FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model

Gaëlle GUSTO[1] and Sophie SCHBATH[1 2]

Abstract

We propose an original statistical method to estimate how the occurrences of a given process along a genome, genes or motifs for instance, may be influenced by the occurrences of a second process. More precisely, the aim is to detect avoided and/or favored distances between two motifs, for instance, suggesting possible interactions at a molecular level. For this, we consider occurrences along the genome as point processes and we use the so-called Hawkes' model. In such model, the intensity at position t depends linearly on the distances to past occurrences of both processes via two unknown profile functions to estimate. We perform a non parametric estimation of both profiles by using B-spline decompositions and a constrained maximum likelihood method. Finally, we use the AIC criterion for the model selection. Simulations show the excellent behavior of our estimation procedure. We then apply it to study (i) the dependence between gene occurrences along the *E. coli* genome and the occurrences of a motif known to be part of the major promoter for this bacterium, and (ii) the dependence between the yeast *S. cerevisiae* genes and the occurrences of putative polyadenylation signals. The results are coherent with known biological properties or previous predictions, meaning this method can be of great interest for functional motif detection, or to improve knowledge of some biological mechanisms.

Keywords: Hawkes' model, AIC criterion, splines, word occurrences, DNA sequence, favored or avoided distances, maximum likelihood.

1 Introduction

Statistical analysis of motif occurrences in DNA sequences is an important task in computational molecular biology. Many motifs, defined here as short sequences of letters in the four DNA-letter alphabet $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, allow interactions between DNA and proteins and then initiate biological processes like gene transcription, restriction, DNA repair, replication, recombination etc. The sequence of letters of some of these functional motifs have been identified, or some consensus patterns have been proposed, but most of them are still unknown. Moreover, they may have very different structures: degenerated words (i.e., several closely related words serving the same function), several words separated by fixed or variable distances, palindromes or stem-loop motifs etc. It is a challenging question to suggest functional motifs to molecular biologists based on statistical properties of their occurrences along a genome. A very common problem is indeed to find words

[1]Unité Mathématique, Informatique et Génome, INRA, 78352 Jouy-en-Josas, France. Email addresses: gaelle.gusto@jouy.inra.fr, sophie.schbath@jouy.inra.fr

[2]To whom correspondence should be addressed

with unexpected frequency; in particular, words with a significantly high frequency compared to a maximal order Markov chain model are of great interest (Rocha *et al.* (1998), Mariño-Ramirez *et al.* (2004) for instance). A huge literature exists on the distribution of word counts in Markovian models (see Reinert *et al.* (2000, 2005) for recent reviews). A related and important problem is to detect genome regions significantly poor or rich in a given motif, related to waiting time problems or scan statistics (Robin and Daudin (1999), Stefanov (2003)). However, it seems that no work has been done so far to study the statistical dependence between motif occurrences. It is clear that constraints on the space between occurrences of one or several motifs can exist if these motifs are part of a common biological process. This can be due, for instance, to the topology of the proteins that bind the DNA, like for the promoters recognized by the RNA polymerase, or to protein interactions.

In this paper, we propose an original statistical approach to study motif occurrences along genomes. Our aim is to model the dependence between motifs in order to identify favored or avoided distances between motifs. This will give additional clues to understand several biological mechanisms. Here, a motif can be replaced by any biological object that occurs along DNA sequences, genes for instance. This method can be used in two different ways: either one is interested in analyzing the influence between two given motifs, or one is interested in identifying motifs that are correlated with the occurrences of a given motif.

Because genomes are long (usually more than one million bases) and motifs of interest are short (3 up to 20 bases long), motif occurrences can be viewed like points along genomes. For convenience, we will work in a continuous framework and then, we will consider the occurrences of a motif along a genome like a point process along the interval $[0, T]$, T being the normalized length of the studied genome. More precisely, let us consider a given motif $\mathbf{u}$ along a genome and let us denote by $(U_i)_{i=1,\dots,I}$ the ordered and normalized positions of this motif in $[0, T]$, and by $(N_t, t \in \mathbb{R})$ the associated counting process. Note that two points cannot occur at the same location t. The intensity $\lambda(t)$ of this point process quantifies the probability that an event occurs at position t. Our framework consists in modeling the occurrences of the motif $\mathbf{u}$ by a linear Hawkes' model (Hawkes (1971)) with the following conditional intensity given the past:

$$\lambda(t \mid \mathcal{H}_t) = \mu + \int_0^T g(t-s)dN_s + \int_0^T g'(t-s)dN'_s$$

where $(N'_t, t \in \mathbb{R})$ is a second point process, $\mathcal{H}_t$ is the history of the processes (N_t, N'_t) at time t^- and $\mu > 0$. The real functions g and g' are such that the conditional intensity is non negative and should satisfy $\int_0^T |g(x)|\, dx < 1$ to ensure the stationarity of the process $(N_t, t \in \mathbb{R})$ (Brémaud and Massoulié (1996)).

Typically, the second counting process N'_t will be associated to the ordered occurrences $(V_j)_{j=1,\dots,J}$ of another motif $\mathbf{v}$ in $[0, T]$. This model means that the probability for a motif $\mathbf{u}$ to occur at position t linearly depends on the distances from t to the past occurrences of $\mathbf{u}$ (function g) and to the past occurrences of $\mathbf{v}$ (function g'). However, we assume that one has a finite memory of the past occurrences; let us denote by D and D' the maximal memories. These quantities can be interpreted like the support of the functions g and g'.

The conditional intensity can then be written like

$$\lambda(t \mid \mathcal{H}_t) = \mu + \sum_{t-D \le U_i < t} g(t - U_i) + \sum_{t-D' \le V_j < t} g'(t - V_j). \tag{1.1}$$

It is in fact reasonable to think that, if there exists a spacing constraint between motif occurrences along genomes, then it is necessarily a local dependency. Long range correlations, if any, should rather be related to the 3D structure of DNA inside the cell and are not what we are interested in. In practice D and D' are chosen by the biologists, depending on the underlying biological process they have in mind. One can also consider large values of D and D' to check if the terms $g(d)$ and $g'(d)$ tend to zero as d grows to infinity.

The functions g and g' will be called profiles. These are really the quantities of interest here because g (respectively g') characterizes the influence of an occurrence of $\mathbf{u}$ (respectively $\mathbf{v}$) on a future occurrence of $\mathbf{u}$. If $g \equiv 0$ then the occurrences of $\mathbf{u}$ are independent from each other, if $g' \equiv 0$ then the occurrences of $\mathbf{u}$ do not depend on the occurrences of $\mathbf{v}$, and if $g \equiv g' \equiv 0$ then the occurrences of $\mathbf{u}$ form an homogeneous Poisson process with rate μ. Now, if $g(d) \gg 0$ (resp. $g'(d) \gg 0$) for some distance d, it means that this distance is favored between two occurrences of $\mathbf{u}$ (resp. between an occurrence of $\mathbf{v}$ and a forward occurrence of $\mathbf{u}$); on the contrary, negative values of the profiles would indicate avoided distances between motifs. Our objective is then to estimate μ, g and g' from the observation of $(U_i)_i$ and $(V_j)_j$.

We propose a non parametric estimation method based on the maximum likelihood method like in Ogata and Akaike (1982). However, we use B-splines to estimate the profiles g and g' which ensures the uniqueness of the maximum likelihood estimate if any. In the present paper, knots are regularly spaced on $[0, D]$ and $[0, D']$ and their numbers are chosen according to the Akaike's Information Criterion (Akaike (1973)). Section 2 is devoted to the description of the estimation procedure. Its validation has been made with various simulations, some of which are presented in Section 3. Section 4 presents two applications, one on the complete *Escherishia coli* genome, the other one on the yeast *Saccharomyces cerevisiae*, and shows promising results. A discussion will end this paper.

2 Estimation procedure

2.1 Class of models

To estimate the profiles g and g', we use two bases of B-splines, one on $[0, D]$ and the other one on $[0, D']$. For this, let d and d' be two integers representing the degrees of the B-splines. For each integers n and n', we define the regularly spaced knots $(\varepsilon_k)_{k=0,\dots,n}$ on $[0, D]$ and $(\varepsilon'_\ell)_{\ell=0,\dots,n'}$ on $[0, D']$ as follows:

$$0 = \varepsilon_0 < \varepsilon_1 < \cdots < \varepsilon_n = D \quad \text{and} \quad \varepsilon_k = \frac{kD}{n},$$
$$0 = \varepsilon'_0 < \varepsilon'_1 < \cdots < \varepsilon'_n = D' \quad \text{and} \quad \varepsilon'_\ell = \frac{\ell D'}{n'};$$

n and n' represent the number of intervals in regular partitions of $[0, D]$ and $[0, D']$.

We then consider the corresponding bases of B-splines $(B_{d,k,\mathcal{E}_n})_{k=-d,\dots,n-1}$ with degree d and set $\mathcal{E}_n = (\varepsilon_k)_{k=-d,\dots,n+d}$ ($\varepsilon_k = 0$ if $k < 0$ and $\varepsilon_k = D$ if $k > n$), and $(B_{d',\ell,\mathcal{E}'_{n'}})_{\ell=-d',\dots,n'-1}$ with degree d' and set $\mathcal{E}'_{n'} = (\varepsilon'_\ell)_{\ell=-d',\dots,n'+d'}$ ($\varepsilon'_\ell = 0$ if $\ell < 0$ and $\varepsilon'_\ell = D'$ if $\ell > n'$). Recall that a B-spline, say $B_{d,k,\mathcal{E}_n}$, is a piecewise polynomial function (d is the degree of the polynomials) whose support is $[\varepsilon_k, \varepsilon_{k+d+1}[$, and $0 \leq B_{d,k,\mathcal{E}_n}(x) \leq 1, \forall x \in [0, D]$ (see de Boor (2001) for a practical presentation of B-splines).

For a given couple (n, n'), we consider the following model $\mathcal{S}_{n,n'}$:

$$\mathcal{S}_{n,n'} = \left\{ s = (\nu, h, h') \ : \ \nu > 0, \ h(x) = \sum_{k=-d}^{n-1} a_k B_{d,k,\mathcal{E}_n}(x), \ a_k \in \mathbb{R}, \right.$$

$$\left. h'(x) = \sum_{\ell=-d'}^{n'-1} a'_\ell B_{d',\ell,\mathcal{E}'_{n'}}(x), \ a'_{k'} \in \mathbb{R} \right\}.$$

The dimension of $\mathcal{S}_{n,n'}$ is $m := 1 + n + d + n' + d'$. For convenience, if $s \in \mathcal{S}_{n,n'}$, define the associated set of parameters $\theta^s_{n,n'} = (\nu, (a_k)_{k=-d,\dots,n-1}, (a'_\ell)_{\ell=-d',\dots,n'-1}) \in \mathbb{R}^m$.

2.2 Maximum likelihood estimators

The partial log-likelihood of the point process $(U_i)_{i=1,\dots,I}$ associated with $N(t)$ is given (Cox (1975)) by:

$$\begin{aligned} L(\mu, g, g') &= \int_0^T \log(\lambda(t \mid \mathcal{H}_t))\, dN(t) - \int_0^T \lambda(t \mid \mathcal{H}_t)\, dt \\ &= \sum_{i=1}^{I} \log(\lambda(U_i \mid \mathcal{H}_t)) - \int_0^T \lambda(t \mid \mathcal{H}_t)\, dt \end{aligned}$$

where the conditional intensity $\lambda(t \mid \mathcal{H}_t)$ is given by (1.1).

We look for an estimator of (μ, g, g') in $\mathcal{S}_{n,n'}$, denoted by $(\widehat{\mu}, \widehat{g}_n, \widehat{g'}_{n'})$ and corresponding to a set of parameters $\widehat{\theta}_{n,n'} = (\widehat{\mu}, (\widehat{a_k})_{k=-d,\dots,n-1}, (\widehat{a'_\ell})_{\ell=-d',\dots,n'-1})$. For convenience, if $s = (\nu, h, h') \in \mathcal{S}_{n,n'}$, then the log-likelihood $L(\nu, h, h')$ will be also denoted by $L(\theta^s_{n,n'})$. Therefore, we propose to consider

$$\widehat{\theta}_{n,n'} = \mathrm{argmax}_{\theta \in \Theta} L(\theta)$$

where Θ is the restriction of $\mathbb{R}^m$ such that $L(\theta)$ is defined namely $\lambda(U_i \mid \mathcal{H}_t) > 0, \forall i = 1, \dots, I$. Since Θ is convex and the Hessian matrix of $L(\theta)$ is strictly negative (Ogata (1978), Ogata and Katsura (1986)), if the likelihood has a local maximum on Θ then it will be a global one. Since the process $(U_i)_{i=1,\dots,I}$ is stationary (we put a condition on the integral of $|g|$), the consistency and the asymptotic normality of the maximum likelihood estimator $\widehat{\theta}_{n,n'}$ as $T \to \infty$, if it exists, hold. Indeed, assumptions of Theorem 5 in Ogata (1978) are satisfied here (for more details, refer to Gusto (2004)). We will then be able to build asymptotic theoretical confidence intervals for μ, $g(x)$ and $g'(x)$; for instance, the 95%

confidence interval of $g(x)$ is $\widehat{g}_n(x) \pm 1.96\, v_n(x)$ where the asymptotic variance $v_n^2(x)$ is obtained by using Fisher information ($g(x)$ is a linear combination of the parameters).

There is no analytical expression for $\widehat{\theta}_{n,n'}$. In practice, the constrained maximization of the likelihood on Θ has been performed using the method of Facchinei *et al.* (1999); this method leads to an unconstrained maximization performed with the Levenberg-Marquart method (Fletcher (1987)).

2.3 Model selection

The polynomial degrees d and d' are definitely set at the begining of the procedure. However, we have to select appropriate values for n and n', the numbers of intervals $[\varepsilon_k, \varepsilon_{k+1}[$ and $[\varepsilon'_\ell, \varepsilon'_{\ell+1}[$ respectively partitioning $[0, D]$ and $[0, D']$. For this we chose to minimize the Akaike's criterion defined by

$$\mathrm{AIC}(n, n') := -2L(\widehat{\theta}_{n,n'}) + 2m,$$

where m is the number of parameters for model $\mathcal{S}_{n,n'}$.

Let $\mathcal{N}$ and $\mathcal{N}'$ be two given sets of integers, typically $\mathcal{N} = \mathcal{N}' = \{1, 2, \dots, n_{\max}\}$, but $\mathcal{N}$ and $\mathcal{N}'$ can be different. We then choose:

$$(\widehat{n}, \widehat{n'}) = \mathrm{argmin}_{n \in \mathcal{N}, n' \in \mathcal{N}'} \mathrm{AIC}(n, n'),$$

and we finally propose the estimator $(\widetilde{\mu}, \widetilde{g}, \widetilde{g'})$ associated to the parameters

$$\widetilde{\theta} = \widehat{\theta}_{\widehat{n}, \widehat{n'}}.$$

In section 3, we will study both the validation of the AIC criterion and the quality of our estimation procedure, with simulations. In particular, we will use the residual analysis and we will compare the risk of the estimator $\widetilde{\theta}$ with the smallest risk of the family $(\widehat{\theta}_{n,n'}, n \in \mathcal{N}, n' \in \mathcal{N}')$.

3 Simulations

In this section, we will simulate 100 processes (U_i) on $[0, 100\,000]$ under model (1.1). For this purpose, we have used the thinning method proposed by Ogata (1981, 1999); This algorithm has the advantage to avoid the heavy numerical computations required by the classical method to simulate multivariate processes (see Daley and Vere-Jones (1988), chapter 7). Several values of (μ, g, g') are considered and the V_j's are drawn, independently of the U_i's, according to an homogeneous Poisson process with rate 0.01 on $[0, 100\,000]$ (1000 points in average).

3.1 Four sets of parameters

We have considered the four following cases for (μ, g, g') (the curves of the profiles are represented on Figure 3). In each case, the supports of the functions g and g' (the maximal memories) are set to $D = D' = 2000$. Moreover, the values of μ have been chosen so that the expected number of points of the U process is equal to 1000.

- Case 1: simple Poisson process

$$\begin{aligned} \mu &= 0.01 \\ g &\equiv g' \equiv 0. \end{aligned}$$

- Case 2: a mixed doubly stochastic Poisson and self-exciting process for which the true model belongs to a class of model $\mathcal{S}_{3,3'}$ (splines of degree 2)

$$\begin{aligned} \mu &= 8.22\,10^{-3} \\ g(t) &= 0.002\,B_{2,-1,\mathcal{E}_3}(t), \qquad \text{with } \mathcal{E}_3 = \{0, \frac{2000}{3}, \frac{4000}{3}, 2000\} \\ g'(t) &= -0.0016\,B_{2,-1,\mathcal{E}_3}(t). \end{aligned}$$

- Case 3: a self-exciting process for which the true model is not in a set $\mathcal{S}_{n,n'}$

$$\begin{aligned} \mu &= 8.44\,10^{-3} \\ g(t) &= e^{-t^*}\left[-0.2t^* + 0.73t^{*2} - t^{*3} + 0.83t^{*4} - 0.12t^{*5}\right]/500, \quad t^* = t/100 \\ g' &\equiv 0. \end{aligned}$$

- Case 4: a mixed doubly stochastic Poisson and self-exciting process for which the true model is not in a set $\mathcal{S}_{n,n'}$

$$\begin{aligned} \mu &= 1.77\,10^{-3} \\ g(t) &= te^{-t/100}/15000 \\ g'(t) &= e^{-t^*}\left[-0.2t^* + 0.73t^{*2} - t^{*3} + 0.83t^{*4} - 0.12t^{*5}\right]/500, \quad t^* = t/100. \end{aligned}$$

3.2 Results

To perform the estimation procedure described in Section 2, Table 1 presents, for each of the 4 cases, the degrees d and d' of the B-splines we have considered together with the sets $\mathcal{N}$ and $\mathcal{N}'$ of possible values for n and n'. Note that, except for case 3, we impose $n = n'$ leading to $\mathcal{N}' = \mathcal{N}$. As for general real-data settings (see the Application section), we chose reasonable polynomial degrees, i.e. $d = d' = 2$. The 6th column of Table 1 gives the expression of the dimension m of the corresponding model $\mathcal{S}_{n,n'}$ with respect to n (and n' for case 3).

Place Table 1 here.

Figure 1 shows, for each of the four cases, the histogram of the 100 dimensions $\widehat{m}$ chosen by the AIC criterion (over the 100 simulations). For cases 1, 2 and 4, $\widehat{n}$ can be deduced from $\widehat{m}$ because $\widehat{m} = 2\widehat{n}+d+d'+1$; in case 3, Figure 2 displays the 2D-histogram for $(\widehat{n}, \widehat{n'})$. One can see that the most often chosen dimension, denoted by $m^\star$, is $m^\star = 7$ for case 1 (thus $n^\star = n'^\star = 1$), $m^\star = 11$ for case 2 (thus $n^\star = n'^\star = 3$), $m^\star = 10$ for case 3 (corresponding to $n^\star = 4$ and $n'^\star = 1$), $m^\star = 23$ for case 4 (thus $n^\star = n'^\star = 9$). Results obtained for cases 1 and 2 are very satisfactory because we have selected the true dimension and both histograms are rather sharp. For cases 3 and 4, there is no true dimension (the true models are not in the class of models considered); the histograms are less sharp, especially for case 4 where the best numbers of intervals seem to be between 8 and 11, but for case 3 it seems that the best dimension is 10 or 11 (note that $n'^\star = 1$ is the correct value for the estimation of $g' \equiv 0$).

Place Figures 1 and 2 here.

Table 2 and Figure 3 present the estimators proposed for μ and for both profiles g and g' in the four cases. These estimators are the empirical averages of the 100 estimators obtained when n and n' are set to the optimal numbers $n^\star$ and $n'^\star$. They are denoted by $\overline{\widehat{\mu}_{n^\star,n'^\star}} := \frac{1}{100}\sum_{i=1}^{100} \widehat{\mu}^{(i)}_{n^\star,n'^\star}$, $\overline{\widehat{g}_{n^\star}} := \frac{1}{100}\sum_{i=1}^{100} \widehat{g}^{(i)}_{n^\star}$ and $\overline{\widehat{g'}_{n^\star}} := \frac{1}{100}\sum_{i=1}^{100} \widehat{g'}^{(i)}_{n'^\star}$, where the index $^{(i)}$ stands for the simulation index. For each quantity, the average over the 100 simulations of the 95% theoretical confidence intervals (i.e. $\widehat{g}^{(i)}_{n^\star}(t) \pm 1.96\, v^{(i)}_{n^\star}(t)$ for $g(t)$) are also reported. More over, the last column of Table 2 gives the empirical mean of the square errors of estimation:

$$(\mu - \widehat{\mu}^{(i)}_{n^\star,n'^\star})^2 + \int_0^D (g(t) - \widehat{g}^{(i)}_{n^\star}(t))^2\, dt + \int_0^{D'} (g'(t) - \widehat{g'}^{(i)}_{n'^\star}(t))^2\, dt.$$

Globally, the estimations are very good. Some parts of the profiles are over- or under-estimated but the shapes are perfectly recovered, especially the x-coordinates of the modes. This is very important for future applications since the method is able to detect quite precisely, ranges of avoided and favored distances between points, if any. Note that the 95% confidence intervals of μ, g and g' generally contain the true values.

Place Table 2 here.

Place Figure 3 here.

Now, to see the variability of the estimation procedure among the simulations, Figure 4 shows the 100 estimated profiles $\widehat{g}^{(i)}_{\widehat{n}^{(i)}}$ and $\widehat{g'}^{(i)}_{\widehat{n'}^{(i)}}$, $i = 1, \ldots, 100$, where $\widehat{n}^{(i)}$ and $\widehat{n'}^{(i)}$ are the numbers of knots selected by AIC for the ith simulation.

In case 1, when considering that $n = n' = 1$, the likelihood ratio test, at level 5%, accepts the hypothesis $\mathrm{H}_0 : \{g \equiv g' \equiv 0\}$ against the general model in 91% of the simulations. In case 3, when considering that $n = 4$ and $n' = 1$, the likelihood ratio test, at level 5%, accepts the hypothesis $\mathrm{H}_0 : \{g' \equiv 0\}$ against the general model in 90% of the simulations, which is quite satisfactory. All the other tests of the form $\mathrm{H}_0 : \{g \equiv 0\}$ or $\mathrm{H}_0 : \{g' \equiv 0\}$ are rejected when setting the model dimension to $m^\star$.

Goodness-of-fit To test the goodness-of-fit of our estimations, we performed classical residual analyzes (Daley and Vere-Jones (1988)). A point process $(T_1, \ldots, T_k)$ with cumulative intensity $\Lambda(x) := \int_0^x \lambda(t)\, dt$ has indeed the property that the residual process defined by $(\Lambda(T_1), \ldots, \Lambda(T_k))$ forms an homogeneous Poisson process with rate 1. Therefore, for cases 1 to 4 and for each of the 100 simulated processes $(U_1, \ldots, U_k)$ (k varies from simulations), we used the Kolmogorov-Smirnov test to check if the processes $(\widehat{\Lambda}_{n,n'}(U_1), \ldots, \widehat{\Lambda}_{n,n'}(U_k))$ could have been drawn from an homogeneous Poisson process with rate 1, for several values of n and n' close to the "optimal" values $n^\star$ and $n'^\star$ selected with AIC. Here, the cumulative intensity $\widehat{\Lambda}_{n^\star,n'^\star}$ is associated to the following estimated intensity

$$\widehat{\lambda}_{n^\star,n'^\star}(t \mid \mathcal{H}_t) := \widehat{\mu}_{n^\star,n'^\star} + \sum_{t-D\le U_i<t} \widehat{g}_{n^\star}(t-U_i) + \sum_{t-D'\le V_j<t} \widehat{g'}_{n'^\star}(t-V_j).$$

Table 3 indicates the percentage of simulations that accept the test at level 5%. The goodness-of-fit of our estimated models are generally too perfect. Indeed, we expect to reject the test 5% of the time but in many cases, we never reject the test. The Kolmogorov-Smirnov test seems to be too conservative in our context.

Place Table 3 here

Validation of the model selection method Model selection is an important area in statistics and lots of recent papers aim at providing better criteria, better penalizations of classical quantities such as log-likelihood or least square criteria (e.g. Barron *et al.* (1999), Castellan (1999), Reynaud-Bouret (2003)). The Akaike Information Criterion is probably the most popular method for model selection: the maximum likelihood is penalized by the model dimension. However, it is now quite well established that AIC is not very adapted as the number of models per dimension is large. In our case, because we have chosen regularly spaced knots, we have few models with a fixed dimension m (if we set $n = n'$, we even have a unique model per dimension). The AIC method should then be suitable. For convenience, let us denote the true model (μ, g, g') by θ. In order to validate the Akaike's criterion, we have compared the risk, for a given loss function ℓ, of the estimator $\widetilde{\theta} = \widehat{\theta}_{\widehat{n},\widehat{n'}}$ of θ obtained by the AIC method with the smallest risk over the collection of estimators $\{\widehat{\theta}_{n,n'}, n \in \mathcal{N}, n' \in \mathcal{N}'\}$, so-called oracle. In other words, we have calculated the ratio

$$F(\widetilde{\theta}) = \frac{\mathbb{E}_\theta(\ell(\theta,\widetilde{\theta}))}{\min_{n,n'} \mathbb{E}_\theta(\ell(\theta,\widehat{\theta}_{n,n'})},$$

where the minimum is taken over $\{n \in \mathcal{N}, n' \in \mathcal{N}'\}$, $\mathbb{E}_\theta$ is the mean under the true model (and will be evaluated thanks to simulations), and the loss function is

$$\ell(\theta,\theta')) = \mathbb{E}_\theta[L(\theta \mid \mathcal{H}_t) - L(\theta' \mid \mathcal{H}_t)];$$

(this loss function is also the Kullback information between the two associated intensities λ). Denote by $\check{m}$ the model dimension associated to the oracle. Table 4 shows that, in each case, we have $\check{m} = m^\star$: AIC then chooses the same dimension as the oracle. Moreover, in cases 1, 2 and 3, the risk ratios are quite close to 1 meaning that AIC gives rather good performance in that cases. Unfortunately, $F(\widetilde{\theta})$ is very large for case 4.

Place Table 4 here.

We then looked at the slope α of the log-likelihood with respect to the model dimension, for high dimensions. Indeed, Lebarbier (2005) recommends to penalize the maximum log-likelihood by $2\alpha\, m$ (slope method); a slope of 0.5 suggests that AIC performs well. In case 4, we have $\alpha \simeq 0.7$ suggesting that AIC does not enough penalize the log-likelihood (not shown here, see Gusto (2004)). In fact, the slope method would choose a model of dimension $13 \neq 23$ (only 4 intervals to estimate both g and g') leading to estimators with a better risk but unfortunately relatively far from the true profiles (not shown here).

Hence, the AIC procedure does not provide the best estimators in terms of the Kullback risk function but seems to select, for the most part, appropriate dimensions and estimated profiles very close to the true ones.

4 Applications

4.1 Influence promoters/genes in *E. coli*

As a first application, we looked at the gene transcription mechanism. Indeed, genes along the chromosomes are generally preceded by several genomic signals to regulate the transcription of the genes or to initiate it. In particular, promoters are usually structured motifs (two or several words separated by variable but constrained distances) located before the genes, not too far, and recognized by the RNA polymerase to bind the DNA and start the transcription of the next gene into RNA. For a given species several different promoters may exist but they are not numerous since they are recognized by the same RNA polymerase (they can be degenerated). Finding the sequence of these motifs is still a huge challenge in Bioinformatics; classical methods aim at finding over-represented words in the sequences just before the genes. Here, we have considered the major promoter of the bacterium *E. coli* and more precisely the word `tataat` contained in its consensus sequence. Most of the genes of *E. coli* should be preceded by this word at a very short distance apart. In order to validate our method, we then looked at the locations U_i of all the genes and at the locations V_j of all the occurrences of `tataat` along the complete genome of the bacterium, hoping to detect short favored distances between genes and previous occurrences of `tataat`.

For this, we have analyzed the sequence composed of both strands of *E. coli* K12's genome[3] (4 639 221 bps), each strand being separated by 10 000 artificial bases, to avoid artificial dependencies between occurrences on one strand and occurrences on the other strand; we took $D = D' = 10\,000$ for the maximal memories. It then represents a sequence of length 9 288 442; there are 4 290 genes[4] and 1 036 occurrences of `tataat`. In average, there is a unique occurrence of `tataat` in a window of size 10 000, for instance in front of a gene.

To estimate the profiles g and g', we chose B-splines of degree $d = d' = 2$ and the numbers n and n' of intervals in the partitions of $[0, 10\,000]$ vary in $\mathcal{N} = \mathcal{N}' = \{1, 2, \ldots, 20\}$. Using the estimation method presented in this paper, we obtained $\widehat{n} = 20$ and $\widehat{n}' = 19$ (almost the largest model), and $\widetilde{\mu} = 3.70\,10^{-4}$ ($\pm\, 2.23\,10^{-5}$ at 95% confidence) which is less than the empirical gene rate of $4.41\,10^{-4}$. The estimated profiles $\widetilde{g}$ and $\widetilde{g}'$, together with their 95% estimates, are given on the top of Figure 5. The tests $\{g \equiv 0\}$ and $\{g' \equiv 0\}$ are rejected at level 5%.

Place Figure 5 here.

The estimator $\widetilde{g}$ (how occurrences of genes are influenced by previous genes) indicates that (i) distances around 1 100 bases are clearly favored, (ii) the smaller the distance between 400 and 0 bps, the higher the avoidance, (iii) long range correlations (distances longer

[3]accession number U00096 in Genbank

[4]we took the positions of the first base of coding sequences (CDS) from Genbank annotations

than 2 500) do not seem to exist. This is quite coherent with the fact that, in bacteria, the genome is very compact leading to short interspaces between genes, and the average length of a gene in *E. coli* is about 1000 bases. The estimator $\widetilde{g'}$ (how occurrences of genes are influenced by previous occurrences of `tataat`) indicates that (i) smaller the distance between 400 and 0 bps, higher the preference, (ii) distances around 1 200 bases are also favored, (iii) these two strong influences seem to be the only ones (the profile fluctuates around 0 for distances greater that 3 000 bases). Here again, the results coincide with the ones we could expect: short distances are favored and the second preference for distances around 1 200 bases can also be a consequence of the average length of a gene and the operon phenomenon; indeed, several consecutive genes may be transcribed together in the same RNA, the promoter and the regulatory elements of the first gene serving for the entire operon. Doing the same study with $D = D' = 4\,000$ does not give much more information (not shown here).

To check that any other pattern as much frequent as `tataat` is not associated with a similar g' profile, we did the same analysis with the two following words: `atatac` (1 057 occurrences) and `tccaac` (1 078 occurrences). The estimated profiles are given on Figure 5. Since the three motifs have quite the same counts, the y-coordinates of the curves can be compared. Thus, the three profiles $\widetilde{g'}$ do not have the same shape, which reinforces the previous observations. Note, however, that the 3 estimated profiles for g (on the left) are very similar. This is not surprising because it models the same influence (between genes), but it also suggest that the influence between genes is stronger than the influences by the motifs, masking some of these latter effects.

We then have considered the following reduced model for the intensity of the gene process:

$$\lambda(t \mid \mathcal{H}_t) = \mu + \sum_{t-D' \leq V_j < t} g'(t - V_j). \tag{4.1}$$

By adapting our method to this embedded model, we get new estimated profiles $\widetilde{g'}$ presented on Figure 6 for the three motifs with $d' = 2$, $D' = 10\,000$, $\mathcal{N}' = \{1, 2, \ldots, 20\}$.

Place Figure 6 here.

The numbers $\widehat{n'}$ of intervals chosen by AIC are $\widehat{n'} = 15$ for `tataat`, $\widehat{n'} = 4$ for `atatac` and $\widehat{n'} = 1$ for `tccaac`. The test for the nullity of g' is only rejected at level 5% for `tataat`. The estimated profiles are different from the one obtained on Figure 5 because the numbers n' are very different, especially for `atatac` and `tccaac`, but it also indicates that, for these two examples, the intensity of the gene process in the general model is mostly governed by the past gene occurrences and do not really depend on the occurrences of `tccaac` nor `atatac`. The profile $\widetilde{g'}$ for `tataat` is on the contrary very similar with the one of Figure 5, in particular short distances are still strongly favored between the promoter `tataat` and the genes. The second peak around 1 200 bases in the general model has disappeared in the reduced model: it seems that it was only the consequence of the corresponding peak of $\widetilde{g}$. Figure 7 shows, for `tataat`, the evolution of $\widehat{g'}_{n'}$ as the number n' of intervals increases (AIC chooses $n' = 15$). As we can see, the preference for short distances is always detected, even for short values of n' (we omit the estimated profiles associated with odd values of n' only to save space).

Place Figure 7 here.

These are very encouraging results in the framework of promoter and regulatory motif detection. We can indeed imagine to combine this method with the ones proposing candidates for promoters or regulatory motifs.

4.2 Putative polyadenylation signals in yeast

The second application concerns the polyadenylation signals in the yeast *Saccharomyces cerevisiae* genome. Several hexamers (`tatata` and some variants) have been proposed *in silico* to be part of the polyadenylation signals in yeast (van Helden *et al.* (2000), Graber *et al.* (2002)). In particular, a position analysis revealed that those words have a high preference to be in 3' flanks of yeast genes with a strong peak around +35 bp after the stop codons. We then applied our method to find eventual preferred distances between the end of genes (stop codons) and the occurrences of the "`tatata`-like motif" (signals E_1 or D_1 in the above references).

We first analyzed the sequence composed of both strands of the yeast chromosome 1 (230 208 bps), each strand being separated by 5 000 artificial bases to avoid dependencies between the two strands. In this sequence of length 465 416 bps, we then built the process U of the positions of the `tatata`-like motif[5] (2392 occurrences) and the process V of the positions of the last base of all the 97 coding sequences from Genbank annotations.

Since we are *a priori* interested by short range influences, we took $D = D' = 1000$ like maximal memories. To estimate the profiles g and g', we chose B-splines of degree $d = d' = 2$ and the numbers n and n' of intervals partitioning $[0, 1\,000]$ vary in $\mathcal{N} = \mathcal{N}' = \{1, 2, \ldots, 20\}$. The AIC criterion chooses $\widehat{n} = 20$ and $\widehat{n}' = 4$ and the corresponding estimated profiles $\widetilde{g}$ and $\widetilde{g'}$, together with their 95% estimates, are given on the top of Figure 8. The tests $\{g \equiv 0\}$ and $\{g' \equiv 0\}$ are rejected at level 5%.

Place Figure 8 here.

The estimator $\widetilde{g}$ indicates that we only have a very local dependence between the occurrences of the `tatata`-like motifs which is confirmed as one decreases the support of the profile g to $D = 100$ (see bottom of Figure 8). In fact we just confirm the fact that the distribution of this overlapping motif can be modeled by a compound Poisson process: it occurs in clumps of overlapping occurrences and the clumps are distributed according to a Poisson process (Schbath (1995), Robin (2002)). The estimator $\widetilde{g'}$ interestingly points out a favored distance around 43 bps which is consistent with the previous work of van Helden *et al.* (2000).

Instead of analyzing separately each of the 17 chromosomes of *S. cerevisiae*, we looked at the sequence obtained by the concatenation of the direct strands of all the 17 chromosomes (including 5 000 artificial bases between each chromosome); this leads to a sequence of length 12 236 590 with 2970 genes and 71 349 occurrences of the motif. We obtained very similar results than for chromosome 1, in particular, a favored distance around 46 bps appears between the stop codon of the genes and following `tatata`-like motif occurrences. As an illustration we only show (see Figure 9) the profile $\widetilde{g'}$ obtained in the reduced model (4.1) for $d' = 2$ and $D' = 300$ (AIC chooses $\widehat{n}' = 10$ among $\{1, 2, \ldots, 20\}$).

[5] `tatata`, `tatgta`, `tacata`, `taaata`, `tgtata`, `tattta`, `taagta`, `tacgta`, `tagata`, `tgtgta`, `taatta`, `atctat`, `tagtta`, `tgtaca`

5 Discussion

We have developed a rigorous statistical method to model the dependencies between motif occurrences along genomes and more precisely to detect avoided and/or favored distances between motifs. Section 4 shows that favored and/or avoided distances detected with our method sound biologically meaningfull. This is of great interest for functional motif detection and to improve knowledge of some biological mechanisms. For instance, we are involved in a French multidisciplinary research project, in which one question concerns the relation between two particular motifs implied in the recombination process in *E. coli*.

This methodology can be applied to more than two interacting motifs, either by adding new profiles in the definition of the intensity, or by using multidimensional profiles (in that case, multidimensional B-splines should be used).

Our method is not symmetric in the direction because we assume that the intensity of a process only depends on the history of the processes. If we are rather interested by studying the dependence with the future, one just has to reverse the genome. However, considering simultaneously both directions is much more complicated, and working with these kind of point processes is an open theoretical problem.

When the distance constraints are strict, leading to sharp peaks for the profiles, regularly spaced knots are not well adapted because selected values of n and/or n' would be necessarily large to point out these strong influences. This was the case for instance in the promoter application when we used the general model (two profiles). If only local influences are concerned (i.e. constraints only for short distances), one can reduce the maximal memories D and/or D'. Otherwise, penalizing more the likelihood (we tried the BIC criterion for instance) is not the solution: smaller numbers of intervals would be chosen but, according to our simulation studies (not shown here), they do not correspond to the true values and the estimated profiles do not look like the true ones. The alternative would be to consider irregular partitions, for instance knots could vary on a fixed grid, in order to decrease the number of knots and then the model dimension. Unfortunately, the AIC is no more adapted in this case because the number of models per dimension is huge (Gusto (2004)). It then requires to consider another penalized criterion. We have been working for now on a new penalized least square criterion, but because of the theoretical difficulties to derive the appropriate penalty, we restrict ourselves to a simple Hawkes' model (no process $(V_j)_j$ and then no profile g'). This theoretical work will be published separately and will not immediately answer our biological question on influences between two motifs.

However, we think regular partitions are not restrictive in practice since the method is quite fast[6] and allowing up to 20 regular intervals for the profile decomposition already gives promising results. Moreover, as shown by Figure 7 the main trend of the profiles are early detected (clues can be obtained with small numbers of intervals). A beta version of the program FADO has been coded in C++ and is available upon request; the inputs simply consist in two sets of positions. This program will lead to an open source software downloadable from the SSB group web site `http://www-mig.jouy.inra.fr/ssb/`.

[6]for instance, 2805 (resp. 4515) seconds were needed to run the $20 * 20$ cases with $D = D' = 200$ (resp. $D = D' = 1000$) for the yeast application (Figure 8), on a PC P4 520, 2.8 GHz, 1GB RAM.

Finally, we recommend to use the reduced model as soon as the number I of occurrences of the process of interest is much larger than the number J of occurrences of the second process. Simply because the masking effect would certainly be important.

Acknowledgment We wish to thank Pascal Massart and Patricia Reynaud-Bouret for helpful discussion, Jacques Van Helden for suggesting the second application and both referees for their advices to improve the manuscript. This work has been supported by the French Action Concertée Incitative IMPBio.

References

Akaike, H. (1973). Information theory and an extension of the maximum likekihood principle. In *2nd International Symposium on Information Theory*, 267–281.

Barron, A. and Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields.* **113** 301–415.

de Boor, C. (2001). *A practical Guide to Splines.* volume 27. Springer-Verlag.

Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *The Annals of Probability.* **24 (3)** 1563–1588.

Castellan, G. (1999). Modified Akaike's criterion for histogram density estimation. Preprint.

Cox, D. R. (1975). Partial likelihood. *Biometrika.* **62 (2)** 269–276.

Daley, D. J. and Vere-Jones, D. (1988). *An introduction to the theory of Point Processes.* (Springer-Verlag, ed.). Springer Series in Statistics.

Davies, D. and Swann, W. H. (1969). Review of constrained optimization. In *Optimization*, (e. R. Fletcher, ed.), chapter 12, 187–201.

Facchinei, F. and Fischer, A. and Kanzow, C. and Peng, J.M. (1999). A simply constrained optimization reformulation of KKT systems arising from variational inequalities. *Applied Mathematics and Optimization.* **40** 19–37.

Fletcher, R. (1987). *Practical methods of optimization.* John Wiley and Sons.

Graber, J.H. and McAllister, G.D. and Smith, T.F. (2002). Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3'-processing sites. *Nucleic Acids Research.* **30** 1851–1858.

Gusto, G. (2004). *Estimation de l'intensité d'un processus de Hawkes généralisé double. Application à la recherche de motifs corépartis le long d'une séquence d'ADN.* PhD thesis, Université Paris Sud.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika.* **58 (1)** 83–90.

Lebarbier, E. (2005). Detecting multiple change-points in the mean of Gaussian process by model selection. *Signal Processing.* To appear.

Ogata, Y. (1978). The asymptotic behaviour of maximum likelihood estimators for stationary point processes. *Ann. Inst. Statist. Math.* **30** 243–261.

Ogata, Y. (1981). On Lewis' simulation method for point processes. *IEEE Transactions on Information Theory.* **27 (1)** 23–31.

Ogata, Y. (1999). Seismicity analysis through point process modeling: a review. *Pure and Applied Geophysics.* **155** 471–507.

Ogata, Y. and Akaike, H. (1982). On linear intensity models for mixed doubly stochastic Poisson and self-exciting point processes. *The Institute of Statistical Mathematics.* **44 (1)** 102–107.

Ogata, Y. and Katsura, K. (1986). Point process models with linearly parametrized intensity for application to earthquake data. *Journal of Applied Probability.* **23A** 291–310.

Mariño-Ramirez, L., Spouge, J., Kanga, G. and Landsman, D. (2004). Statistical analysis of over-represented words in human promoter sequences. *Nuc. Acid. Research* **32** 949–958.

Reynaud-Bouret, P. (2003). Adaptative estimation of the intensity of inhomogeneous Poisson Processes via concentration inequalities. *Probab. Theory Relat. Fields* **126** 103–153.

Reinert, G., Schbath, S. and Waterman, M. (2000). Probabilistic and statistical properties of words: an overview. *J. Comp. Biol.* **7** 1–46.

Reinert, G., Schbath, S. and Waterman, M. (2005). Statistics on words with applications to biological sequences. In *Applied Combinatorics on Words*. Encyclopedia of Mathematics and its Applications **105**. Cambridge University Press.

Robin, S. (2002). A compound Poisson model for words occurrences in DNA sequences. *J. R. Statist. Soc. C.* **51** 437–451.

Robin, S. and Daudin, J. (1999). Exact distribution of word occurrences in a random sequence of letters. *Journal of Applied Probability.* **36 (1)** 179–193.

Rocha, E.P.C., Viari, A. and Danchin, A. (1998). Oligonucleotide bias in *Bacillus subtilis*: general trends and taxonomic comparisons. *Nucl. Acids Res.* **26** 2971–2980.

Schbath, S., (1995) Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics* **1** 1–16.

Stefanov, V.T. (2003) The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach, *J. Appl. Probab.* **40** 881–892.

van Helden, J. and del Olmo, M. and Pérez-Ortin, J.E. (2000). Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. *Nucleic Acids Research.* **28** 1000–1010.

	d	d'	$\mathcal{N}$	$\mathcal{N}'$	m
case 1	2	2	$\{1,\ldots,20\}$	$(n = n')$	$2n + 5$
case 2	2	2	$\{1,\ldots,20\}$	$(n = n')$	$2n + 5$
case 3	2	2	$\{1,\ldots,20\}$	$\{1,\ldots,5\}$	$n + n' + 5$
case 4	2	2	$\{1,\ldots,20\}$	$(n = n')$	$2n + 5$

Table 1: Spline degrees (d, d') and possible values for the numbers of intervals (n, n') for the estimation procedure in cases 1 to 4. The dimension of the associated model is given on the last column.

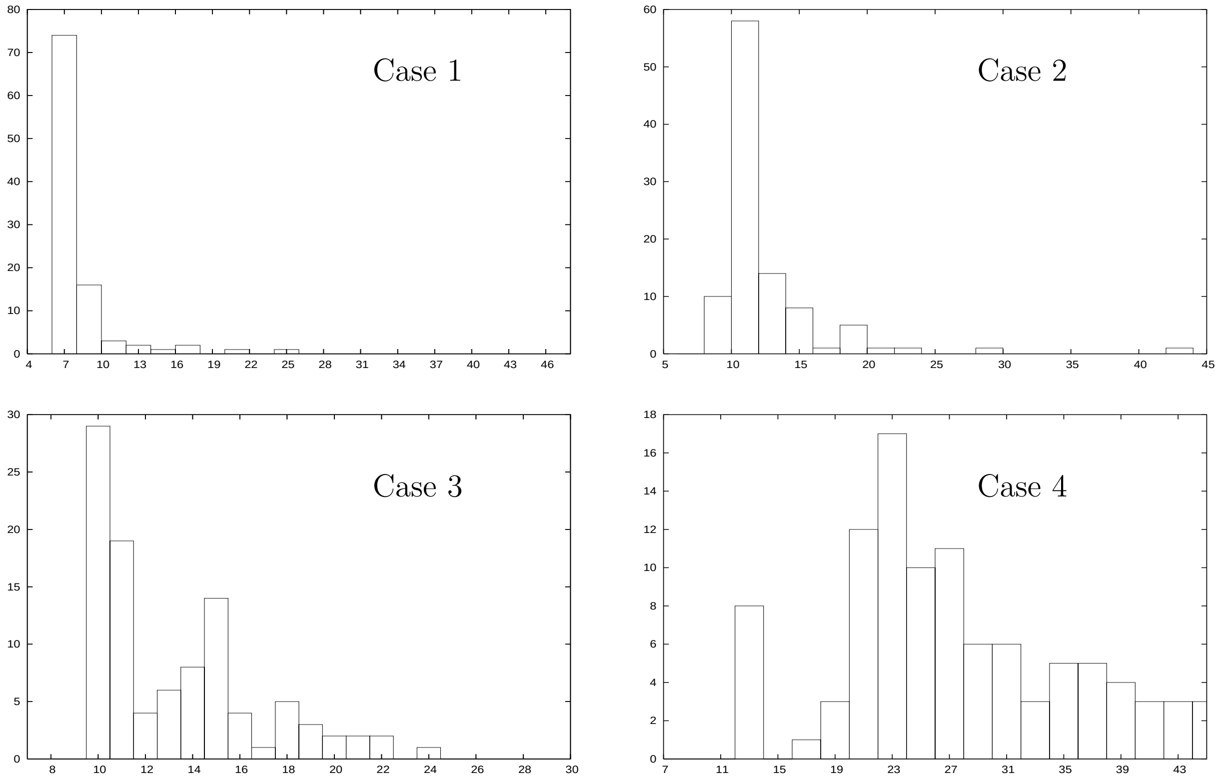

Figure 1: Histograms of the model dimensions selected by the AIC criterion over 100 simulations.

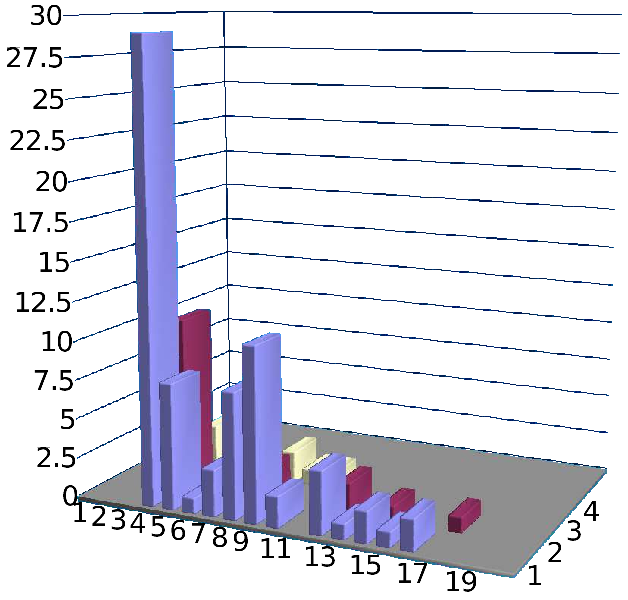

Figure 2: Histogram of the couples $(\widehat{n}, \widehat{n}')$ selected by the AIC criterion over 100 simulations for case 3. Colors correspond to the values of $\widehat{n}'$.

	μ	$\overline{\widehat{\mu}_{n^\star,n'^\star}}$	95% CI	MSE
case 1	$1.00\,10^{-2}$	$9.67\,10^{-3}$	$\pm\, 3.65\,10^{-3}$	$7.68\,10^{-5}$
case 2	$8.22\,10^{-3}$	$1.11\,10^{-2}$	$\pm\, 2.88\,10^{-3}$	$1.89\,10^{-3}$
case 3	$8.44\,10^{-3}$	$8.30\,10^{-3}$	$\pm\, 3.23\,10^{-3}$	$1.96\,10^{-2}$
case 4	$1.77\,10^{-3}$	$1.17\,10^{-3}$	$\pm\, 2.17\,10^{-3}$	$1.18\,10^{-3}$

Table 2: True values of μ, averaged estimators $\overline{\widehat{\mu}_{n^\star,n'^\star}}$ and averaged 95% confidence intervals for the most selected couple $(n^\star, n'^\star)$ over the 100 simulations. The last column gives the mean square error of estimation when $n = n^\star$ and $n' = n'^\star$.

$n = n'$					
case 1			**1**	2	3
			100	100	100
case 2	1	2	**3**	4	5
	96	99	**100**	100	100
case 4	7	8	**9**	10	11
	94	96	**95**	96	94

			n'		
			1	2	3
case 3		2	90	90	89
		3	100	100	100
	n	**4**	**100**	100	100
		5	100	100	100
		6	100	100	100

Table 3: Percentage of simulations for which the goodness-of-fit test is accepted at level 5%. For each of the four cases, results are only reported for estimations with $n \in \{n^\star \pm 2\}$ and $n' \in \{n'^\star \pm 2\}$ (results for $n = n^\star$ and $n' = n'^\star$ are boldfaced). Recall that for cases 1, 2 and 4, we set $n = n'$.

	$F(\widetilde{\theta})$	$m^\star$	$\check{m}$
case 1	1.44	3	3
case 2	1.32	11	11
case 3	1.72	8	8
case 4	4.23	23	23

Table 4: Risk ratio to evaluate the AIC criterion, and model dimensions chosen by AIC ($m^\star$) and by the oracle ($\check{m}$).

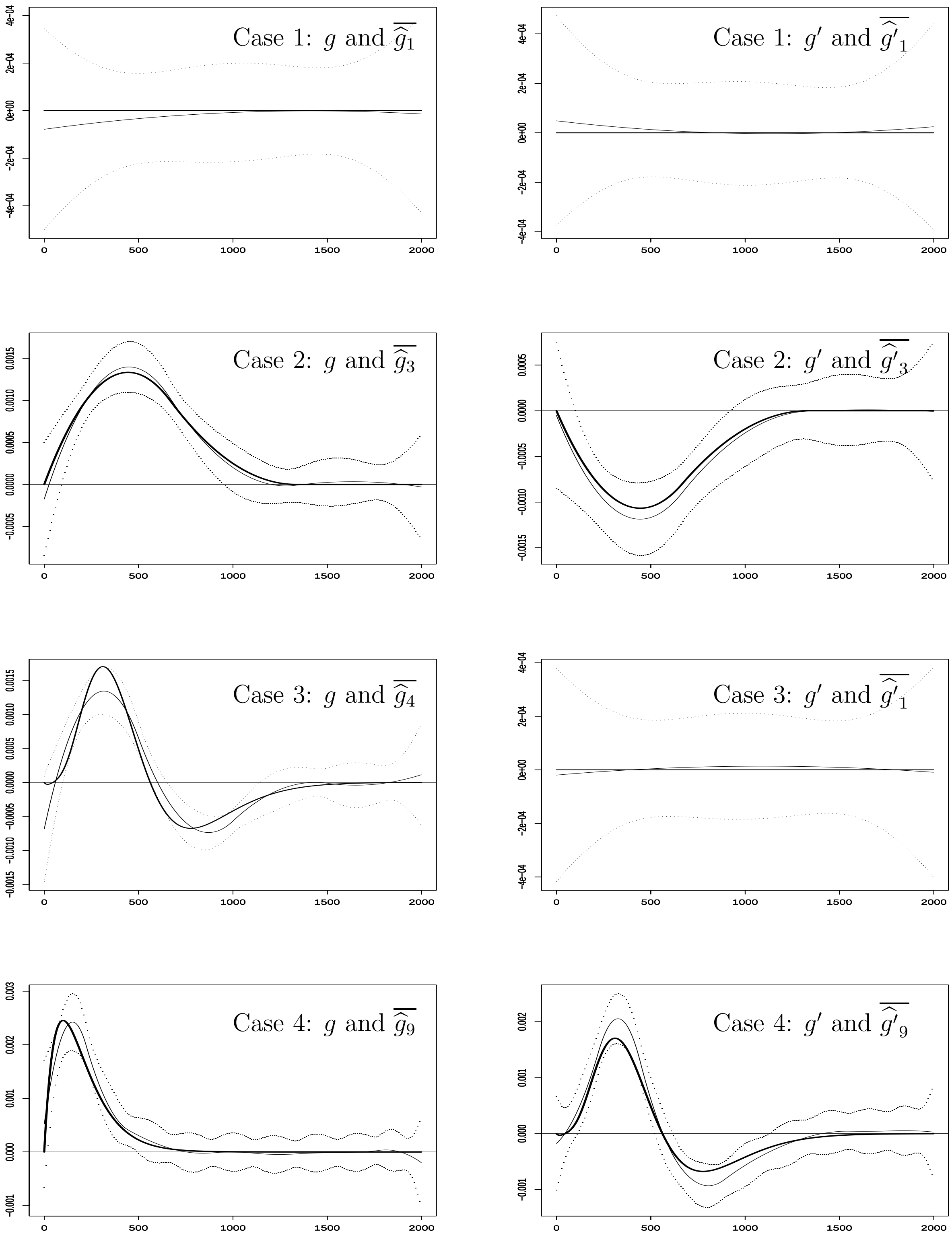

Figure 3: True profiles (g, g') [thick lines] and averaged estimated profiles $(\overline{\widehat{g}_{n^\star}}, \overline{\widehat{g'}_{n'^\star}})$ [thin lines] for the most often selected couple $(n^\star, n'^\star)$ over the 100 simulations for case 1 (top) to case 4 (bottom). The dotted lines correspond to the 95% theoretical confidence intervals.

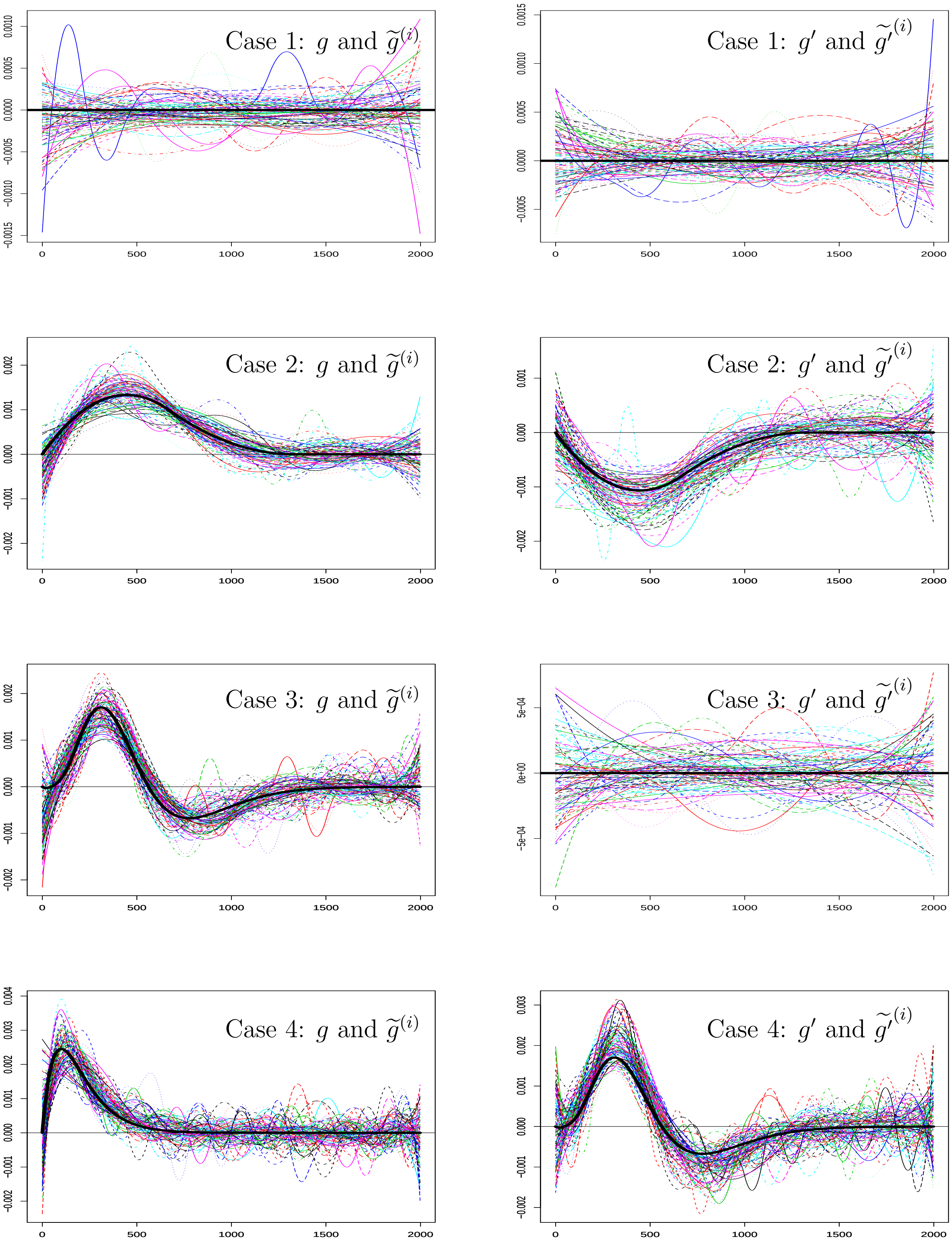

Figure 4: True profiles (g, g') [thick lines] and estimated profiles $(\widetilde{g}^{(i)}, \widetilde{g'}^{(i)})$ for each of the 100 simulations (one "color" per simulation), for case 1 (top) to case 4 (bottom).

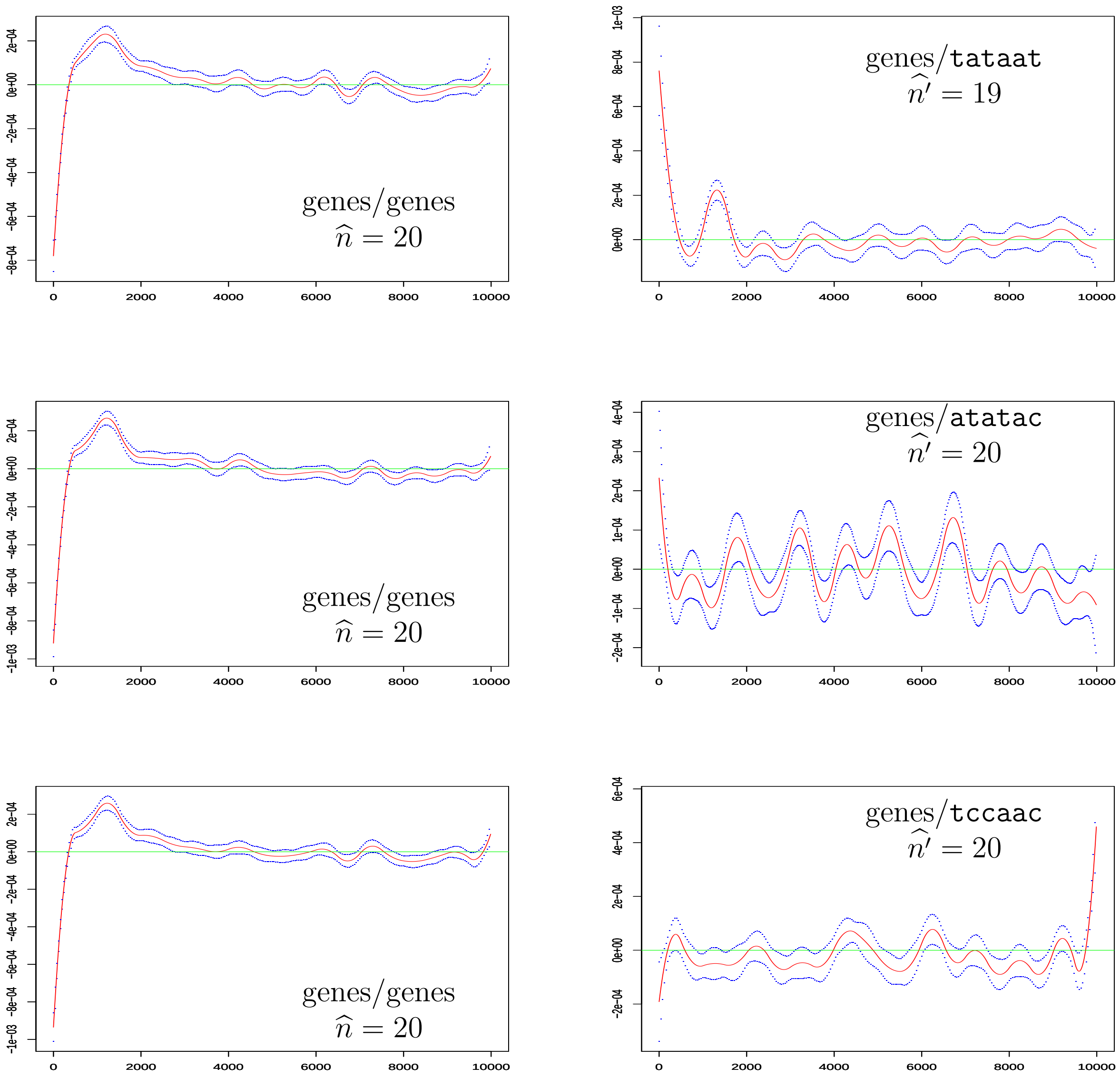

Figure 5: Estimated profiles $\widetilde{g}$ (left) and $\widetilde{g'}$ (right) obtained while analyzing the *E. coli* gene process with respect to the occurrences of `tataat` (top), `atatac` (middle) or `tccaac` (bottom). The dotted blue lines represent the 95% estimates.

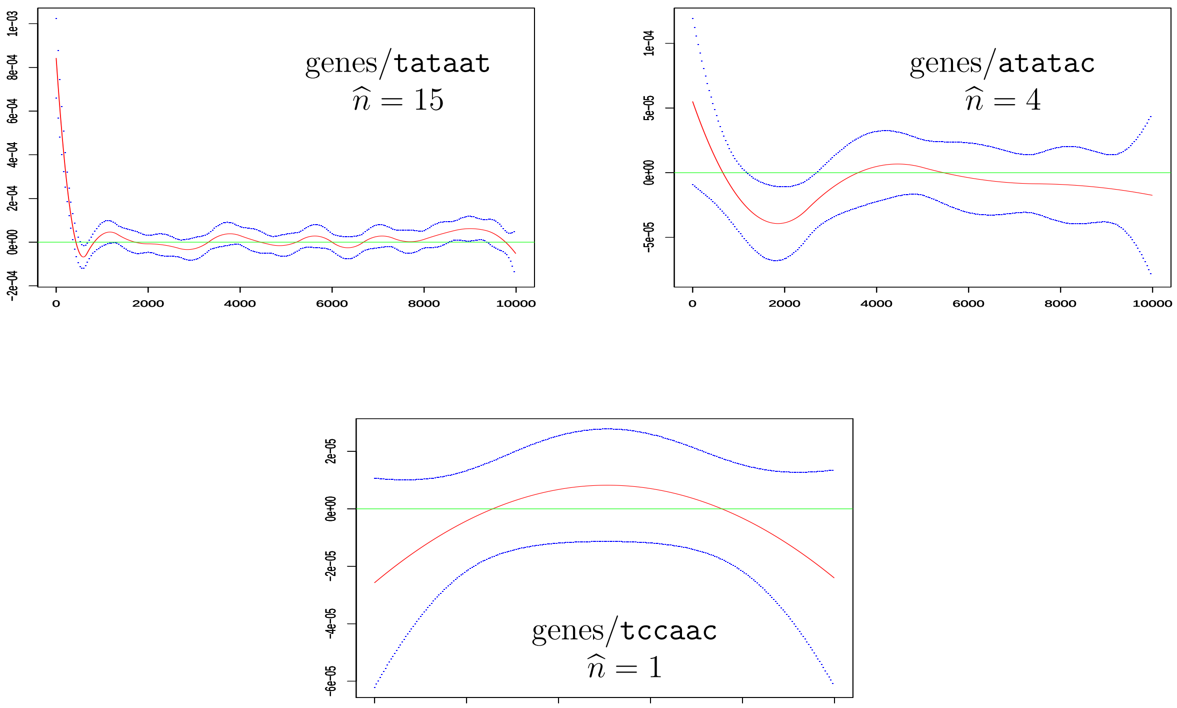

Figure 6: Estimated profiles $\widetilde{g}'$ obtained while analyzing the *E. coli* gene process with respect to the occurrences of `tataat`, `atatac` or `tccaac` under the reduced model ($g \equiv 0$). The dotted blue lines represent the 95% estimates.

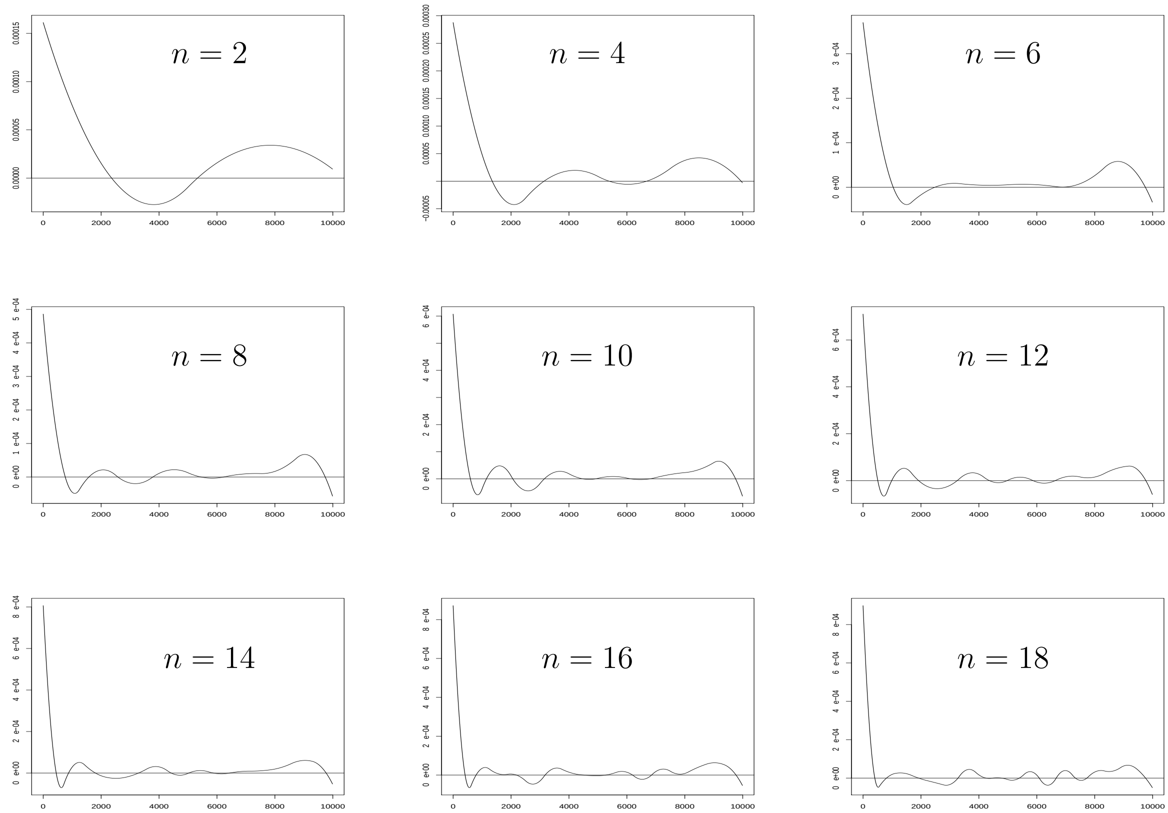

Figure 7: Estimated profiles $\widehat{g'}_{n'}$ for $n' \in \{2, 4, 6, \ldots, 18\}$ obtained while analyzing the *E. coli* gene process with respect to the occurrences of `tataat` under the reduced model ($g \equiv 0$). (AIC chooses $n' = 15$.)

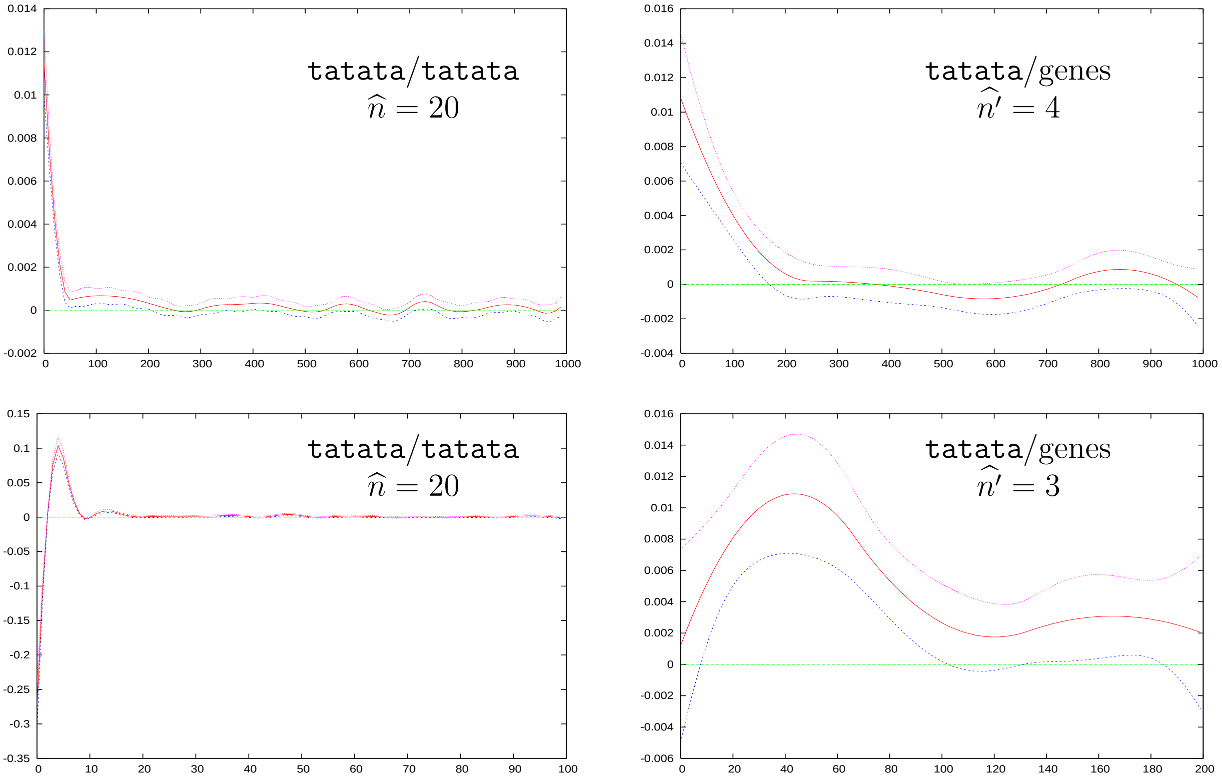

Figure 8: Estimated profiles $\widetilde{g}$ (left) and $\widetilde{g'}$ (right) obtained while analyzing the dependence between the occurrences of the `tatata`-like motif and the end of preceding genes in the first chromosome of *S. cerevisiae*. Two values of the profiles' support have been considered : $D = D' = 1000$ (top) and $D = 100$, $D' = 200$ (bottom). Profiles are drawn with their 95% estimates (dotted lines).

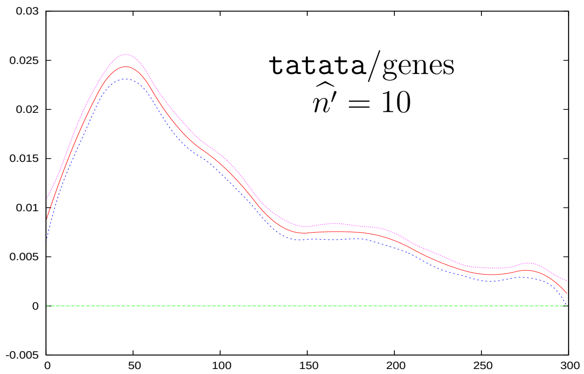

Figure 9: Estimated profile $\widetilde{g'}$, with its 95% estimate, obtained while analyzing the dependence between the occurrences of the `tatata`-like motif and the end of preceding genes in the 17 chromosomes of *S. cerevisiae* (direct strands).