Journal of Computational Biology
Volume 7, Numbers 1/2, 2000
Mary Ann Liebert, Inc.
Pp. 193–201

An Overview on the Distribution of Word Counts in Markov Chains

SOPHIE SCHBATH[1]

Abstract

In this paper, we give an overview about the different results existing on the statistical distribution of word counts in a Markovian sequence of letters. Results concerning the number of overlapping occurrences, the number of renewals and the number of clumps will be presented. Counts of single words and also multiple words are considered. Most of the results are approximations as the length of the sequence tends to infinity. We will see that Gaussian approximations switch to (compound) Poisson approximations for rare words. Modeling DNA sequences or proteins by stationary Markov chains, these results can be used to study the statistical frequency of motifs in a given sequence.

Key words: word count distribution, Markovian random sequence, overlapping occurrences, renewals, clumps.

1 Introduction

Probability on pattern occurrences in a random sequence of letters has been widely studied and has applications in many areas such as reliability, games, cryptography and molecular biology. The random sequences are mostly Bernoulli trials or Markovian sequences on a two-letter alphabet, but one can also find results on more general alphabets. Quantities of interest can be the length of the longest run, waiting times for first occurrences of runs and patterns, distances between occurrences and number of occurrences of patterns. Here are some references. The asymptotic behavior of the length of the longest run is studied in Erdös and Rényi (1970), Guibas and Odlyzko (1980), Samarova (1981), Nemetz and Kusolitsch (1982), Kusolitsch (1982), Arratia and Waterman (1985, 1989), Deheuvels *et al.* (1986) and Novak (1994, 1995). Mean, variance, or exact distribution of the waiting times for first occurrences of patterns are usually derived using generating functions as in Feller (1968), Li (1980), Guibas and Odlyzko (1981b), Blom (1982), Blom and Thorburn (1982), Thorburn (1983), Breen *et al.* (1985) for independent trials, and in Rajarshi (1974), Gerber and Li (1981), Schwager (1983), Benevento (1984), Banjevic (1988), Rudander (1996) and Stefanov and Pakes (1997) under Markovian hypothesis. Whereas Dembo and Karlin (1992) present asymptotic results on the distances between occurrences of

[1]Institut National de la Recherche Agronomique, Unité de Biométrie, 78352 Jouy-en-Josas, France, Sophie.Schbath@jouy.inra.fr

patterns using Poisson approximations, Robin and Daudin (1999) give the corresponding exact distributions.

We will now focus on probability and statistics on word counts in a random sequence of letters on a general alphabet $\mathcal{A}$. Let us consider overlapping occurrences for the moment. The challenging questions are "how many occurrences of a given word W should we expect in such random sequences of letters?" and "is this given word W significantly over- or underrepresented in this observed sequence?" The first question assumes we know the probabilistic model and its parameters, under which the random sequence has been drawn; it involves probability on word counts. The second question, however, involves statistics on word counts since the true model is unknown and one has to fit a model on the observed sequence; in particular, parameters will be estimated.

Three different ways of counting occurrences will be considered: the number of overlapping occurrences of a word W takes into account all occurrences of W, whereas the declumped count only takes into account the occurrences of W that do not overlap a previous occurrence of W. The latter is also the number of clumps of W in the sequence given that a clump is a maximal set of overlapping occurrences of W. For instance, in the sequence `ACGAATAATAAATAAGGCAATAAA`, there are four occurrences of `AATAA` (starting at positions 4, 7, 11 and 19) and two clumps of `AATAA` (starting at positions 4 and 19); the first clump is composed of three overlapping occurrences of `AATAA` whereas the second one is composed of a unique occurrence. In between, a renewal of W, as introduced by Feller (1968), is either the first occurrence of W in the sequence or an occurrence of W that does not overlap a previous renewal of W. In the above example, there are three renewals of `AATAA` (starting at positions 4, 11 and 19).

In Section 2, we will focus on the number of overlapping occurrences of words. We will review probabilistic and statistical results on the single word count distribution (Sections 2.1 and 2.2), with a special attention to rare words (Section 2.3); Section 2.4 is devoted to counts of multiple words. Section 3 concerns the renewal count (Section 3.1) and the declumped count (Section 3.2): two ways of counting nonoverlapping occurrences of a word in a sequence.

2 Count of overlapping occurrences of a word

In this section, we consider the number of overlapping occurrences of a h-letter word $W = w_1 \cdots w_h$ on the alphabet $\mathcal{A}$, denoted by $N(W)$, in a random sequence of length n on $\mathcal{A}$. By defining the random variables Y_i by

$$Y_i = \mathbb{1}\{\text{an occurrence of } W \text{ starts at position } i \text{ in the sequence}\},$$

the count $N(W)$ is simply given by

$$N(W) = \sum_{i=1}^{n-h+1} Y_i.$$

Denote by $\mu_{\mathrm{model}}(W)$ the probability that an occurrence of W starts at a given position in the random sequence. If the sequence is a stationary m-order Markov chain with transition matrix $\Pi = (\pi(a_1 \cdots a_m, a_{m+1}))_{a_1,\ldots,a_{m+1} \in \mathcal{A}}$ and stationary distribution $(\mu(a_1 \cdots a_m))_{a_1,\ldots,a_m \in \mathcal{A}}$, the so called model M$m$, then

$$\mu_m(W) = \mu(w_1 \cdots w_m) \prod_{j=1}^{h-m} \pi(w_j \cdots w_{j+m-1}, w_{j+m}).$$

Independent trials with probability letters $(\mu(a))_{a \in \mathcal{A}}$, called model M0, is a particular case of model M1: the transition matrix Π has identical rows $(\mu(a))_{a \in \mathcal{A}}$ and the stationary distribution is $(\mu(a))_{a \in \mathcal{A}}$.

2.1 Probability on word counts

Let us suppose we know the true model, independent trials (M0) or one-order Markov chain model (M1), used to draw the random sequence of length n. One can calculate the mean and variance of the count $N(W)$ of the word W (Lundstrom (1990), Kleffe and Borodovsky (1992)) and obtain the asymptotic normality of $n^{-1/2}N(W)$ as the length n of the sequence grows to infinity (Lundstrom (1990)). In M1 the expressions are as follows:

$$\begin{aligned}
\mathbb{E}_1 N(W) &= (n-h+1)\mu_1(W), \\
\mathrm{Var}_1 N(W) &= (n-h+1)\mu_1(W) + 2 \sum_{\substack{p \in \mathcal{P}(W) \\ p \le h-2}} (n-h-p+1)\mu_1(w_1 \cdots w_p w_1 \cdots w_h) \\
&\quad + \mu_1^2(W) \left(-(n-h+1)^2 + \frac{2}{\mu(w_1)} \sum_{d=1}^{n-2h+1} (n-2h+2-d)\Pi^d(w_h, w_1) \right)
\end{aligned}$$

where $\mathcal{P}(W)$ is the period[2] set of W. Note that, except in M0, the variance formula contains an "infinite" sum that penalizes its computation. Régnier (1999) proposes another formula that can be calculated through a symbolic computation system such as Maple. The author also provides the moment-generating function using a combinatorial method; similar results are presented in Robin and Daudin (1999) (see also Régnier and Szpankowski (1997)). The variance and more generally the distribution of $N(W)$ strongly depend on the overlapping structure of the word.

2.2 Statistics on word counts

As soon as the true model is unknown and the parameters of the fitted model (the transition probabilities of a Markov chain for instance) are estimated according to the observed sequence of letters, the exact distribution results vanish. Even if the Gaussian approximation of $n^{-1/2}(N(W) - \widehat{N}(W))$, where $\widehat{N}(W)$ is a natural estimator of the expected count, is

[2] A period of a word W is here an integer $p \in \{1, \ldots, h-1\}$ such that $w_i = w_{i+p}$, $\forall i \in \{1, \ldots, h-p\}$; a period then corresponds to a possible lag between two overlapping occurrences of W.

still valid, the previous asymptotic results have to be modified. In particular, the variance of $(N(W)-\widehat{N}(W))$ is different from the variance of $N(W)$. Whereas Pevzner *et al.* (1989) and Stűckle *et al.* (1990) did not take into account this remark, Brendel *et al.* (1986), Avery (1987) and Phillips *et al.* (1987) approximated it with $\widehat{N}(W)$, as for a Poisson random variable, neglecting the overlapping structure of the word. Using the δ-method, Lundstrom (1990) derived the asymptotic variance of $(N(W)-\widehat{N}(W))$, under Markovian hypothesis, clearly different from the asymptotic variance of $N(W)$ (see Waterman (1995) for an exposition). However, the δ-method is quite tedious. Prum *et al.* (1995) gave an easier way to calculate this asymptotic variance, based on the asymptotic conditional variance of $N(W)$ given the sufficient statistic of the Markov chain model. Their method has been generalized to m-order Markov chain models, with homogeneous transition probabilities or with 3-periodic transition probabilities (Schbath (1995b), Schbath *et al.* (1995)). In the model Mm and for the following natural estimator

$$\widehat{N}_m(W) = N(w_1\cdots w_{m+1})\prod_{j=2}^{h-m}\frac{N(w_j\cdots w_{j+m})}{N(w_j\cdots w_{j+m-1})},$$

the asymptotic variance of $n^{-1/2}(N(W)-\widehat{N}_m(W))$ is

$$\begin{aligned}\sigma_m^2(W) &= \mu_m(W) + 2\sum_{\substack{p\in\mathcal{P}(W)\\ p\leq h-m-1}}\mu_m(w_1\cdots w_p w_1\cdots w_h)\\ &\quad +\mu_m(W)^2\left(\sum_{a_1,\dots,a_m}\frac{n(a_1\cdots a_m\bullet)^2}{\mu(a_1\cdots a_m)}-\sum_{a_1,\dots,a_{m+1}}\frac{n(a_1\cdots a_{m+1})^2}{\mu_m(a_1\cdots a_{m+1})}+\frac{1-2n(w_1\cdots w_m\bullet)}{\mu(w_1\cdots w_m)}\right),\end{aligned}$$

where $n(\cdot)$ denotes the number of occurrences inside W and $n(a_1\cdots a_m\bullet)$ stands for $\sum_{b\in\mathcal{A}} n(a_1\cdots a_m b)$ (see also Schbath (1997)).

The asymptotically standard Gaussian statistic $(N(W)-\widehat{N}_m(W))/\sqrt{n\widehat{\sigma}_m^2(W)}$, where $\widehat{\sigma}_m^2(W)$ is the natural plug-in estimator of $\sigma_m^2(W)$, can then be used to find words with exceptional frequency in DNA sequences as it is done in the *R'MES* software (Gélis and Schbath (1996)).

2.3 Probability and statistics on rare word occurrences

Other approaches have been used when W is a rare word, meaning that the expected count $\mathbb{E}N(W)$ is too small. Rare words can be assimilated to long words. For independent trials, Poisson or compound Poisson approximations of $N(W)$ have been obtained using generating functions (Chryssaphinou and Papastavridis (1988), Hirano and Aki (1993)) or using the Chen-Stein method (Arratia *et al.* (1990), Godbole (1991), Godbole (1994)). The Chen-Stein method, as stated in Arratia *et al.* (1989), is a powerful tool since it gives a bound on the Poisson or compound Poisson approximation error. One can then control the goodness of the approximation (cf. Figure 1), contrarily to the Gaussian approximation. Since there is a tendency towards a clustering of the occurrences of periodic words[3], in the sense of Guibas and Odlyzko (1981a),

[3] A word having at least one nonzero period is called a periodic word

a compound Poisson approximation is more appropriate than a Poisson approximation. For Markovian-dependent trials, the difficulty comes from the periodic structure of the words. Karlin and Ost (1987) first proved, using generating functions, that the count of success runs in a Markov chain on $\{0,1\}$ has a limiting compound Poisson distribution. Then, using the Chen-Stein method, Geske *et al.* (1995) proved a compound Poisson approximation for the count of words having a unique principal[4] period like `ACTACTAC`. However, parameters of the limiting distributions were not provided. The result of Geske *et al.* has been generalized to general patterns in Schbath (1995a) and parameters of the compound Poisson distribution were derived. Finally, the word count $N(W)$ can then be approximated by a sum $\sum_{k\geq 1} kZ_k$ with independent Poisson variables Z_k whose expectation $(n-h+1)\widetilde{\mu}_m^{(k)}(W)$ is given in Mm by

$$\widetilde{\mu}_m^{(k)}(W) = (1 - A_m(W))^2 A_m(W)^{k-1} \mu_m(W)$$

and

$$A_m(W) = \sum_{\substack{p \in \mathcal{P}'(W) \\ p \leq h-m}} \prod_{j=1}^{p} \pi(w_j \cdots w_{j+m-1}, w_{j+m});$$

$\mathcal{P}'(W)$ is the set of the principal periods of W. The parameters $\widetilde{\mu}_m^{(k)}(W)$ are exactly the expected frequencies of clumps of exactly k occurrences of W. The error bounds obtained in the two previous papers seem to be improved using the Stein method for compound Poisson distribution (Roos (1994), Erhardsson (1999b)). Related results have been announced by Erhardsson (1999a).

All these works concern probability on word counts. The advantage of Poisson and compound Poisson approximations is their validity when parameters of the model are estimated with the maximum likelihood estimators (MLE). Indeed, since the Chen-Stein method gives the total variation distance between the count $N(W)$ and an ad-hoc (compound) Poisson variable, it is easy to check that the total variation distance between two (compound) Poisson variables converges to zero as soon as the parameters of these two distributions are sufficiently asymptotically equivalent. This is the case using MLEs (Schbath (1995b)). Thus, the compound Poisson approximation obtained by Schbath (1995a) can be used to study the statistical frequency of long words in DNA sequences (Gélis and Schbath (1996)).

2.4 Probability and statistics on counts of multiple words

Rather than considering a single word, one may be interested in multiple words or in word families (so called generalized words). Lundstrom (1990) first derived a joint central limit theorem for the vector $n^{-1/2}(N(W_j) - \widehat{N}(W_j))_{j=1,\dots,s}$ associated to s words W_1, W_2, ..., W_s and the asymptotic covariance matrix is obtained using the δ-method. However, the method developed in Prum *et al.* (1995) allows to obtain directly the asymptotic covariance matrix.

[4]A principle period of a word is either the minimal period of the word or a period that is not a multiple of the minimal period

In the model Mm, the asymptotic covariance corresponding to the words $W = w_1 \cdots w_h$ and $W' = w'_1 \cdots w'_h$ is

$$\begin{aligned}
\sigma_m(W,W') \quad = \quad & \sum_{\substack{p \in \mathcal{P}(W,W') \\ p \leq h-m-1}} \mu_m(w_1 \cdots w_p w'_1 \cdots w'_{h'}) + \sum_{\substack{p \in \mathcal{P}(W',W) \\ p \leq h'-m-1}} \mu_m(w'_1 \cdots w'_p w_1 \cdots w_h) \\
& + \mu_m(W)\mu_m(W') \left(\sum_{a_1,\ldots,a_m} \frac{n(a_1 \cdots a_m \bullet) n'(a_1 \cdots a_m \bullet)}{\mu(a_1 \cdots a_m)} \right. \\
& \qquad - \sum_{a_1,\ldots,a_{m+1}} \frac{n(a_1 \cdots a_{m+1}) n'(a_1 \cdots a_{m+1})}{\mu_m(a_1 \cdots a_{m+1})} \\
& \left. + \frac{\mathbb{1}\{w_1 \cdots w_m = w'_1 \cdots w'_m\}}{\mu(w_1 \cdots w_m)} - \frac{n'(w_1 \cdots w_m \bullet)}{\mu(w_1 \cdots w_m)} - \frac{n(w'_1 \cdots w'_m \bullet)}{\mu(w'_1 \cdots w'_m)} \right),
\end{aligned}$$

where $\mathcal{P}(W,W')$ is the set of integers $0 < p < h$ such that an occurrence of W' can start p letters after an occurrence of W (see Figure 1 for an example), $n(\cdot)$ denotes the number of occurrences inside W and $n'(\cdot)$ denotes the number of occurrences inside W'. One can then derive the asymptotic normality of

$$N(\mathcal{W}) = \sum_{W \in \mathcal{W}} N(W)$$

where the word family $\mathcal{W}$ is the set of words $\{W_1, W_2, \ldots, W_s\}$; it can be used to detect exceptional families of words in biological sequences (Bouvier *et al.* (1999)).

For rare words, Reinert and Schbath (1998) provide a compound Poisson process approximation of the vector $(N(W_1), N(W_2), \ldots, N(W_s))$, by following and improving the method in Schbath (1995a). The fewer overlaps between W_1, W_2, $\ldots$, W_s the better the approximation; this guarantees the asymptotic independence of the counts and allows to derive a compound Poisson approximation for $N(\mathcal{W})$. Figure 1 gives an example of the goodness of the approximation. Erhardsson (1997) obtains a similar compound Poisson approximation for $N(\mathcal{W})$ by using a Markov chain coupling rather than a local approach, leading to a different type of error bound. A comparison could be of interest.

Concerning the exact distribution, Régnier (1999) gives the moment-generating function of the vector $(N(W_1), N(W_2), \ldots, N(W_s))$ when the model parameters are supposed to be known.

`Length of the sequence:` $n = 48502$ `Alphabet=`$\{A, C, G, T\}$

$$\texttt{Transition matrix:}\Pi = \begin{pmatrix} 0.2994 & 0.2086 & 0.2215 & 0.2705 \\ 0.2830 & 0.2198 & 0.2740 & 0.2232 \\ 0.2540 & 0.2820 & 0.2480 & 0.2160 \\ 0.1813 & 0.2232 & 0.3164 & 0.2791 \end{pmatrix}$$

$$\begin{array}{lll} & & \mathcal{P}(W_1, W_1) = \{5, 10, 11\} \\ W_1 = \texttt{GGACTGGACTGG} & \mathbb{E}_1 N(W_1) = 0.0027 & \mathcal{P}(W_1, W_2) = \{6, 11\} \\ W_2 = \texttt{GACTGGTAGGAC} & \mathbb{E}_1 N(W_2) = 0.0013 & \mathcal{P}(W_2, W_1) = \{8\} \\ & & \mathcal{P}(W_2, W_2) = \{9\} \end{array}$$

• Error bounds of the one-dimensional compound Poisson approximations of $N(W_1)$ and $N(W_2)$

$$\begin{aligned} \forall t > 0,\ |\mathbb{P}(N(W_1) \geq t) - \mathbb{P}(\mathrm{CP}_1 \geq t)| &\leq 3.6\text{e-}08 \\ \forall t > 0,\ |\mathbb{P}(N(W_2) \geq t) - \mathbb{P}(\mathrm{CP}_2 \geq t)| &\leq 8.6\text{e-}09 \end{aligned}$$

• Error bound of the joint compound Poisson approximation of $(N(W_1), N(W_2))$

$$\sup_{B \subset \mathbb{N}^2} |\mathbb{P}((N(W_1), N(W_2)) \in B) - \mathbb{P}((\mathrm{CP}_1, \mathrm{CP}_2) \in B)| \leq 7.5\text{e-}07$$

Figure 1: Consider a sequence of length 48502 generated by a one-order Markov chain with transition matrix Π chosen according to the phage *Lambda* genome, and the two 12-letter words `GGACTGGACTGG` and `GACTGGTAGGAC`. These words are rare in the sequence since their expected counts are respectively 0.0027 and 0.0013. They are periodic words. A period of a word is a possible lag between two overlapping occurrences of this word; for instance, `GGACTGGACTGG` has three periods. In the same way, a period of the ordered pair of words (W_1, W_2) is the lag between an occurrence of W_2 that may overlap a previous occurrence of W_1. Here, an occurrence of W_1 may overlap an occurrence of W_2 in three different ways in the sequence. Denoting respectively by CP_1 and CP_2 the limiting compound Poisson distributions associated with the counts $N(W_1)$ and $N(W_2)$, the Chen-Stein method allows to control the approximation error. We can see that the one-dimensional approximations of $N(W_1)$ and $N(W_2)$, and the two-dimensional approximation of $(N(W_1), N(W_2))$ are good.

3 Counts of non-overlapping occurrences of a word

3.1 Probability on renewal counts

The count $R(W)$ of renewals of W in a random sequence of letters, as defined in the Introduction, has also been widely studied. The model is supposed to be known. For rare words, Poisson approximations were obtained: some with an underlying model of independent trials (Chryssaphinou and Papastavridis (1988), Godbole (1991), Godbole and Schaffner (1993), Fu (1993)), others in the Markovian case first for runs (Rajarshi (1974), Godbole (1991), Godbole and Schaffner (1993)) and for general patterns recently in Reinert *et al.* (1999). For nonrare words, results of Breen *et al.* (1985) and Biggins and Cannings (1987) on the mean recurrence time between competing renewals, respectively under independent and Markovian-dependent trials, have been extended to derive joint central limit theorems for the vector of counts of nonoverlapping occurrences of s words as competing renewals (Tanushev and Arratia (1997), Tanushev (1996)). We state the result for a unique word W in the model M1. Let Γ be the $|\mathcal{A}| \times |\mathcal{A}|$ matrix where each row is the vector $(\mu(a), a \in \mathcal{A})$, let Π be the transition matrix and put $Z := \sum_{k\geq 1} (\Pi - \Gamma)^k$. By denoting $Q(z)$ the overlap-matching polynomial associated with W:

$$Q(z) := \sum_{p \in \mathcal{P}(W) \cup \{0\}} \prod_{j=1}^{p} \pi(w_j, w_{j+1}) z^p,$$

one has

$$n^{-1/2} \left(R(W) - \frac{\mathbb{E}_1 N(W)}{Q(1)} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, v)$$

with

$$v = \left(\frac{\mu_1(W)}{Q(1)} \right)^2 \left(1 - 2h + 2\frac{Q'(1)}{Q(1)} + \frac{2Z(w_h, w_1)}{\mu(w_1)} \right);$$

$Q'(\cdot)$ is the derivative of $Q(\cdot)$.

In regard to exact results, the generating function of the renewal count was first derived for runs in a stationary $\{0,1\}$-valued Markov chain (Hirano and Aki (1993)) and recently by Régnier (1999) for general patterns in a Markov chain. This paper also provides the correct mean in M1, namely

$$\mathbb{E}_1 R(W) = \frac{\mathbb{E}_1 N(W)}{Q(1)} + \frac{\mu_1(W) Q'(1)}{Q^2(1)}.$$

3.2 Probability and statistics on declumped counts

With the convention that $Y_i = 0$ if $i < 1$, the number of clumps of W, denoted by $\widetilde{N}(W)$, is given by

$$\widetilde{N}(W) = \sum_{i=1}^{n-h+1} Y_i (1 - Y_{i-1})(1 - Y_{i-2}) \cdots (1 - Y_{i-h+1}).$$

As far as we know, no results on the exact distribution exist in the literature.

From Eq. (1) in Schbath (1995a), there exists a word family $\mathcal{W}$ depending on W such that

$$\widetilde{N}(W) = N(W) - N(\mathcal{W}).$$

Therefore, one can easily derive a Gaussian approximation of $n^{-1/2}\widetilde{N}(W)$ using results on multiple words described in Section 2.4, especially the asymptotic covariance formulae.

Finally, because clumps do not overlap in the sequence, the declumped count of rare words is well approximated by a Poisson variable (Karlin and Ost (1987), Arratia *et al.* (1990), Schbath (1995a)). This approximation has only been achieved in M0 and M1, and the mean of the limiting Poisson distribution is $(n-h+1)\widetilde{\mu}_m(W)$ with (in M1):

$$\widetilde{\mu}_1(W) = \mu_1(W) - \sum_{p \in \mathcal{P}'(W)} \mu_1(w_1 \cdots w_p w_1 \cdots w_h).$$

4 Conclusion

Exact distributions of word counts (overlapping occurrences and renewals) have been derived in Markov chains with known transition probabilities. It is particularly of interest for short sequences. For very long sequences, it is often easier to use the limiting distribution. Depending on the word and on the method of counting word occurrences, three limiting distribution have been obtained:

- Gaussian approximations for counts of nonrare words (overlapping occurrences, renewals and clumps),
- Poisson approximations for renewal counts of rare words, for declumped counts of rare words and for numbers of "overlapping" occurrences of nonperiodic rare words, and finally,
- compound Poisson approximations for numbers of overlapping occurrences of rare words.

Note that, for rare words, the difference between Poisson and compound Poisson approximations comes simply from the fact that counted occurrences may or may not overlap in the sequence: by definition, renewals, clumps and occurrences of nonperiodic words do not overlap in the sequence, leading to a simple Poisson approximation.

In practice, DNA sequences and proteins are not Markov chains. To study the significance of word counts in a biological sequence, one usually considers a Markov chain whose transition probabilities are not known but are estimated according to the observed sequence. Using an m-order Markov chain consists in comparing what occurs in the biological sequence with what should occur in a random sequence having on average the same 1-, 2-, ..., $(m+1)$-letter words. In this case, one can only provide asymptotic results; the asymptotic restriction here ensures that estimated transition probabilities are very close to true transition probabilities. This requires a long sequence, which is often the case with DNA sequences. In this framework, limiting distributions for the number of overlapping occurrences and the number of clumps are those obtained under the known model (summarized in the previous paragraph) by plugging in the MLE estimators of the transition probabilities, except for the asymptotic variances of the Gaussian distributions that are really different ones.

Therefore, these results allow to identify words or word families for which the number of overlapping occurrences or the number of clumps are not expected under a suitable model for the observed sequence.

Acknowledgment

I really thank the referee for hepful remarks improving this overview. This work was partially supported by the French CNRS Genome Project 34.

References

Arratia, R., Goldstein, L. and Gordon, L. 1989. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Prob.* 17, 9–25.

Arratia, R., Goldstein, L. and Gordon, L. 1990. Poisson approximation and the Chen-Stein method. *Statistical Science.* 5, 403–434.

Arratia, R. and Waterman, M. S. 1985. Critical phenomena in sequence matching. *Ann. Prob.* 13, 1236–1249.

Arratia, R. and Waterman, M. S. 1989. The Erdös-Rényi strong law for pattern matching with a given porportion of mismatches. *Ann. Prob.* 17, 1152–1169.

Avery, P. J. 1987. The analysis of intron data and their use in the detection of short signals. *J. Mol. Evol.* 26, 335–340.

Banjevic, D. 1988. On some statistics connected with runs in Markov chains. *J. Appl. Prob.* 25, 815–821.

Benevento, R. 1984. The occurrence of sequence patterns in ergodic Markov chains. *Stoch. Proc. Appl.* 17, 369–373.

Biggins, J. D. and Cannings, C. 1987. Markov renewal processes, counters and repeated sequences in Markov chains. *Adv. Appl. Prob.* 19, 521–545.

Blom, G. 1982. On the mean number of random digits until a given sequence occurs. *J. Appl. Prob.* 19, 136–143.

Blom, G. and Thorburn, D. 1982. How many random digits are required until given sequences are obtained? *J. Appl. Prob.* 19, 518–531.

BOUVIER, A., GÉLIS, F. and SCHBATH, S. 1999. *R'MES : Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 2.* Guide de l'utilisateur. INRA, Biométrie, F78352 Jouy-en-Josas. (`http://www-bia.inra.fr/J/AB/genome/RMES/welcome.html`).

BREEN, S., WATERMAN, M. S. and ZHANG, N. 1985. Renewal theory for several patterns. *J. Appl. Prob.* 22, 228–234.

BRENDEL, V., BECKMANN, J. S. and TRIFONOV, E. N. 1986. Linguistics of nucleotide sequences: Morphology and comparison of vocabularies. *J. Biomol. Struct. Dynamics.* 4, 11–21.

CHRYSSAPHINOU, O. and PAPASTAVRIDIS, S. 1988. A limit theorem on the number of overlapping appearances of a pattern in a sequence of independent trials. *Prob. Theory Rel. Fields.* 79, 129–143.

DEHEUVELS, P., DEVROYE, L. and LYNCH, J. 1986. Exact convergence rate in the limit theorems of Erdös-Rényi and Shepp. *Ann. Prob.* 14, 209–223.

DEMBO, A. and KARLIN, S. 1992. Poisson approximations for r-scan processes. *Ann. Appl. Prob.* 2, 329–357.

ERDÖS, P. and RÉNYI, A. 1970. On a new law of large numbers. *J. Analyse Math.* 23, 103–111.

ERHARDSSON, T. 1997. *Compound Poisson approximation for Markov chains.* PhD thesis, Royal Institute of Technology, Stockholm.

ERHARDSSON, T. 1999a. Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. Preprint.

ERHARDSSON, T. 1999b. Compound Poisson approximation for Markov chains using Stein's method. *Ann. Prob.* 27, 565–596.

FELLER, W. 1968. *An introduction to Probability Theory and its Applications.* Wiley, New York.

FU, J. C. 1993. Poisson convergence in reliability of a large linearly connected system as related to coin tossing. *Statistica Sinica.* 3, 261–275.

GERBER, H. U. and LI, S.-Y. 1981. The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain. *Stoch. Proc. Appl.* 11, 101–108.

GESKE, M. X., GODBOLE, A. P., SCHAFFNER, A. A., SKOLNICK, A. M. and WALLSTROM, G. L. 1995. Compound Poisson approximations for word patterns under Markovian hypotheses. *J. Appl. Prob.* 32, 877–892.

GÉLIS, F. and SCHBATH, S. 1996. *R'MES : Recherche de Mots Exceptionnels dans les Séquences d'ADN – Version 1.* Guide de l'utilisateur. INRA, Biométrie, F78352 Jouy-en-Josas. (`http://www-bia.inra.fr/J/AB/genome/RMES/welcome.html`).

GODBOLE, A. P. 1991. Poisson approximations for runs and patterns of rare events. *Adv. Appl. Prob.* 23, 851–865.

GODBOLE, A. P. and SCHAFFNER, A. A. 1993. Improved poisson approximations for word patterns. *Adv. Appl. Prob.* 25, 334–347.

GODBOLE, A. 1994. Applications of the Chen-Stein method to the theory of patterns and runs: an annotated bibliography, 253–261. In Godbole, A. and Papastavridis, S., eds., *Runs and Patterns in Probability*, Kluwer.

GUIBAS, L. J. and ODLYZKO, A. M. 1980. Long repetitive patterns in random sequences. *Z. Wahrscheinlichkeitsth.* 53, 241–262.

GUIBAS, L. J. and ODLYZKO, A. M. 1981a. Periods in strings. *J. Combinatorial Theory A.* 30, 19–42.

GUIBAS, L. J. and ODLYZKO, A. M. 1981b. String overlaps, pattern matching, and nontransitive games. *J. Combinatorial Theory A.* 30, 183–208.

HIRANO, K. and AKI, S. 1993. On number of occurrences of sucess runs of specified length in a two-state Markov chain. *Statistica Sinica.* 3, 313–320.

KARLIN, S. and OST, F. 1987. Counts of long aligned word matches among random letter sequences. *Ann. Prob.* 19, 293–351.

KLEFFE, J. and BORODOVSKY, M. 1992. First and second moment of counts of words in random texts generated by Markov chains. *Comp. Applic. Biosci.* 8, 433–441.

KUSOLITSCH, N. 1982. Longest runs in Markov chains. *Prob. Statist. Inference.* 223–230.

LI, S. 1980. A martingale approach to the study of occurrence of sequences patterns in repeated experiments. *Ann. Prob.* 8, 1171–1176.

LUNDSTROM, R. 1990. *Stochastic models and statistical methods for DNA sequence data.* PhD thesis, University of Utah.

NEMETZ, T. and KUSOLITSCH, N. 1982. On the longest run of coincidences. *Z. Wahrscheinlichkeitsth.* 61, 59–73.

NOVAK, S. Y. 1994. Poisson approximation for the number of long match patterns in a random sequences. *Theory Prob. Appl.* 39.

NOVAK, S. Y. 1995. Long match patterns in random sequences. *Siberian Adv. Math.* 5, 128–140.

PEVZNER, P. A., BORODOVSKY, M. Y. and MIRONOV, A. A. 1989. Linguistics of nucleotides sequences I: The significance of deviations from mean statistical characteristics and prediction of the frequencies of occurrence of words. *J. Biomol. Struct. Dynamics.* 6, 1013–1026.

PHILLIPS, G. J., ARNOLD, J. and IVARIE, R. 1987. The effect of codon usage on the oligonucleotide composition of the *e. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. *Nucl. Acids Res.* 15, 2627–2638.

PRUM, B., RODOLPHE, F. and TURCKHEIM, É. DE 1995. Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* 57, 205–220.

RAJARSHI, M. B. 1974. Success runs in a two-state Markov chain. *J. Appl. Prob.* 11, 190–192.

REINERT, G., SCHBATH, S. and WATERMAN, M. 1999. Probabilistic and statistical properties of words. In Berstel, J. and Perrin, D., eds. *Applied Combinatorics on Words*, to appear.

REINERT, G. and SCHBATH, S. 1998. Compound Poisson and Poisson process approximations for occurrences of multiple words in Markov chains. *J. Comp. Biol.* 5, 223–253.

RÉGNIER, M. 1999. A unified approach to word occurrence probabilities. *To appear in Discrete Applied Mathematics, Special Issue on Computational Biology.*

RÉGNIER, M. and SZPANKOWSKI, W. 1997. On pattern frequency occurences in a markovian sequence. Purdue University, CSD-TR-96-043; preliminary draft at ISIT'97; *submitted.*

ROBIN, S. and DAUDIN, J.-J. 1999. Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Prob.* 36, 179–193.

ROOS, M. 1994. Stein's method for compound Poisson approximation : the local approach. *Ann. Appl. Prob.* 4, 1177–1187.

RUDANDER, J. 1996. *On the first occurrence of a given pattern in a semi-Markov process.* PhD thesis, University of Uppsala.

SAMAROVA, S. 1981. On the length of the longest head run for a Markov chain with two states. *Theory Prob. Appl.* 26, 498–509.

SCHBATH, S. 1995a. Compound poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics.* 1, 1–16. (http://www.emath.fr/ps/).

SCHBATH, S. 1995b. *Étude asymptotique du nombre d'occurrences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquence exceptionnelle dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V.

SCHBATH, S. 1997. An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comp. Biol.* 4, 61–82.

SCHBATH, S., PRUM, B. and TURCKHEIM, É. DE 1995. Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comp. Biol.* 2, 417–437.

SCHWAGER. 1983. Run probabilities in sequences of Markov-dependent trials. *J. Amer. Statist. Assoc.* 78, 168–175.

STEFANOV, V. and PAKES, A. G. 1997. Explicit distributional results in pattern formation. *Ann. Appl. Prob.* 7, 666–678.

STŰCKLE, E. E., EMMRICH, C., GROB, U. and NIELSEN, P. J. 1990. Statistical analysis of nucleotide sequences. *Nucl. Acids Res.* 18, 6641–6647.

TANUSHEV, M. S. 1996. Central limit theorem for several patterns in a Markov chain sequence of letters. Preprint.

TANUSHEV, M. S. and ARRATIA, R. 1997. Central limit theorem for renewal theory for several patterns. *J. Comp. Biol.* 4, 35–44.

THORBURN, D. 1983. On the mean number of trials until the last trials satisfy a given condition. *Stoch. Proc. Appl.* 16, 211–217.

WATERMAN, M. S. 1995. *Introduction to computational biology.* Chapman & Hall.