Algorithms for Molecular Biology

BioMed Central

Research

Open Access

Pattern statistics on Markov chains and sensitivity to parameter estimation

Grégory Nuel*

Address: Laboratoire Statistique et Génome, University of Evry, CNRS (8071), INRA(1152), 523, place des terrasses de l'Agora, 91034 Evry CEDEX, France

Email: Grégory Nuel* - nuel@genopole.cnrs.fr

* Corresponding author

Published: 17 October 2006

Algorithms for Molecular Biology 2006, 1:17 doi:10.1186/1748-7188-1-17

Received: 07 April 2006
Accepted: 17 October 2006

This article is available from: http://www.almob.org/content/1/1/17

Abstract

Background: In order to compute pattern statistics in computational biology a Markov model is commonly used to take into account the sequence composition. Usually its parameter must be estimated. The aim of this paper is to determine how sensitive these statistics are to parameter estimation, and what are the consequences of this variability on pattern studies (finding the most over-represented words in a genome, the most significant common words to a set of sequences,...).

Results: In the particular case where pattern statistics (overlap counting only) computed through binomial approximations we use the delta-method to give an explicit expression of σ, the standard deviation of a pattern statistic. This result is validated using simulations and a simple pattern study is also considered.

Conclusion: We establish that the use of high order Markov model could easily lead to major mistakes due to the high sensitivity of pattern statistics to parameter estimation.

Background

In order to study pattern occurrences in biological sequences, simple frequencies are not relevant in most cases because of pattern overlapping structure as well as composition bias in the sequences. A common workaround consists to compute the significance of an observation assuming the sequence $X = X_1 \dots X\ell$ over the finite alphabet $\mathcal{A}$. (size k) is generated according to an order $m \geq 1$ homogeneous, stationary and ergodic Markov model. Let π (size k^{m+1}) defined by

$$\pi(w, a) = (X_{m+1} = a | X_1 \dots X_m = w) \qquad \forall (w, a) \in \mathcal{A}^m \times \mathcal{A} \quad (1)$$

be the parameter of this Markov model, Π its transition matrix (note that we have $\Pi = \pi$ only if $m = 1$) and μ its stationary distribution (defined by $\mu \times \Pi = \mu$).

We then introduce the pattern statistic defined by

$$S = \begin{cases} -\log_{10} \mathbb{P}(N \geq N_{\text{obs}}) & \text{if } N_{\text{obs}} \geq \mathbb{E}[N] \\ \log_{10} \mathbb{P}(N \leq N_{\text{obs}}) & \text{if } N_{\text{obs}} < \mathbb{E}[N] \end{cases} \qquad (2)$$

where N is the random number of overlapping occurrences (*i. e.* X = aabababa contains three overlapping occurrences of aba but only two non-overlapping ones) of a given fixed pattern on the random sequence X and N_{obs} is an observation.

When π is known (and hence μ), several statistical methods are available to compute S: exact computations [1-4], Gaussian [5,6], binomial [7,8], compound Poisson [9-11] or large deviations approximations [12]. But in general, the parameter π is not available and must be estimated. Let us denote by $\mathbf{N}_0$ (resp. $\mathbf{N}_1$) the (overlap) frequencies of all words of size m (resp. $m + 1$) in the sequence $Y = Y_1 \ldots Y_n$, then the Maximum-Likelihood Estimator (MLE) of π is given by

$$\hat{\pi}(w,a) = \frac{\mathbf{N}_1(wa)}{\sum_{b\in\mathcal{A}} \mathbf{N}_1(wb)} \quad \forall(w,a) \in \mathcal{A}^m \times \mathcal{A} \qquad (3)$$

and the MLE of μ (as a function of π) is therefore defined by $\hat{\mu} \times \hat{\Pi} = \hat{\mu}$ where $\hat{\Pi}$ is the transition matrix associated to $\hat{\pi}$

We introduce now the following estimators

$$\mu_{\mathbf{N}}(w) = \frac{\mathbf{N}_0(w)}{n-m+1} \quad \text{and} \quad \pi_{\mathbf{N}}(w,a) = \frac{\mathbf{N}_1(wa)}{\mathbf{N}_0(w)} \quad \forall(w,a) \in \mathcal{A}^m \times \mathcal{A} \qquad (4)$$

which are known to be asymptotically equivalent with the MLE when n is large.

The quality of parameter estimation depends both on the number of parameters to estimate (k^{m+1} for an order m Markov model) and of the length (n) of the homogeneous sequence used for their estimation. When the same sequence (or set of sequences) is used both for observed frequencies and parameter estimation, m should not be greater than $h - 2$ for a pattern of length h (as else, the observed frequency of the pattern will be included in the model). As literature often suggests to use the highest possible order, it is hence common to consider $m = 6$ or more (for a DNA pattern of size $h \geq 8$). Moreover, because of the homogeneity assumption of the model, the considered genomes have often to be segmented first. As a result, the sequences length used for parameter estimations are often dramatically reduced by such segmentation (*e. g.* $n = 10^5$ to $n = 10^6$ at the very best for DNA sequences). It is hence quite common to encounter high order Markov models estimated on rather short sequences which could result in high sensitivity to parameter estimation.

Considering that Y is generated through a Markov model of parameter π, the main goal of this paper is to study the distribution of $S_{\mathbf{N}}$, the statistic S computed using the estimators $\mu_{\mathbf{N}}$ and $\pi_{\mathbf{N}}$, and the consequences of its variability in projects using pattern statistics. We first present in details how the delta-method can be used to get a Gaussian approximation for the distribution of $S_{\mathbf{N}}$ (using a binomial approximation to compute the pattern statistics). Then these approximations are validated through simulations and, at last, we consider a classical pattern study (finding the most over-represented patterns of a given size) and we evaluate the detrimental effect of parameter estimations both in terms of true positive rate and rank accordance.

Materials and methods

Distribution of N = (N$_0$, N$_1$)

As the estimators defined in (4) are expressed as functions of $\mathbf{N}_0$ and $\mathbf{N}_1$ we first study their distribution. Using a Gaussian approximation, we have

$$\mathcal{L}\left(\underbrace{\begin{bmatrix}\mathbf{N}_0\\ \mathbf{N}_1\end{bmatrix}}_{\mathbf{N}}\right) \simeq \mathcal{N}\left(\underbrace{\begin{bmatrix}\mathbf{E}_0\\ \mathbf{E}_1\end{bmatrix}}_{\mathbf{E}}, \underbrace{\begin{bmatrix}\mathbf{C}_{0,0} & \mathbf{C}_{0,1}\\ \mathbf{C}_{1,0} & \mathbf{C}_{1,1}\end{bmatrix}}_{\mathbf{C}}\right) \qquad (5)$$

where, for $i, j \in \{0, 1\}$, $\mathbf{E}_i \in \mathbb{R}^{d_i}$, and $\mathbf{C}_{i,j} \in \mathbb{R}^{d_i} \times \mathbb{R}^{d_j}$ with $d_i = k^{m+i}$. One can note that $\mathbf{C}_{0,0}$ and $\mathbf{C}_{1,1}$ are symmetric, and ${}^t(\mathbf{C}_{1,0}) = \mathbf{C}_{0,1}$ (where ^{t} is the matrix transpose operator).

In the stationary case, exact expression of $\mathbf{E}$ and $\mathbf{C}$ can be computed according to [5].

Expectation is simply given $\forall w \in \mathcal{A}^m$ by

$$\mathbf{E}_0(w) = (n - m + 1)\, \mu(w) \qquad \mathbf{E}_1(wa) = (n - m)\, \mu(w)\Pi(w, a) \qquad \forall(w, a) \in \mathcal{A}^m \times \mathcal{A} \qquad (6)$$

In order to give more fluidity to this paper, the expression of the covariance matrix $\mathbf{C}$ have been moved in appendix A. Let us remark, before going forward that substituting $\mathbf{N}$ by $\mathbf{E}$ in (4) immediately gives

$$\mu_{\mathbf{E}} = \mu \quad \text{and} \quad \pi_{\mathbf{E}} = \left(1 - \frac{1}{n-m+1}\right)\pi \qquad (7)$$

Delta method

Let us start with a simple case. We consider a single pattern which is over-represented (seen more than expected) so we have

$$S_{\mathbf{N}} = -\log_{10} F^{+}(\mathbf{N}) \quad \text{with} \quad F^{+}(\mathbf{N}) \triangleq \mathbb{P}_{\mu_{\mathbf{N}},\pi_{\mathbf{N}}}(N \geq N_{\text{obs}}) \qquad (8)$$

where the function F^+ also depends on the sequence length ℓ and the considered pattern.

If F^+ is differentiate, the delta-method (a simple first order Taylor expansion around $\mathbf{N} = \mathbf{E}$, see [13]) provides the following approximation:

$$S_{\mathbf{N}} \simeq -\log_{10} F^{+}(\mathbf{E}) - \frac{{}^{t}(\mathbf{N}-\mathbf{E})\nabla F^{+}(\mathbf{E})}{\ln(10)F^{+}(\mathbf{E})} \qquad (9)$$

and hence, using (7) we have

$$S_{\mathbf{N}} \simeq S - \frac{{}^{t}(\mathbf{N}-\mathbf{E})\nabla F^{+}(\mathbf{E})}{\ln(10)F^{+}(\mathbf{E})} \qquad (10)$$

for n large enough. The distribution of $\hat{S}$ is therefore approximated by

$\mathcal{L}\ (S_{\mathbf{N}}) \simeq \mathcal{N}\ (S, \sigma^2)$ (11)

with

$$\sigma = \frac{\sqrt{{}^{t}\nabla F^{+}(\mathbf{E}) \times \mathbf{C} \times \nabla F^{+}(\mathbf{E})}}{\ln(10)F^{+}(\mathbf{E})} \qquad (12)$$

In consequence, computing σ requires both to compute **C** (done in appendix A) and ∇F^{+} (**E**).

Single pattern

The exact expression of F^{+} is computable through many different methods [1-4] but is too much complicated to derive explicitly ∇F^{+}. To overcome this problem, we propose to consider an approximation of F^{+}. As said in introduction, many kind of approximations are available (Gaussian, binomial, compound Poisson or large deviations). In this paper, we have chosen to use a binomial approximation as it provides an expression which is analytically differentiable and is known to be a good heuristic to the problem [8].

For a single non-degenerate pattern (*i.e.* a simple word) $W = w_1 \dots w_h$ ($w_i \in \mathcal{A}$) with $h \geq m - 1$ we first denote by

$P(\mathbf{N}) = \mu_{\mathbf{N}}\ (w_1 \dots w_m) \times \pi_{\mathbf{N}}\ (w_1 \dots w_m, w_{m+1}) \times \dots \times \pi_{\mathbf{N}}\ (w_{h-m} \dots w_{h-1}, w_h)$ (13)

the probability for W to occur at a given position in the sequence and then we get

$$F^{+}(\mathbf{N}) \simeq \mathbb{P}(\mathcal{B}(\ell_h, P(\mathbf{N})) \geq N_{\text{obs}}) = \frac{\beta(P(\mathbf{N}), N_{\text{obs}}, \ell_h - N_{\text{obs}} + 1)}{\beta(N_{\text{obs}}, \ell_h - N_{\text{obs}} + 1)} \qquad (14)$$

where $\mathcal{B}$ denotes the binomial distribution, with $\ell_h = \ell - h + 1$ and where the β functions (complete and incomplete) and their relation to the binomial cumulative distribution function are described in appendix B.

Note that if we consider non-overlapping occurrences instead of overlapping ones, we can still use a binomial approximation for the distribution of N, but the expression of $P(\mathbf{N})$ is more complicated as it involves the autocorrelation polynome of the pattern [14]. This point is not developed in this paper.

Replacing $\mu_{\mathbf{N}}$ and $\pi_{\mathbf{N}}$ by their expression easily gives

$$P(\mathbf{N}) = \frac{1}{n-m+1} \prod_{w \in \mathcal{A}^m} \frac{\prod_{a \in \mathcal{A}} \mathbf{N}_1(wa)^{A_1(wa)}}{\mathbf{N}_0(w)^{A_0(w)}} \qquad (15)$$

where $A_1(wa)$ counts occurrences of the word wa in $W = w_1 \dots w_h$ and $A_0\ (w)$ counts occurrences of the word w in $w_2 \dots w_{h-1}$. Note that in the particular case where $h = m - 1$, all $A_0\ (w)$ are null and we simply get $(n - m + l) \times P\ (\mathbf{N}) = \mathbf{N}_1\ (W)$.

Using the derivative properties of the incomplete beta function (see appendix B for more details) we hence get

$$\nabla F^{+}(\mathbf{N}) \simeq \frac{P(\mathbf{N})^{N_{\text{obs}}-1}(1-P(\mathbf{N}))^{\ell_h - N_{\text{obs}}}}{\beta(N_{\text{obs}}, \ell_h - N_{\text{obs}} + 1)} \times \nabla P(\mathbf{N}) \qquad (16)$$

so all we need is to compute $\nabla P(\mathbf{N})$.

For all $(w, a) \in \mathcal{A}^m \times \mathcal{A}$ we have

$$\frac{\partial P(\mathbf{N})}{\partial \mathbf{N}_0(w)} = -\frac{A_0(w)}{\mathbf{N}_0(w)} \times P(\mathbf{N}) \qquad (17)$$

and

$$\frac{\partial P(\mathbf{N})}{\partial \mathbf{N}_1(w)} = -\frac{A_1(wa)}{\mathbf{N}_1(wa)} \times P(\mathbf{N}) \qquad (18)$$

If we denote by

$P = \mu\ (w_1 \dots w_m) \times \pi\ (w_1 \dots w_m, w_{m+1}) \times \dots \times \pi\ (w_{h-m} \dots w_{h-1}, w_h)$ (19)

the *true* probability for W to occur at a given position in the sequence X then we get, using (7) in (13), that

$$P(\mathbf{E}) = p \times \left(1 - \frac{1}{n-m+1}\right)^{h-m} \simeq p \qquad (20)$$

for n large enough. We hence get

$$\nabla F^{+}(\mathbf{E}) \simeq \frac{p^{N_{\text{obs}}}(1-p)^{\ell_h - N_{\text{obs}}}}{\beta(N_{\text{obs}}, \ell_h - N_{\text{obs}} + 1)} \times \mathbf{G} \qquad (21)$$

where ${}^{t}\mathbf{G} = [{}^{t}\mathbf{G}_0\ {}^{t}\mathbf{G}_1]$ is defined by

$$\mathbf{G}_0(w) = -\frac{A_0(w)}{\mathbf{E}_0(w)} \quad \text{and} \quad \mathbf{G}_1(wa) = -\frac{A_1(wa)}{\mathbf{E}_1(wa)} \qquad (22)$$

Using equation (12) we finally get

$$\sigma \simeq Q^{+}\sqrt{{}^{t}\mathbf{G}\times\mathbf{C}\times\mathbf{G}} \qquad (23)$$

where

$$Q^{+} = \frac{p^{N_{\text{obs}}}(1-p)^{\ell_h - N_{\text{obs}}}}{\ln(10)\beta(p, N_{\text{obs}}, \ell_h - N_{\text{obs}} + 1)} \qquad (24)$$

and then, a computation of σ is possible by plug-in. Without considering the computation of **E** and **C**, the complexity of this approach is $O(h)$ (where h is the size of the pattern).

When a degenerate pattern (finite set of words) is considered instead of a single word, it is easy to adapt this method by summing the contribution p of each word belonging to the pattern. This point is left to the reader.

Under-represented pattern

In the case of an under-represented pattern we have

$$S_{\mathbf{N}} = \log_{10} F^{-}(\mathbf{N}) \quad \text{with} \quad F^{-}(\mathbf{N}) \triangleq \mathbb{P}_{\mu_{\mathbf{N}},\pi_{\mathbf{N}}}(N \le N_{\text{obs}}). \qquad (25)$$

Using a binomial approximation we get

$$F^{-}(\mathbf{N}) \simeq \mathbb{P}(\mathcal{B}(\ell_h, P(\mathbf{N})) \le N_{\text{obs}}) = \frac{\beta^{-}(P(\mathbf{N}), N_{\text{obs}} + 1, \ell_h - N_{\text{obs}})}{\beta(N_{\text{obs}} + 1, \ell_h - N_{\text{obs}})} \qquad (26)$$

and, by the same method than in the over-represented case we finally have

$$\sigma \simeq Q^{-}\sqrt{{}^{t}\mathbf{G}\times\mathbf{C}\times\mathbf{G}} \qquad (27)$$

where

$$Q^{-} = \frac{p^{N_{\text{obs}}+1}(1-p)^{\ell_h - N_{\text{obs}} - 1}}{\ln(10)\beta^{-}(p, N_{\text{obs}} + 1, \ell_h - N_{\text{obs}})} \qquad (28)$$

Two distinct patterns

We consider now two patterns V and W instead of one and want to study the joint distribution of $S_{\mathbf{N}}$ (V) and $S_{\mathbf{N}}$ (W) their corresponding pattern statistics.

With a similar argument as in section "delta method", it is easy to show that

$$\mathcal{L}\left(\begin{bmatrix} S_{\mathbf{N}}(V) \\ S_{\mathbf{N}}(W) \end{bmatrix}\right) \simeq \mathcal{N}\left(\begin{bmatrix} S(V) \\ S(W) \end{bmatrix}, \begin{bmatrix} \sigma_V^2 & \sigma_{V,W} \\ \sigma_{V,W} & \sigma_W^2 \end{bmatrix}\right) \qquad (29)$$

where σ_V (resp. σ_W) is the standard deviation σ for the pattern V (resp. W) and where

$$\sigma_{V,W} = \frac{{}^{t}\nabla F_V^{\varepsilon}(\mathbf{E})\times\mathbf{C}\times\nabla F_W^{\eta}(\mathbf{E})}{\ln(10)F_V^{\varepsilon}(\mathbf{E})\times\ln(10)F_W^{\eta}(\mathbf{E})} \qquad (30)$$

where

$$\varepsilon \text{ (resp. } \eta) = \begin{cases} + & \text{if pattern } V \text{ (resp. } W) \text{ is over-represented} \\ - & \text{if pattern } V \text{ (resp. } W) \text{ is unter-represented} \end{cases}. \qquad (31)$$

And after using results of sections "single pattern" and "under-represented pattern" we finally get

$$\sigma_{V,W} = \left(Q_V^{\varepsilon} Q_W^{\eta}\right)\times\left({}^{t}\nabla\mathbf{G}_V \times\mathbf{C}\times\nabla\mathbf{G}_W\right) \qquad (32)$$

where Q_V^{ε} (resp. W) and $\mathbf{G}_V$ (resp. W) are the constant Q (Q^{+} and Q^{-}) and the vector **G** for the pattern V (resp. W).

Simulations

It is also possible to study the empirical distribution of a $S_{\mathbf{N}}$ (for one or more patterns) through simulations.

In order to do so, we first draw M independent sequences $Y^j = Y_1^j \ldots Y_n^j$ using an order m stationary Markov model of parameters π. Complexity of this step is $O(M \times n)$.

For each j we get the frequencies $\mathbf{N}^j = (\mathbf{N}_0^j, \mathbf{N}_1^j)$ (with complexity $O(n)$ for each sequence) of the words of size m and $m + 1$ in the sequence Y^j and use it to compute $S^j = S_{\mathbf{N}^j}$ (exact value or approximation). Complexity here depends on the statistical method used to compute S^j (*e.g.* $O(h)$ using a binomial approximation).

We now have a M – sample $S^1, \ldots, S^M$ of $S_{\mathbf{N}}$ from which we can easily estimate σ and thus, valid or invalid the approximation through the delta-method.

When used with large value of n (*e.g.* several millions or more), the complexity of this approach is slowed by the drawn of the sequences Y_j. It is therefore possible to improve the method by simulating directly the frequencies **N** through (5). As this approximation has a very small impact on the distribution of $S_{\mathbf{N}}$ (data not shown) it may dramatically speed-up the computations when considering large n or M. It is nevertheless important to point out that drawing a Gaussian vector size L requires to precompute the Choleski decomposition of its covariance matrix which could be a limiting factor when considering large L.

Results and discussion

Validation

Simple case

Let us start with a simple case: a binary alphabet $\mathcal{A}$ = {a, b} (k = 2) with an order m = 1 Markov model

$$\pi = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix} \qquad (33)$$

which stationary distribution is μ = (6/13,7/13) and we work on a sequence of length n = 10 000.

The first thing to do is to compute **E** and **C** (see appendix A for details).

Now, we consider the pattern W = ababa occurring N_{obs} = 1221 times in a sequence of length $\ell = n$ = 10 000. We have

$$p = \mu(\text{a})\ \Pi\ (\text{a,b})^2\ \Pi\ (\text{b,a})^2 = 8.142 \times 10^{-2} \qquad (34)$$

so $\mathbb{E}\,[N(\text{ababa})] = (\ell - 4)p = 813.8 \simeq 0.66 \times N_{\text{obs}}$ and hence the pattern is over-represented. Its statistic (using binomial approximation) is

$$S \simeq -\log_{10} \mathbb{P}(\mathcal{B}(\ell - 5 + 1, p) \geq N_{\text{obs}}) = 43.74285 \qquad (35)$$

We have

$$Q^+ = \frac{p^{N_{\text{obs}}-1}(1-p)^{\ell-4-N_{\text{obs}}}}{\ln(10)\beta(p, N_{\text{obs}}, \ell - 3 - N_{\text{obs}})} = 193.3258 \qquad (36)$$

and

$${}^t\mathbf{G}_0 = \begin{bmatrix} \dfrac{-1}{\mathbf{E}_0(\text{a})} & \dfrac{-2}{\mathbf{E}_0(\text{b})} \end{bmatrix} = \begin{bmatrix} -2.17\times10^{-5} & -3.71\times10^{-5} \end{bmatrix} \qquad (37)$$

and

$${}^t\mathbf{G}_1 = \begin{bmatrix} 0 & \dfrac{2}{\mathbf{E}_1(\text{ab})} & \dfrac{2}{\mathbf{E}_1(\text{ba})} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 6.19\times10^{-5} & 6.19\times10^{-5} & 0 \end{bmatrix} \qquad (38)$$

Finally, we get

$$\sigma = Q^+ \sqrt{{}^t\mathbf{G}\times\mathbf{C}\times\mathbf{G}} = 6.1020774 \qquad (39)$$

As our pattern statistics is the decimal logarithm of the p-value, σ = 6 means that the ratio of the estimated p-value over the true one could easily range from 10^{-12} ($10^{-2 \times \sigma}$) to 10^{12} ($10^{2 \times \sigma}$) which is huge.

We can see on fig. 1 the empirical distribution of S_N compared to the theoretical distribution. Even if the two distributions are closely related, an adjustment test (Kolmogorov-Smirnov) shows that they are different.

In the fig. 2 we compare σ to its estimator $\hat{\sigma}$ for several values of N_{obs}. We can see that our theoretical values of σ fits very well to the empirical ones.

The equation (39) gives an explicit expression of σ as a product of two terms. Once the pattern and the true parameter π are fixed, the first term (Q) depends only on ℓ and N_{obs} while the second one only depends on the length n of the sequence used for the parameter estimation (see appendix C for an explicit expression of σ in the particular case of an order 0 Markov model).

To study the variations of $\sigma(n)$ as a function of n we therefore need to study $\mathbf{G}(n)$ and $\mathbf{C}(n)$. Using equations (6) and (22) we get that

$$\mathbf{E}(n) = O(n) \quad \text{and} \quad \mathbf{G}(n) = O\left(\frac{1}{n}\right) \qquad (40)$$

Using equations (57) and (58) in appendix A we also get that $\mathbf{C} = \mathbf{M} + \mathbf{O} + {}^t\,\mathbf{E}\mathbf{E}$ with

$$\mathbf{M}(n) = O(n^2) \text{ and } \mathbf{O}(n) = O(n) \qquad (41)$$

so finally

$$\sigma(n) \simeq \tilde{\sigma}(n) = Q^+ \times \sqrt{A + \frac{B}{n}} \qquad (42)$$

for large n, with

$$A = \lim_{n\to+\infty} {}^t\mathbf{G}(\mathbf{C} - \mathbf{O})\mathbf{G} \qquad (43)$$

and

$$B = \lim_{n\to+\infty} n\times\ {}^t\mathbf{GOG} \qquad (44)$$

We can see on fig. 3 that $\tilde{\sigma}$ is not a very good approximation of σ for small n, but, as the approximation is far easier to compute (and trivial to invert) than the true value, this can be useful when we need to compute a minimum length n to obtain a given σ.

We also see on the same figure that σ grows rapidly when n decreases. For example, we get $\sigma \simeq 20$ for n = 5000 (while equation (35) gives $S \simeq 264.4$).

As we consider here a binary alphabet (k = 2) and a first order Markov model (m = 1) we have only $k^m(k - 1)$ = 2 parameters to estimate with a sample of size n = 5000 (so we have 2500 sample per parameter). Although this situ-

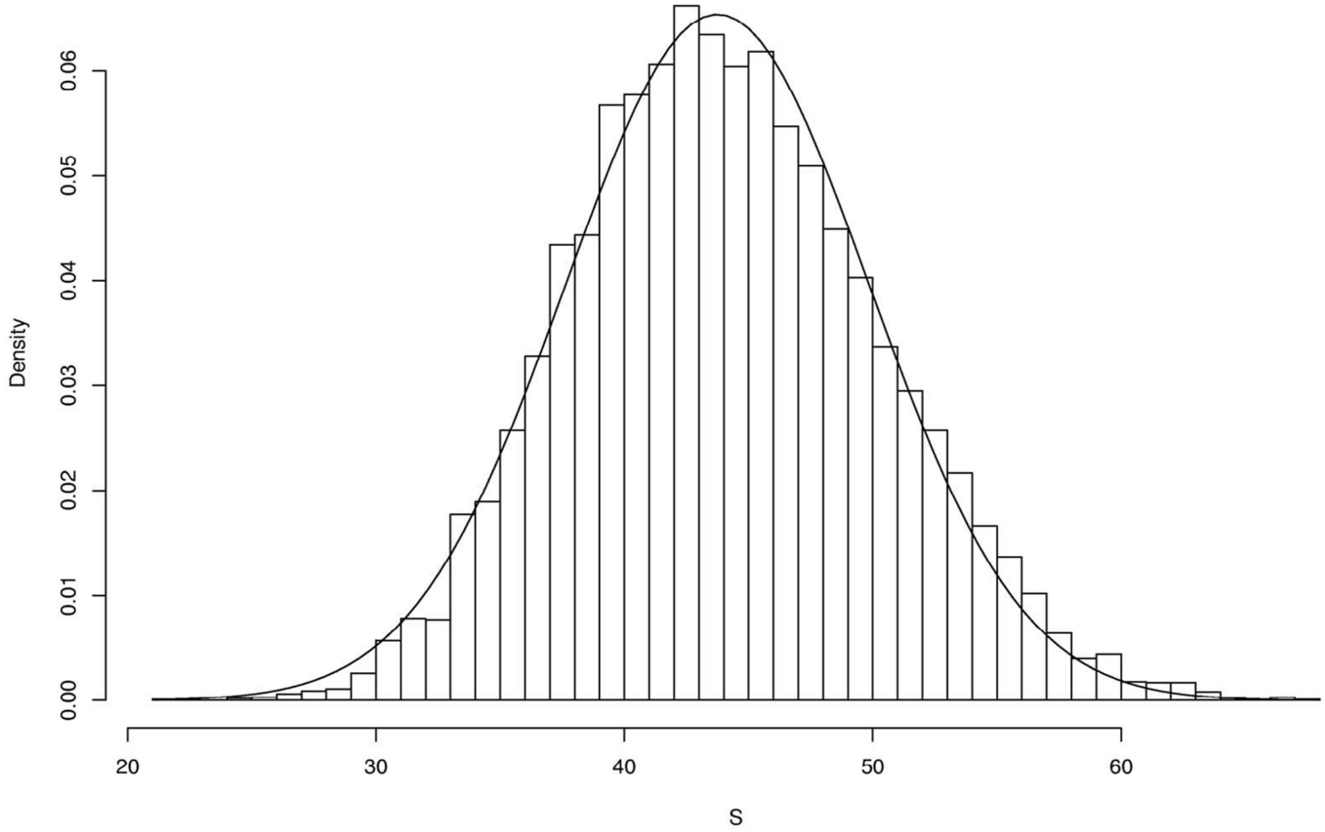

Figure 1 $\hat{S}$

Empirical and theoretical distributions of $\hat{S}$. A sample of size 10 000 have been used to get the empirical distribution. The solid line represents the density of $\mathcal{N}$ (S, σ^2). The adjustment test of Kolmogorov-Smirnov give D = 0.023 which corresponds to a p-value of $p = 5.3 \times 10^{-5}$. N_{obs} = 1221 and $n = \ell = 10\ 000$.

ation seems quite comfortable, the sensitivity to parameter estimation appears in fact to be so large that we could have a factor 10^{40} between the true p-value and its estimate.

Practical case

We have seen with our first example that our approximation works very well in a simple case. Will this hold with more practical cases?

To answer this question, let us consider the following experimental design:

- one pattern: W = acgtacgt;

- two genomes: *Escherichia coli* K12 ($\ell = n = 4639675$) and *Mycoplasma genitalium* ($\ell = n = 580076$);

- five Markov orders: $m = 1$ to $m = 5$ (larger m are not considered since the computation of **C** becomes then intractable).

As the sequence lengths and compositions of the two considered genomes differ a lot, we have to take a different value of N_{obs} for each organism: $N_{obs} = 30$ for *M. genitalium* and $N_{obs} = 150$ for *E. coli*. Proceeding as indicated in section "simulations", we use the algorithm 1 for each experiment.

Algorithm 1 simulations for one experiment in the practical case

1: estimate the order m parameter π (and μ) from the original sequence. Although these parameters are estimated, they are considered as the true parameters;

2: compute $S = -\log_{10} (N \geq N_{obs})$;

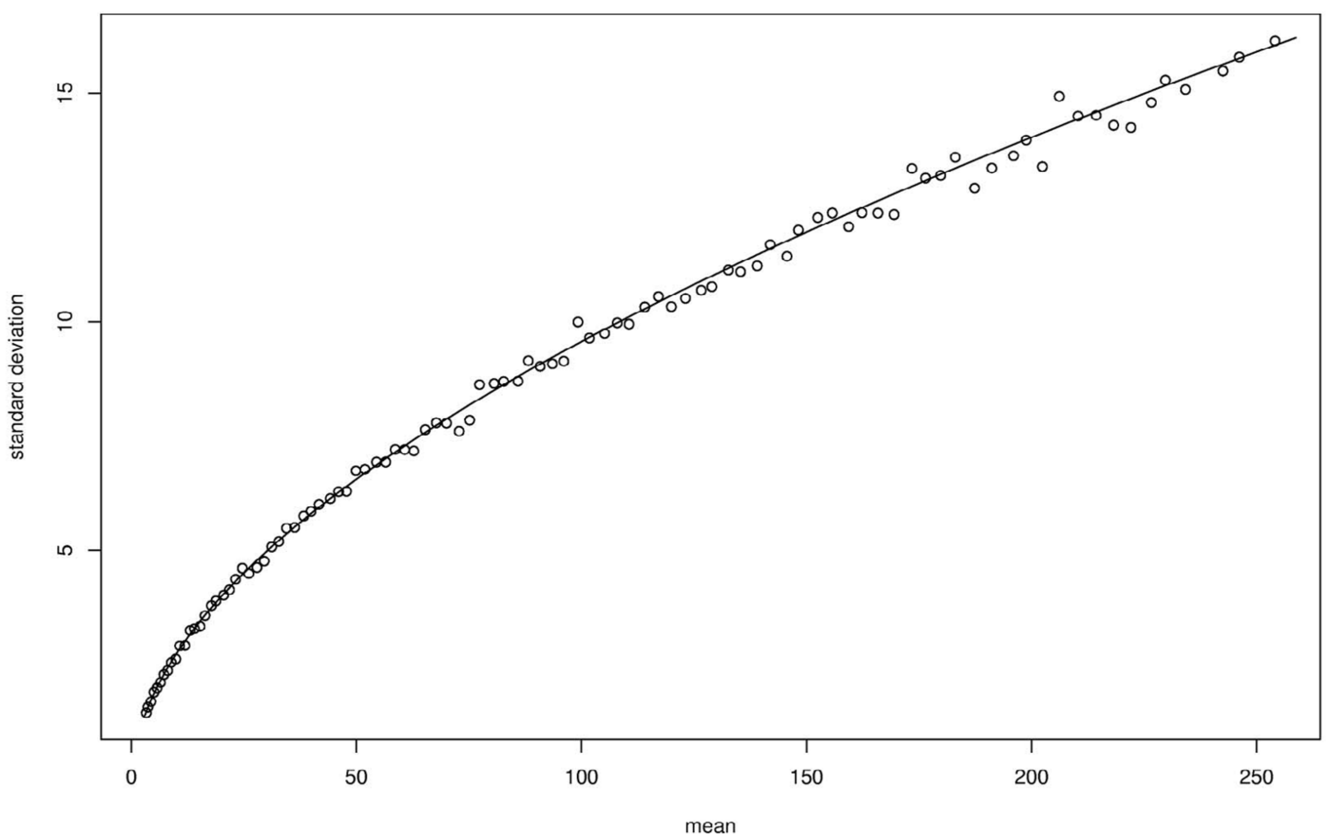

Figure 2 $\hat{\sigma}$

Comparison of σ and $\hat{\sigma}$. $\hat{\sigma}$ is estimated with a sample of size 1 000 and N_{obs} takes its values from 900 to 1 900. The solid line represents the theoretical values and the circles the empirical ones. The statistic S is used on the x-axis. $n = \ell = 10\ 000$.

```
compute $\sigma$ using approximation (23)
for $j = 1 \dots 1\ 000$ do
    draw a random sequence $Y = Y_1 \dots Y_n$ according to and order $m$ stationary Markov model of parameter $\pi$;
    compute $\mathbf{N}$ the frequency vector of all size $m$ and size $m + 1$ words in $Y$;
    compute $S^j = S_{\mathbf{N}} = -\log_{10} (N \geq N_{obs})$;
end for
compute $\hat{S}$ (resp. $\hat{\sigma}$) the mean (resp. standard deviation) of the sample $S^1, \dots, S^j$.
```

We can see on table 1 the results for *E. coli*. For each Markov model considered, our approximation of σ is very close to the empiric ones and, as with figure 1, the Gaussian distribution fit well to the empiric one (data not shown). Table 2 shows the same behaviour with *M. genitalium* except for $m = 5$ where $\hat{\sigma}$ differs slightly more than in the other cases from its theoretical value. To understand this phenomenon, let us first recall the expression of $P(\mathbf{N})$ for $m = 5$ using equation (15):

$$P(\mathbf{N}) = \frac{\mathbf{N}_1(\text{agctac}) \times \mathbf{N}_1(\text{gctacg}) \times \mathbf{N}_1(\text{ctacgt})}{(\ell - m + 1) \times \mathbf{N}_0(\text{gctac}) \times \mathbf{N}_0(\text{ctacg})}$$

and as $(\mathbf{N}_1\ (\text{agctac}) = 0) \simeq 2.26 \times 10^{-6}$, $(\mathbf{N}_1\ (\text{gctacg}) = 0) \simeq 1.35 \times 10^{-1}$ and $(\mathbf{N}_1\ (\text{ctacgt}) = 0) \simeq 1.24 \times 10^{-4}$ we will have $P(\mathbf{N}) = 0$ roughly 14% of the time. This happened 123 times in our sample of size 1 000, each time preventing to compute $S_{\mathbf{N}}$. The sample is hence biased and $\hat{S}$ and $\hat{\sigma}$ are therefore not accurate.

What happen now if we use another statistical method to compute the pattern statistics. As the binomial approximation is supposed to be close to the exact solution, we expect the standard deviation obtained with other statisti-

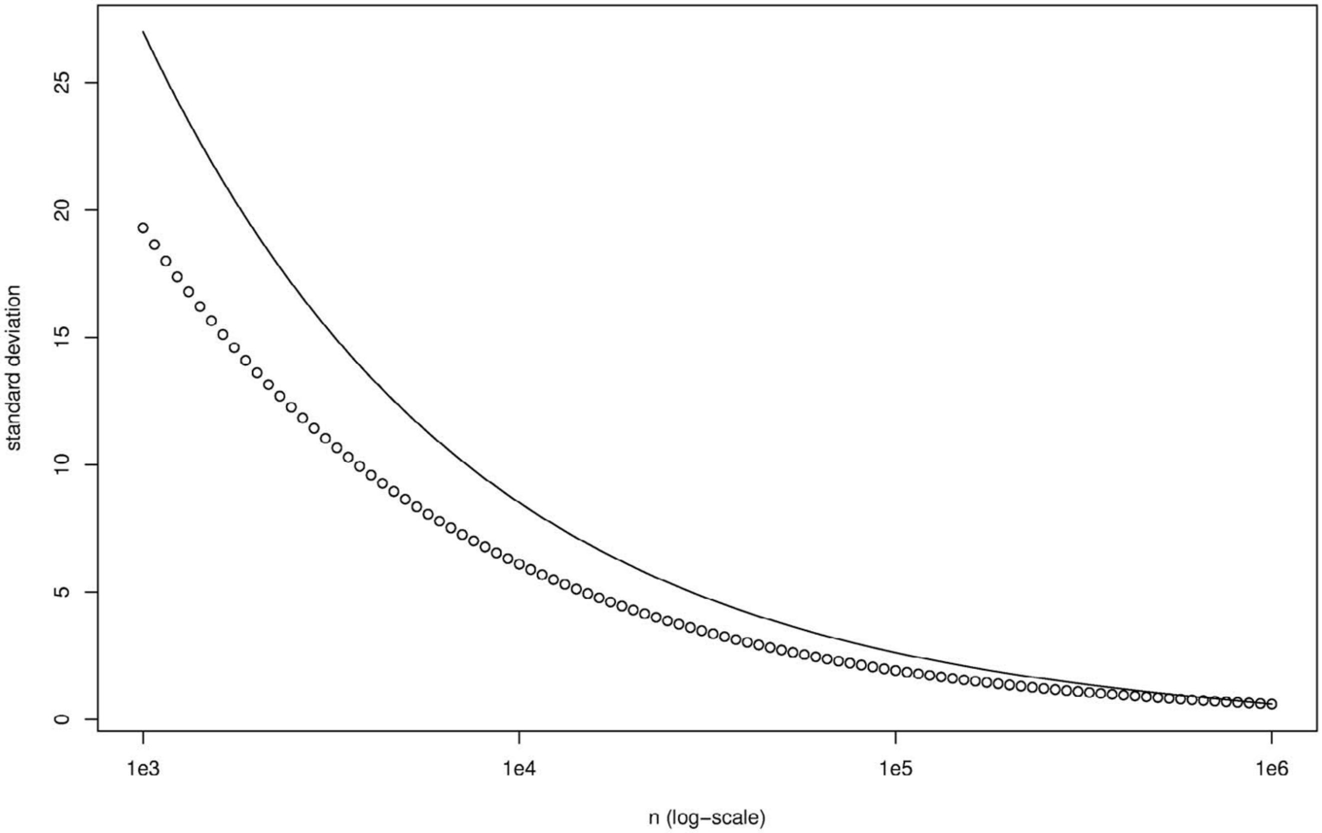

Figure 3

Comparison of $\sigma(n)$ and $\tilde{\sigma}(n)$. The circles reprensent $\sigma(n)$ and the solid line $\tilde{\sigma}(n)$. $n_\infty = 10^6$ have been used to compute the value of *A* and *B*. $N_{\text{obs}} = 1221$ and $\ell = 10\,000$.

cal methods to remain close to σ. In table 3, we compare the empirical results using binomial approximations (like above) but also compound Poisson or large deviations approximations. Both empirical means and standard deviations are close to the theoretical ones thus validating the method.

Choice of a Markov model order

Through the computation of σ we can measure the sensitivity of pattern statistics to parameter estimations. A very natural question is then, how this variability could affect a pattern statistic study, and, as this variability grows with the Markov model order, how to choose this parameter.

Table 1: Comparison of theoretical and empirical pattern statistic mean and standard deviation on *Escherichia coli* K12.

m	S	σ	$\hat{S}$	$\hat{\sigma}$
1	35.57	0.28	35.57	0.27
2	31.61	0.49	31.60	0.50
3	46.75	1.04	46.77	1.03
4	45.33	1.74	45.32	1.81
5	62.27	3.45	62.36	3.34

We consider the pattern W = acgtacgt with $N_{\text{obs}} = 150$. The sequence length is $\ell = 4639675$, we use an order m Markov model and a sample of size $M = 1\,000$.

Table 2: Comparison of theoretical and empirical pattern statistic mean and standard deviation on *Mycoplasma genitalium*.

m	S	σ	$\hat{S}$	$\hat{\sigma}$
1	42.48	0.38	42.47	0.40
2	44.62	0.78	44.62	0.81
3	55.96	1.49	56.02	1.52
4	55.06	3.39	55.48	3.48
5	56.49	10.35	57.21*	9.09*

We consider the pattern W = acgtacgt with $N_{\text{obs}} = 30$. The sequence length is ℓ = 580 076, we use an order m Markov model and a sample of size M = 1 000. (*) for 123 terms in the sample we got P (**N**) = 0 and hence, $S_{\mathbf{N}}$ was not computed.

We propose here to consider the case of a very simple pattern study: we want to find the 100 most over-represented octamers (DNA words of size 8) in a given genome. Assuming the true parameter π (and hence μ) is known, we can compute REF = $\{W_1,..., W_{100}\}$, the list of these words (ordered by decreasing statistics, so that the most over-represented one is the first one).

For each estimates $\hat{\mu}$ and $\hat{\pi}$, we can compute $\widehat{\text{REF}}$ the 100 most over-represented octamers in the genome using the statistic $\hat{S}$ and compare it to the truth. In order to do so, we first compute the true positive rate (TP rate) defined by the rate of common words in $\widehat{\text{REF}}$ and REF, and the rank accordance rate (RA rate) defined by the Kendall's tau [[15], Chapter 13] between S and $\hat{S}$ ranks of $\{\widehat{\text{REF}} \cup \text{REF}\}$. Such statistic is in the range [-1,1] and has the value 1 for the complete rank accordance and the value -1 for the complete rank discordance.

As in the section "practical case", we consider two genomes: *Escherichia coli* K12 ($\ell = n = 4639675$) and *Mycoplasma genitalium* ($\ell = n = 580076$). For each Markov model order m from 1 to 6, we estimate π on the sequence (by maximum of likelihood), compute the REF list and then draw a sample of $\widehat{\text{REF}}$ from which we get estimates for the expectation of TP and RA rates.

Results are given in tables 4 and 5. We can see that, surprisingly, the TP rate could be very low even for long genome such as *E. coli* when high order Markov model (m = 6) are used. Of course, these rates are even worse on *M. genitalium* whose genome is ten times smaller than the first one. It is also clear that the RA rate is more affected by the variability induced by parameter estimation than the TP rate.

Based on these results, we conclude that our pattern study requires a sample size per free parameter of at least a few thousands if we want reliable results. In our examples this has for consequence that the Markov order should not be greater than 4 (or 5 at the very most) for *E. coli* and 3 (or 4 at the very most) on *M. genitalium* without resulting in important errors.

Conclusion

The delta-method allows us to approximate the distribution of $\hat{S}$ by a Gaussian distribution. This first requires to compute the expectation and covariance matrix of frequencies and then to study the derivative of a function which is specific of the method used to compute the pattern statistics. In the case of the binomial approximations, we have found an explicit expression of σ the standard deviation of $\hat{S}$.

It is clear that our approximation of σ using the delta-method relies one two major assumptions: 1) the distribution of **N** is Gaussian; 2) F^+ is regular enough (*e.g.* not

Table 3: Comparison of theoretical and empirical pattern statistics mean and deviation on *Mycoplasma genitalium*.

theoretical		binomial		compound Poisson		large deviations	
S	σ	$\hat{S}$	$\hat{\sigma}$	$\hat{S}$	$\hat{\sigma}$	$\hat{S}$	$\hat{\sigma}$
55.96	1.49	56.05	1.47	55.42	1.45	54.27	1.43

We consider the pattern W = acgtacgt with $N_{\text{obs}} = 30$. The sequence length is ℓ = 580076, we use an order m = 3 Markov model and a sample of size M = 1 000. The pattern statistics are computed (from left to right) through binomial, compound Poisson or large deviations approximations.

Table 4: Mean true positive rate and rank accordance rate in *Escherichia coli* K12.

Markov order	1	2	3	4	5	6
TP rate	99.0%	98.0%	97.9%	94.4%	82.1%	47.6%
RA rate	99.0%	95.5%	91.5%	83.9%	68.0%	36.5%
× 10^3	383.33	95.83	23.96	5.99	1.50	0.37

Both quantities are estimated with 1 000 simulations. We consider the 1 00 most over-represented octamers, the sequence length is ℓ = 4639675. The last row gives the sample size per free parameter (length *n* of the sequence divided by the number $k^m(k - 1)$ of parameters).

too steep) around **E**. When m grows, **E** closes to the boundary of the definition range of F^+ hence degrading assumption 2. Moreover, it is well known that Gaussian approximations for word frequencies become weaker when the expected numbers of their occurrences become smaller, thus degrading assumption 1. It is therefore obvious that our approximation of σ will get less and less reliable as m grows.

However, the approximation of σ has been validated through simulations and appears to be very reliable (even for m = 5 or 6). As pattern statistics computed through binomial approximations are close to the exact statistics [8], the value of σ should not differ a lot when another statistical method is used. We have compared our approximations to the empiric distribution obtained using compound Poisson and large deviations approximations and, as expected, our approximations remains quite reliable even for these statistical methods.

The variability due to parameter estimation is of course related to the Markov model order m and to the size k of the alphabet (as we have k^{m+1} parameters for this model) and to the length n of the sequence used for this estimation. For example, considering an order m = 6 model with n = 4639675 (*Escherichia coli* K12 complete genome) requires to estimate 3 × 4^6 = 4096 free parameters which results roughly in 400 observation per free parameter. Although this situation seems quite comfortable, we have seen with our simulations that it leads an unacceptable variability for pattern statistics.

As literature often advices to use the highest possible Markov order for a given pattern problem (which means $m = h - 2$ for pattern of size h) it is easy to understand that such a practice could have very detrimental effects on the computed statistics unless huge data are available for estimation purpose. Even if we consider the more reasonable attitude to choose m using the classical framework of model selection (*e.g.* using the Akaike Information Criterion – AIC –) we get m = 5 for *Mycoplasma genitalium* and m = 6 for *Escherichia coli* K12 hence resulting in both cases in the same catastrophic results in terms of false positive and even worse ones in terms of ranking.

Moreover, we assumed here that our model was homogeneous all along the considered sequences. This is obviously completely false when complete genomes are considered. So it is more likely that the sample size n would be far smaller than a million on classical pattern studies (even of human genomes for example). As a result, the variability we pointed out in this paper will have a considerable detrimental effect on most studies unless the Markov order is carefully set.

In order to do so, we advice to compute our approximation of σ each time a pattern statistic is produced and then to evaluate, either by simulation (like in this paper) or by a theoretical work the impact of this variability on the considered study.

Competing interests

The author declares that he has no competing interests.

Appendix A

We give here the expression of the covariance matrix **C** introduced in section "distribution of $\mathbf{N} = (\mathbf{N}_0, \mathbf{N}_1)$". The sequence Y (of length n) is generated by an homogeneous, stationary and ergodic order m Markov model of parameter π and stationary distribution μ. We want to compute the covariance of the vector **N** of random frequencies of size m and m + 1 words.

Table 5: Mean true positive rate and rank accordance rate in *Mycoplasma genitalium*.

Markov order	1	2	3	4	5	6
TP rate	95.5%	93.6%	90.4%	81.8%	66.0%	25.0%
RA rate	92.6%	85.4%	79.8%	66.5%	45.1%	11.0%
× 10^3	48.33	12.08	3.02	0.76	0.19	0.05

Both quantities are estimated with 1 000 simulations. We consider the 1 00 most over-represented octamers, the sequence length is ℓ = 580076. The last row gives the sample size per free parameter (length *n* of the sequence divided by the number $k^m(k - 1)$ of parameters).

For any word w (of size h_w), we introduce the following notation for $h_w \leq i \leq n$

$$I_i(w) = \mathbb{I}_{\{w \text{ end in position } i\}} = \mathbb{I}_{\{Y_{i-h_w+1}^{i}=w\}} \qquad (45)$$

where $Y_i^j = Y_i \ldots Y_j$ for all $i \leq j$. If $h_w \geq m$, we denote by

$$p(w) = \mu(w_1^m)\Pi(w_1^m, w_{m+1})\ldots\Pi(w_{h_w-m}^{h-1}, w_h) \qquad (46)$$

the probability to see one occurrence of w at a given position in the sequence. At last, if we consider another word v (of size $h_v = m$) and if $h_w = m$, we denote by

$$\Pi_\delta(v,w) = \sum_{x \in \mathcal{A}^\delta} p(vxw) \qquad (47)$$

the probability to see occurrences of v and w separated by a gap of length δ.

For any words v an w (to simplify, we suppose that $h_v \geq h_w$) then, for all $\delta \in \mathbb{Z}$ and

$\max(h_v, h_w - \delta) \leq i \leq \min(n, n - \delta)$ we have

$$\mathbb{E}\ [I_i(v)\ I_{i+\delta}(w)] = D_\delta(v, w) \qquad (48)$$

which do not depend on i.

It is therefore easy to show that

$$\begin{aligned}\mathbb{E}[\mathbf{N}(v)\mathbf{N}(w)] &= \sum_{i=h_v}^{n}\sum_{\delta=h_w-i}^{n-i} D_\delta(v,w) && (49)\\ &= \sum_{\delta=h_w-n}^{n-h_v} N_\delta D_\delta(v,w) && (50)\\ &= \mathbf{M}(v,w) + \mathbf{O}(v,m) && (51)\end{aligned}$$

where the main part ($2n - h_v - h_w + 2$ terms) is given by

$$\mathbf{M}(v,w) = \sum_{\delta=h_v}^{n-h_w} N_{-\delta}D_{-\delta}(v,w) + \sum_{\delta=h_w}^{n-h_v} N_\delta D_\delta(v,w) \qquad (52)$$

and the overlapping part ($h_v + h_w - 1$ terms) by

$$\mathbf{O}(v,w) = \sum_{\delta=-h_v+1}^{h_w-1} N_\delta D_\delta(v,w) \qquad (53)$$

and with

$$N_\delta = \begin{cases} n - h_w + 1 + \delta & \delta \in [h_w - n, h_w - h_v[\\ n - h_v + 1 & \delta \in [h_w - h_v, 0] \\ n - h_v + 1 - \delta & \delta \in]0, n - h_v] \end{cases} \qquad (54)$$

As we have

$$\mathbf{C}(v, w) = \mathbf{M}\ (v, w) + \mathbf{O}\ (v, w) - \mathbf{E}\ (v)\ \mathbf{E}\ (w) \qquad (55)$$

the problem is hence to compute **M** and **O** for all pairs of size m or $m + 1$ words. In order to simplify, we will just treat here the case of a pair of size m words (other cases can be derived from this special case).

For the main part we obtain

$$\begin{aligned}\mathbf{M}(v,w) = &\sum_{\delta=m}^{n-m} N_{-\delta}\mu(w)\Pi_{\delta-m+1}(w,v)\\ &+\sum_{\delta=m}^{n-m} N_\delta \mu(v)\Pi_{\delta-m+1}(v,w)\end{aligned} \qquad (56)$$

($2n - 2m + 2$ terms). As $P_k(v, w)$ quickly converges toward $\mu(w)$ when k grows (convergence speed is given by λ^k where λ is the magnitude of the second eigenvalue of the transition matrix Π). So there exists a rank $r \geq m$ such as

$$\begin{aligned}\mathbf{M}(v,w) \simeq\ &\mu(v)\mu(w)\sum_{\delta=r}^{n-m}(N_{-\delta} + N_\delta)\\ &+\sum_{\delta=m}^{r-1} N_{-\delta}\mu(w)\Pi_{\delta-m+1}(w,v)\\ &+\sum_{\delta=m}^{r-1} N_\delta \mu(v)\Pi_{\delta-m+1}(v,w)\end{aligned} \qquad (57)$$

which has only $2r - 2m + 1$ terms.

And for the overlapping part we get

$$\begin{aligned}\mathbf{O}(v,w) = &N_0 \times \mu(v) \times \mathbb{I}_{\{v=w\}}\\ &+\sum_{\delta=1}^{m-1} N_{-\delta} \times p(wv_{m-\delta+1}^{m}) \times \mathbb{I}_{\{v_1^{m-\delta}=w_{1+\delta}^{m}\}}\\ &+\sum_{\delta=1}^{m-1} N_\delta \times p(vw_{m-\delta+1}^{m}) \times \mathbb{I}_{\{v_{1+\delta}^{m}=w_1^{m-\delta}\}}\end{aligned} \qquad (58)$$

which has $2m + 1$ terms.

So the overall complexity for the computation of one term of **C** is hence $O(r)$ where the value of r is directly connected to the magnitude λ of the second eigenvalue of the transition matrix.

In the particular case of an order one Markov model (m = 1), we give here the complete expressions of **M** and **O**.

For all $a, b, c, d \in \mathcal{A}$, we have

$$\mathbf{M}(a,b) \simeq (n-r+1)(n-r)\mu(a)\mu(b) + \sum_{\delta=1}^{r-1}(n-\delta)\Big(\mu(b)\Pi^{\delta}(b,a)+\mu(a)\Pi^{\delta}(a,b)\Big) \quad (59)$$

$$\mathbf{O}\ (a,b) = n\mu(a)\ \mathbb{I}_{\{a=b\}} \quad (60)$$

$$\frac{\mathbf{M}(ab,c)}{\Pi(a,b)} \simeq (n-r)(n-r-1)\mu(a)\mu(c) + \sum_{\delta=1}^{r-1}(n-\delta-1)\Big(\mu(c)\Pi^{\delta}(c,a)+\mu(a)\Pi^{\delta}(b,c)\Big) \quad (61)$$

$$\frac{\mathbf{O}(ab,c)}{\Pi(a,b)} = (n-1)\mu(a)(\mathbb{I}_{\{a=c\}}+\mathbb{I}_{\{b=c\}}) \quad (62)$$

$$\frac{\mathbf{M}(ab,cd)}{\Pi(a,b)\Pi(c,d)} \simeq (n-r-1)(n-r-2)\mu(a)\mu(c) + \sum_{\delta=1}^{r-1}(n-\delta-2)\Big(\mu(c)\Pi^{\delta}(d,a)+\mu(a)\Pi^{\delta}(b,c)\Big) \quad (63)$$

$$\frac{\mathbf{O}(ab,cd)}{\Pi(a,b)} = (n-1)\mu(a)\mathbb{I}_{\{ab=cd\}} + (n-2)\Pi(c,d)\Big(\mu(c)\mathbb{I}_{\{a=d\}}+\mu(a)\mathbb{I}_{\{b=c\}}\Big) \quad (64)$$

With the example given in section "validation" we get for the expectation

$$\mathbf{E}_0^t = [4615.4\ 5384.6] \quad (65)$$

and

$$\mathbf{E}_1^t = [1384.5\ 3230.4\ 3230.4\ 2153.6] \quad (66)$$

The magnitude of the second eigenvalue of Π is $\lambda = 0.3$, then rank $r = 19$ give a relative error $< 10^{-10}$ and we get for the covariance

$$\mathbf{C}_{0,0} = \begin{bmatrix} 1338.28 & -1338.28 \\ -1338.28 & 1338.28 \end{bmatrix} \quad (67)$$

$$\mathbf{C}_{1,0} = \begin{bmatrix} 1146.9 & 191.2 & 191.2 & -1529.2 \\ -1146.9 & -191.2 & -191.2 & 1529.2 \end{bmatrix} \quad (68)$$

and

$$\mathbf{C}_{1,1} = \begin{bmatrix} 1536.8 & -390.0 & -390.0 & -756.9 \\ -390.0 & 581.2 & 581.0 & -772.2 \\ -390.0 & 581.0 & 581.2 & -772.2 \\ -756.9 & -772.2 & -772.2 & 2301.4 \end{bmatrix} \quad (69)$$

Appendix B

The beta function is defined by

$$\beta(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt \quad (70)$$

for all $a, b > 0$. The incomplete beta function for all $x \in [0,1]$ is then defined by

$$\beta(x,a,b) = \int_0^x t^{a-1}(1-t)^{b-1}dt \quad (71)$$

and

$$\beta^{-}(x,a,b) = \beta(a,b) - \beta(x,a,b) \quad (72)$$

$$= \int_x^1 t^{a-1}(1-t)^{b-1}dt \quad (73)$$

Using a continued fraction representation, these functions can be quickly numerically evaluated in $O(\sqrt{\max(a,b)})$ in the worst case [15, Chapter 6].

A great interest of this function is that it is connected to the cumulative distribution function of a binomial distribution by the following relation:

$$\mathbb{P}(\mathcal{B}(n,p) \geq k) = \frac{\beta(p,k,n-k+1)}{\beta(k,n-k+1)} \quad (74)$$

with $(n, k) \in {}^* \times$, $0 \leq k \leq n$ and $p \in [0,1]$.

Finally, let us remark that the incomplete beta function is differentiable in x and that

$$\frac{\partial\beta(x,a,b)}{\partial x} = x^{a-1}(1-x)^{b-1} \quad (75)$$

Appendix C

We give here the complete expression of σ for a single pattern in the special case of an order m = 0 homogeneous Markov model of parameter μ.

The MLE of μ is given by

$$\mu_{\mathbf{N}} = \frac{\mathbf{N}_1}{n} \quad (76)$$

where $\mathbf{N}_1$ is the frequency of all letters.

A Gaussian approximation gives

$$\mathcal{L}\,(\mathbf{N}_1) \simeq \mathcal{N}\,(\mathbf{E}_1, \mathbf{C}_{1,1}) \qquad (77)$$

with $\mathbf{E}_1 = n\mu$ and, for all $a, b \in \mathcal{A}$,

$$\mathbf{C}_{1,1}\,(a, b) = n\mu\,(a)\,\mathbb{I}_{a=b} - n\mu(a) \times n\mu(b) \qquad (78)$$

We have also

$$P(\mathbf{N}) = \frac{1}{n^h} \prod_{a \in \mathcal{A}} \mathbf{N}_1(a)^{A_1(a)} \qquad (79)$$

which implies for all $a \in \mathcal{A}$ that

$$\frac{\partial P(\mathbf{N})}{\partial \mathbf{N}_1(a)} = \frac{A_1(a)}{\underbrace{\mathbf{N}_1(a)}_{\mathbf{G}_1(a)}} \times P(\mathbf{N}) \qquad (80)$$

So finally we get

$$\sigma \simeq Q\sqrt{{}^t\mathbf{G}_1 \times \mathbf{C}_{1,1} \times \mathbf{G}_1} \qquad (81)$$

where Q is either defined by equation (24) if the pattern is over-represented or by equation (28) if under-represented.

References

1. Atteson W: **Calculating the exact probability of language-like patterns in biomolecular sequences.** *Pro 6th Int Conf on Intelligent Systems for Molecular Biology* 1998:17-24.
2. Régnier M, Szpankowski W: **On pattern frequency occurrences in a Markovian sequence.** *Algorithmica* 1998, **22(4):**631-649.
3. Robin S, Daudin JJ: **Exact distribution of word occurrences in a random sequence of letters.** *J App Prob* 1999, **36:**179-193.
4. Nuel G: **Effective p-value computations using Finite Markov Chain Imbedding (FMCI): application to local score and to pattern statistics.** *Algorithms Mol Biol* 2006, **1(1):**5.
5. Kleffe J, Borodovski M: **First and second moment of counts of words in random text generated by Markov chains.** *Comp Applic Biosci* 1992, **8:**443-441.
6. Prum B, Rodolphe F, de Turckheim E: **Finding words with unexpected frequencies in DNA sequences.** *J R Statist Soc B* 1995, **11:**190-192.
7. van Helden J, André B, Collado-Vides J: **Extracting Regulatory Sites from the Upstream Region of Yeast Genes by Computational Analysis of Oligonucleotide Frequencies.** *J Mol Biol* 1998, **281:**827-842.
8. Nuel G: **S-SPatt: Simple Statististices for Patterns on Markov chains.** *Bioinformatics* 2005, **21(13):**3051-3052.
9. Chrysaphinou O, Papastavridis S: **A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials.** *Proba Theory Relat Fields* 1988, **79(1):**129-143.
10. Arratia R, Goldstein L, Gordon L: **Poisson approximation and the Chen-Stein method.** *Stat Sci* 1990, **5(4):**403-434.
11. Schbath S: **Compound Poisson approximation of word counts in DNA sequences.** *ESAIM Probab Stat* 1995, **1:**1-16.
12. Nuel G: **LD-SPatt: Large Deviations Statistics for Patterns on Markov Chains.** *J Comput Biol* 2004, **11(6):**1023-1033.
13. Oehlert GW: **A note on the delta method.** *American Statistician* 1992, **46:**27-29.
14. Reinert G, Schbath S, Waterman M: **Chapter 6: Statistics on words with applications to biological sequences.** In *Applied Combinatorics on Words* Cambridge Universtity Press; 2005.
15. Press WH, Teukolsky SA, Vetterling WT, Flannery BP: *Numerical Recipes in C* 2nd edition. Cambridge Universtity Press; 1988.