\Pi_{\frac{t}{n}}$ are similar. The drifting transition matrix has the following form

$$\Pi_{\frac{t}{n}}(u,v) = \sum_{i=0}^{d} p_i(t)\Pi_{\frac{i}{d}}(u,v),$$

where p_i are the polynomial functions of degree d such that

$$\forall (i,j) \in \{0,\ldots,d\}^2,\ p_i\left(\frac{nj}{d}\right) = \mathbf{1}_{\{i=j\}}.$$

Polynomials p_i are chosen to have stochastic matrices $\Pi_{\frac{i}{d}}$. Hence, for $t = ni/d$, we have $\Pi_{\frac{t}{n}} = \Pi_{\frac{i}{d}}$ and for all integer $0 \leq t \leq n$, we have $\sum_{v\in\mathcal{A}} \Pi_{\frac{t}{n}}(u,v) = 1$.

For $d = 1$, we intuitively obtain $p_0(t) = 1 - t/n$ and $p_1(t) = t/n$ in order to have $\Pi_{\frac{t}{n}} = (1 - t/n)\,\Pi_0 + (t/n)\Pi_1$. We give their expression for degree $d = 2$ to illustrate that polynomial functions p_i have not a so simple expression than for degree 1. Indeed,

$$\Pi_{\frac{t}{n}} = p_0(t)\Pi_0 + p_1(t)\Pi_{\frac{1}{2}} + p_2(t)\Pi_1$$

leads to

$$\Pi_{\frac{t}{n}} = \left(2\frac{t^2}{n^2} - 3\frac{t}{n} + 1\right)\Pi_0 + \left(-4\frac{t^2}{n^2} + 4\frac{t}{n}\right)\Pi_{\frac{1}{2}} + \left(2\frac{t^2}{n^2} - \frac{t}{n}\right)\Pi_1.$$

Note that such a system is easy to solve for any degree because it is a simple linear system of $(d+1)(d+1)$ independent equations with $(d+1)(d+1)$ variables. Nonetheless, we cannot give a general explicit expression for p_i with any degree d. At degree 3, we have

$$\begin{aligned}\Pi_{\frac{t}{n}} &= \left(-\frac{9}{2}\frac{t^3}{n^3} + 9\frac{t^2}{n^2} - \frac{11}{2}\frac{t}{n} + 1\right)\Pi_0 + \left(\frac{27}{2}\frac{t^3}{n^3} - \frac{45}{2}\frac{t^2}{n^2} + 9\frac{t}{n}\right)\Pi_{\frac{1}{3}} \\ &+ \left(-\frac{27}{2}\frac{t^3}{n^3} + 18\frac{t^2}{n^2} - \frac{9}{2}\frac{t}{n}\right)\Pi_{\frac{2}{3}} + \left(\frac{9}{2}\frac{t^3}{n^3} - \frac{9}{2}\frac{t^2}{n^2} + \frac{t}{n}\right)\Pi_1.\end{aligned}$$

2.2.1 Estimation by a matrix regression method.

As for the linear drift, we minimize the following function

$$\sum_{\ell=1}^{N}\sum_{u\in\mathcal{A}^k}\sum_{v\in\mathcal{A}}\left(\widehat{\Pi}_{S_\ell}(u,v) - \sum_{i=0}^{d} p_i(n\tau_\ell)\Pi_{\frac{i}{d}}(u,v)\right)^2.$$

Hence, for each (u, v) in $\mathcal{A}^k \times \mathcal{A}$, solving system $AX = B$ where A, X and B are defined below, gives us $\widehat{\Pi}_{\frac{i}{d}}$, estimators of $\Pi_{\frac{i}{d}}$.

- $A_{ij} = \sum_{\ell=1}^{N} p_i(n\tau_\ell)p_j(n\tau_\ell), \quad 0 \le i, j \le d;$
- $X_i = \widehat{\Pi}_{\frac{i}{d}}(u,v), \quad 1 \le i \le d;$
- $B_i = \sum_{\ell=1}^{N} p_i(n\tau_\ell)\widehat{\Pi}_{S_\ell}(u,v), \quad 1 \le i \le d.$

2.2.2 Estimation by a point by point method.

Once again, as in the linear drift case, we need to minimize the following function

$$\sum_{t=1}^{n} \sum_{u \in \mathcal{A}^k} \sum_{v \in \mathcal{A}} \mathbf{1}_{\{X_{t-k} \dots X_{t-1} = u\}} \left(\Pi_{\frac{t}{n}}(u, v) - \mathbf{1}_{\{X_t = v\}} \right)^2. \tag{1}$$

As in the matrix regression method, we need to solve for each (u, v) in $\mathcal{A}^k \times \mathcal{A}$, a system $AX = B$ where

- $A_{ij} = \sum_{t=k}^{n} \mathbf{1}_{\{X_{t-k} \dots X_{t-1} = u\}} p_i(t) p_j(t), \quad 0 \leq i, j \leq d;$
- $X_i = \widehat{\Pi}_{\frac{i}{d}}(u, v), \quad 1 \leq i \leq d;$
- $B_i = \sum_{t=k}^{n} p_i(t) \mathbf{1}_{\{X_{t-k} \dots X_{t-1} = u, X_t = v\}}, \quad 1 \leq i \leq d.$

Hence, we obtain $\widehat{\Pi}_{\frac{i}{d}}$, estimators of $\Pi_{\frac{i}{d}}$.

2.3 Comparison of the methods

There are some differences between matrix regression method and point by point method. Matrix regression method uses preliminary estimations on each segment S_ℓ and the global estimators are computed at a unique point of each segment (the center τ_ℓ). Point by point method enables a direct estimation on all the points of the sequence.

We use the log-likelihood to compare the two methods of estimation of drifting Markov models (see Table 1). We estimate models on the *phage Lambda* complete genome (see Wu & Taylor (1971)) and we consider these models as the true models. Then, we simulate a sequence with each one of these models and we compute the log-likelihood for the two estimation methods. We notice that whatever the order, point by point method always gives better likelihood than the regression method. That is a little bit more apparent when we compute the log-likelihood on the real *phage Lambda* complete genome (see Table 2). Thus, point by point method presents the advantages of a more practical implementation and better results and it is the method which we will privilege thereafter.

Table 1: Log-likelihood of drifting Markov models computed on sequence simulated by each one of these models (R means regression method and P means point by point method).

Degree		0	1	2	3	4	5
Order	R	−67191	−66999	−66962	−66910	−66909	−66907
0	P	**−67191**	**−66999**	**−66962**	**−66910**	**−66909**	**−66907**
Order	R	−66718	−66504	−66448	−66382	−66376	−66368
1	P	**−66710**	**−66501**	**−66445**	**−66380**	**−66374**	**−66366**
Order	R	−66706	−66482	−66407	−66321	−66295	−66275
2	P	**−66693**	**−66477**	**−66402**	**−66317**	**−66290**	**−66270**
Order	R	−66630	−66331	−66186	−66038	−65938	−65883
3	P	**−66612**	**−66320**	**−66169**	**−66014**	**−65898**	**−65817**

Table 2: Log-likelihood of drifting Markov models on *phage Lambda* (R means regression method and P means point by point method).

Degree		0	1	2	3	4	5
Order	R	−67191	−66973	−66934	−66873	−66760	−66680
0	P	**−67191**	**−66973**	**−66934**	**−66873**	**66760**	**−66680**
Order	R	−66743	−66500	−66439	−66362	−66234	−66146
1	P	**−66714**	**−66483**	**−66419**	**−66345**	**−66220**	**−66135**
Order	R	−66052	−65657	−65577	−65438	−65281	−65160
2	P	**−66005**	**−65631**	**−65544**	**−65410**	**−65255**	**−65139**
Order	R	−65661	−65168	−65033	−64809	−64597	−64432
3	P	**−65579**	**−65116**	**−64951**	**−64746**	**−64497**	**−64329**

2.4 Consistence of the estimators

We assume that our estimators are asymptotically unbiased and their variances converge to zero (these theoretical results are not presented here). Thus, our estimators are consistent. In order to show this consistence, we simulate some data where the true model is known. Firstly, we estimate a model on the *phage Lambda* complete genome and we consider this model as the true model. Then, we simulate some sequences with this model and estimate a mean model on all these sequences. In Appendix B, we give an example with both true and estimated models. For each parameter of the model, we note the absolute value of the difference between the true

parameter and the estimated one. In Table 3, we give the mean of these differences for some different drifting Markov models to assume consistence.

Table 3: Comparison between true models and estimated ones. We give the mean of absolute values of differences between true parameters and estimated ones. The number of simulated sequences is given by N.

Degree		0	2	4	6
	N				
Order 1	1	0.0026691	0.00943604	0.00920864	0.00996372
	10	0.0018065	0.00323508	0.00381477	0.00338738
	100	0.0001906	0.00081519	0.00092941	0.00114088
Order 2	1	0.0356951	0.0499792	0.0484183	0.0549545
	10	0.0346129	0.0447842	0.0421838	0.0500320
	100	0.0333321	0.0442818	0.0406379	0.0482310

3 Implementation and results

We developed a program, called DRIMM (as drifting Markov model), dedicated to the estimation of drifting Markov models. This software is written in ANSI `C++` and developed on x86 GNU/Linux systems with GCC 3.4, and successfully tested with GCC latest versions on Sun and Apple Mac OSX systems. It relies on `seq++` library (Miele et al. 2005) and will soon be integrated on `seq++` library. Compilation and installation are compliant with the GNU standard procedure. It is available on request from the author. The software is licensed under the GNU General Public License (`http://www.gnu.org`).

3.1 Marginal distributions

DMMs offer models which describe faithfully real sequences. This fact is particularly highlighted by the study of the probability distributions of nucleotides in the present section. Indeed, analyzing μ_t, the probability distribution of nucleotides at position t associated with our models, is the main way to evaluate their quality. At order 1, the distribution μ_t is recursively defined as follows:

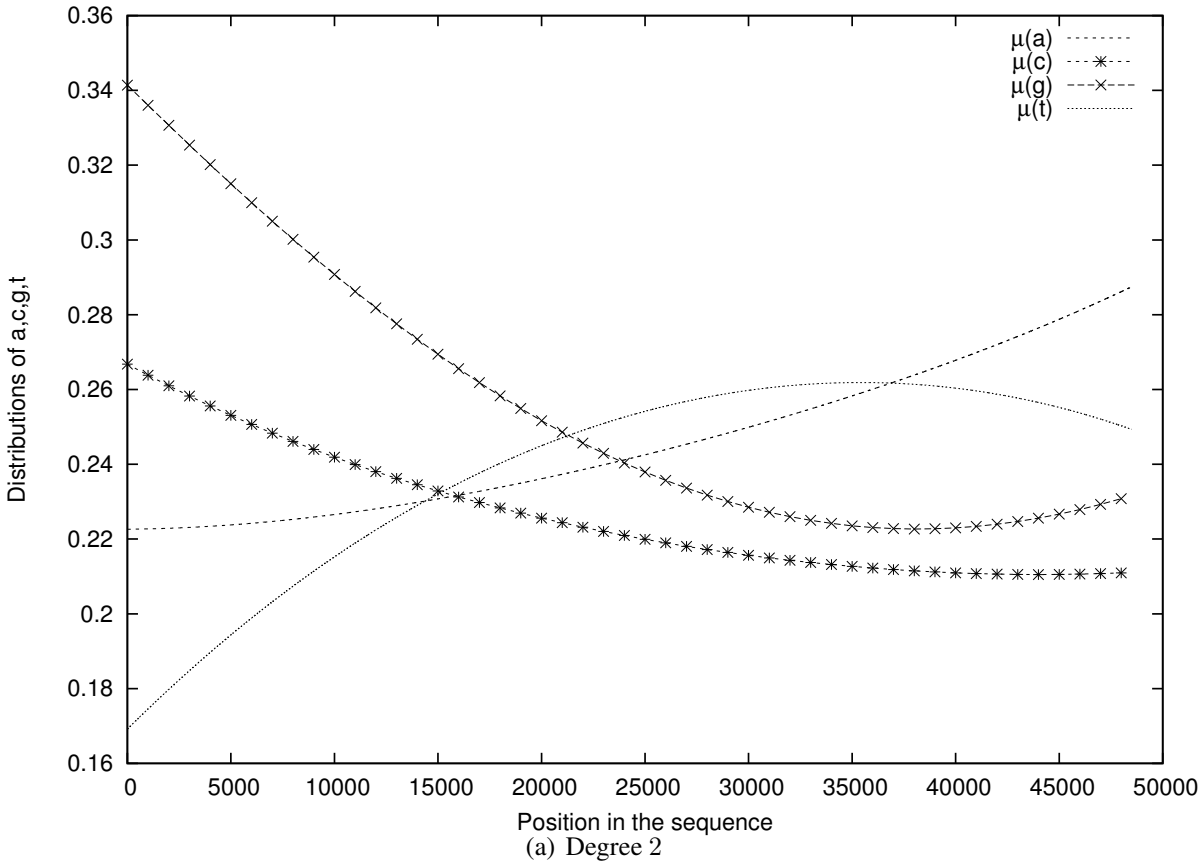
$$\mu_{t+1}(v) = \sum_{u \in \mathcal{A}} \mu_t(u) \Pi_{\frac{t}{n}}(u, v) \quad \forall v \in \mathcal{A}.$$

There are similar definitions for order greater than 1. We recall that an ergodic Markov chain Π on a finite state space has a unique stationary probability distribution ν (such that $\nu\Pi = \nu$). The transition matrix Π_0 is ergodic thus, we choose μ_0 as the stationary probability distribution ν_0 of Π_0. We compute the probability distribution μ_t for each position t to analyze the composition of *phage Lambda* complete genome. First, we draw the evolution, with respect to t, of these distributions of `a`, `c`, `g` and `t` to observe the differences of composition in the sequence.

We present in Figures 1(a), 1(b), 1(c) and 1(d) these distributions for a modeling of the *phage Lambda* sequence by a DMM respectively of degree 2, 4, 6 and 8. For degree $d = 2$, we already notice that `gc`-rate decreases with respect to the position in the sequence. For degree $d = 8$, we observe the first `gc`-rich segment already obtained by an HMM algorithm developed by Muri (1997).

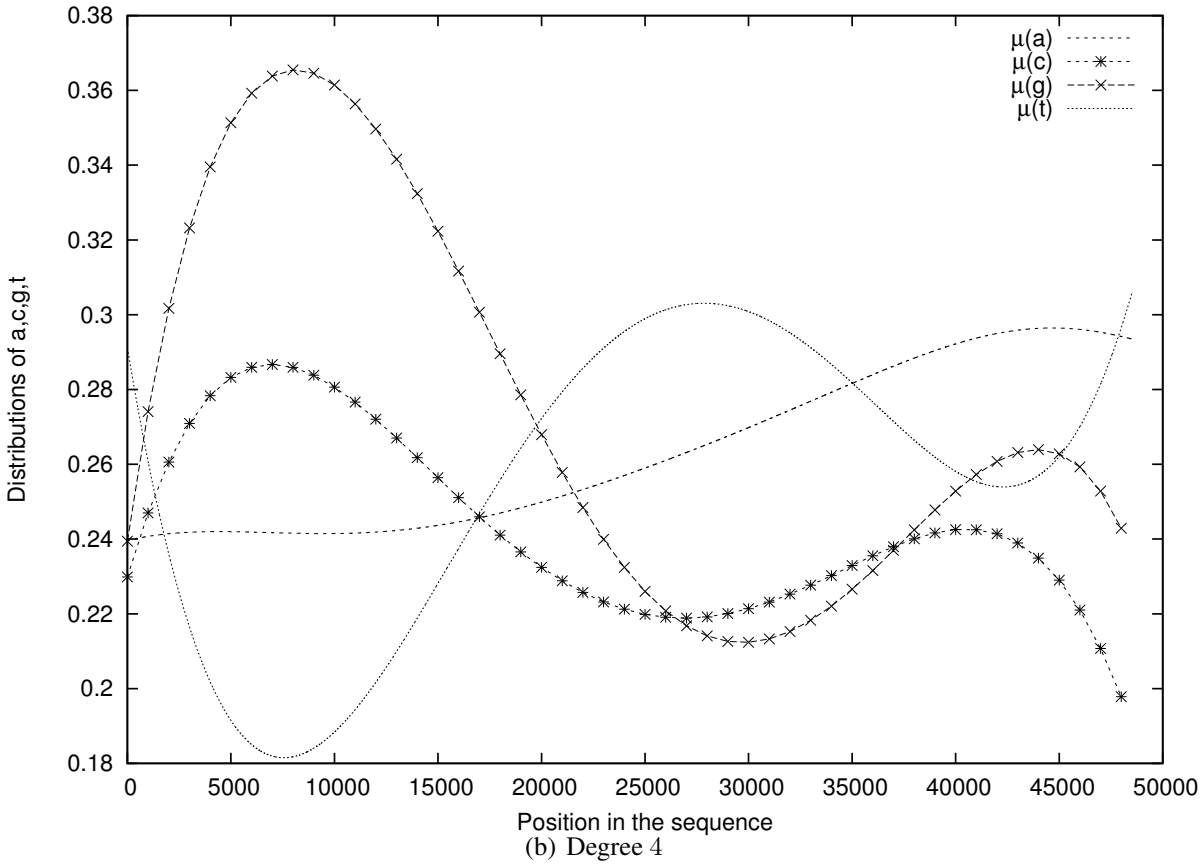
Moreover, comparing the HMM segmentation in Muri (1997) and the DMM evolution of distributions, we observe similarities. Looking at Figure 1(d), we observe very reliable probability distributions of letters which correspond to the HMM segmentation (see Figure 2). This comparison is interesting because it shows the limits of HMM. Although the first long `gc`-rich segment is well known and provided by HMM, other parts of the HMM segmentation are not really convincing in view of the evolution of transition probability. Moreover, DMMs are more tractable numerically and they provide a soft evolution contrary to the sudden segmentation of HMMs.

To establish reliability of drifting Markov models, we compare evolutions of probability distributions with nucleotide frequencies. To compute these frequencies, we use sliding windows of size 2000 nucleotides. Figures 3(a), 3(b), 3(c) and 3(d) show that probability distributions of nucleotides of our degree 8 models are very close to the real distribution of nucleotides in the sequence (respectively nucleotide `a`, `c`, `g` and `t`). Degree 8 is sufficient to observe a good similarity between the curves. In order to compare our polynomial DMM to other Markov models, we draw in Figures 3(a), 3(b), 3(c) and 3(d) the evolutions of probability distributions under an order 1 DMM of degree 0 (it corresponds to a classical homogeneous order 1 Markov model). It turns out that the distance between the two curves is smaller in the case of degree 8 DMM. In the HMM case, we do not observe only one constant probability for each letter as in the Markov model, but few regions with constant probability corresponding to the HMM segmentation. From Figure 4(a) to Figure 4(h), we compare evolutions of probability distributions for an order 1 HMM with 3 hidden states and an order 1 degree 3 DMM on the *Phage T4* complete genome (see Miller et al. (2003)).



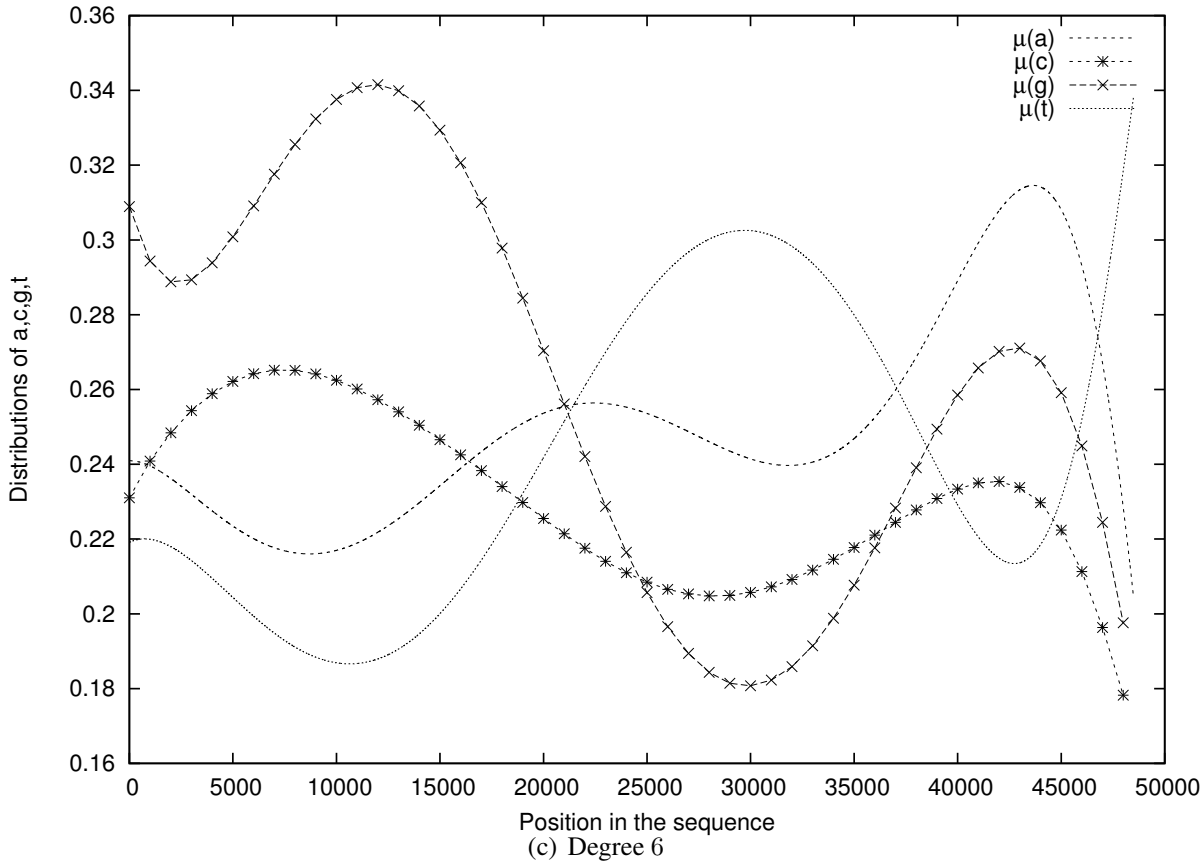
(a) Degree 2

Figure 1: Probability distributions μ of the 4 nucleotides `a`, `c`, `g` and `t` for degree 2 DMM in *Phage Lambda* genome.



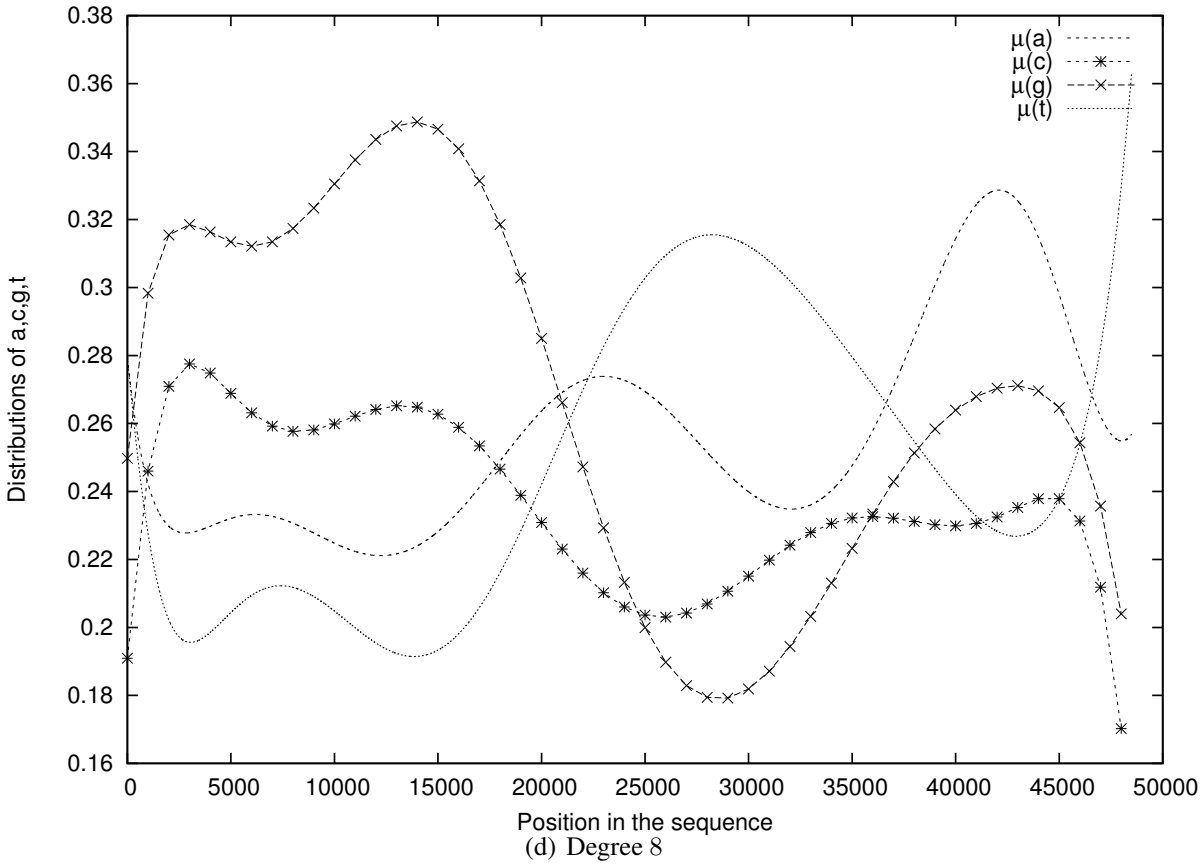
(b) Degree 4

Figure 1: Probability distributions μ of the 4 nucleotides a, c, g and t for degree 4 DMM in *Phage Lambda* genome.



(c) Degree 6

Figure 1: Probability distributions μ of the 4 nucleotides `a`, `c`, `g` and `t` for degree 6 DMM in *Phage Lambda* genome.



(d) Degree 8

Figure 1: Probability distributions μ of the 4 nucleotides a, c, g and t for degree 8 DMM in *Phage Lambda* genome.

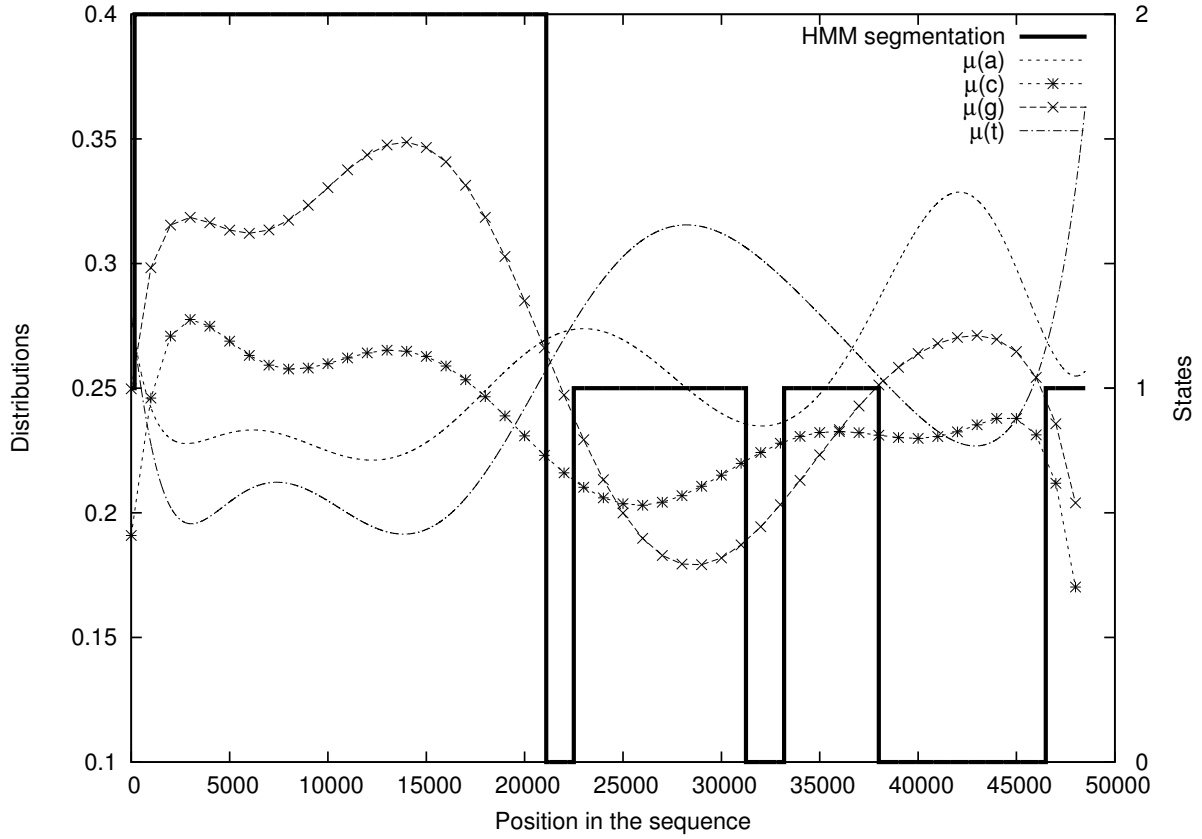
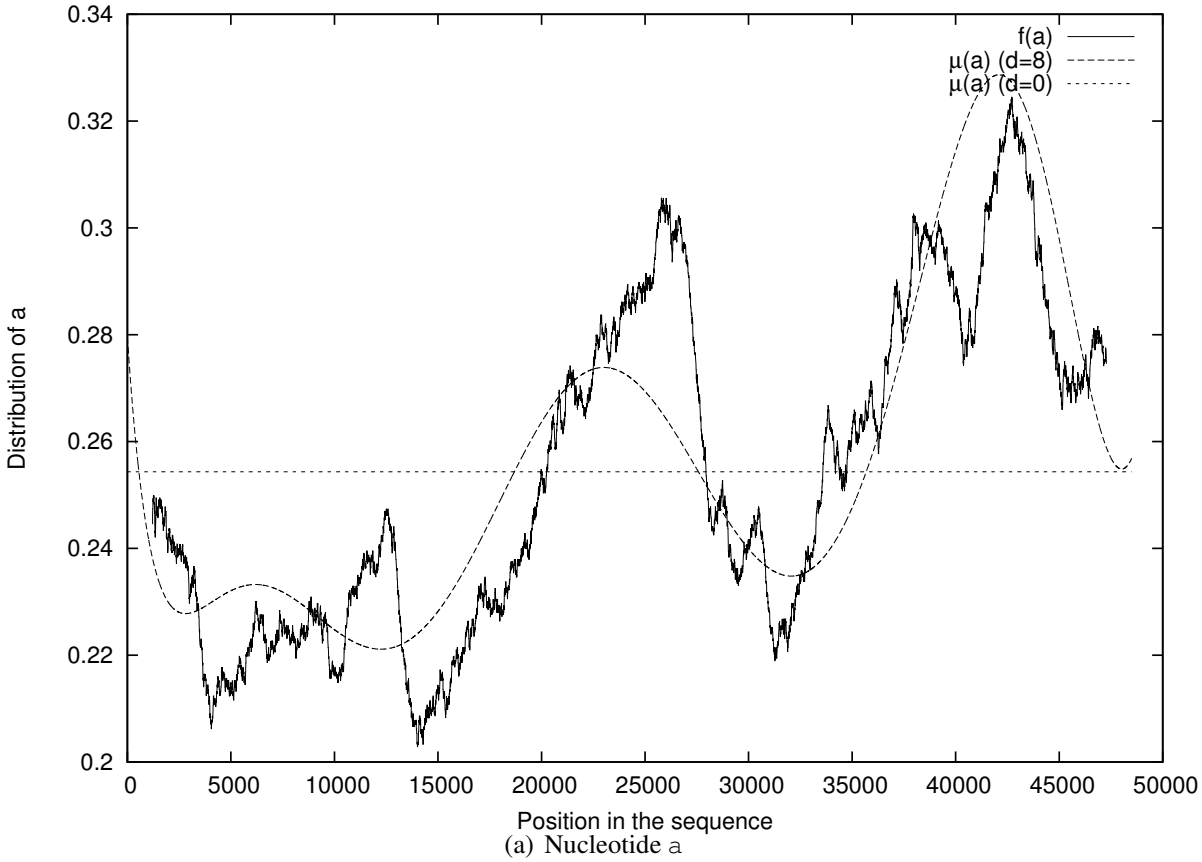
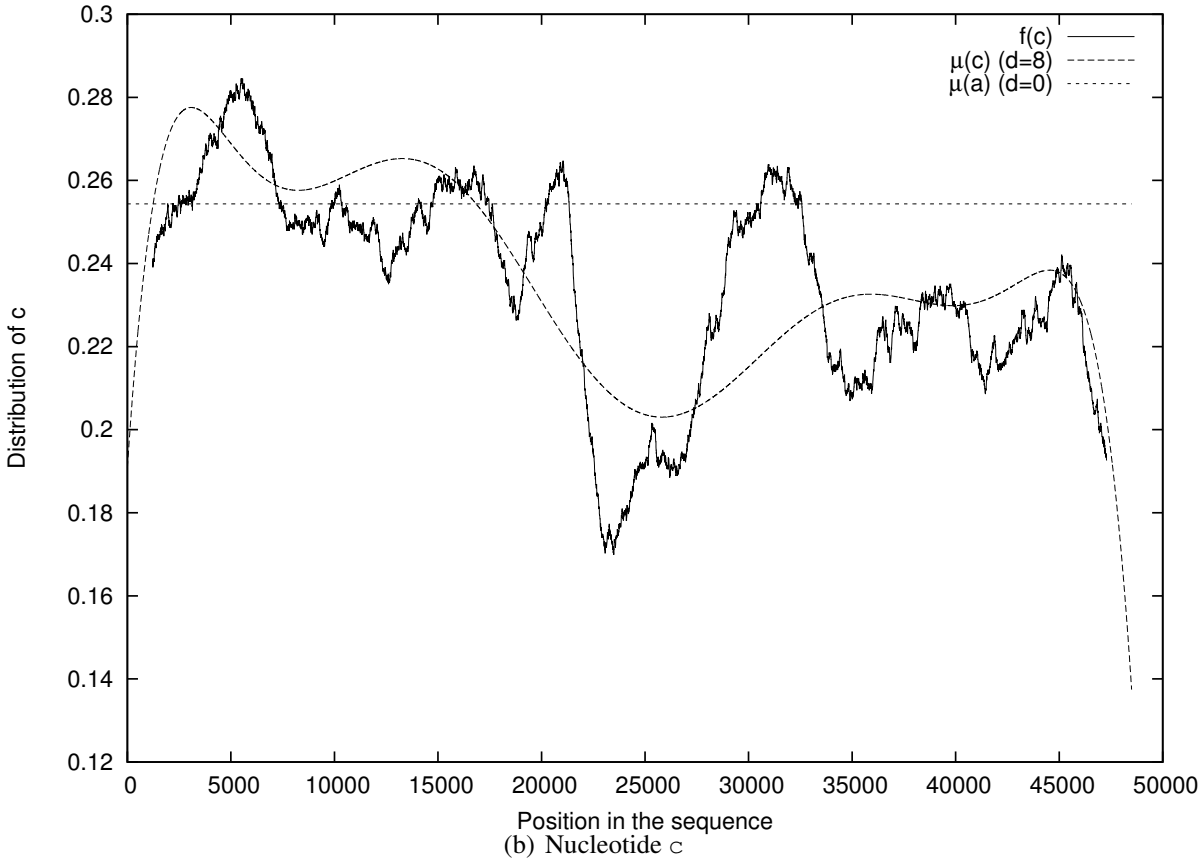


Figure 2: Probability distributions of nucleotides for a DMM of degree 8 in *Phage Lambda* genome and HMM segmentation with three hidden states (0, 1 and 2, marked on the vertical right-hand side axis).



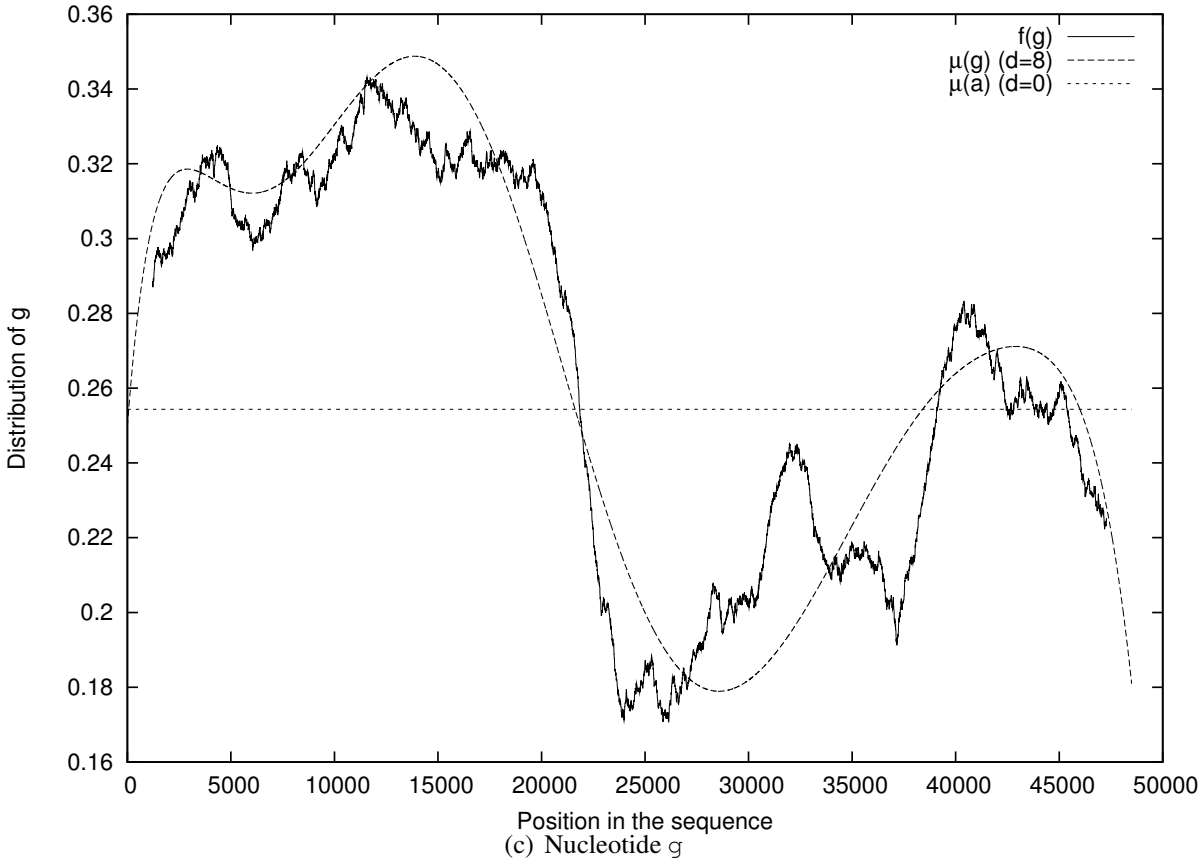
(a) Nucleotide a

Figure 3: Frequency f and probability distribution μ of a for degrees $d = 0$ and $d = 8$ DMM in *Phage Lambda* genome.



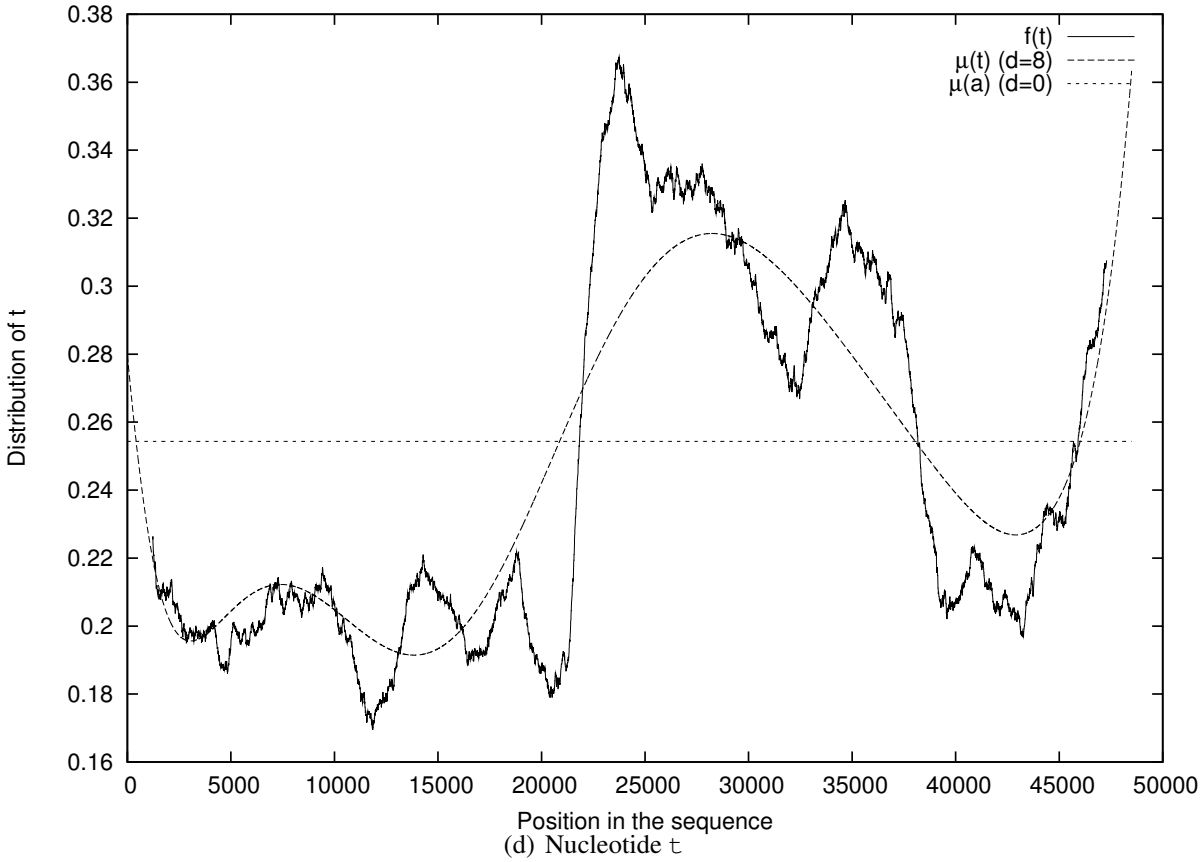
(b) Nucleotide c

Figure 3: Frequency f and probability distribution μ of c for degrees $d = 0$ and $d = 8$ DMM in *Phage Lambda* genome.



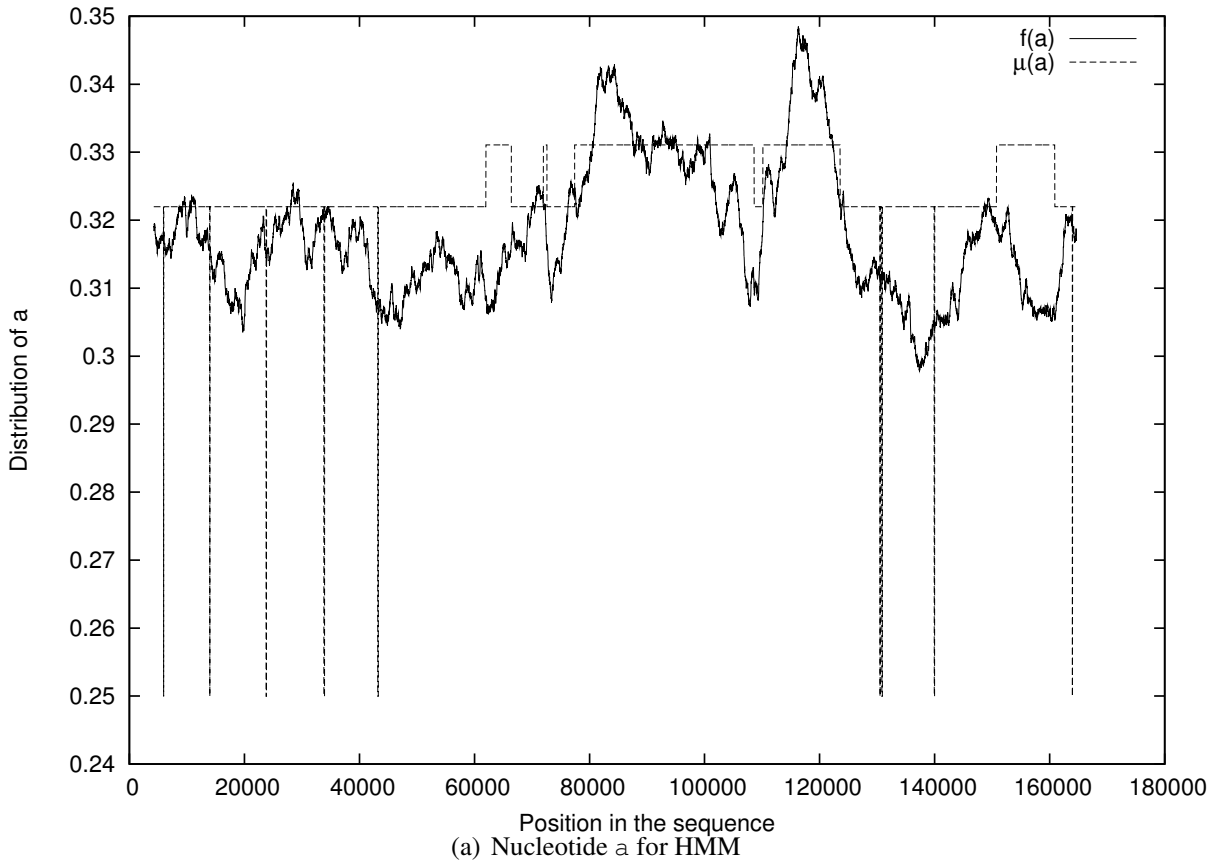
(c) Nucleotide $\mathtt{g}$

Figure 3: Frequency f and probability distribution μ of $\mathtt{g}$ for degrees $d = 0$ and $d = 8$ DMM in *Phage Lambda* genome.



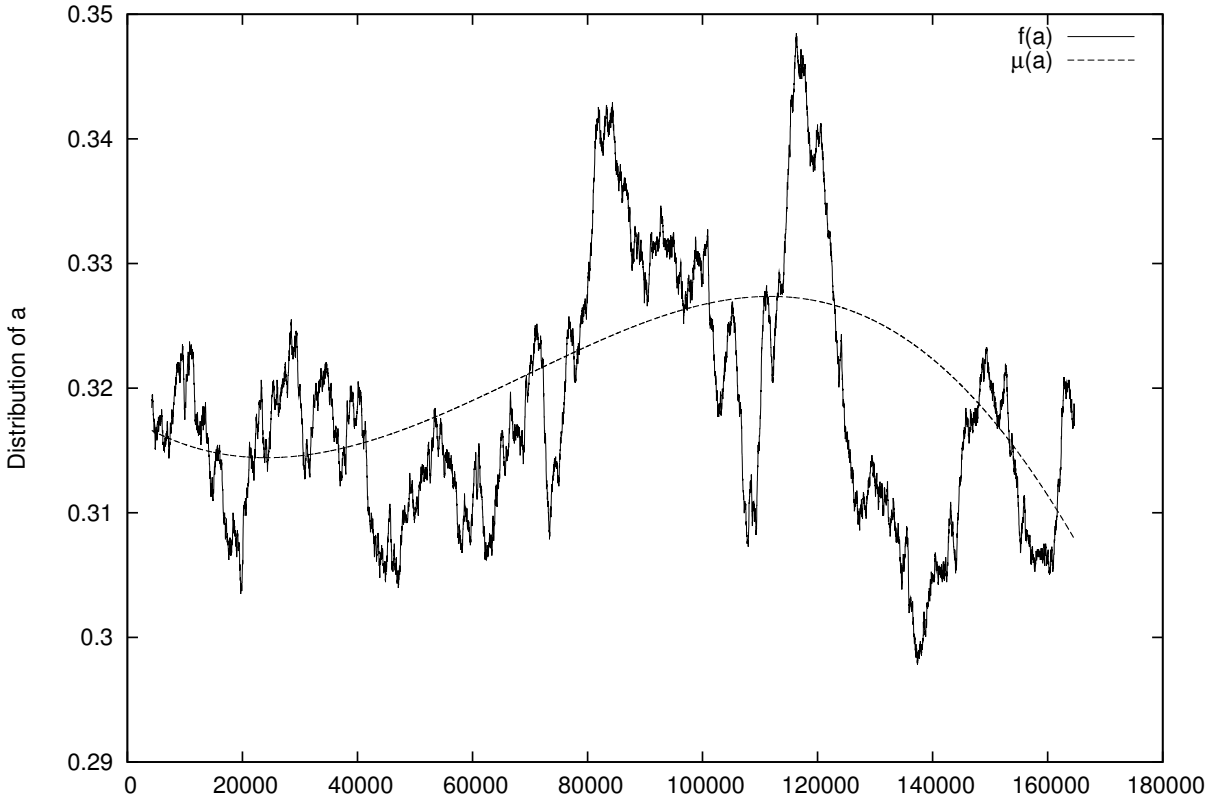
(d) Nucleotide t

Figure 3: Frequency f and probability distribution μ of t for degrees $d = 0$ and $d = 8$ DMM in *Phage Lambda* genome.



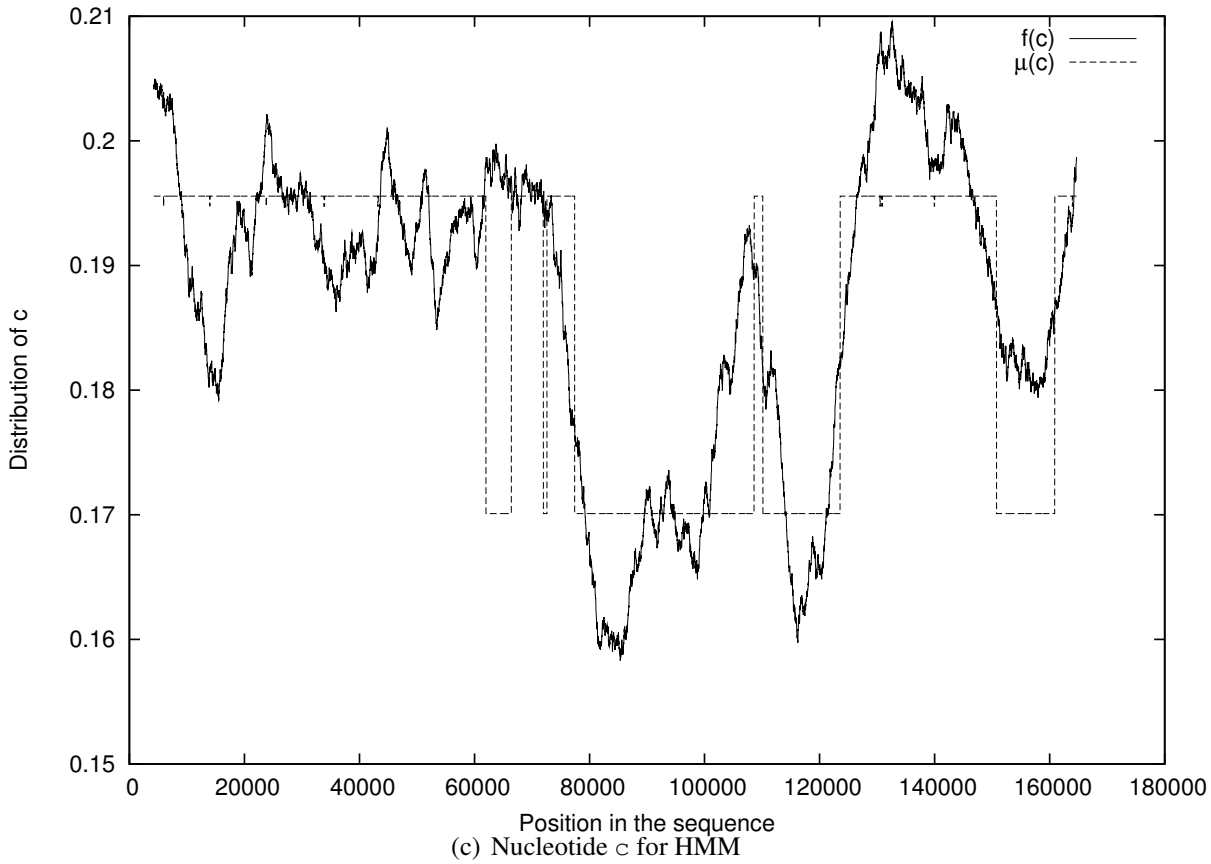
(a) Nucleotide a for HMM

Figure 4: Frequency f and probability distribution μ of a for a 3-states HMM in *Phage T4* genome.



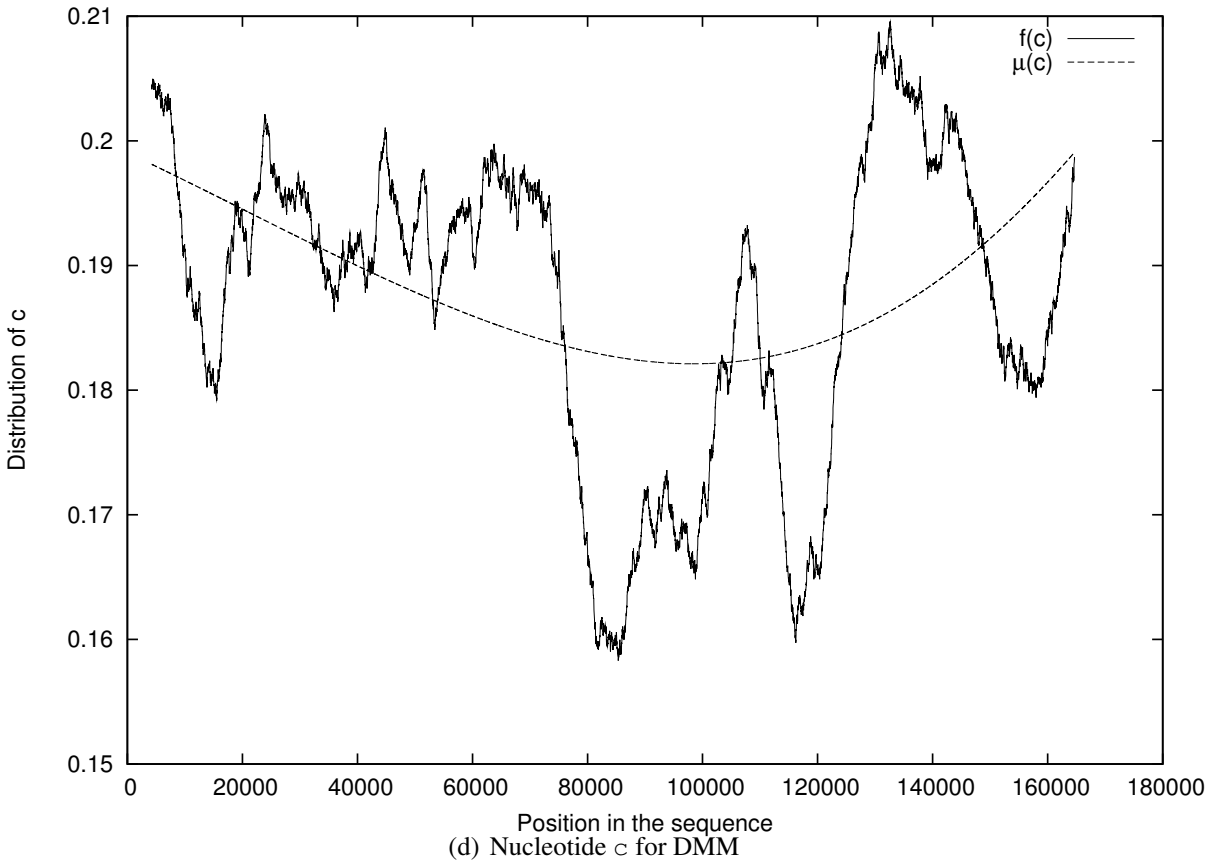
(b) Nucleotide a for DMM

Figure 4: Frequency f and probability distribution μ of a for a degree 3 DMM in *Phage T4* genome.



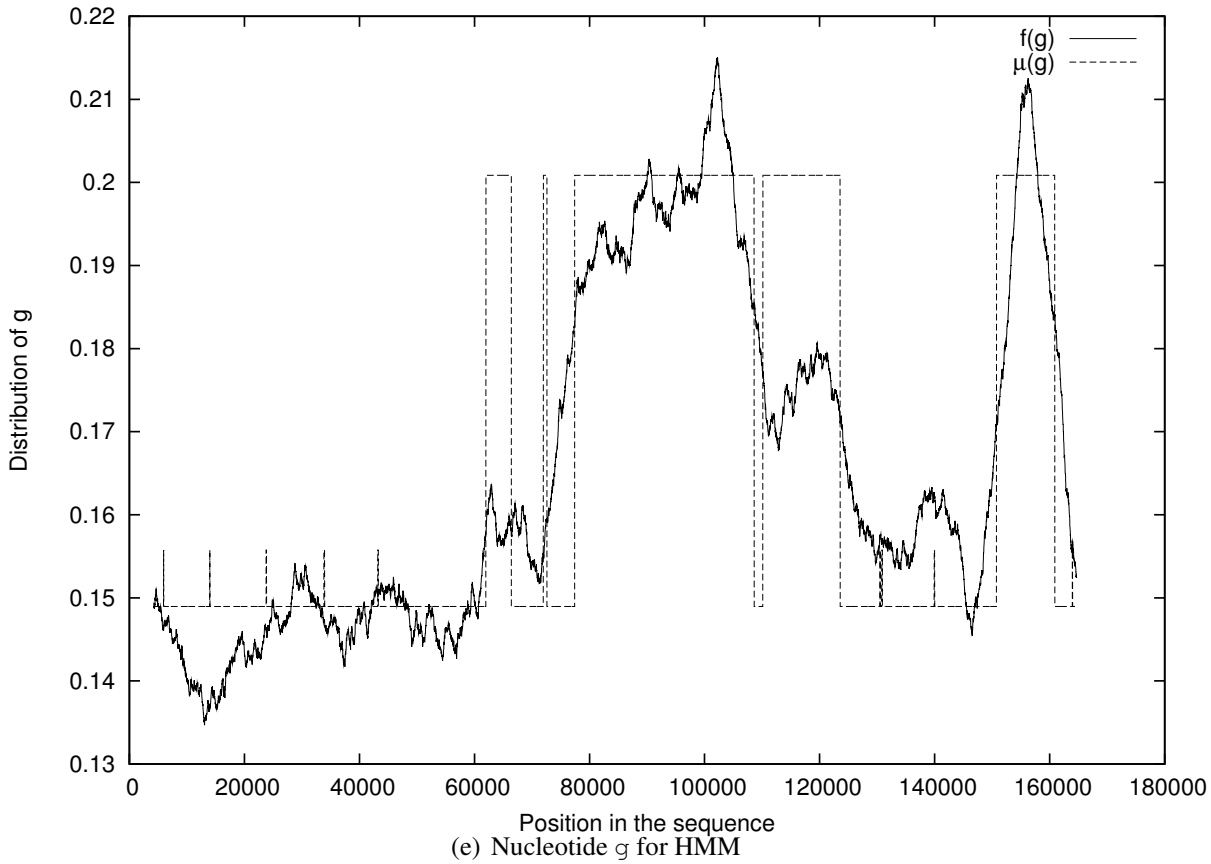
(c) Nucleotide c for HMM

Figure 4: Frequency f and probability distribution μ of c for a 3-states HMM in *Phage T4* genome.



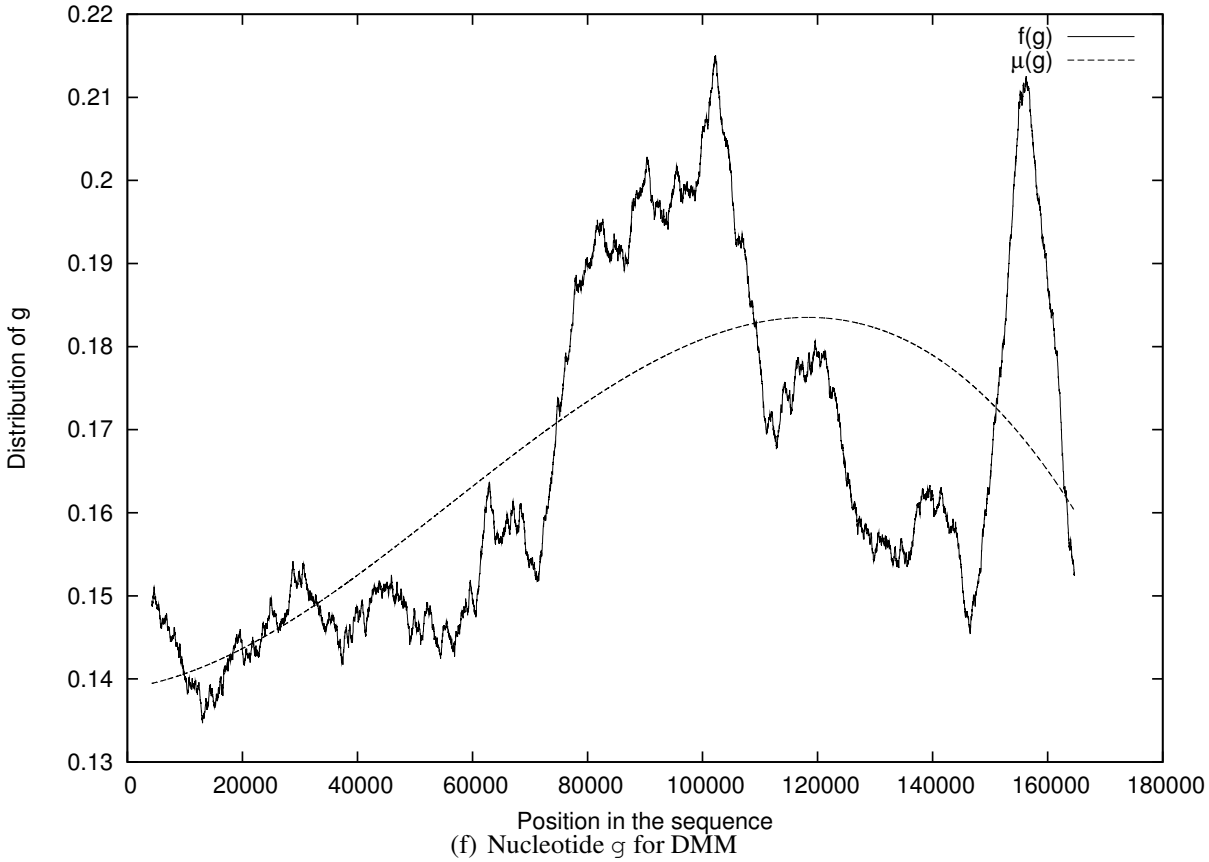
(d) Nucleotide c for DMM

Figure 4: Frequency f and probability distribution μ of c for a degree 3 DMM in *Phage T4* genome.



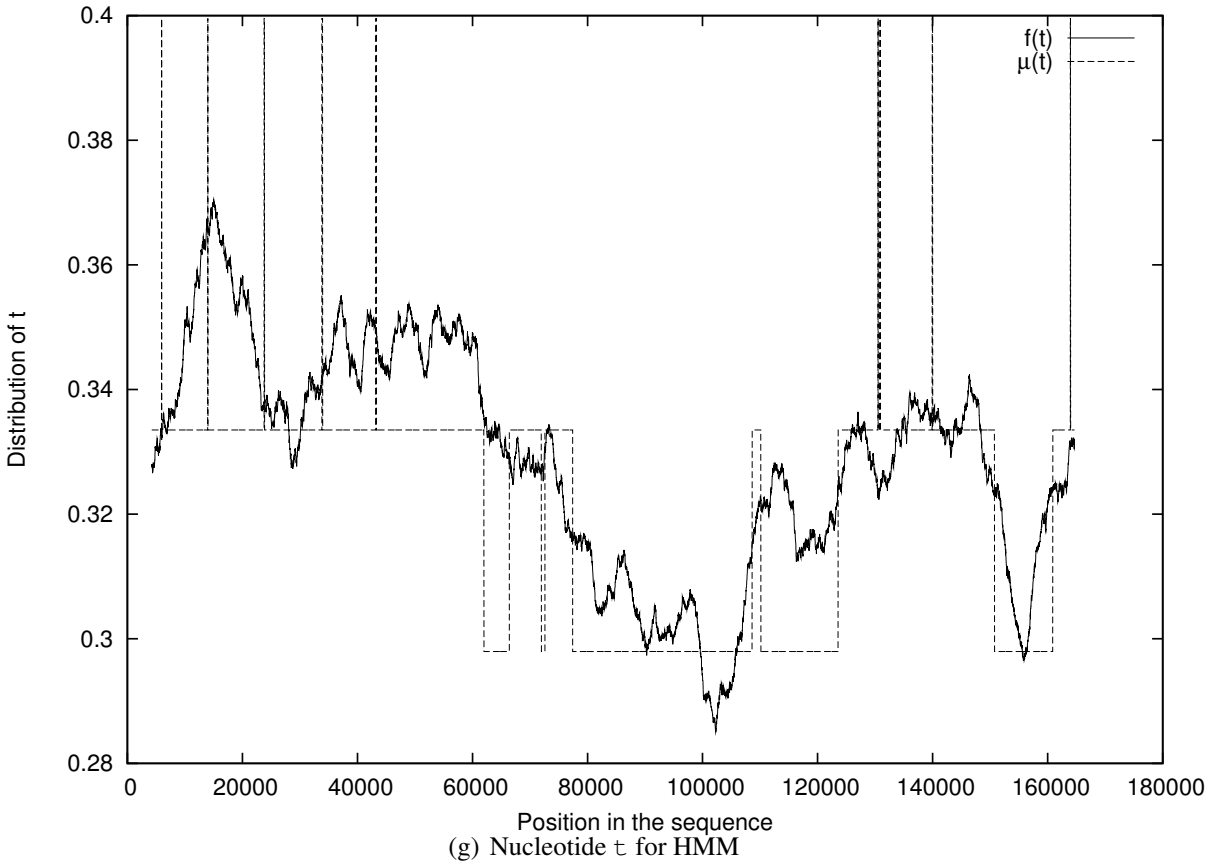
(e) Nucleotide $\mathtt{g}$ for HMM

Figure 4: Frequency f and probability distribution μ of $\mathtt{g}$ for a 3-states HMM in *Phage T4* genome.



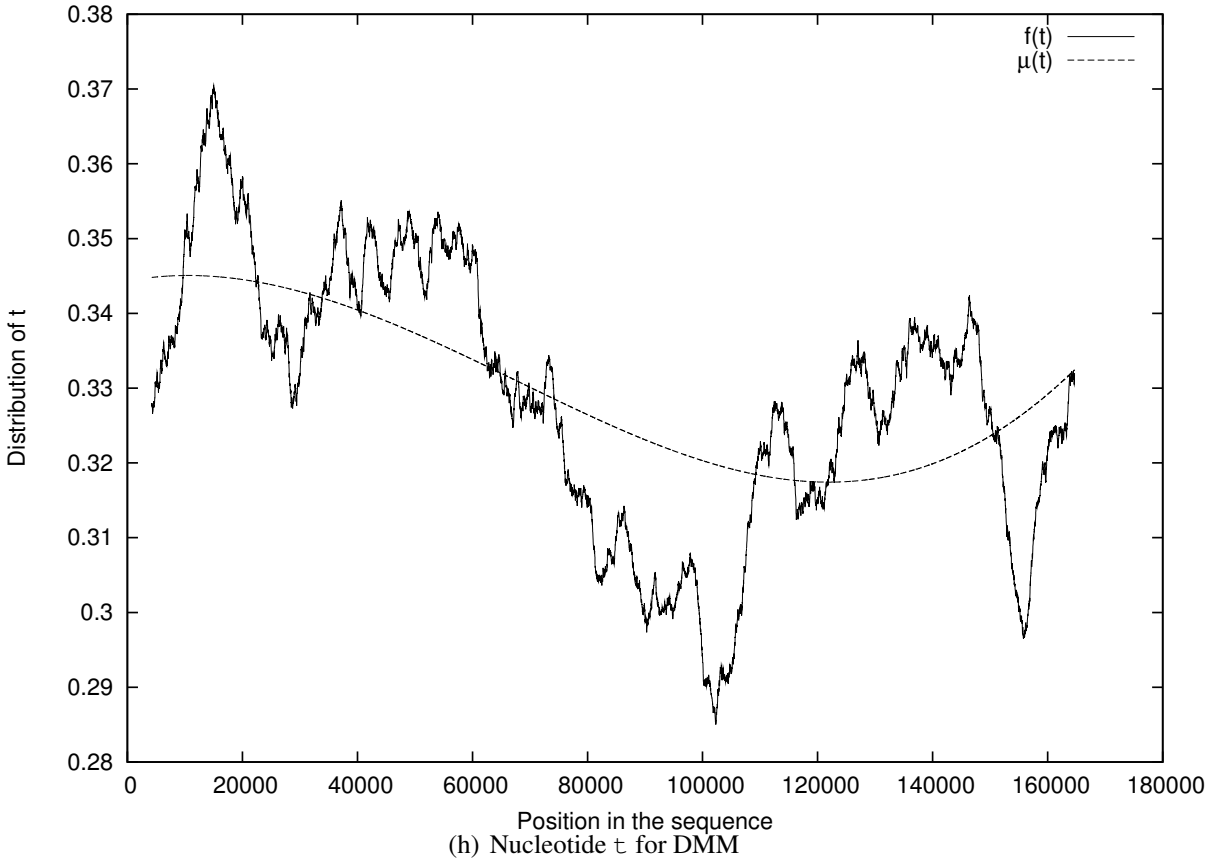
(f) Nucleotide $\mathtt{g}$ for DMM

Figure 4: Frequency f and probability distribution μ of $\mathtt{g}$ for a degree 3 DMM in *Phage T4* genome.



(g) Nucleotide t for HMM

Figure 4: Frequency f and probability distribution μ of t for a 3-states HMM in *Phage T4* genome.



(h) Nucleotide t for DMM

Figure 4: Frequency f and probability distribution μ of t for a degree 3 DMM in *Phage T4* genome.

We also compute a distance d_{df} between these evolutions and nucleotide frequencies:

$$d_{df} = \sum_{v \in \mathcal{A}} \sum_{t} (f_t(v) - \mu_t(v))^2,$$

where $f_t(v)$ is the frequency of the nucleotide v at position t and $\mu_t(v)$ the probability distribution of v at position t. In the way to avoid long computations, we do not take into account all positions t. Then we have no more than 10000 positions, uniformly distributed. It is sufficient in order to compute d_{df}. An order 1 HMM with 3 hidden states and an order 1 DMM with degree 3 have approximately the same number of parameters (in fact, this number is 42 for the HMM and 48 for the DMM). However, we already note that d_{df} is lightly smaller for the DMM: 5.865 versus 5.873. Obviously, this distance is still smaller for a degree 8 DMM ($d_{df} = 3.391$). In that sense, we show that DMMs represent a new class of flexible models for DNA sequences that can be hoped to provide better fits than HMMs in many cases.

In order to illustrate this fact in another way, we draw in Figure 5, the frequency of gc in the complete genome of *phage Lambda*. As we said in the introduction, biologists are very concerned in the gc-percent because it may induce presence of genes. They consider five families of isochores: two gc-poor families (L1 and L2) and three gc-rich families (H1, H2 and H3) (Bernardi 1993, Oliver et al. 2001). But the transition between two families is often judged to be too sudden when modeled by HMMs. DMM, with its continuous evolution, is a good way to model these transitions. For instance, in Figure 5, between the position 26000 and 32000, we observe a linear increase of the gc content that we model with a degree 1 DMM. Thus DMMs are useful for modeling of heterogeneous phenomena, in particular the linear evolution of gc content, whereas HMM would predict a constant evolution or an abrupt change.

3.2 AIC and BIC: comparisons between different models.

In order to analyze drifting Markov models, we compute AIC and BIC values of these models. First of all, we recall the definition of AIC and BIC values (Akaike Information Criterion and Bayesian Information Criterion, introduced respectively by Akaike (1974) and Schwarz (1978)):

$$\begin{aligned} AIC &= -(2L(\theta) - 2K) \\ BIC &= -(2L(\theta) - K \log n) \end{aligned}$$

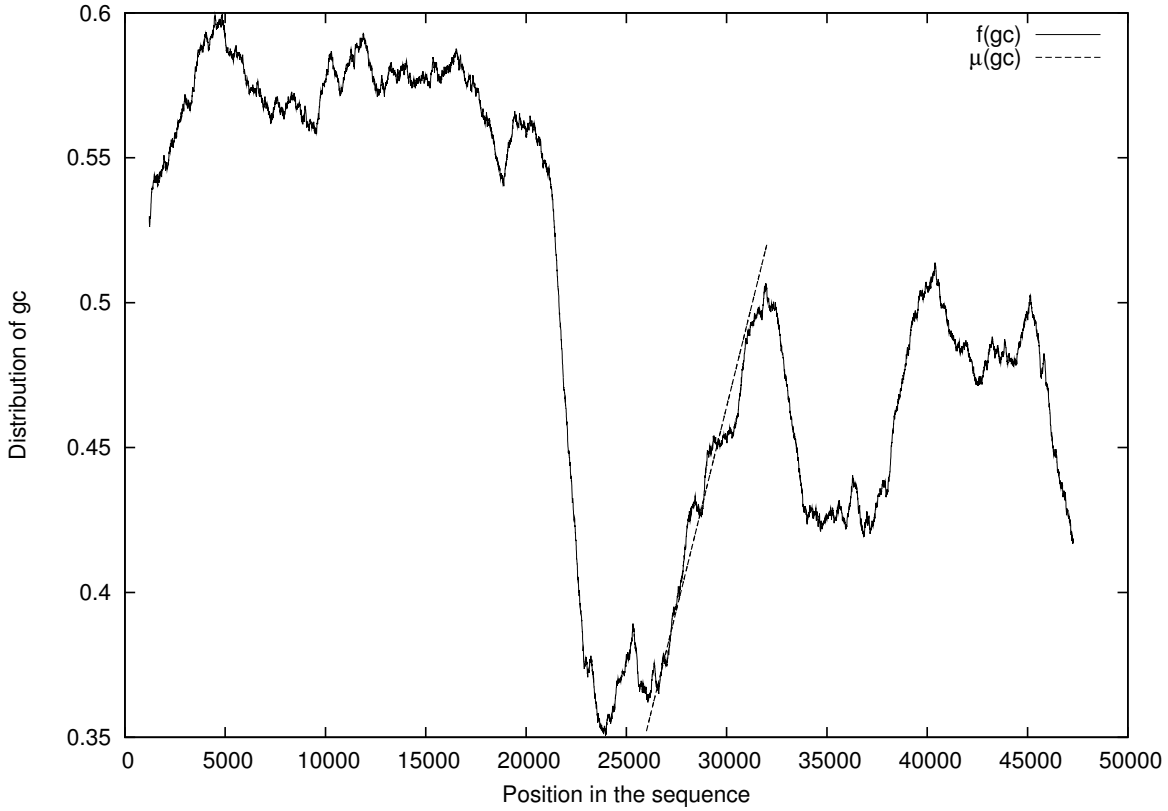


Figure 5: Frequency f and probability distribution μ of $\mathtt{gc}$ for a degree 1 DMM between the position 26000 and 32000 in *Phage Lambda* genome.

where $L(\theta)$ is the log-likelihood of the model, K the number of parameters and n the sample size. The model which has the smallest AIC or BIC is considered as the "best" model according to this criterion.

AIC and BIC are usually built by adding a penalization to the log-likelihood evaluated at the maximum of likelihood. Although least squares estimator is the only disposal, we use here adapted criteria obtained by a penalization of the log-likelihood taken in this estimator. This can be justified by the fact that for Markov chain the mean-square estimation (based on a formula similar to our formula (1)) is asymptotically equivalent to the maximum of likelihood one.

In Table 4, we compute BIC for order 0, 1, 2 and 3 a DMM of degrees from 0 to 5, estimated by the point by point method. These results have been obtained on the *Haemophilus influenzae* complete genome (see Fleischmann et al. (1995)).

Table 4: BIC of drifting Markov models on *Haemophilus influenzae*.

Degree	0	1	2	3	4	5
Order 0	4970473	4970494	4969534	4969471	4969472	**4969358**
Order 1	4907845	4907947	**4907011**	4907051	4907108	4907117
Order 2	4892907	4893442	4892907	**4890996**	4893807	4894224
Order 3	**4868040**	4870422	4871721	4874079	4876395	4878650

Whereas AIC prefers models with a lot of parameters (results are not presented here, but AIC-values generally decrease with order and degree), BIC prefers models with a small number of parameters. That is why a DMMs with high degrees are partially ignored by BIC. Indeed, for an order k DMM of degree d, the number of parameters K is equal to $(d+1)|\mathcal{A}|^k(|\mathcal{A}|-1)$. You can choose to select order and degree of a DMM with BIC, but higher-dimensional parameterization of DMM provides better fits to the real sequence, as you can see in the precedent section.

Moreover, in order to compare DMMs with other currents methods, the main way is to see in Figures 3(a), 3(b), 3(c) and 3(d) that variations of nucleotides are continuous whatever the position. According to BIC, DMMs are better models than classical Markov models (whatever the order and the degree), but HMMs are better models than DMMs. As DMM can be adjusted to a whole sequence, DMM can be seen as competitive model to HMM. However note that a perspective of this work is to introduce DMMs in HMMs: a DMM could be a hidden state of an HMM. Even if this difference of BIC is not very large, the essential thing to remember is that we provide the first models including the possibility of a continuous variation of the transition matrix. Combined to the quality of HMM, DMM provides powerful tools for sequences analysis.

3.3 Replication origin

An example of application of our new models is the search of replication origins on the bacteria. This application draws its inspiration from the program ORILOC (Lobry 2000), which has been developed for the prediction of bacterial replications origins. DNA replication is the process of copying a double-stranded DNA strand in a cell, prior to cell division. The two resulting double strands are identical, and each of them consists of one original and one newly synthesized strand. The replication origin is a unique DNA sequence at which DNA replication is initiated. DNA replication may proceed from this point bidirectionally or unidirectionally. Based on the compositional asymmetries between the leading and the lagging strand of replication, the program performs a DNA walk (see Lobry (1999)) to obtain the position of the replication origin. A curve is drawn by this program and the peak of this curve corresponds to the replication origin. The values allowing to draw the curve are computed as follows. The first value is 0, and during the walk along the DNA sequence, ORILOC adds 1 each time letter $\mathtt{g}$ is found and subtracts 1 each time letter $\mathtt{c}$ is found. Thus ORILOC does not rely on a probabilistic model, it draws a curve by running along the real sequence.

We use the same properties of asymmetries in bacterial genomes to perform a detection of the replication origins based on DMMs. Indeed, thanks to the computation of probability distributions of nucleotides at each position t in the sequence, we draw a curve similar to ORILOC. The values of our curve are computed as follows. The first value is 0, and at each position in the sequence, we add the probability of letter $\mathtt{g}$ and subtract the probability of letter $\mathtt{c}$.

This work was done on the complete genome of *Chlamydia trachomatis* (see Stephens et al. (1998)). Note on Figure 6 the great similarity between the curve obtained by the software ORILOC and the one obtained by DMMs. Note also that our curve is softer than the one of ORILOC because the aim of DMM is to model soft transitions. Although search of replication origins is a break-point detection problem, our method works in the sense that it offers to biologists a window which permits to find the replication origin "in vivo". Then soft transitions do not prevent us to locate the origin of replication. The advantage of our method is to be able to compute analytically a maximum.

3.4 Rare words

A second and important example of application of DMMs is the search of rare words in DNA sequences. Many DNA sequence analysis are based on the distribution of the occurrences of patterns having some special biological function.

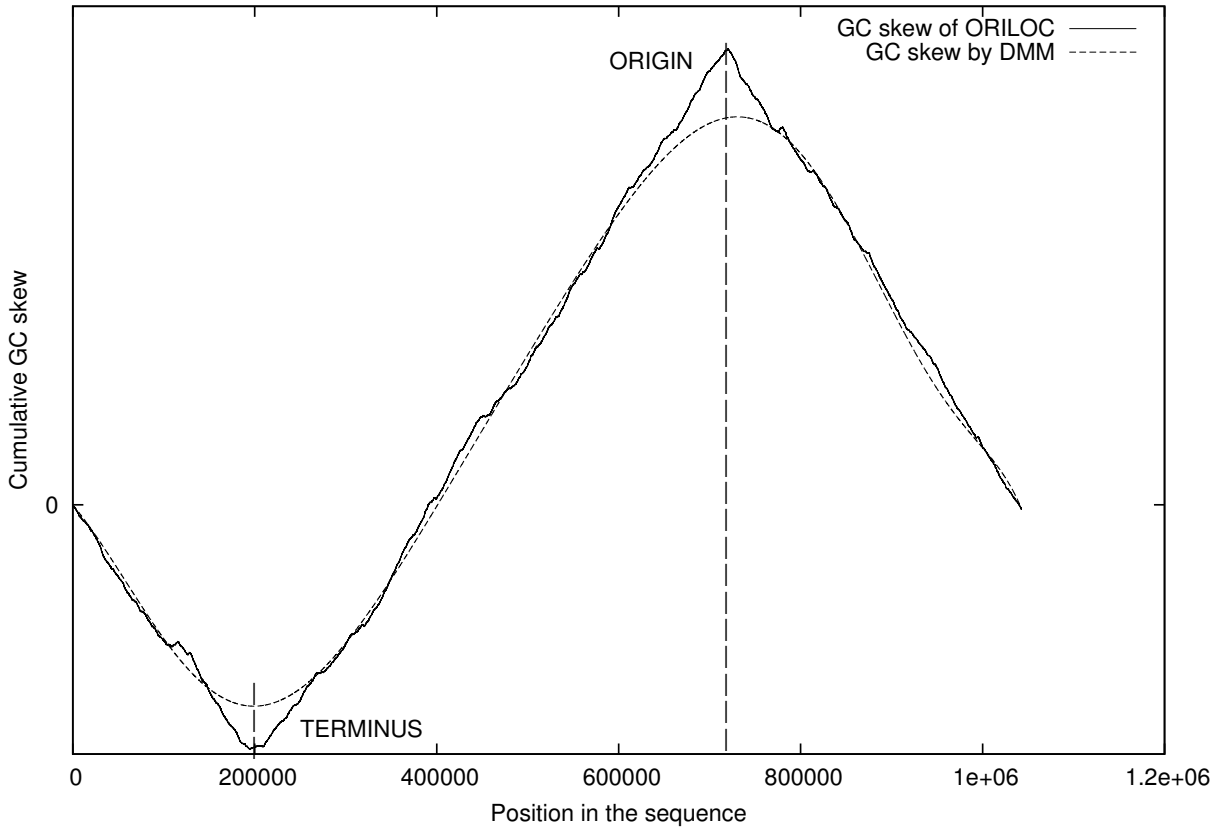


Figure 6: Search of replication origin in *C. trachomatis*. Comparison between ORILOC (Lobry 2000) and Drifting Markov Models (DMM).

An important problem is to determine the statistical significance of a word frequency in a DNA sequence. Nicodème et al. (2002) discuss this relevance of finding over- or under-represented words. The naive idea is the following: a word may have a significant low frequency in a DNA sequence because it disrupts replication or gene expression, whereas a significantly frequent word may have a fundamental activity with regard to genome stability. Well-known examples of words with exceptional frequencies in DNA sequences are biological palindromes corresponding to restriction sites avoided for instance in *E. coli* (Karlin et al. 1992), the Cross-over Hotspot Instigator sites in several bacteria, in *E. coli* for example (Smith et al. 1981, El Karoui et al. 1999), and uptake sequences (Smith et al. 1999) or polyadenylation signals (van Helden et al. 2000). The most popular approach consist in fitting a Markov model on the sequence and computing the p-value which is $\mathbb{P}(N > N_{obs})$ for an over-represented word or $\mathbb{P}(N < N_{obs})$ for an under-represented word, where N is the random variable of the number of occurrences of the studied word and N_{obs} the number of observed occurrences. We define the pattern statistic associated to any number N_{obs} by:

$$S = \begin{cases} -\log_{10} \mathbb{P}(N > N_{obs}) & \text{if } N \geq \mathbb{E}(N) \\ +\log_{10} \mathbb{P}(N < N_{obs}) & \text{if } N < \mathbb{E}(N) \end{cases} .$$

This way, a pattern has a positive statistic if it is seen more than expected, a negative statistic if seen less than expected, in both cases, the corresponding p-value is given (in log scale) by the magnitude of the statistic. See Nuel (2006) for a review of the methods available to compute pattern statistics on text generated by a Markov source.

As these probabilities are computed under a model, small p-value can be provided for some words without biological interest if the model is not reliable. That is why it is preferable to rely on a background model the most possible close to the real sequence. DMMs provide it. It always will be more convincing to obtain p-values for the most realistic models. In that way, considering a DMM for searching rare words in sequences seems to be a better approach than using Markov models (see on Figures 3(a), 3(b), 3(c) and 3(d) that DMM offers a model closer to the reality than Markov model).

Numerical complexities appear when we want to compute exact p-value of inhomogeneous Markov models but a new approach proposed by Nuel (2004), using finite Markov chain imbedding (FMCI, see Lou (1996)), provides solutions to this problem. We refer to Nuel & Prum (2007) for a detailed description of this method.

Table 5: Classification of words of size 5 of *Phage lambda* complete genome, for different models according to their pattern statistic S. N_{obs} is the observed number of occurrences of the word. Exp means Expected value. We only give the five first under-represented words and the five first over-represented words.

MM				HMM 3 states				DMM degree 1			
Words	N_{obs}	Exp	S	Words	N_{obs}	Exp	S	Words	N_{obs}	Exp	S
aattg	32	88.22	−11.41	aattg	32	83.38	−10.07	aattg	32	86.53	−10.94
ttggg	20	65.12	−10.33	acttg	13	47.59	−8.57	ttgga	21	64.94	−9.76
ttgga	21	66.70	−10.29	tctag	2	24.60	−8.19	ttggg	20	62.94	−9.66
acttg	13	50.74	−9.59	ttgga	21	59.47	−8.15	acttg	13	50.27	−9.44
taggg	3	29.60	−9.21	tcgag	9	39.01	−8.11	tcgag	9	40.69	−8.68
gccgg	114	53.97	12.13	gctgg	127	65.44	14.23	gctgg	127	64.80	11.77
ctgaa	124	61.02	12.16	ctgaa	124	61.34	14.90	ctgaa	124	60.85	12.21
tccgg	100	39.98	15.08	ccgga	112	44.00	20.58	tccgg	100	38.81	16.18
ccgga	112	43.11	17.93	tccgg	100	36.50	20.65	ccgga	112	43.57	18.10
gcaga	141	57.51	20.20	gcaga	141	58.35	22.66	gcaga	141	57.59	20.31

We just give here one example of search of rare words. We choose the most popular word in this domain, the Chi site of *Escherichia coli K12* (see Blattner et al. (1997)). We consider the complete sequence of the bacteria where the Chi site `gctggtgg` appears 499 times. As can be seen in Table 6, the Chi site was expected to appear 70.10 times by an order 1 DMM of degree 0 and 175.31 times by an order 2 DMM of order 8. In a more realistic model such as DMM, Chi sites are more expected than in other models. As already said, we cannot compare p-value of different models between them. But we could compare the different classification provided by the different models. Which classification do you prefer? That one given by HMM and its segmentation or that one given by DMM and its soft evolution? Obviously, it is more reliable to consider p-values in the model which provide a better fit to the data even if it is higher-dimensional parameterized. Thus, polynomial DMMs are very useful for the search of rare words in DNA sequences. In Table 5, we give classification of words of size 5, for classical Markov model, 3-states HMM and degree 1 DMM, at order 1.

Table 6: Pattern statistic S (log p-value) of the over-represented word `gctggtgg` for DMMs of different orders and degrees: the Chi of *E. coli* which appears 499 times in the sequence. Note that a DMM of degree 0 corresponds to a classical Markov model.

Order	Degree	Expected value	S
1	0	70.10	240.814
1	1	70.26	240.398
1	2	71.88	238.766
1	3	71.87	238.774
1	8	71.94	238.605
2	0	173.84	88.902
2	1	174.03	88.747
2	2	175.16	87.837
2	3	175.10	87.881
2	8	175.31	87.717

4 Discussion and conclusion

We introduce a new class of inhomogeneous Markov models, the drifting Markov models. These new models allow the transition matrix to vary along the sequence.

Notwithstanding the fact that classical Markov models are homogeneous, hidden Markov models cannot model every heterogeneity structures. Heterogeneity of sequences encourages us to consider more flexible models such as drifting Markov models and the continuous variation of their transition matrix. An important illustration of these models concerns the gc-content of a DNA sequence. It is commonly accepted that a high gc-content may induce presence of genes (Zoubak et al. 1996). Since they provide a soft evolution and a different transition matrix at each position in the sequence, DMM provides a better fit to the gc-content than HMM with its sudden changes of state. Other applications such as the search of replication origins and especially the search of rare words are very relevant examples of the possibilities of DMMs. We conclude that DMMs are convenient tools for the statistical analysis of sequences. They provide detailed description of the sequence and can be used for structural analysis or direct biological applications. Moreover, it would be interesting not to limit our studies to polynomial drift. Future prospects are to fit new models with co-variables such as the gc-content, the degree of hydrophobicity or an indicator of the protein structure (α-helix, β-sheet...).

Appendix A: Estimation by maximum of likelihood

We give here an example of the systems we would need to solve to provide estimation by maximum likelihood. It corresponds to one of the 4 systems of 6 equations with 6 variables obtained for an order 1 DMM of degree 1 for the nucleotide alphabet $\mathcal{A} = \{a, c, g, t\}$. Knowing that n is very high (as the length of the DNA sequence) and that all the parameters $\Pi_0(u, v)$ and $\Pi_1(u, v)$ are in the denominator of each equality, you can note the numerical complexity which precludes the use of this natural method. Obviously, complexity still is a problem for DMM of higher order and degree.

$$\left\{\begin{array}{rcl} \displaystyle\sum_{t=1}^{n}\left(1-\frac{t}{n}\right)\frac{\mathbf{1}_{\{X_{t-1}=u,X_t=v\}}}{\Pi_{\frac{t}{n}}(u,v)} & = & \displaystyle\sum_{t=1}^{n}\left(1-\frac{t}{n}\right)\frac{\mathbf{1}_{\{X_{t-1}=u,X_t=u\}}}{\Pi_{\frac{t}{n}}(u,u)} \\ \displaystyle\sum_{t=1}^{n}\left(\frac{t}{n}\right)\frac{\mathbf{1}_{\{X_{t-1}=u,X_t=v\}}}{\Pi_{\frac{t}{n}}(u,v)} & = & \displaystyle\sum_{t=1}^{n}\left(\frac{t}{n}\right)\frac{\mathbf{1}_{\{X_{t-1}=u,X_t=u\}}}{\Pi_{\frac{t}{n}}(u,u)} \end{array}\right. \iff$$

$$
\left\{
\begin{array}{l}
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=c\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,c)+\left(\frac{t}{n}\right)\Pi_1(a,c)} = \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=a\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\frac{t}{n}(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=g\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,g)+\left(\frac{t}{n}\right)\Pi_1(a,g)} = \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=a\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\frac{t}{n}(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=t\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,t)+\left(\frac{t}{n}\right)\Pi_1(a,t)} = \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=a\}}\left(\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\frac{t}{n}(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=c\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,c)+\left(\frac{t}{n}\right)\Pi_1(a,c)} = \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=a\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\frac{t}{n}(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=g\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,g)+\left(\frac{t}{n}\right)\Pi_1(a,g)} = \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=a\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\frac{t}{n}(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))} \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=t\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)\Pi_0(a,t)+\left(\frac{t}{n}\right)\Pi_1(a,t)} = \\
\displaystyle\sum_{t=1}^{n} \frac{\mathbf{1}_{\{X_{t-1}=a,X_t=a\}}\left(1-\frac{t}{n}\right)}{\left(1-\frac{t}{n}\right)(1-\Pi_0(a,c)-\Pi_0(a,g)-\Pi_0(a,t))+\frac{t}{n}(1-\Pi_1(a,c)-\Pi_1(a,g)-\Pi_1(a,t))}
\end{array}
\right.
$$

Appendix B: Consistence of the estimators

In order to show consistence of our estimators, we give here one example of the results of some simulations. In this example, we estimate an order 1 degree 2 drifting Markov model on the *phage Lambda* complete genome (see Wu & Taylor (1971)) and we consider this model as the true model. Then, we simulate 10 sequences with this model and estimate a mean model on all these sequences. Both true and estimated models are shown in Table 7. You can see the great similarity between matrices.

Table 7: Comparison between transition matrices of true and estimated models

Matrix	True model				Estimated model			
Π_0	0.2637	0.2753	0.2623	0.1986	0.2634	0.2811	0.2596	0.1958
	0.2559	0.2889	0.3090	0.1462	0.2573	0.2839	0.3114	0.1472
	0.2107	0.3712	0.2582	0.1600	0.2077	0.3656	0.2662	0.1606
	0.1201	0.4873	0.2011	0.1914	0.1240	0.4884	0.1967	0.1910
$\Pi_{0.5}$	0.2931	0.2156	0.2044	0.2870	0.2913	0.2145	0.2065	0.2878
	0.2424	0.2445	0.2760	0.2371	0.2422	0.2469	0.2733	0.2376
	0.2976	0.2533	0.2127	0.2364	0.2954	0.2558	0.2127	0.2361
	0.1856	0.2891	0.2301	0.2953	0.1838	0.2883	0.2320	0.2959
Π_1	0.3480	0.2017	0.1833	0.2670	0.3559	0.1972	0.1781	0.2689
	0.3061	0.2032	0.2686	0.2221	0.3047	0.2034	0.2682	0.2238
	0.3076	0.2457	0.2037	0.2433	0.3047	0.2460	0.2017	0.2476
	0.2073	0.2997	0.2132	0.2798	0.2092	0.3044	0.2119	0.2744

References

Akaike, H. (1974), 'A new look at the statistical identification model', *IEEE Transactions on Automatic Control* **19**, 716–723.

Almagor, H. (1983), 'A Markov analysis of DNA sequences', *J.Theor. Biol.* **104**, 633–645.

Bernardi, G. (1993), 'The vertebrate Genome: Isochores and Evolution', *Mol. Biol. Evol.* **10**, 186–204.

Blaisdell, B. (1985), 'Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucaryotic nuclear DNA sequences both protein-coding and noncoding', *J. Mol. Evol.* **21**, 278–288.

Blattner, F., Plunkett, G., Bloch, C., Perna, N., Burland, V., Riley, M., Collado-Vides, J., Glasner, J., Rode, C., G.F., M., Gregor, J., N.W., D., Kirkpatrick, H., Goeden, M., Rose, D., Mau, B. & Shao, Y. (1997), 'The complete genome sequence of escherichia coli k-12', *Science* **277**, 1453–74.

Churchill, G. (1989), 'Stochastic models for heterogeneous DNA sequences', *Bull. Math. Biol.* **268**, 8–14.

El Karoui, M., Biaudet, V., Schbath, S. & Gruss, A. (1999), 'Characteristics of Chi distribution on different bacterial genomes', *Res. Microbiol.* **150**, 579–587.

Fickett, J. W., Torney, D. C. & Wolf, D. R. (1992), 'Base compositional Structure of Genomes', *Genomics* **13**, 1056–1064.

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J., Scott, J., Shirley, R., Liu, L., Glodek, A., Kelley, J., Weidman, J., Phillips, C., Spriggs, T., Hedblom, E., Cotton, M., Utterback, T., Hanna, M., Nguyen, D., Saudek, D., Brandon, R., Fine, L., Fritchman, J., Fuhrmann, J., Geoghagen, N., Gnehm, C., McDonald, L., Small, K., Fraser, C., Smith, H. & Venter, J. (1995), 'Whole-genome random sequencing and assembly of haemophilus influenzae rd', *Science* **269**, 496–512.

Gelfand, M., Kozhukhin, C. & P.A., P. (1992), 'Extendable words in nucleotide sequences', *Bioinformatics* **8**, 129–135.

Karlin, S., Burge, C. & Campbell, A. (1992), 'Statistical analyses of counts and distributions of restriction sites in dna sequences', *Nucl. Acids Res.* **20**, 1363–1370.

Krogh, A., Mian, L. & Haussler, D. (1994), 'A hidden Markov model that finds genes in *escherichia coli* DNA', *Nucl. Acids Res.* **22**, 4768–4778.

Lobry, J. (1999), 'Genomic landscapes', *Microbiol. Today* **26**, 164–165.

Lobry, J. (2000), 'Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes', *Bioinformatics* **16**, 560–561.

Lou, W. (1996), 'On runs and longest run tests: A method of finite markov chain imbedding', *J. Am. Statis. Assoc.* **91**, 373–380.

Miele, V., Bourguignon, P., Robelin, D., Nuel, G. & Richard, H. (2005), 'seq++ : analyzing biological sequences with a range of Markov-related models', *Bioinformatics* **21**, 2783–2784.

Miller, E., Kutter, E., Mosig, G., Arisaka, F., Kunisawa, T. & Rger, W. (2003), 'Bacteriophage T4 genome', *Microbiology and molecular biology reviews* **67**(1), 86–156.

Muri, F. (1997), Comparaisons d'algorithmes d'identification de chaînes de Markov cachées et application à la détection de régions homogènes dans les séquences d'ADN, PhD thesis, Université Paris V. 156–194.

Nicodème, P., Doerks, T. & Vingron, M. (2002), 'Proteome analysis based on motif statistics', *Bioinformatics* **18**(Suppl. 2), 5161–5171.

Nicolas, P., Bize, L., Muri, F., Hoebeke, M., Rodolphe, F., Ehrlich, S., Prum, B. & Bessières, P. (2002), 'Mining *bascillus subtilis* chromosome heterogeneity using hidden Markov models', *Nucl. Acids Res.* **30**, 1418–1426.

Nuel, G. (2001), Grandes déviations et chaînes de Markov pour l'étude des occurrences de mots dans les séquences biologiques, PhD thesis, Université d'Evry Val d'Essonne.

Nuel, G. (2004), 'Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics', *Journal of Computational Biology* **11**, 1023–1033.

Nuel, G. (2006), 'Numerical Solutions for Patterns Statistics on Markov Chains', *Statistical Applications in Genetics and Molecular Biology* **5**.

Nuel, G. & Prum, B. (2007), *Analyse statistique des séquences biologiques: modélisation markovienne, alignements et motifs*, Hermes.

Oliver, J., Bernaola-Galván, P., Carpena, P. & Román-Roldán, R. (2001), 'Isochore chromosome maps of eukaryotic genomes', *Gene* **276**, 47–56.

Phillips, G., Arnold, J. & Ivarie, R. (1987), 'The effect of codon usage on the oligonucleotide composition of the e. coli genome and identification of over- and underrepresented sequences by Markov chain analysis', *Nucl. Acids Res.* **15**, 2627–2638.

Reinert, G. & Schbath, S. (1998), 'Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains', *J. Comput. Biol.* **5**, 223–253.

Schbath, S., Prum, B. & de Turckheim, E. (1995), 'Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences', *Journal of Computational Biology* **2**, 417–437.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Ann. Statist.* **6**, 461–464.

Simons, G., Yao, Y. & Morton, G. (2005), 'Global Markov models for eukaryote nucleotide data', *J. Statist. Plann. Inference* **130**, 251–275.

Smith, G., Kunes, S., Schultz, D., Taylor, A. & Triman, K. (1981), 'Structure of chi hotspots of generalized recombination', *Cell* **24**, 429–36.

Smith, H., Gwinn, M. & Salzberg, S. (1999), 'DNA uptake signal sequences in naturally transformable bacteria', *Res. Microbiol.* **150**, 603–616.

Stanke, M. & Waack, S. (2003), 'Gene prediction with a hidden Markov model and a new intron submodel', *Bioinformatics* **19**, 215–225.

Stephens, R., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R., Zhao, Q., Koonin, E. & Davis, R. (1998), 'Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis.', *Science* **282**, 754–759.

van Helden, J., del Olmo, M. & Pérez-Ortín, J. (2000), 'Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals', *Nucl. Acids Res.* **28**, 1000–1010.

Wu, R. & Taylor, E. (1971), 'Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA', *J Mol Biol.* **57**, 491–511.

Zoubak, S., Clay, O. & Bernardi, G. (1996), 'The gene distribution of the human genome', *Gene* **174**, 95–102.