Moments of the Count of a Regular Expression in a Heterogeneous Random Sequence

G. Nuel

Moments of the Count of a Regular Expression in a Heterogeneous Random Sequence

G. Nuel

February 12, 2019

Abstract

We focus here on the distribution of the random count N of a regular expression in a multi-state random sequence generated by a heterogenous Markov source. We first briefly recall how classical Markov chain embedding techniques allow reducing the problem to the count of specific transitions in a (heterogenous) order 1 Markov chain over a deterministic finite automaton state space. From this result we derive the expression of both the mgf/pgf of N as well as the factorial and non-factorial moments of N. We then introduce the notion of evidence-based constraints in this context. Following the classical forward/backward algorithm in hidden Markov models, we provide explicit recursions allowing to compute the mgf/pgf of N under the evidence constraint. All the results presented are illustrated with a toy example.

Keywords: probability generating function; moment generating function; probabilistic graphical model, Bayesian network; sum-product algorithm; forward/backward algorithms.

1 Introduction

Let $\mathcal{R}$ be a regular expression on the finite alphabet Σ. For any sequence $X \in \Sigma^n$ we define N as the number of occurrences of $\mathcal{R}$ in X. Our main purpose is to study the distribution and moments of N under the assumption that X is generated according to a heterogeneous Markovian source. The distribution of N under various assumptions (ex: binary sequences, memory-less random sources, homogeneous Markovian source, etc.) is known to have many applications in a wide range of fields (reliability, linguistic, data compression, bioinformatics, etc.) and as been extensively studied by the literature [30, 11, 43, 23, 46, 38, 3, 21].

Most of this research mainly focus on homogeneous models because this assumption usually allows to obtain simpler and computationally more efficient formulas. Heterogenous models are nevertheless often encountered, either directly as continuous process [9, 49] or, more often, as discrete process though hidden Markov models – HMMs – [42, 15, 47]. In the context of HMMs the forward/backward algorithm [4, 18] allows computing efficiently the posterior distribution of the hidden states given the observations which is a heterogenous Markov model. The distribution of N in such hidden sequences as been for example extensively studied in [1, 27] and others like [20] explicitly consider the case of heterogenous Markov models.

In the present work, we focus on the distribution and moments of N specifically under the heterogeneous assumption. We start by recalling how classical Markov chain embedding techniques allow reducing the problem to the count of specific transitions in a (heterogenous) order 1 Markov chain over a deterministic finite automaton state space. Next we recall how to obtain efficiently the factorial moments of N through the derivatives of the probability generating function (pgf).

We then present two results: 1) we establish the non-factorial moments of N by introducing a formal (partial) computation of the moment generating function (mgf) of N; 2) inspired by the probabilistic graphical models [14], we focus on the constrained distribution of N given a generic *evidence* by introducing modified Forward-Backward recursions both for the pgf and mgf of N. This last result is clearly the most original and innovative part of the present work.

All results are illustrated through a simple example over a binary alphabet and the complete R source code of the computations is provided as a supplementary material.

2 Recalls and Notations

Let $\mathcal{R}$ be a regular expression on the finite alphabet Σ. For any random sequence $X \in \Sigma^n$ we define an occurrence of $\mathcal{R}$ in position i as the event $\{X_1 \dots X_i \in \Sigma^\star\mathcal{R}\}$ (*i.e.* $\mathcal{R}$ is a suffix of $X_1 \dots X_n$). The random (overlapping[1]) count of $\mathcal{R}$ in X is defined by $N = \sum_{i=1}^{n} \mathbf{1}\{X_1 \dots X_i \in \Sigma^\star\mathcal{R}\}$. Our main interest is to study the distribution of N when X is generated by a random Markov source (ex: i.i.d. sequence, homogeneous or heterogeneous finite order Markov chain, variable length Markov chain, etc.).

We call *Markov chain embedding*[2] (MCE) of the problem any first-order Markov

[1] Non-overlapping counts – also called renewal counts – are sometimes considered instead. For simplification purpose we consider here only overlapping counts but the extension to non-overlapping counts is straightforward (see [34] for example).

[2] Also called *finite Markov chain imbedding* or *auxiliary Markov chain* by some authors.

sequence $Y \in \mathcal{Q}^n$ of the finite state space $\mathcal{Q}$ such as $\{X_1 \dots X_i \in \Sigma^\star \mathcal{R}\} = \{Y_i \in \mathcal{F}\}\ \forall i \in \{1, \dots, n\}$ where $\mathcal{F} \subset \mathcal{Q}$ is the subset of final states. For any $p, q \in \mathcal{Q}$ we denote by $T_i(p,q) = \mathbb{P}(Y_i = q |\ Y_{i-1} = p)$ the transitions in position i and we decompose the transition matrix into $\boldsymbol{T}_i = \boldsymbol{P}_i + \boldsymbol{Q}_i$ where $P_i(p,q) = \mathbf{1}_{q \notin \mathcal{F}} T_i(p,q)$ are the non-counting transitions and $Q_i(p,q) = \mathbf{1}_{q \in \mathcal{F}} T_i(p,q)$ are the counting transitions.

This notion, initially introduced by [12] has been extensively used by many authors [16, 10, 8, 33, 26, 40, 3]. For many years, MCE was constructed in an *ad hoc* manner for each considered problem (ex: runs, urn problems, scan statistics, sparse-seed, etc.) until the connexion with the finite automaton theory [13] was pointed out by several authors [30, 5, 17] and the notion of "optimal" MCE finally emerged [19, 34, 22, 45, 25, 29, 28].

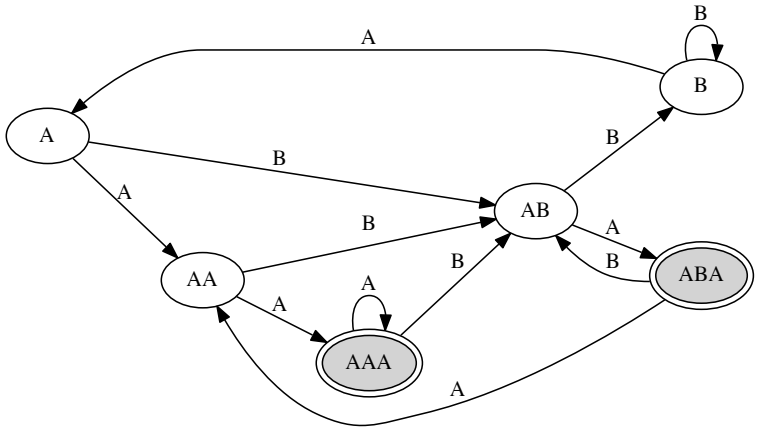
Theorem 1 ([34])**.** *We can use Deterministic Finite Automata (DFA) to build an* optimal *(in $|\mathcal{Q}|$) Markov chain embedding of any problem. The probability generating function (pgf) of N can therefore be written as:*

$$G(z) \stackrel{def}{=} \mathbb{E}\left[z^N\right] = \sum_{k=0}^{\infty} \mathbb{P}(N=k) z^k = \boldsymbol{u} \left[\prod_{i=2}^{n} (\boldsymbol{P}_i + z\boldsymbol{Q}_i) \right] \boldsymbol{v} \tag{1}$$

where $\boldsymbol{u}$ is a starting (row-)vector and $\boldsymbol{v} = (1\ 1\ \dots\ 1)^T$.

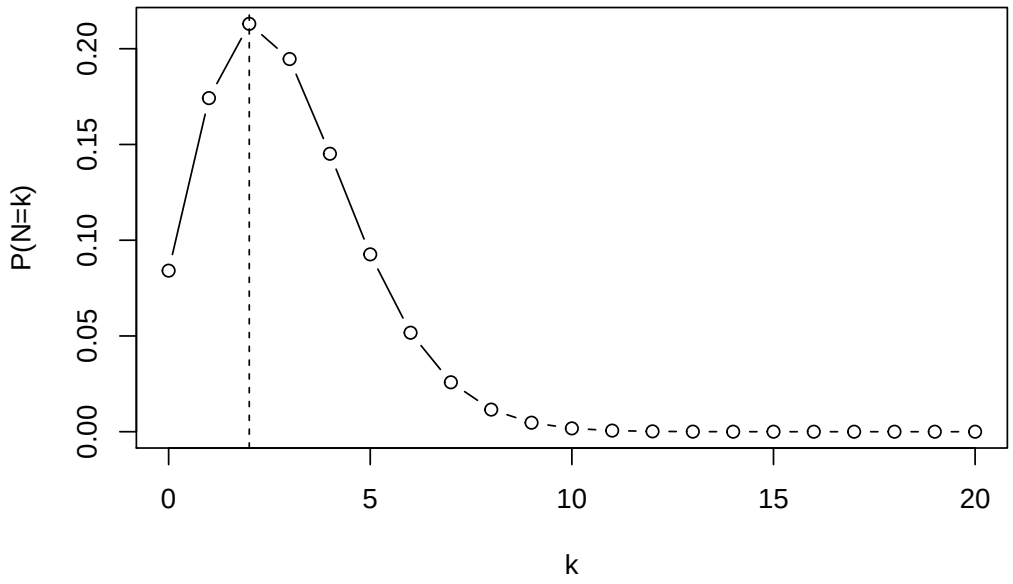
It is therefore possible to use Eq. (1) to compute the pgf of N and hence the full distribution of N with complexity $\mathcal{O}(n \times N_{\max} \times |\mathcal{Q}| \times |\Sigma|)$, where $N_{\max}$ is the maximum number of occurrences of N in a arbitrary sequence of length n. Note that in the particular case where the transition of the MCE is homogeneous (*i.e.* $\boldsymbol{T}_i \equiv \boldsymbol{T}\ \forall i$), the computation of $G(z)$ as a rational function can be done efficiently using formal matrix inversions, the expression of $G(z)$ then being used to obtain distributions or moments using Taylor-Expansion techniques. For more details, see [30, 44, 39]. These interesting approaches will not be further developed in this article since they assume homogeneity and we do not want to restrict ourselves to this particular case in the present work.

In Figure 1 we can see a toy example of MCE for a simple problem over the binary alphabet $\Sigma = \{\mathtt{A}, \mathtt{B}\}$. In Figure 2 we can see the result of the mgf computation for this problem. Note that we deliberately decided to use a homogeneous model in the illustrative example for the sake of simplicity. This choice might appear to be in contradiction with the main purpose of the paper (to focus on heterogeneous models), but all the computations presented obviously remain valid in the homogeneous case and, as we will see in Section 4, even homogeneous background model like the one we consider here can generate heterogeneity when considering constrained distributions. From now on, this example will be used to illustrate all the results presented.



$$\boldsymbol{P}_i + z\boldsymbol{Q}_i = \left(\begin{array}{c|cccccc} & \mathtt{A} & \mathtt{B} & \mathtt{AA} & \mathtt{AB} & \mathtt{AAA} & \mathtt{ABA} \\ \hline \mathtt{A} & 0 & 0 & \pi_i(\mathtt{A}) & \pi_i(\mathtt{B}) & 0 & 0 \\ \mathtt{B} & \pi_i(\mathtt{A}) & \pi_i(\mathtt{B}) & 0 & 0 & 0 & 0 \\ \mathtt{AA} & 0 & 0 & 0 & \pi_i(\mathtt{B}) & z\pi_i(\mathtt{A}) & 0 \\ \mathtt{AB} & 0 & \pi_i(\mathtt{B}) & 0 & 0 & 0 & z\pi_i(\mathtt{A}) \\ \mathtt{AAA} & 0 & 0 & 0 & \pi_i(\mathtt{B}) & z\pi_i(\mathtt{A}) & 0 \\ \mathtt{ABA} & 0 & 0 & \pi_i(\mathtt{A}) & \pi_i(\mathtt{B}) & 0 & 0 \end{array}\right)$$

Figure 1: Example of (optimal) MCE for $\mathcal{R} = \mathtt{A}[\mathtt{AB}]\mathtt{A}$ over the binary alphabet $\Sigma = \{\mathtt{A}, \mathtt{B}\}$, and where X generated by a possibly heterogeneous memory-less source. Top: the DFA corresponding to the problem with $\mathcal{Q} = \{\mathtt{A}, \mathtt{B}, \mathtt{AA}, \mathtt{AB}, \mathtt{AAA}, \mathtt{ABA}\}$ and $\mathcal{F} = \{\mathtt{AAA}, \mathtt{ABA}\}$. Bottom: the transition Matrix and denote by $\pi_i(\cdot) = \mathbb{P}(X_i = \cdot)$ where counting transitions are marked with the dummy variable z.



$$G(z) = \boldsymbol{u}(\boldsymbol{P} + z\boldsymbol{Q})^{n-1}\boldsymbol{v} = 0.084z^0 + 0.17z^1 + 0.21z^2 + 0.19z^3 + 0.15z^4 + 0.093z^5 + 0.052z^6 + 0.026z^7 + 0.012z^8 + 0.0047z^9 + 0.0017z^{10} + 0.00058z^{11} + 0.00018z^{12} + 4.9 \times 10^{-5}z^{13} + 1.2 \times 10^{-5}z^{14} + 2.6 \times 10^{-6}z^{15} + 5.1 \times 10^{-7}z^{16} + 6.6 \times 10^{-8}z^{17} + 1.1 \times 10^{-8}z^{18}$$

Figure 2: Distribution of N assuming that $\pi_i(\texttt{A}) = 0.4$, $\pi_i(\texttt{B}) = 0.6$ for all i and $n = 20$. Top: graphical representation of the distribution (mode indicated by the dashed line). Bottom: the pgf of N.

The reader interested by computations with heterogeneous sources might refer to the supplementary material where we consider the present example both with homogenous and heterogenous sources with a simple adaptation of the same formulas.

Obviously, since the mgf provides the full distribution of N, it can therefore be used to compute any moment of N. However, it is well known that various moments (typically: the expectation and variance of N) can be obtained with a dramatically lower computational cost usually under the homogeneous assumption [6, 41]. In [37], all order lower than k factorial moments of N are established in the heterogeneous case with complexity $\mathcal{O}(n \times k \times |\mathcal{Q}| \times |\Sigma|)$ using a simple modification of Eq. (1):

Corollary 2 ([37]). *For any $k \geqslant 0$ we have:*

$$\mathbb{E}\left[(N)_k\right] \overset{def}{=} \mathbb{E}\left[\frac{N!}{(N-k)!}\right] = G^{(k)}(1) = k! \left[\boldsymbol{u}\left[\prod_{i=2}^{n}(\boldsymbol{P}_i + \boldsymbol{Q}_i + z\boldsymbol{Q}_i)\right]\boldsymbol{v}\right]_{z^k} \tag{2}$$

where $[\cdot]_{z^k}$ denote the extraction of the z-polynomial coefficient of degree k.

We can see in Figure 3 the factorial moments computed for $k = 1, 2, \ldots, 5$ in our example using both the complete distribution (pgf-based) and the faster Eq. (2). Without surprise, the results are identical up to the machine precision.

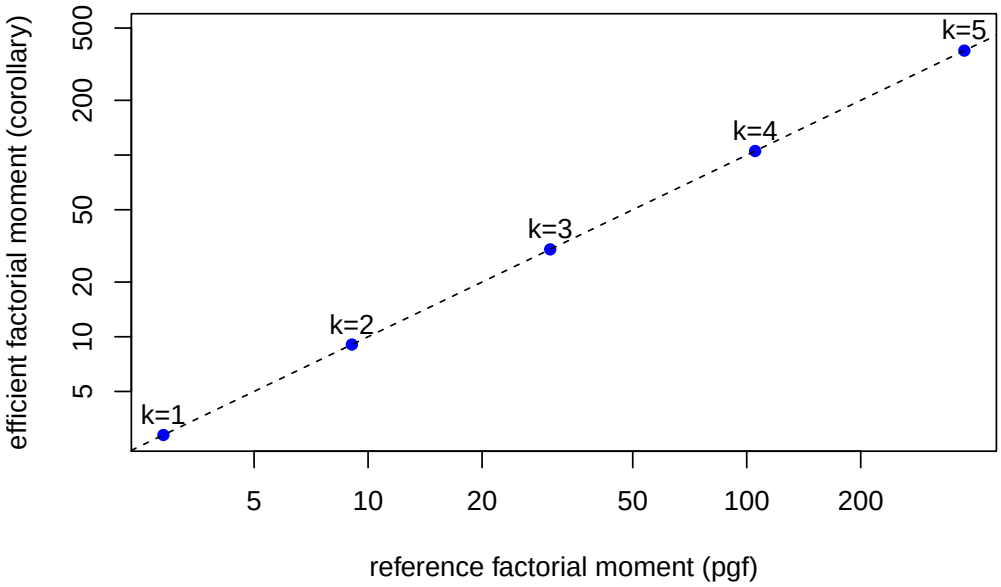
3 Non-Factorial Moments

As explained in [37], factorial moments can be used to compute cumulants and/or non-factorial moments using various polynomial relationships. But [38] pointed out the limitations of such approach in floating-point arithmetic. In practice, high order non-factorial moments (ex: $k = 5, 6, 7, \ldots$) are not reliable using factorial moments unless using arbitrary-precision computations or sophisticated modular rational approaches [39]. It is therefore interesting to find an alternative way to compute directly non-factorial moments.

For that purpose, we simply derive from Eq. (1) the general expression of the moment-generating function (mgf) of N:

$$M(t) \overset{\text{def}}{=} \mathbb{E}\left[e^{tN}\right] = G(e^t) = \boldsymbol{u}\left[\prod_{i=2}^{n}(\boldsymbol{P}_i + e^t\boldsymbol{Q}_i)\right]\boldsymbol{v}. \tag{3}$$

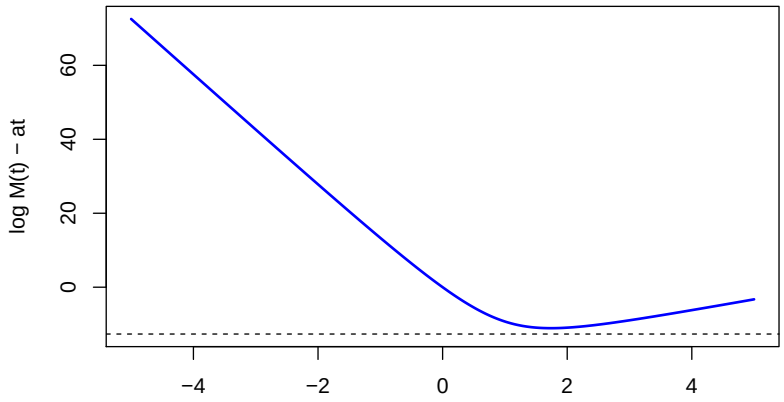
Interestingly, the value $M(t)$ of the mgf for any real number t can be computed at a modest computational cost of $\mathcal{O}(n \times |\mathcal{Q}| \times |\Sigma|)$. As pointed out in [38] it is hence



$$\boldsymbol{u}(\boldsymbol{P}+\boldsymbol{Q}+z\boldsymbol{Q})^{n-1}\boldsymbol{v} = 1z^0 + 2.88z^1 + 4.5312z^2 + 5.044224z^3 + 4.38534144z^4 + 3.1302746112z^5 + \ldots$$

$$\begin{aligned}\mathbb{E}\left[(\boldsymbol{N})_{\mathbf{5}}\right] &= 5! \times 3.1302746112 = \mathbf{375.632953344} \\ &= \underbrace{\sum_{k=0}^{18} \frac{k!}{(k-5)!}\mathbb{P}(N=k)}_{\text{reference (pgf)}} = \mathbf{375.632953344}\end{aligned}$$

Figure 3: Factorial moments of order $\leqslant k = 5$ for our illustrative example. Top: comparison of the reference values (full pgf) to the efficient computation. Bottom: details of the efficient computation.



$$-12.66459 = \log \mathbb{P}(N \geqslant 15) \leqslant \min_t \{\log M(t) - 15 \times t\} = -11.08537$$

Figure 4: Chernoff's bound for our illustration example with $n = 20$ and $a = 15$. Top: graphical representation of the bound $\log M(t) - at$, the dashed line corresponds to the exact value of $\log \mathbb{P}(N \geqslant a)$. Bottom: the inequality corresponding to the sharpest bound.

possible to compute numerically the cumulant-generating function $\Lambda(t) = \log M(t)$ or N which can be used to obtain the following Chernoff's bound [7]:

$$\log \mathbb{P}(N \geqslant a) \leqslant \min_t \{\Lambda(t) - ta\} \tag{4}$$

for any (large) $a > 0$. In the homogeneous case, the results can be extended to establish large deviation results like in [32], but the present bound has the great interest to be valid for any finite n and without any homogeneity assumption. Moreover, as explained in [38], the first two derivatives of $\Lambda(t)$ can be easily computed both allowing to speed up the numerical optimization of Eq. (4), and providing precise-type large deviation approximation using Bahadur-Rao results [2]. We can see in Figure 4 that the Chernoff's bound can be surprisingly sharp even with $n = 20$.

We are now ready to introduce our new result. Instead of using Eq. (3) for a given $t \in \mathbb{R}$, we can simply replace e^t by its generating function and we immediately establish:

Theorem 3.

$$M(t) = \sum_{k=0}^{\infty} \mathbb{E}\left[N^k\right] \frac{t^k}{k!} = \boldsymbol{u} \prod_{i=2}^{n} \left(\boldsymbol{P}_i + \sum_{k=0}^{\infty} \frac{t^k}{k!} \boldsymbol{Q}_i \right) \boldsymbol{v}$$

$$\boldsymbol{u}(\boldsymbol{P}+e^t\boldsymbol{Q})^{n-1}\boldsymbol{v} = 1t^0 + 2.88t^1 + 5.9712t^2 + 10.055424t^3 + 14.71487744t^4 + 19.3630374912t^5 + \dots$$

$$\begin{aligned}\mathbb{E}\left[\boldsymbol{N^5}\right] &= 5! \times 19.3630374912 = \mathbf{2323.564498944001} \\ &= \underbrace{\sum_{k=0}^{18} k^5 \mathbb{P}(N=k)}_{\text{reference (pgf)}} = \mathbf{2323.564498944001}\end{aligned}$$

$$\mathbb{E}[(N)_5] = 24\mathbb{E}[N^1] - 50\mathbb{E}[N^2] + 35\mathbb{E}[N^3] - 10\mathbb{E}[N^4] + \mathbb{E}[N^5]$$
$$\mathbb{E}[N^5] = \mathbb{E}[(N)_1] + 15\mathbb{E}[(N)_2] + 25\mathbb{E}[(N)_3] + 10\mathbb{E}[(N)_4] + \mathbb{E}[(N)_5]$$

Figure 5: Computation of the mgf of N for our illustration example up to the order $k = 5$ terms. Explicit relationships between order $k = 5$ factorial and non-factorial moments are also given.

and hence, for all $k \geqslant 0$ we have:

$$\mathbb{E}\left[N^k\right] = k! M^{(k)}(t) = k! \left[\boldsymbol{u}\left[\prod_{i=2}^{n}\left(\boldsymbol{P}_i + \sum_{j=0}^{k}\frac{t^j}{j!}\boldsymbol{Q}_i\right)\right]\boldsymbol{v}\right]_{t^k}.$$

Proof. The result is simply obtained by interchanging the finite product with the infinite summation. The polynomial terms of order $> k$ are omitted since they simply do not contribute of the terms of degree $\leqslant k$. □

Using Theorem 3, it is possible to compute all order lower than k non-factorial moments of N with complexity $\mathcal{O}(n \times k^2 \times |\mathcal{Q}| \times |\Sigma|)$. Note that the complexity in k is here quadratic while it was linear in k for the factorial moments. This is due to the fact that each term of the product now requires multiplying two order k polynomials (it was a degree k by degree 1 product for the factorial moments). We can see in Figure 5 the application of Theorem 3 to obtain the order $k = 5$ moment of N in our example.

4 Conditional Distributions

4.1 Notion of Evidence

The distribution of N under-constraints has been also studied [48, 35, 24]. The most common constraint is of the form $N = a$. We want here to consider constraints of the form $Y_i \in \mathcal{Y}_i$ for all i, where $\mathcal{Y}_i \subset \mathcal{Q}$ is the subset of acceptable values for Y_i. Inspiring from the probabilistic graphical model theory [14] we call *evidence* the event $\text{ev} = \cup_i \{Y_i \in \mathcal{Y}_i\}$ and our purpose is now to establish the distribution and moments of N conditionally to the evidence.

For example, in the particular case of our illustration example, if $\mathcal{Y}_i = \mathcal{Q}$ (neutral evidence) for all $i \in \{1, \ldots, n\}$ but $\mathcal{Y}_5 = \mathcal{Y}_{15} = \mathcal{F}$ and $\mathcal{Y}_{10} = \mathcal{Q} \setminus \mathcal{F}$ we have the resulting evidence is: "$\mathcal{R}$ occurs in $i = 5, 15$ but not in $i = 10$". Obviously with such evidence, we know that we must have $N \geqslant 2$ (there is at least two occurrences) and $N \leqslant N_{\max} - 1 = 17$.

For any $\mathcal{I} \subset \{1, \ldots, n\}$ let $\text{ev}_{\mathcal{I}} = \cup_{i \in \mathcal{I}} \text{ev}_i$ with $\text{ev}_i = \{Y_i \in \mathcal{Y}_i\}$. Let $\text{ev}_{<i} = \text{ev}_{\{1,\ldots,i-1\}}$, $\text{ev}_{\leqslant i} = \text{ev}_{\{1,\ldots,i\}}$, $\text{ev}_{>i} = \text{ev}_{\{i+1,\ldots,n\}}$; hence we get $\text{ev} = \text{ev}_{\{1,\ldots,n\}} = \text{ev}_{<n+1} = \text{ev}_{\leqslant n} = \text{ev}_{>0} = \text{ev}_{\geqslant 1}$.

4.2 Forward and Backward

For any $i \in \{1, \ldots, n\}$ and for all $q \in \mathcal{Q}$ we introduce the following forward and backward polynomials:

$$F_i(q) = \sum_{Y_{<i}} z^{N_i} \mathbb{P}(Y_{<i}, Y_i = q, \text{ev}_{\leqslant i}) \quad B_i(q) = \sum_{Y_{>i}} z^{N - N_i} \mathbb{P}(Y_{>i}, \text{ev}_{>i} | Y_i = q) \tag{5}$$

where $Y_{<i} = (Y_1, \ldots, Y_{i-1})$ ($\emptyset$ for $i = 1$), $Y_{>i} = Y_{i+1}, \ldots, Y_n$ ($\emptyset$ for $i = n$), and with $N_i = \sum_{j=1}^{i} \mathbf{1}_{Y_j \in \mathcal{F}}$ ($N_0 = 0$).

Theorem 4. *For all $i \in \{1, \ldots, n\}$ and $q \in \mathcal{Q}$ we have :*

$$F_i(q) B_i(q) = \sum_{Y_{<i-1}} \sum_{Y_{>i}} z^N \mathbb{P}(Y_{<i}, Y_i = q, Y_{>i}, \text{ev}) \tag{6}$$

and for all $i \in \{2, \ldots, n\}$ and $p, q \in \mathcal{Q}$ we have

$$F_{i-1}(p)[P_i(p,q) + zQ_i(p,q)]\mathbf{1}_{q \in \mathcal{Y}_i} B_i(q) = \sum_{Y_{<i-1}} \sum_{Y_{>i}} z^N \mathbb{P}(Y_{<i-1}, Y_{i-1} = p, Y_i = q, Y_{>i}, \text{ev}). \tag{7}$$

Proof. We only give here the proof of Eq. (6) since the proof of Eq. (7) is almost identical. We take advantage of the Markov property, in particular the independence of $Y_{<i}$ and $Y_{>i}$ conditionally to $Y_i = q$. We get:

$$\begin{aligned}
&z^N \mathbb{P}(Y_{<i}, Y_i = q, Y_{>i}, \mathrm{ev}) = z^{N_i} z^{N-N_i} \mathbb{P}(Y_{<i}, Y_i = q, Y_{>i}, \mathrm{ev}_{\leq i}, \mathrm{ev}_{>i}) \\
&= z^{N_i} z^{N-N_i} \mathbb{P}(Y_{<i}, Y_{>i}, \mathrm{ev}_{\leq i}, \mathrm{ev}_{>i} | Y_i = q) \times \mathbb{P}(Y_i = q) \\
&= z^{N_i} z^{N-N_i} \mathbb{P}(Y_{<i}, \mathrm{ev}_{\leq i} | Y_i = q) \times \mathbb{P}(Y_{>i}, \mathrm{ev}_{>i} | Y_i = q) \times \mathbb{P}(Y_i = q) \\
&= z^{N_i} \mathbb{P}(Y_{<i}, Y_i = q, \mathrm{ev}_{\leq i}) \times z^{N-N_i} \mathbb{P}(Y_{>i}, \mathrm{ev}_{>i} | Y_i = q)
\end{aligned}$$

from which the summation over $Y_{<i}$ and $Y_{>i}$ immediately gives the results since the only common term between the two sums is precisely the fixed $Y_i = q$. □

Corollary 5. *The forward polynomials can be computed recursively for $i = 2, \ldots, n$ with*

$$F_i(q) = \sum_p F_{i-1}(p)[P_i(p,q) + zQ_i(p,q)]\mathbf{1}_{q \in \mathcal{Y}_i} \tag{8}$$

with the initialization $F_1(q) = \mathbf{1}_{q \in \mathcal{Y}_1} z^{\mathbf{1}_{q \in \mathcal{F}}} u_q$, and the backward polynomials can be computed recursively for $i = n, \ldots, 2$ with:

$$B_{i-1}(p) = \sum_q [P_i(p,q) + zQ_i(p,q)]\mathbf{1}_{q \in \mathcal{Y}_i} B_i(q) \tag{9}$$

and with the convention that $B_n(\cdot) \equiv 1$.

Proof. We prove only Eq. (8) since the proof of Eq. (9) is almost identical. It is clear that:

$$\begin{aligned}
&\sum_{Y_{<i-1}} \sum_{Y_{>i}} z^N \mathbb{P}(Y_{<i}, Y_i = q, Y_{>i}, \mathrm{ev}) \\
&\qquad = \sum_{p \in \mathcal{Q}} \sum_{Y_{<i-1}} \sum_{Y_{>i}} z^N \mathbb{P}(Y_{<i-1}, Y_{i-1} = p, Y_i = q, Y_{>i}, \mathrm{ev})
\end{aligned}$$

by simply using Eq. (6) and Eq. (7) of Theorem 4 we hence get:

$$F_i(q) \cancel{B_i(q)} = \sum_{p \in \mathcal{Q}} F_{i-1}(p)[P_i(p,q) + zQ_i(p,q)]\mathbf{1}_{q \in \mathcal{Y}_i} \cancel{B_i(q)}$$

which gives Eq. (8). □

Performing the complete forward/backward recursion hence results, like for Theorem 1 in a complexity of $\mathcal{O}(n \times N_{\max} \times |\mathcal{Q}| \times |\Sigma|)$, where $N_{\max}$ is the maximum number of occurrences of N in a arbitrary sequence of length n. Once the forward/backward quantities available, it is possible to compute the forward and backward quantities and we can easily derive from Theorem 4 quantities of interest:

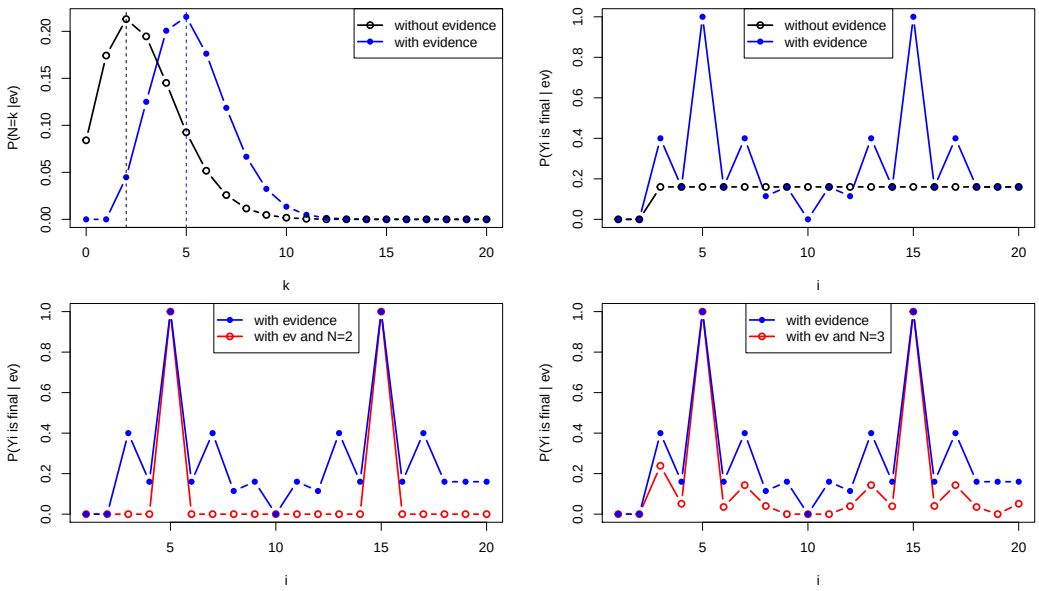


Figure 6: Distribution of N and position-specific probabilities of occurrence for our illustration example with ev = "$\mathcal{R}$ occurs in $i = 5, 15$ but not in $i = 10$" and under various constraints. Top-left: pgf with or without ev. Top-right: position-specific occurrence probability with or without ev. Bottom-left: same thing with ev and the additional constraint that $N = 2$. Bottom-right: same thing with $N = 3$.

Corollary 6. *By observing that* $\mathbb{P}(Y_i = q, N = k, \text{ev}) = [F_i(q)B_i(q)]_{z^k}$ *for all* i *we derive:*

$$\mathbb{P}(N = k, \text{ev}) = \left[\sum_{q\in\mathcal{Q}} F_i(q)B_i(q)\right]_{z^k} \quad \mathbb{P}(\text{ev}) = \left[\sum_{q\in\mathcal{Q}} F_i(q)B_i(q)\right]_{z=1} \tag{10}$$

$$\mathbb{P}(N = k|\text{ev}) = \frac{\mathbb{P}(N = k, \text{ev})}{\mathbb{P}(\text{ev})} \quad \mathbb{P}(Y_i = q|N = k, \text{ev}) = \frac{[F_i(q)B_i(q)]_{z^k}}{\mathbb{P}(N = k, \text{ev})} \tag{11}$$

Proof. Immediate from Eq. (6). □

Note that if the marginal computation of $\mathbb{P}(N = k, \text{ev})$ or $\mathbb{P}(\text{ev})$ can be done using only the forward recursion (with $i = n$) or the backward recursion (with $i = 1$), the position specific $\mathbb{P}(Y_i = q|N = k, \text{ev})$ does require the computation of both $F_i(q)$ (and hence $F_1, \ldots, F_{i-1}$) and $B_i(q)$ (and hence $B_{i+1}, \ldots, B_n$).

We can see in Figure 6 the distribution of N with various evidence. In the top-left graph, we see since the evidence ev = "$\mathcal{R}$ occurs in $i = 5, 15$ but not in $i = 10$" forces $\mathcal{R}$ to occur at least two times (in positions $i = 5, 15$), the distribution under

ev is logically shifted on the right by two units in comparison with the unconstrained distribution. In the top-right plot, the marginal occurrence of $\mathcal{R}$ with no evidence is straightforward in our memory-less model: occurrence is impossible for the first two positions (since $\mathcal{R}$ as a length of 3) and the marginal probability of occurrence is identical for the remaining positions. When adding the evidence, these probabilities are dramatically altered: first, the marginal probability in positions $5, 10, 15$ directly reflects the evidence but neighbor positions are also modified. In the bottom-left plot, by adding "$N = 2$" to the evidence, the only possible configuration of the system is exactly two occurrences of $\mathcal{R}$ in positions 5 and 15, and the marginal position-specific probability of occurrence perfectly reflects this constraint. Finally, in the bottom-right plot, by adding "$N = 3$" to the evidence, the resulting marginal distribution indicates the posterior position of the only remaining occurrence of $\mathcal{R}$ which is not fixed by the constraint.

4.3 Moments

By replacing z by e^t in the definition of the forward and backward quantities, we obtain the mgf versions F_i^{mgf} and B_i^{mgf} of these quantities.

Theorem 7. *The mgf-forward/backward quantities can be computed recursively for $i = 2, \dots, n$ with*

$$F_i^{mgf}(q) = \sum_{q \in \mathcal{Q}} F_{i-1}^{mgf}(p) \left[P_i(p,q) + \sum_{k \geqslant 0} \frac{t^k}{k!} Q_i(p,q) \right] \mathbf{1}_{q \in \mathcal{Y}_i} \tag{12}$$

and for $i = n, \dots, 2$ with:

$$B_{i-1}^{mgf}(p) = \sum_{q \in \mathcal{Q}} \left[P_i(p,q) + \sum_{k \geqslant 0} \frac{t^k}{k!} Q_i(p,q) \right] \mathbf{1}_{q \in \mathcal{Y}_i} B_i^{mgf}(q) \tag{13}$$

and we derive from these quantities the following expressions:

$$\mathbb{P}(\text{ev}) = \left[\sum_{q \in \mathcal{Q}} F_i^{mgf}(q) B_i^{mgf}(q) \right]_{t^0} \qquad \mathbb{E}\left[N^k \middle| \text{ev} \right] = \frac{k!}{\mathbb{P}(\text{ev})} \left[\sum_{q \in \mathcal{Q}} F_i^{mgf}(q) B_i^{mgf}(q) \right]_{t^k}. \tag{14}$$

Proof. Simply replacing z by e^t in Corollary 5 immediately allows to establish Eq. (12) and Eq. (13). From the analog of Theorem 4 with the same substitution we easily get:

$$\sum_{q \in \mathcal{Q}} F_i^{\text{mgf}}(q) B_i^{\text{mgf}}(q) = \mathbb{E}\left[e^{tN} \mathbf{1}_{\text{ev}} \right] = \sum_{k=0}^{\infty} \frac{t^k}{k!} E\left[N^k \mathbf{1}_{\text{ev}} \right].$$

□

	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$
$\mathbb{E}[N^k]$	2.88	11.94	60.33	353.16	2323.56
$\mathbb{E}[(N+2)^k]$	4.88	27.46	174.55	1230.60	9486.23
$\mathbb{E}[N^k\|\text{ev}]$	5.27	31.17	203.55	1446.37	11060.46

Table 1: Various moments of N in our illustration example.

The Theorem 7 hence provide an efficient way to compute all order $\leqslant k$ moments of N conditionally to the evidence with complexity $\mathcal{O}(n \times k^2 \times |\mathcal{Q}| \times |\Sigma|)$ which is exactly the complexity we had without any evidence.

In Table 1 we can see various moments of N for our illustration example. Without surprise, the moments of N are dramatically changed by adding the evidence ev = "$\mathcal{R}$ occurs in $i = 5, 15$ but not in $i = 10$".

5 Conclusion

The main contribution of this paper are the explicit computation of the pgf and of the mgf of the random count N of a regular expression $\mathcal{R}$ in a multi-state sequence generated by a heterogenous Markovian source conditionally to an evidence. The complexity for computing all terms of degree $\leqslant k$ is linear in k for the pgf (Theorem 4) and quadratic for the mgf (Theorem 7)[3]. These results allow to compute the marginal distribution of the occurrence conditionally to the evidence and/or constraints of the form $\{N = a\}$, as well as conditional non-factorial moments of any order.

We considered here an evidence based on constrained values of Y_i (*i.e.* MCE state in position i). Since, like in [27] our approach is based on the sum-product algorithm (also called forward/backward for Markovian models), we can obviously easily extend our evidence to constraints on the sequence itself (*i.e.* X_i) and/or on the counting process (*i.e.* on N_i, the number of occurrences up to position i). This should allow for example to consider occurrences of regular expressions in degenerated sequences like in [36, 31], or to consider subtle constraints like ev $= \{N_{500} \geqslant 10, N_{1000} \leqslant 20\}$.

As explained in Section 4.1, here we only consider evidence of the form ev $= \cup_i \{Y_i \in \mathcal{Y}_i\}$. Of course, more complex evidence can be considered. For example, one might condition as well on the number of occurrences in a sliding window. However, more sophisticated constraints like the simultaneous occurence or non-occurrence in two given position would turn the Markov model in a true Bayesian

[3]The base complexity – *i.e.* for $k = 1$ – for both approaches is $\mathcal{O}(n \times |\mathcal{Q}| \times |\Sigma|)$ where n is the sequence length, $|\Sigma|$ the alphabet size, and $|\mathcal{Q}|$ the number of states of the MCE.

network. In that case, the methods presented here can certainly be generalized by considering sum-product computations in junction tree (see [14] for more details) with various extensions in perspective: Markov-trees (ex: phylogeny), Markovian sequences structural constraints (ex: $X_1 = X_n$ for a circular sequence, stem-loop constraints for structured RNA), Markov random fields (social networks or image segmentation), etc. t

these evidence should always be expressed as a weighted distribution of the support of a Markovian random variable

References

[1] John AD Aston and Donald EK Martin. Distributions associated with general runs and patterns in hidden Markov models. *The Annals of Applied Statistics*, pages 585–611, 2007.

[2] Raghu Raj Bahadur and R Ranga Rao. On deviations of the sample mean. *The Annals of Mathematical Statistics*, 31(4):1015–1027, 1960.

[3] Narayanaswamy Balakrishnan and Markos V Koutras. *Runs and scans with applications*, volume 764. John Wiley & Sons, 2011.

[4] Leonard E Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The annals of mathematical statistics*, 41(1):164–171, 1970.

[5] Valentina Boeva, Julien Clément, Mireille Régnier, and Mathias Vandenbogaert. Assessing the significance of sets of words. In *Annual Symposium on Combinatorial Pattern Matching*, pages 358–370. Springer, 2005.

[6] Richard Cowan. Expected frequencies of dna patterns using whittle's formula. *Journal of applied probability*, pages 886–892, 1991.

[7] Frank Den Hollander. *Large deviations*, volume 14. American Mathematical Soc., 2008.

[8] Morteza Ebneshahrashoob, Tangan Gao, and Mengnien Wu. An efficient algorithm for exact distribution of discrete scan statistics. *Methodology and Computing in Applied Probability*, 7(4):459–471, 2005.

[9] James W Fickett, David C Torney, and David R Wolf. Base compositional structure of genomes. *Genomics*, 13(4):1056–1064, 1992.

[10] James C Fu. Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, pages 957–974, 1996.

[11] James C Fu and WY Wendy Lou. *Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach*. World Scientific, 2003.

[12] JC Fu and MV Koutras. Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association*, 89(427):1050–1058, 1994.

[13] John E Hopcroft, Rajeev Motwani, and Jeffrey D Ullman. Introduction to automata theory, languages, and computation. *ACM SIGACT News*, 32(1): 60–65, 2001.

[14] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[15] Timo Koski. *Hidden Markov models for bioinformatics*, volume 2. Springer Science & Business Media, 2001.

[16] MV Koutras and VA Alexandrou. Runs, scans and urn model distributions: a unified Markov chain approach. *Annals of the Institute of Statistical Mathematics*, 47(4):743–766, 1995.

[17] Gregory Kucherov, Laurent Noé, and Mikhail Roytberg. Subset seed automaton. In *International Conference on Implementation and Application of Automata*, pages 180–191. Springer, 2007.

[18] Brian G Leroux. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

[19] Manuel E Lladser. Minimal Markov chain embeddings of pattern problems. In *2007 Information Theory and Applications Workshop*, pages 251–255. IEEE, 2007.

[20] Manuel E Lladser. Markovian embeddings of general random strings. In *2008 Proceedings of the Fifth Workshop on Analytic Algorithmics and Combinatorics (ANALCO)*, pages 183–190. SIAM, 2008.

[21] Manuel E Lladser and Stephen R Chestnut. Approximation of sojourn-times via maximal couplings: motif frequency distributions. *Journal of mathematical biology*, 69(1):147–182, 2014.

[22] Manuel E Lladser, Meredith D Betterton, and Rob Knight. Multiple pattern matching: A Markov chain approach. *Journal of mathematical biology*, 56 (1-2):51–92, 2008.

[23] M Lothaire. *Applied combinatorics on words*, volume 105. Cambridge University Press, 2005.

[24] Frosso S Makri and Zaharias M Psillakis. Exact distributions of constrained (k, ℓ) strings of failures between subsequent successes. *Statistical Papers*, 54 (3):783–806, 2013.

[25] Tobias Marschall and Sven Rahmann. Probabilistic arithmetic automata and their application to pattern matching statistics. In *Annual Symposium on Combinatorial Pattern Matching*, pages 95–106. Springer, 2008.

[26] Donald EK Martin. Application of auxiliary Markov chains to start-up demonstration tests. *European Journal of Operational Research*, 184(2):574–583, 2008.

[27] Donald EK Martin and John AD Aston. Distribution of statistics of hidden state sequences through the sum-product algorithm. *Methodology and Computing in Applied Probability*, 15(4):897–918, 2013.

[28] Donald EK Martin and Laurent Noé. Faster exact distributions of pattern statistics through sequential elimination of states. *Annals of the Institute of Statistical Mathematics*, pages 1–18, 2015.

[29] Donald EK Martin, Deidra A Coleman, et al. Distribution of clump statistics for a collection of words. *Journal of Applied Probability*, 48(4):1049–1059, 2011.

[30] Pierre Nicodeme, Bruno Salvy, and Philippe Flajolet. Motif statistics. *Theoretical Computer Science*, 287(2):593–617, 2002.

[31] G Nuel and V Delos. Counting regular expressions in degenerated sequences through lazy Markov chain embedding. In *Forging Connections between Computational Mathematics and Computational Geometry*, pages 235–246. Springer, 2016.

[32] Grégory Nuel. Ld-spatt: large deviations statistics for patterns on Markov chains. *Journal of Computational Biology*, 11(6):1023–1033, 2004.

[33] Grégory Nuel. Effective p-value computations using finite Markov chain imbedding (fmci): application to local score and to pattern statistics. *Algorithms for molecular biology*, 1(1):1, 2006.

[34] Gregory Nuel. Pattern Markov chains: optimal Markov chain embedding through deterministic finite automata. *Journal of Applied Probability*, pages 226–243, 2008a.

[35] Grégory Nuel. Waiting time distribution for pattern occurrence in a constrained sequence: an embedding Markov chain approach. *Discrete Mathematics and Theoretical Computer Science*, 10(3), 2008b.

[36] Grégory Nuel. Counting patterns in degenerated sequences. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 222–232. Springer, 2009.

[37] Grégory Nuel. On the first k moments of the random count of a pattern in a multistate sequence generated by a Markov source. *Journal of Applied Probability*, 47(4):1105–1123, 2010.

[38] Grégory Nuel. *Significance score of motifs in biological sequences*. INTECH Open Access Publisher, 2011.

[39] Grégory Nuel and Jean-Guillaume Dumas. Sparse approaches for the exact distribution of patterns in long state sequences generated by a Markov source. *Theoretical Computer Science*, 479:22–42, 2013.

[40] Gregory Nuel, Leslie Regad, Juliette Martin, and Anne-Claude Camproux. Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. *Algorithms for Molecular Biology*, 5(1):1, 2010.

[41] Bernard Prum, François Rodolphe, and Élisabeth de Turckheim. Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 205–220, 1995.

[42] Lawrence R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[43] Mireille Régnier and Alain Denise. Rare events and conditional events on random strings. *Discrete Mathematics & Theoretical Computer Science*, 6(2): 191–214, 2004.

[44] G Reinert, S Schbath, and MS Waterman. Probabilistic and statistical properties of finite words in finite sequences. *Lothaire: Applied Combinatorics on Words*, 2005.

[45] Paolo Ribeca and Emanuele Raineri. Faster exact Markovian probability functions for motif occurrences: a dfa-only approach. *Bioinformatics*, 24(24): 2839–2848, 2008.

[46] Stéphane Robin, François Rodolphe, and Sophie Schbath. *DNA, words and models: statistics of exceptional words*. Cambridge University Press, 2005.

[47] Christopher A Sims and Tao Zha. Were there regime switches in us monetary policy? *The American Economic Review*, 96(1):54–81, 2006.

[48] Valeri Stefanov and Wojciech Szpankowski. Waiting time distributions for pattern occurrence in a constrained sequence. *Discrete Mathematics and Theoretical Computer Science*, 9(1), 2007.

[49] Nicolas Vergne. Drifting Markov models with polynomial drift and applications to dna sequences. *Statistical applications in genetics and molecular biology*, 7(1), 2008.