JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 4, 2001
Mary Ann Liebert, Inc.
Pp. 349–359

Numerical Comparison of Several Approximations of the Word Count Distribution in Random Sequences

STÉPHANE ROBIN and SOPHIE SCHBATH[1]

Abstract

The exact distribution of word counts in random sequences and several approximations have been proposed in the past few years. The exact distribution has no theoretical limit but may require prohibitive computation time. On the other hand, approximate distributions can be rapidly calculated but, in practice, are only accurate under specific conditions. After making a survey of these distributions, we compare them according to both their accuracy and computational cost. Rules are suggested for choosing between Gaussian approximations, compound Poisson approximation and exact distribution. This work is illustrated with the detection of exceptional words in the phage *Lambda* genome.

Key words: approximate distributions, exact distribution, exceptional words, simulations, word count.

1 Introduction

Counts of words or motifs are part of the elementary statistics used in biological sequence analysis, for example to find significantly under- or overrepresented words. Knowledge of the statistical distribution of these counts is necessary to assess the significance of the observed results.

The exact distribution of the count of a word is known under the hypothesis that the letters are independent (Bernoulli) or under a Markov model: see Régnier (1999) or Robin and Daudin (1999). It is theoretically possible to calculate this distribution for any word in any sequence. In practice, it is impossible to compute it in a reasonable time for long sequences or for very frequent words. On the other hand, two kinds of approximations exist: Gaussian approximations (Kleffe and Borodovsky (1992), Prum *et al.* (1995)) and Poisson approximations (Schbath (1995), Geske *et al.* (1995)). The asymptotic framework in which these approximations are valid are different but they both require that the length of the sequence tends to infinity. Their main advantage is that, in most cases, they require much shorter computation times.

This work consists then in comparing these different methods considering both the accuracy of the approximations and the computation time. We aim to answer the question: *Which method should I use to study the frequency of this word in a sequence of this length?*. We distinguish here the case where the parameters of the model are known (probabilistic approach) from the case where they need to be estimated on the analyzed sequence (statistical

[1]INRA, Unité Mathématique, Informatique & Génome, F78026 Versailles.
e-mail: Stephane.Robin@versailles.inra.fr, Sophie.Schbath@versailles.inra.fr.

approach). In the second case, the estimated expected count is not independent from the observed count.

We recall in Section 2 the different results existing on the distribution of the count of a given word (more technical details are given in the appendix). We describe in Section 3 the simulations design (factors and comparison criteria). Section 4 presents the results of the probabilistic approach; those of the statistical approach form Section 5. As a concrete situation, we finally study in Section 6 the influence of the different existing methods on the detection of under- or overrepresented 3-, 6- and 9-nucleotides in the complete genome of the phage *Lambda*.

2 Exact distribution and approximations

We are concerned with the occurrences of a k-letter word $w = w_1 \cdots w_k$ in a random sequence S of ℓ letters from a finite alphabet $\mathcal{A}$. The exact distribution of the word count and the asymptotic distributions presented here are available under a Markov model for the sequence (see Schbath (2000) or Reinert *et al.* (2000) for an overview). Using Markov models is a way to take into account the observed frequencies of short words: nucleotides for Markov model of order zero (Bernoulli model), dinucleotides for first order Markov model, etc. However, we restrict ourselves to the Bernoulli model because the distributions, especially the exact distribution, are more easily computable. Increasing the order of the Markov model does not affect the quality of the approximation but increases the computation time. We then assume that the letters $\{S_i\}_{i=1,\dots,\ell}$ are independent and identically distributed according to a given distribution μ on $\mathcal{A}$. Under this model, the probability that the word w occurs at a given position in S is

$$\pi(w) = \mu(w_1) \times \cdots \times \mu(w_k).$$

Let $N = N_\ell$ denote the number of overlapping occurrences of w in the sequence S. For instance $w =$ `aaa` occurs four times in `aaatgaaaaac`, starting at positions 1, 6, 7 and 8. We will focus on the distribution of N defined by

$$p(n) = \Pr\{N = n\}, \quad \forall n \geq 0,$$

or by its cumulative distribution function $P(n) = \Pr\{N \leq n\}$.

2.1 Exact distribution

The exact distribution of the count N is given in Régnier (1999) through its probability generating function, which is a rational function. The probabilities $p(n)$ are the coefficients of the Taylor expansion of this generating function; in the case of a rational function, this expansion can be obtained with a finite recurrence.

The exact distribution of the count N_ℓ can also be obtained via the distribution of the position X_n of the n-th occurrence of w in S, thanks to the equality $\Pr\{X_n \leq \ell - k + 1\} = \Pr\{N_\ell \geq n\}$, for all n. Because the generating function of X_n given in Robin and Daudin (1999) is explicit and the recurrence is easy to program (see Appendix A.1), we shall use it to calculate the $p(n)$'s. In the most favorable cases, the computation time is in $O(kn \log \ell)$ but, for the sake of numerical stability, an elementary algorithm in $O(kn\ell)$ is often more satisfying.

2.2 Gaussian approximations (KB and PRT).

If the distribution μ of the letters is known, the mean $\mathbb{E}_\mu(N)$ and variance $\mathbb{V}_\mu(N)$ of the count N can be calculated (see Kleffe and Borodovsky (1992) or Appendix A.2 for the formulas). Using a central limit theorem, the distribution of N converges to a Gaussian distribution as the length of the sequence grows to infinity.

For the case where the distribution μ of the letters is unknown but estimated with the vector $\widehat{\mu}$ of the observed frequencies of the letters in a given sequence, Prum *et al.* (1995) propose another Gaussian approximation. This approximation takes the dependency between the observed count and the expected count $\mathbb{E}_{\widehat{\mu}}(N)$ into account: in this, the normalization factor is different from the standard deviation of the count (see Appendix A.3).

Both of these approximations are based on the fact that the expected count is a linear function of the length of the sequence, which is valid only for frequent words. Let p_{KB} and p_{PRT} denote the corresponding Gaussian distributions.

2.3 Compound Poisson approximation (CP).

For rare words, Schbath (1995) and Geske *et al.* (1995) consider a sequence of length ℓ growing to infinity and a word of length k_ℓ growing as $\log \ell$. In this asymptotic framework, the approximate distribution of the count, denoted p_{CP}, is a compound Poisson distribution, the parameters of which are given explicitly in Appendix A.4. For nonoverlapping words (a word is said to be nonoverlapping if and only if two of its occurrences can never overlap in a sequence), the compound Poisson distribution is an ordinary Poisson distribution.

A bound of the approximation error is known from the Chen-Stein method (see Schbath (1995) or Reinert and Schbath (1998)): it is in $\ell k \pi(w)^2 \simeq \mathbb{E}_\mu(N) \times k\pi(w)$ and tends to zero when ℓ goes to infinity (see also Erhardsson (1999, 2000)). This convergence still holds when using $\widehat{\mu}$ instead of μ.

3 Design

In the remainder we shall consider sequences of letters taken in the alphabet of the nucleotides $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

The parameters of the p_{PRT} and p_{CP} distributions are calculated using C++ subroutines of the *R'MES* software (Bouvier *et al.* (1999)). The exact distribution and its mean and variance used in the p_{KB} distribution are calculated using a C subroutine of the EXCEP software (Klaerr-Blanchard *et al.* (2000)).

Factors. To compare the quality of the approximations in a sufficiently wide range of situations, we considered the following factors:

- *The length ℓ of the sequence* (10 levels): $\ell = 10^i$ for $i = 1.5, 2.0, 2.5, \ldots, 6.0$.
- *The distribution μ of the nucleotides* $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ (2 levels): uniform μ = (.25, .25, .25, .25) and HI (adjusted to the genome of *H. Influenzae* in the replication direction, known to be particularly unbalanced) $\mu \simeq$ (.31, .18, .20, .31).

- *The word w*: this factor gathers eight levels corresponding to all the possible combinations of three binary subfactors: the length of the word ($k = 3$ or 9 nucleotides), the frequency of the nucleotides present in the word (rare or frequent with respect to the HI distribution) and its overlapping structure (overlapping or nonoverlapping). The eight selected words are: `ttt`, `att`, `ccc`, `ccg`, `atatatata`, `ttatatata`, `cgcgcgcgc` and `cgcgcgcgg`.

Quality criteria. For each situation corresponding to a particular combination (ℓ, μ, w), we calculate several criteria measuring the quality of each approximation. We present here only the main ones.

- *The total variation distance $d_{\mathrm{TV}}(p, \hat{p})$* between the exact distribution p and its approximation $\hat{p}$:

$$d_{\mathrm{TV}}(p, \hat{p}) = \sup_{A \subset \mathbb{N}} |p(A) - \hat{p}(A)| = \frac{1}{2} \sum_{n \geq 0} |p(n) - \hat{p}(n)|$$

  which measures the greatest error one can make, in terms of probability, when using $\hat{p}$ instead of p (Barbour *et al.* (1992)). This distance will be our main criterion to evaluate the global quality of an approximation.

- *The Kolmogorov distance $d_K(p, \hat{p})$*:

$$d_{\mathrm{K}}(p, \hat{p}) = \max_{n \geq 0} |P(n) - \widehat{P}(n)|$$

  This distance is less severe than the total variation distance since it considers only events of the form $A = (-\infty; n]$. Yet these events are usually the ones we are interested in, especially when checking over- or underrepresentation of words.

- *The maximal difference between point probabilities*: $\max_{n \geq 0} |\hat{p}(n) - p(n)|$.

- *The effective levels $\tilde{\alpha}$ of six quantiles of the distribution $\hat{p}$.* For the six levels $\alpha = 0.05\%$, 0.5%, 5%, 95%, 99.5% and 99.95%, we calculate first the quantile $\hat{u}_\alpha$ of $\hat{p}$ such that $\widehat{P}(\hat{u}_\alpha) = \alpha$ and then the exact level $\tilde{\alpha}$ such that $P(\hat{u}_\alpha) = \tilde{\alpha}$. These levels give an indication of the quality of the approximation for the tail of the distribution.

4 Probabilistic approach

Goal. We consider here the case where the parameters of the model (i.e., the distribution μ of the nucleotides) is known. In this case, the exact distribution $p = p^\mu$ of the count is known and the only two relevant approximate distributions are p_{CP} and p_{KB} as the vector μ of the frequencies does not need to be estimated. The exact distribution and the two approximate distributions are then calculated in a deterministic way for every of the 160 possible combinations of the factors (ℓ, μ, w).

Quality of the approximation. For every couple (μ, w) and for both approximations, the maximal difference between point probabilities (see Section 3) tends to zero when the length of the sequence raises to infinity (not shown here). However, this does not guarantee good behavior of the approximation of the whole distribution for which the total variation distance is a much more relevant criterion. Figure 1 shows the evolution of this distance as a function of the expected count $\mathbb{E}(N)$, for $\mu = \mathrm{HI}$.

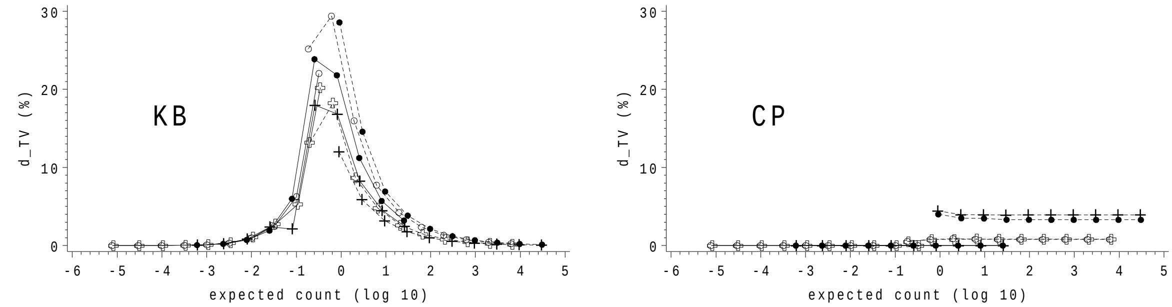

Figure 1: Total variation distance between p and, successively, p_{KB} and p_{CP} as a function of the expected count for the HI distribution of the nucleotides. The 8 curves correspond to the 8 considered words: solid line for long words, dashed for short ones; filled symbols for frequent words, empty for rare ones; circles for overlapping words, crosses for nonoverlapping ones.

The Gaussian approximation KB gives satisfying results ($d_{\mathrm{TV}} \leq 1\%$, an acceptable threshold for tests of level 5% or 10%) when the expected count exceeds 500. The good quality of this approximation for small expected counts ($\mathbb{E}(N) \leq 10^{-2}$) is due to the fact that the distributions, exact and approximate, are concentrated on the event $\{N = 0\}$. The bad quality of that for intermediate expected counts is especially due to the nonsymmetrical shape of the exact distribution.
The results regarding long and short words do not cover the same range of value for the expected count $\mathbb{E}(N)$, but the continuity of the results issued from these two groups leads us to conclude the expected count is the main parameter that determines the quality of the KB approximation.

The CP approximation gives good results for the long words ($k = 9$) and the short rare words (`ccc` and `ccg`). For the short but frequent words (`ttt` and `att`) this approximation is less satisfactory. It appears here the expected count is not the only parameter that determines the quality of the approximation: the probability $\pi(w)$ has also an influence on this quality. This observation agrees with the Chen-Stein bound (see Section 2.3). Moreover, the total variation distance does not tend to zero when the length of the sequence tends to infinity. To observe the announced convergence, the length of the word would need to increase with the length of the sequence. It has to be pointed out that the Chen-Stein bound gives a pessimistic measure of the quality of the approximation: it exceeds 100% for $w = \texttt{ttt}$ and $\ell \geq 100$, and is equal to 14% for $w = \texttt{ccg}$ and $\ell = 100,000$, whereas the real distance in total variation is only 0.8%.

The above remarks are still valid looking at the effective levels or at the Kolmogorov distance that varies here between 50% and 70% of the total variation distance (not shown here).

The results obtained with the uniform distribution μ are coherent with the ones observed with HI, except for the distinction between rare and frequent words which has no meaning in this case.

$\mathbb{E}(N)$	word length	ℓ	KB d_{TV}	KB time (ms)	CP d_{TV}	CP time (ms)	Exact time (ms)
$\simeq 0.01$	3	< 1 bp	*out of the design*				
	9	3.2 kb	1.12	0.07	$< 10^{-3}$	**0.08**	**5.65**
$\simeq 1$	3	0.1 kb	20.7	$< 10^{-3}$	2.24	0.01	**0.02**
	9	316 kb	**15.7**	**0.40**	**0.002**	**0.25**	**1334**
$\simeq 50$	3	3.2 kb	2.70	0.04	1.81	7.69	53.7
	9	> 1 Mb	*out of the design*				
$\simeq 500$	3	32 kb	**0.87**	**0.10**	**1.81**	**302.7**	**3434**
	9	> 1 Mb	*out of the design*				

Table 1: Total variation distance (%) and computation time (ms) for the p_{KB}, p_{CP} and exact distributions for four magnitudes of the expected count. For a given word length, the total variation distance and the computation time are averaged on the two corresponding overlapping words. Computation times have been obtained with a SUN Entreprise 450 (Ultra SPARC II 300MHz), 1.5 Go RAM.

Computation time. Because the numerical computations of the three distributions p, p_{KB} and p_{CP} do not required the same amount of time, we have listed in Table 1 these computation times for different word lengths and different sequence lengths. We have used the uniform distribution for the letters, so no distinction is made between rare and frequent words. For a given word length, 3 or 9, the computation time is averaged on the two overlapping words defined in Section 3. Instead of considering the 2×10 possible couples for (k, ℓ), we have considered five cases with respect to the order of magnitude of the expected count. Finally, by computation time for the distribution p (or p_{KB} or p_{CP}), we mean the time needed to calculate all the point probabilities $p(n)$ for n between 0 and $N_{\max}$. The quantity $N_{\max}$ depends on the word, on ℓ, and on μ and is chosen in such a way that the exact tail probability is lower than 10^{-6}; it is identical for all the distributions, exact or approximates.

We clearly see from Table 1 that the approximate distributions can be calculated more rapidly than the exact distribution. As soon as one wants to process a large number of words, the exact distribution becomes quite unreachable in long sequences or for frequent words.

It is important to note that the computation times given in Table 1 are the ones for overlapping words. For nonoverlapping words, it makes quite no difference except for the CP approximation; indeed the compound Poisson distribution reduces to a Poisson distribution that is much faster to calculate. Moreover, if one is interested only in the tail distribution of the count $\Pr\{N \geq n\}$, there is no need to calculate all the point probabilities of the Gaussian and Poisson distributions.

Suggested rule. To choose which distribution to use, we suggest balancing the quality of the approximation and the computation time required to compute the different distributions (see Table 1). The rule seems to be as follows:

- to use the exact distribution when $\mathbb{E}(N) \leq 0.01$ (however, CP is 70 times faster for an excellent quality of approximation and could be preferred for many iterations);
- to use the exact distribution for short words ($k = 3$) but p_{CP} for long words ($k = 9$) when $\mathbb{E}(N) \simeq 1$;
- to use the exact distribution when $\mathbb{E}(N) \simeq 50$ or p_{KB} for many iterations; and
- to use p_{KB} when $\mathbb{E}(N) \geq 500$: the CP approximation is both slower and less accurate, and the exact distribution is too time consuming.

5 Statistical approach

Simulations. We consider now the case where the distribution μ of the nucleotides is estimated with the observed frequencies in a DNA sequence, as is commonly done in practice. For each letter distribution μ (uniform and HI) and each sequence length, we simulate 1,000 sequences from which we successively

- estimate the empirical distribution $\widehat{\mu}$;
- compute the four approximate distributions given $\widehat{\mu}$: $p^{\widehat{\mu}}$, $p^{\widehat{\mu}}_{\mathrm{KB}}$, $p^{\widehat{\mu}}_{\mathrm{PRT}}$ and $p^{\widehat{\mu}}_{\mathrm{CP}}$; and
- calculate the different criteria described in Section 3 using the exact distribution p^{μ}.

Of course, in practice, the exact distribution p^{μ} is not available since μ is unknown and we can only use the four approximate distributions $p^{\widehat{\mu}}$, $p^{\widehat{\mu}}_{\mathrm{KB}}$, $p^{\widehat{\mu}}_{\mathrm{PRT}}$ and $p^{\widehat{\mu}}_{\mathrm{CP}}$. However, these simulations allow us to get the exact distribution and to compare it to each of its approximations. Moreover, we are able to determine the influence of the estimation of the parameters by comparing $p^{\widehat{\mu}}$ to p^{μ}. From here on, the approximate distribution $p^{\widehat{\mu}}$ is denoted by "exact" distribution.

Because $p^{\widehat{\mu}}$ is long to compute, especially for long sequences (cf. Table 1), we reduced the number of simulations from 1000 to 100 for sequences longer than 10^5. The results presented in Figure 2 are then obtained from $65,000$ simulations.

Total variation distance. From Figure 2, one observes the following main trends:

1. Both Gaussian approximations KB and PRT have similar behaviors; the compound Poisson distribution p_{CP} looks like the "exact" one $p^{\widehat{\mu}}$.
2. All four approximations are of good quality when the expected count is small ($\mathbb{E}(N) \leq 1$) simply because the distributions are all concentrated on the event $\{N = 0\}$. This has been already observed in the probabilistic approach when no estimation is done (cf. Section 4).

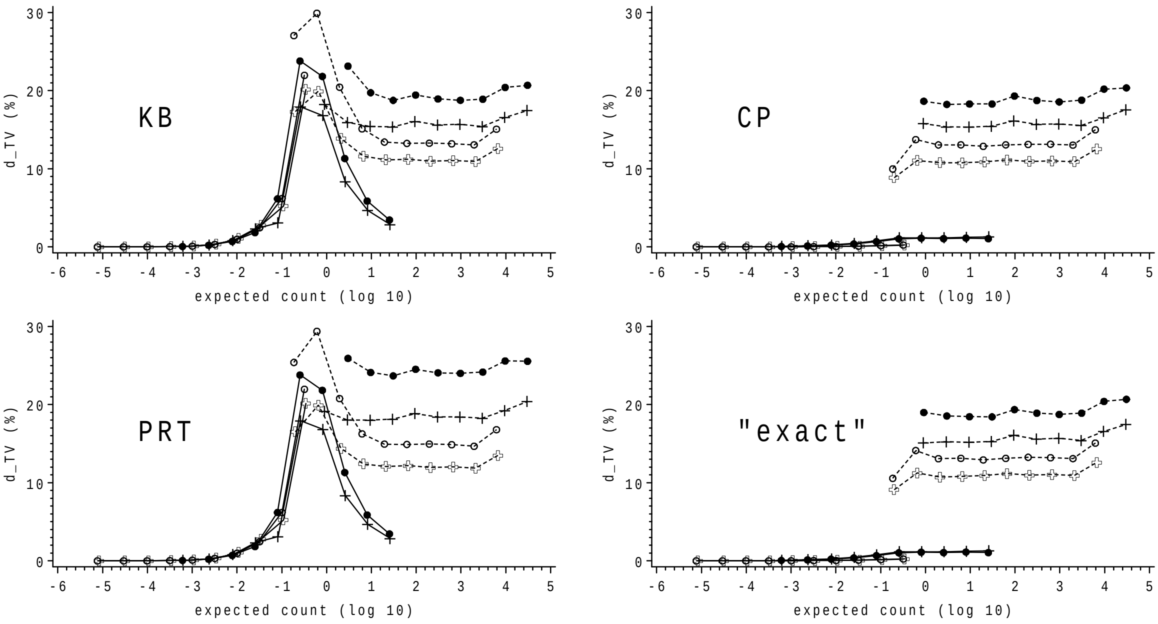

Figure 2: Total variation distance between p^μ and, successively, $p_{\text{KB}}^{\widehat{\mu}}$, $p_{\text{PRT}}^{\widehat{\mu}}$, $p_{\text{CP}}^{\widehat{\mu}}$ and $p^{\widehat{\mu}}$ (`"exact"`) as a function of the expected count for the HI distribution of the nucleotides (same legend as Figure 1). These values are mean calculated over 1,000 simulations, except the last two of each curve that are calculated on 100 simulations.

3. Surprisingly, when the expected count is high ($\mathbb{E}(N) \gg 1$), all four approximations provide the same very bad quality. The total variation distances remain quite constant as $\mathbb{E}(N)$ grows to infinity, contrarily to the maxima of the point probabilities that converge to zero (not shown here). The bad behavior of the "exact" distribution $p^{\widehat{\mu}}$ compared to p^μ shows that the penalization due to the estimation of the parameters is very important. By analogy with the triangular inequality

$$d_{\text{TV}}(\widehat{p}^{\widehat{\mu}}, p^\mu) \leq d_{\text{TV}}(\widehat{p}^{\widehat{\mu}}, p^{\widehat{\mu}}) + d_{\text{TV}}(p^{\widehat{\mu}}, p^\mu),$$

   the approximation cost $d_{\text{TV}}(\widehat{p}^{\widehat{\mu}}, p^{\widehat{\mu}})$ seems negligible compared to the estimation cost $d_{\text{TV}}(p^{\widehat{\mu}}, p^\mu)$. In the case of the CP approximation, the three distances of the above inequality are, respectively illustrated by Figure 2 (CP), Figure 1 (CP) and Figure 2 ("exact").

4. For a fixed high expected count, the bigger the word probability $\pi(w)$ the worse the quality of the approximations. This quality is also worse for overlapping words.

The Kolmogorov distance here is very close to the total variation distance. Hence, the four previous trends are not artifacts of the total variation distance.

Effective levels. In most sequence analyses, the main interest concerns the tail distribution of the count, so the global criterion of the total variation distance may be too strong; the

same applies for the Kolmogorov distance. To study the quality of the four approximations in the tail distribution, we then look at the effective levels described in Section 3.
We get the result that all four approximations overestimate the effective levels associated with 95%, 99.5% and 99.95% but, as the expected count grows, they get closer to the announced levels even if no convergence is guaranteed. As an example, Table 2 gives, for an expected count close to a thousand, the effective levels associated with the announced levels of 95% and 99.95%. For this criterion, the compound Poisson approximation CP gives the best results.

The effective levels associated with 5%, 0.5% or 0.05% are all higher than those announced and do not seem to get closer to them as the expected count increases.

Announced	Effective levels (%)			
levels (%)	$p^{\widehat{\mu}}_{\mathrm{KB}}$	$p^{\widehat{\mu}}_{\mathrm{PRT}}$	$p^{\widehat{\mu}}_{\mathrm{CP}}$	$p^{\widehat{\mu}}$
95.00	93.03	90.04	*93.38*	93.02
99.50	98.93	97.81	*99.05*	98.35
99.95	99.83	99.46	*99.86*	99.84

Table 2: Effective levels of the quantiles of each of the approximate distributions associated with the announced levels 95%, 99.5% and 99.95%.

Computation time. Times required to compute the distributions $p^{\widehat{\mu}}$, $p^{\widehat{\mu}}_{\mathrm{KB}}$ and $p^{\widehat{\mu}}_{\mathrm{CP}}$ are the same using $\widehat{\mu}$ instead of μ and are given in Table 1. Since the computation times of the variances of $p^{\widehat{\mu}}_{\mathrm{KB}}$ and $p^{\widehat{\mu}}_{\mathrm{PRT}}$ are equivalent in the Bernoulli model, we can also refer to Table 1 for the Gaussian distribution $p^{\widehat{\mu}}_{\mathrm{PRT}}$.

Conclusion. The suggested rule to choose between the approximations is the one stated in Section 4 putting together the Gaussian approximations. The crucial point is that the estimation of the parameters has a disastrous effect on the quality of the four approximations.

6 Application to unexpected oligonucleotides in DNA sequences

We focus now on a practical problem in DNA sequence analysis: the detection of oligonucleotides that occur with an unexpected frequency in a given sequence. The problem consists in calculating for each word w (let us say of a given length) the critical probability – the so called p-value – $\Pr\{N(w) \geq n(w)\}$, where $n(w)$ is the observed count of w in the DNA sequence. A high p-value corresponds to an underrepresented word, whereas a small p-value corresponds to an overrepresented word. However, with the typical case where the model parameters are not known, one can only use the distribution $\widehat{\mu}$ estimated according to the DNA sequence. One then has four methods to approximate the p-value and the challenging question is “which one to use ?”.

For this purpose, we analyzed the phage *Lambda* (48,502bps), and we calculated successively for each oligonucleotide of length 3, 6 and 9 the four approximate p-values using the "exact" distribution $p^{\widehat{\mu}}$, the Gaussian distributions $p^{\widehat{\mu}}_{\mathrm{KB}}$ and $p^{\widehat{\mu}}_{\mathrm{PRT}}$, and the compound Poisson distribution $p^{\widehat{\mu}}_{\mathrm{CP}}$.

Figure 3 displays the significance of the counts of the words in the genome of *Lambda*. Each point corresponds to a word: its x-coordinate is the (normalized) p-value of w under p^{μ} and the y-coordinate is its (normalized) approximate p-value using one of the distributions $p^{\widehat{\mu}}_{\mathrm{KB}}$, $p^{\widehat{\mu}}_{\mathrm{PRT}}$ or $p^{\widehat{\mu}}_{\mathrm{CP}}$. For convenience, we use the probit normalization: denoting ϕ the cumulative distribution function of the standard Gaussian distribution, the normalized p-values given on the graph are $\phi^{-1}(p\text{-value})$.

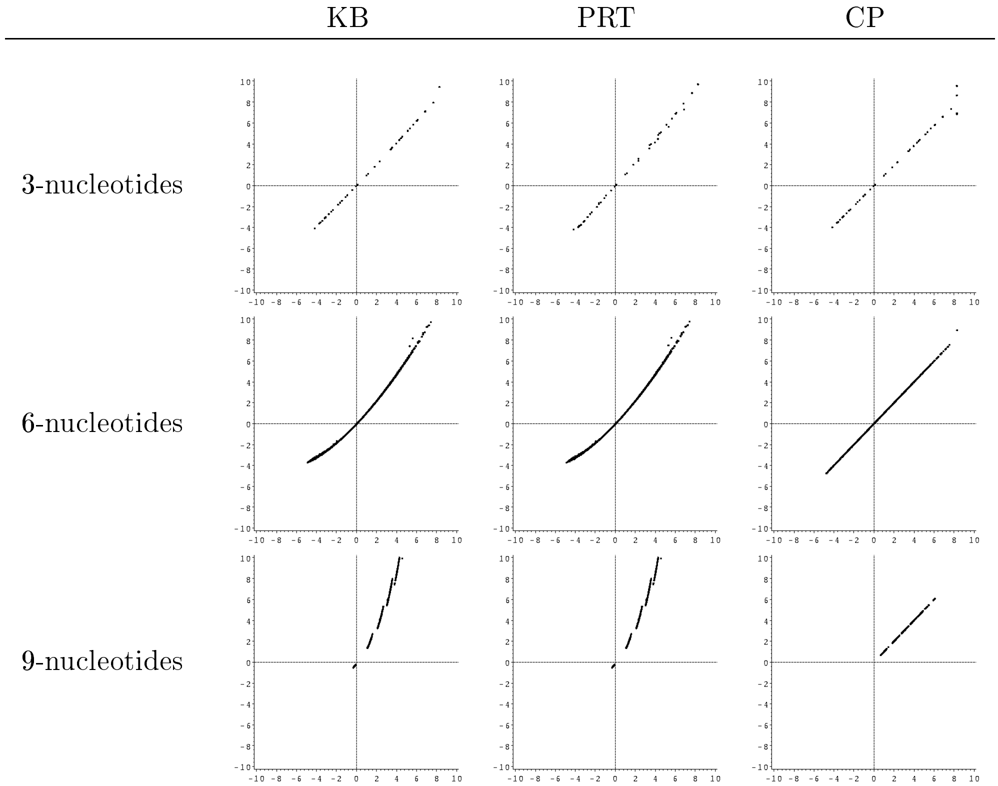

Figure 3: The p-values of the counts of all the oligonucleotides of lengths 3, 6 and 9 in the genome of the phage *Lambda* (48,502bps). Comparison of the KB, PRT and CP approximations (y-axis) with the "exact" significancy (x-axis). All the p-values are normalized according to the probit transformation.

All the graphs show a very good coherence between the approximations. Despite the nonlinear shape observed for the Gaussian approximations KB and PRT, the ordering of the words is very well conserved using any of the four methods. The evolution of the shape of the curves for KB and PRT, as the length of words increases, confirms that the rarer the word the worse the Gaussian approximations (see Section 2). Surprisingly, the CP approximation gives p-values very similar to the "exact" ones, even for short words. Kendall's correlation coefficient, based on the number of inversions between the orders, is greater than 0.98 for all the approximations for the 3-nucleotides and greater than 0.998 for the oligonucleotides of lengths 6 and 9.

Note that the calculation of the "exact" p-values of the 262,144 9-nucleotides took 6 hours while it only required around 20 seconds using the Gaussian or the compound Poisson approximation.

Therefore, we strongly recommend first using one of the KB, PRT or CP approximation to detect the correct set of most under- and most overrepresented oligonucleotides. This set of exceptional words can then be analyzed more accurately using the "exact" distribution.

Conclusion

This work has been performed under the Bernoulli model (Markov model of order zero). Increasing the order of the Markov model does not affect the quality of the approximations but increases the computation time. In the first order Markov model, times are slightly greater than in the Bernoulli model. However, considering higher orders implies increasing the alphabet and then manipulating matrices with big dimensions. The computations of the exact distribution and the Gaussian distribution p_{KB} are then difficult since they require one to diagonalize such matrices.

As expected, the Gaussian approximations are the most efficient for frequent words. The present work shows that words with expected count exceeding hundreds can be studied with these approximations. On the contrary, for rarer words, they give very bad results. For these words, the compound Poisson approximation is, unsurprisingly, adapted; surprisingly, it also performs well for frequent words.

The catastrophic influence of the estimation of the parameters really was not expected at the beginning of this work. The comparison of p^{μ} and $p^{\widehat{\mu}}$ is the most demonstrative example. This raises the problem of the existence of a true distribution μ different from the empirical one $\widehat{\mu}$, when analyzing biological sequences. Biologically, it could seem reasonable to consider that $\widehat{\mu}$ is in fact the true distribution and then to work conditionally to it. But in this case, the induced model is not a Markovian model anymore. We can think of a sampling model without replacement for which the only existing method is the conditional approach of Prum *et al.* (1995): the PRT Gaussian approximation of this paper. At the present time, there is no literature either on the exact distribution of the count or on a Poisson approximation for rare words.

When no specific words are a priori determined according to some biological information, we need to exhaustively analyze a large set of words and we suggest the following strategy. Using any of the four distributions ($p^{\widehat{\mu}}$ $p^{\widehat{\mu}}_{\mathrm{KB}}$, $p^{\widehat{\mu}}_{\mathrm{PRT}}$ and $p^{\widehat{\mu}}_{\mathrm{CP}}$) give the same ordering of the words with respect to the exceptionality of their frequency. It is then natural to use the fastest one to calculate. This preliminary analysis reveals a subset of exeptionnal words that has to be reanalyzed more accurately with the exact distribution.

Acknowledgment

The authors want to thank E. Coward for lending his programs calculating the exact distribution of the count.

A Calculating the distributions

The exact distribution and the parameters of the limiting distributions depend on the overlapping structure of the word w, in particular on the periods of w (Guibas and Odlyzko (1981)).

Definition 1 *A period of the word* $w = w_1 \cdots w_k$ *is an integer* $p \in \{1, \dots, k-1\}$ *such that* $w_i = w_{i+p}$, $\forall i = 1, \dots, k-p$. *Let* $\mathcal{P}(w)$ *be the set of periods of* w.

Let $\mathcal{P}'(w)$ be the set of the *principal* periods of w, namely, the periods that are not strictly multiple of the minimal period of w, and define the quantities

$$\begin{aligned} A_i(w) &= \mu(w_1) \times \cdots \times \mu(w_i) \qquad \text{for } i = 1, \dots k, \\ A(w) &= \sum_{p \in \mathcal{P}'(w)} A_p(w). \end{aligned}$$

A.1 Exact distribution

The recurrences given here can be found with different notations in Robin and Daudin (1999): they give a simple way to calculate the probability

$$q(x, n) = \Pr\{\text{the } n\text{-th occurrence of } w \text{ starts exactly at position } x\}.$$

We use the convention that $q(x, n)$ is null as soon as x or n is negative. The probability at the first position $q(1, n)$ is equal to $\pi(w)$ if $n = 1$ and null otherwise. The following $q(x, n)$'s can be calculated according to the recurrence:

$$\begin{aligned} q(x, n) = \\ & q(x-1, n) - \mu(w)\left[q(x-k, n) - q(x-k, n-1)\right] \\ & - \sum_{p \in \mathcal{P}(w)} A_p(w)\left[q(x-p, n) - q(x-p-1, n) - q(x-p, n-1) + q(x-p-1, n-1)\right]. \end{aligned}$$

In a sequence of length ℓ, the cumulative distribution function of the count $P(n)$ is equal to the probability that the $(n+1)$-th occurrence of w starts after position $(\ell - k + 1)$:

$$P(n) = \sum_{x=\ell-k+2}^{\infty} q(x, n+1) = 1 - \sum_{x=1}^{\ell-k+1} q(x, n+1).$$

A.2 Gaussian approximation (KB)

When the letter distribution μ is known, the expectation of N is simply given by $\mathbb{E}_\mu(N) = (\ell - k + 1)\pi(w)$. The variance given by Kleffe and Borodovsky (1992) reduces in the Bernoulli model to

$$\begin{aligned} \mathbb{V}_\mu(N) &= \mathbb{E}_\mu(N) + 2 \sum_{p \in \mathcal{P}(w)} (\ell - k - p + 1)\pi(w_1 \cdots w_p w_1 \cdots w_k) \\ &\quad + \pi^2(w)\Bigg[\ell(1 - 2k) + 3k^2 - 4k + 1\Bigg]. \end{aligned}$$

The distribution of N is then approximated by the Gaussian distribution with mean $\mathbb{E}_\mu(N)$ and variance $\mathbb{V}_\mu(N)$.

A.3 Gaussian approximation (PRT)

When the letter distribution μ is unknown but estimated by $\widehat{\mu}$, the distribution of N is then approximated by the Gaussian distribution with mean $\mathbb{E}_{\widehat{\mu}}(N) = (\ell - k + 1)\widehat{\pi}(w) = (\ell - k + 1)\widehat{\mu}(w_1) \times \cdots \times \widehat{\mu}(w_k)$ and variance $\widehat{v}$. The variance given by Prum *et al.* (1995) is $\mathbb{E}_{\widehat{\mu}}[N - \mathbb{E}_{\widehat{\mu}}(N)]^2$ and reduces in the Bernoulli model to

$$\begin{aligned} \widehat{v} &= \mathbb{E}_{\widehat{\mu}}(N) + 2 \sum_{p \in \mathcal{P}(w)} (\ell - k - p + 1)\widehat{\pi}(w_1 \cdots w_p w_1 \cdots w_k) \\ &\quad + (\ell - k + 1)\widehat{\pi}^2(w)\left[(k-1)^2 - \sum_{a \in \mathcal{A}} \frac{\nu^2(a)}{\widehat{\mu}(a)}\right], \end{aligned}$$

where $\nu(a)$ is the number of a in the word w. Note that the variance $\hat{v}$ is asymptotically smaller than $\mathbb{V}_\mu(N)$: $\lim_{\ell \to \infty} \frac{1}{\ell}\left[\widehat{v} - \mathbb{V}_\mu(N)\right] = \pi^2(w)\left[k^2 - \sum_{a \in \mathcal{A}} \frac{\nu^2(a)}{\mu(a)}\right] \leq 0.$

A.4 Compound Poisson approximation (CP)

The compound Poisson distribution $\mathrm{CP}(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots)$ that approximates the distribution of the number of occurrences N of w in the Bernoulli sequence $\{S_i\}_{i=1\ldots,\ell}$ (Schbath (1995)) is the distribution of $\sum_{j \geq 1} j\, Z_j$ where the Z_j's are independent random Poisson variables with parameters $\widetilde{\lambda}_j$ given by

$$\widetilde{\lambda}_j = (\ell - k + 1)\pi(w)\left[1 - A(w)\right]^2 A^{j-1}(w).$$

Note that if w is not periodic ($\mathcal{P}(w) = \emptyset$) then $A(w) = 0$, $\widetilde{\lambda}_j = 0$ for $j \geq 2$, and the compound Poisson variable reduces to a single Poisson variable Z_1.

Since the expression of $\widetilde{\lambda}_j$ is geometric, it turns out the compound Poisson distribution $\mathrm{CP}(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots)$ is exactly the Pólya-Aeppli distribution with parameters $(\widetilde{\lambda} := \sum_j \widetilde{\lambda}_j, A(w))$; the Pólya-Aeppli distribution with parameters (λ, v) is defined like the distribution of $\sum_{c=1}^{C} K_c$ where C is a random Poisson variable with mean λ and the random variables K_c's are independent and identically distributed according to the geometric distribution with parameter v. With $\widetilde{\lambda} = (\ell - k + 1)\pi(w)[1 - A(w)]$, we then have (see Johnson *et al.* (1992), p378):

$$\begin{cases} p_{\mathrm{CP}}(0) = \exp(-\widetilde{\lambda}), \\ p_{\mathrm{CP}}(n) = \exp(-\widetilde{\lambda})A^n(w) \displaystyle\sum_{c=1}^{n} \frac{1}{c!} \binom{n-1}{c-1} \left[\frac{1 - A(w)}{A(w)}\, \widetilde{\lambda}\right]^c, \ n \geq 1, \end{cases}$$

where $\binom{a}{b}$ denotes the binomial coefficient $a!/(b!(a-b)!)$. The mean and the variance of $\mathrm{CP}(\widetilde{\lambda}_1, \widetilde{\lambda}_2, \ldots)$ are respectively $(\ell - k + 1)\pi(w) = \mathbb{E}_\mu(N)$ and $\mathbb{E}_\mu(N)[1 + A(w)]/[1 - A(w)]$.

References

Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson approximation.* Oxford-University Press.

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.inra.fr/bia/J/AB/genome/RMES/welcome.html`).

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

Erhardsson, T. (2000). Compound Poisson approximation for counts of rare patterns in Markov chains and extreme sojourns in birth-death chains. Preprint.

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.

Guibas, L. J. and Odlyzko, A. M. (1981). Periods in strings. *J. Combinatorial Theory A.* **30** 19–42.

Johnson, N. L., Kotz, S. and Kemp, A. W. (1992). *Univariate discrete distributions.* Wiley: New-York.

Klaerr-Blanchard, M., Chiapello, H. and Coward, E. (2000). Detecting localized repeats in genomic sequences: A new strategy and its application to *B. subtilis* and *A. thaliana* sequences. *Computer Chem.* **24 (1)** 57–70.

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.

Prum, B., Rodolphe, F. and Turckheim, E. de (1995). Finding words with unexpected frequencies in DNA sequences. *J. R. Statist. Soc. B.* **57** 205–220.

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

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, *preliminary version at RECOMB'98.*

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

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

Schbath, S. (2000). An overview on the distribution of word counts in Markov chains. *J. Comp. Bio.* **7** 193–202.