JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 2, Number 3, 1995
Mary Ann Liebert, Inc.
Pp. 417–437

Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences

Sophie SCHBATH[1], Bernard PRUM[2] and Élisabeth de TURCKHEIM[1]

Current title : Motifs and models for DNA sequences

Abstract

Identifying exceptional motifs is often used for extracting information from long DNA sequences. The two difficulties of the method are the choice of the model that defines the expected frequencies of words and the approximation of the variance of the difference $T(W)$ between the number of occurences of a word W and its estimation. We consider here different Markov chain models, either with stationary or periodic transition probabilities. We estimate the variance of the difference $T(W)$ by the conditional variance of the number of occurences of W given the oligonucleotides counts that define the model. Two applications show how to use asymptotically standard normal statistics associated with the counts to describe a given sequence in terms of its outlying words. Sequences of *Escherichia coli* and of *Bacillus subtilis* are compared with respect to their exceptional tri- and tetranucleotides. For both bacteria, exceptional 3-words are mainly found in the coding frame. *E. coli* palindrome counts are analyzed in different models, showing that many overabundant words are one-letter mutations of avoided palindromes.

Key words: DNA sequences, unexpected frequencies, statistical models, Markov chains, asymptotic variance.

[1]INRA, Département de Biométrie et Intelligence Artificielle, F78352 Jouy-en-Josas Cedex, France.
[2]Université Paris V and CNRS, UA 1323, 75270 Paris Cedex 06, France.

1 Introduction

Modeling DNA sequences with stochastic models and developing statistical methods to analyze the enormous set of data that results from the multiple projects of DNA sequencing are challenging questions for statisticians and biologists.

Because a DNA sequence is naturally represented as a finite but long sequence of discrete variables $X_1, X_2, \ldots, X_n$ where X_i belongs to the 4-letter alphabet $\mathcal{A} = \{A, C, G, T\}$, simple models such as m-order Markov chains have been widely considered (Blaisdell 1985, Brendel et al. 1986, Phillips et al. 1987b, Avery 1987, Gelfand et al. 1992) . Special Markov chains have also been considered. For instance, the Mixture Transition Distribution Model (Raftery, 1985 and Raftery and Tavaré , 1994), which specifies the m-order transition probability as a linear combination on the m previous observations of one order transition probabilities, allows a high order of dependencies with a small number of parameters.

A rather wide use of Markov chain models consists in predicting the frequency of oligonucleotides (which we call here words) to identify those that show an important deviation between their observed frequency and their frequency predicted by the model M. Such words are called contrast words in Brendel *et al.* (1986), meaningful words in Pevzner *et al.* (1989), and anomalous words in Pevzner (1992). Many relevant studies show that these words are useful, either because they have some interesting biological property, such as being restriction sites (for phages lambda and T7 of *E.coli*, see Brendel *et al.* 1986, for instance), or because the set of exceptional words is a useful characteristic to, say, compare introns and exons (Beckmann *et al.* 1986), or to detect outlying sequences (Pietrokovski and Trifonov, 1992). But, in general, these studies use an approximate test statistic that is not asymptotically correct when the length of the sequence increases.

In a given model M for the sequence $X_1, X_2, \ldots, X_{n+h-1}$, we denote $\mu_M(W)$, the probability of occurrence of the word $W = w_1 w_2 \cdots w_h$ at a given position,

$$\mu_M(W) = \Pr\{X_i = w_1, X_{i+1} = w_2, \ldots, X_{i+h-1} = w_h\} .$$

The observed count of occurrences of W is

$$N(W) = N(w_1 w_2 \cdots w_h) = \sum_{i=1}^{i=n} \mathbb{I}\{X_i = w_1, X_{i+1} = w_2, \ldots, X_{i+h-1} = w_h\} ,$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function. This count includes overlapping occurences of W when W overlaps itself. The expectation of $N(W)$ is $n\mu_M(W)$.

To standardize the difference between the observed count $N(W)$ and its predicted value in the model M, $n\hat{\mu}_M(W)$, we need the variance of the difference, which is not the variance of $N(W)$ since $\hat{\mu}_M(W)$ is random and converges to $\mu_M(W)$ at the same rate as $N(W)/n$. This fact has not been considered by Pevzner *et al.* (1989) or Stückle *et al.* (1990). Other authors like Brendel *et al.* (1986) or Avery (1987), approximate the desired variance with $n\hat{\mu}_M(W)$. Phillips *et al.* (1987b) make the same approximation because they use residuals that are asymptotically equivalent to $(n\hat{\mu}_M(W))^{-1/2}(N(W)-n\hat{\mu}_M(W))$. But when the conditions for a central limit theorem are fulfilled, namely a large n and a short length for the words, $\hat{\mu}_M(W)$ does not converge to the variance of the limit of $T_M(W) = n^{-1/2}(N(W) - n\hat{\mu}_M(W))$.

A Poisson or a compound Poisson approximation can be used when we consider a sequence of words W_n, the length of which increases with n so that $n\mu(W_n)$ converges and the Chen-Stein technique gives a bound for the approximation error (Chryssaphinou and Papastravidis 1988, Godbole 1991, Schbath 1995a). We are not in this case considering a fixed given word with an increasing length of the sequence as the asymptotic frame for this study.

The question of the model choice can be considered as a nuisance problem when biologically significant words are sought. But, it could also be an interesting stepwise method to describe the statistical distribution of letters in a given sequence. Starting from a simple model, it is relevant to regard the lack of fit of the data to the model as biological information. The next step is to describe this lack of fit and to use it, for example to compare different sequences. Then, a more general model should be chosen, which includes this information. The deviation of the data from this new model is a possible new information, orthogonal, in a loose sense, to the preceding one. This idea is consistent with the presence of multiple codes, as explained by Trifonov (1989), all necessarily degenerate because these codes overlap on the same stretches of DNA. For this reason too, stochastic models are appropriate.

Chi-square or loglikelihood statistics globally measure the difference of fit between two nested models. For instance, to compare the m-order Markov chain to the p-order Markov chain, $m < p$, the chi-square statistics is just

$$\sum_W \frac{(N(W) - \widehat{N}(W))^2}{\widehat{N}(W)},$$

where the sum is taken over the words W of length $p+1$ and $\widehat{N}(W)$ is an estimation of the expectation of $N(W)$ in the m-order Markov chain model. It turns out that this estimator is not satisfactory as we will see later. Each term of such as a statistic could be understood as a residue that tells how far from the model the frequency of W is. But these terms do not have the same asymptotic variances. On the contrary, the statistics used later in this paper are normalized residues associated with words.

In section 2, we define the different models, including models where the transition probabilities depend on the position in the codon. In section 3, we solve the problem of finding the asymptotic variance of $T_M(W)$ for these models. Section 4 shows two examples of an application. A short discussion is proposed in section 5.

2 Markov chain models

2.1 Stationary models

Markov chain models appear in a natural way to integrate the observed frequencies of short words in the model. For instance, the first-order Markov chain model has been used after noting that the observed frequencies of dinucleotides are not concordant with those predicted with the model where the X_i are modeled as independent, identically distributed variables (Nussinov 1981, Burge *et al.* 1992).

We denote by Mm the m-order stationary Markov chain model for the sequence $X_1, X_2, \ldots, X_n$ and $\pi(a_1 a_2 \cdots a_m, a)$ the transition probability

$$\pi(a_1 a_2 \cdots a_m, a) = \Pr\{X_i = a \mid X_{i-m} = a_1, X_{i-m+1} = a_2, \ldots, X_{i-1} = a_m\} .$$

The maximum likelihood estimators are

$$\hat{\pi}(a_1 a_2 \cdots a_m, a) = \frac{N(a_1 a_2 \cdots a_m a)}{N(a_1 a_2 \cdots a_m)} .$$

Therefore, Mm is exactly fitted to the counts of the $(m+1)$-words (words with $m+1$ letters) and the shortest words to study as possible exceptional words are the $(m+2)$-words. Starting from a given word length h, if we are interested in identifying exceptional h-words, the largest Markov chain model to consider is of order $h-2$. In general, for $m \leq h-2$,

$$\mu_m(W) = \mu(w_1 w_2 \cdots w_m) \prod_{j=1}^{h-m} \pi(w_j w_{j+1} \cdots w_{j+m-1}, w_{j+m}) , \tag{1}$$

and the natural estimator of $\mu_m(W)$ is

$$\hat{\mu}_m(W) = \frac{1}{n} N(w_1 w_2 \cdots w_m) \prod_{j=1}^{h-m} \frac{N(w_j w_{j+1} \cdots w_{j+m-1} w_{j+m})}{N(w_j w_{j+1} \cdots w_{j+m-1})} . \tag{2}$$

The estimator appearing in the chi-square or the loglikelihood statistic to compare the Markov chains of order m and $h-1$ is

$$\hat{\hat{\mu}}_m(W) = \frac{1}{n} N(w_1 w_2 \cdots w_{h-1}) \frac{N(w_{h-m} \cdots w_{h-1} w_h)}{N(w_{h-m} \cdots w_{h-1})} .$$

The two estimators are equal if $m = h-2$. The second measures a possible error only on the last transition $\pi(w_{h-m+1} \cdots w_{h-1}, w_h)$ and a statistic based on this estimator will not take into account a deviation of the model for the prefixes of W, $W^p = w_1 w_2 \cdots w_p$, $p < h$.

2.2 Markov chains with periodic transitions

Periodicities of the coding sequences have been widely observed and described (Pevzner *et al.* 1989, Avery 1987, Lagunez-Otero and Trifonov 1992, Arquès and Michel 1990b) with a period of 3 nucleotides. Other periodicities such as 2, 6 or 9 nucleotides have also been identified in introns or in 5' and 3' regions of genes (Arquès and Michel 1990b). It is therefore appropriate to consider Markov chains with periodic transitions. In the case of a 3 nucleotide period, we define a model where the transition probability of a letter is different when the letter has the first, second or third position in the codon. The general m-order Markov chain with 3-letter periodic transition denoted *Mm_3* is such that

$$\pi_k(a_1 a_2 \cdots a_m, a) = \Pr\{X_{3j+k} = a \mid X_{3j+k-m} = a_1, X_{3j+k-m+1} = a_2, \ldots, X_{3j+k-1} = a_m\} \tag{3}$$

for any j, where $k = 1, 2, 3$ is called the phase of the expected letter a. We also define the phase of a word occurrence as the position of its last letter modulo 3. In such a model we can consider 3 phased counts associated with a given word W

$$N(W,k) = \sum_{j,\,3j+k-h+1 \geq 1}^{3j+k \leq n} \mathbb{I}\{X_{3j+k-h+1} = w_1, X_{3j+k-h+2} = w_2, \ldots, X_{3j+k} = w_h\} \quad k = 1,2,3$$

and a global count

$$N(W) = \sum_{k=1,2,3} N(W,k)\,.$$

Another model is useful in the case where the codon usage is fixed, that is when the counts $N(abc, 3)$ or their expectations are given. For coding sequences, this is suggested by the degeneracy of the genetic code with the idea that the distribution of the codons is a characteristic of sequences (Médigue *et al.* 1991). A parsimonious model denoted MC is defined by

$$\begin{aligned} \Pr\{X_{3j} = c \mid X_1, X_2, \ldots, X_{3j-2} = a, X_{3j-1} = b\} &= \pi_3(ab, c) \\ \Pr\{X_{3j+k} = c \mid X_1, X_2, \ldots, X_{3j+k-1} = b\} &= \pi_k(b, c) \quad \text{for} \quad k = 1, 2\,. \end{aligned}$$

This model fits exactly the codon counts $N(abc, 3)$ and the dinucleotide counts $N(ab, k)$. This model MC has 72 parameters. It is a submodel of *M2_3* (144 parameters) and a supermodel of *M1_3* (36 parameters). It is related to the conditional model considered by Altschul and Erickson (1985).

Other periodic models could be considered such as the compound Markov model of Arquès and Michel (1990a) or a model corresponding to the hypothesis of a primitive code (RNY)n, as proposed in Shepherd (1981). Other codes, such as the framing code

(G-nonG-N)n of ribosomes (Trifonov 1987, Lagunez-Otero and Trifonov 1992) are also relevant for such modeling.

In the following section, we propose an expression for the asymptotic variance $\sigma^2_M(W)$ of $T_M(W) = n^{-1/2}(N(W) - n\hat{\mu}_M(W))$ for the models *M1*, *M1_3*, *MC*, *Mm*, *Mm_3*. This variance is used to define a normalized statistic $U_M(W) = T_M(W)/\hat{\sigma}_M(W)$. The set of these statistics describes the discrepancy between the model and the data just as residuals usually do.

3 Variance of $n^{-1/2}(N(W) - n\hat{\mu}(W))$

3.1 Model M1

The model *M1* has been considered in Prum *et al.* (1995). We summarize their method before extending it to other models. Under the model *M1*, the natural estimator of $\mu_1(W) = \Pr\{X_i = w_1, \ldots, X_{i+h-1} = w_h\}$ given by (2) is

$$\hat{\mu}_1(W) = \frac{1}{n}\frac{\prod_{j=1}^{h-1} N(w_j w_{j+1})}{\prod_{j=2}^{h-1} N(w_j)} .$$

A direct calculation of $\mathrm{Var}(N(W))$ is not simple (Stückle *et al.* 1990, Kleffe and Borodovsky 1992) and the covariance matrix is even more complicated. The calculation of $\mathrm{Var}(N(W) - n\hat{\mu}_1(W))$ can be solved by Taylor expanding $T(W)$ with respect to $N(W)$ and the $N(U)$ for all m- and $(m+1)$-words U. Using the variances and covariances of these counts is possible, but complicated for long words and almost intractable for higher order models.

Another possible estimator of $n\mu_1(W)$ is the conditional expectation of $N(W)$ given the dinucleotide counts $N(ab)$ and the first nucleotide of the sequence X_1. Asymptotic expansions show that this conditional expectation is asymptotically equivalent to $n\hat{\mu}_1(W)$ and that the conditional variance of $N(W)$ is a consistent estimate of the limit of $\mathrm{Var}(N(W) - n\hat{\mu}_1(W))$.

Let $\mathcal{S}$ denote the set of all sequences of length n that have the same X_1 and the same $N(ab)$'s as the given sequence. Cowan (1991) shows that the expression of $\mathbf{E}(N(W) \mid \mathcal{S})$ is a very simple consequence of a formula of Whittle (1955), given in the appendix, which gives the cardinal of the set $\mathcal{S}$. In *M1*, all the sequences of $\mathcal{S}$ have the same probability equal to $1/\mathrm{card}(\mathcal{S})$. The nice idea of Cowan is to remark that there is a one-to-one correspondence between the set of sequences in $\mathcal{S}$ that have a word W at position i and the set of sequences of $\mathcal{S} \setminus W$ that have a word $w_1 w_h$ at position i, where the set $\mathcal{S} \setminus W$ is obtained from sequences of $\mathcal{S}$ by deletion of the inner letters $w_2, w_3, \ldots, w_{h-1}$ of a word W. Then, the counts $N(ab)$ are changed into $N(ab) - n(ab)$

for $ab \neq w_1 w_h$ and $N(w_1 w_h)$ is changed to $N(w_1 w_h) - n(w_1 w_h) + 1$, where

$$n(ab) = \sum_{j=1}^{h-1} \mathbb{I}\{w_j = a, w_{j+1} = b\} .$$

Because of the one-to-one correspondence,

$$\mathbf{E}(N(W) \mid \mathcal{S}) \operatorname{card}(\mathcal{S}) = \mathbf{E}(N(w_1 w_h) \mid \mathcal{S} \setminus W) \operatorname{card}(\mathcal{S} \setminus W) .$$

As $\mathbf{E}(N(w_1 w_h) \mid \mathcal{S} \setminus W)$ is just $N(w_1 w_h) - n(w_1 w_h) + 1$, Cowan's formula follows

$$\mathbf{E}(N(W) \mid \mathcal{S}) = (N(w_1 w_h) - n(w_1 w_h) + 1) \frac{\operatorname{card}(\mathcal{S} \setminus W)}{\operatorname{card}(\mathcal{S})} .$$

Expanding Whittle's formula when n tends to infinity shows that

$$\mathbf{E}(N(W) \mid \mathcal{S}) = n\hat{\mu}_1(W) + O(1) ,$$

so that the two statistics $T_1 = n^{-1/2}(N(W) - n\hat{\mu}_1(W))$ and $T_1' = n^{-1/2}(N(W) - \mathbf{E}(N(W) \mid \mathcal{S}))$ have their difference converging to zero in probability. The variance of their limit distribution is approached by the conditional variance of $n^{-1/2}N(W)$ given $\mathcal{S}$. The value of $\operatorname{Var}(N(W) \mid \mathcal{S})$ is easily obtained with the same method, provided

a) the case of words overlapping by more than one letter is separately considered; for instance, if $W = CAGCA$ there are two occurrences of W in $CAGCAGCA$,

b) a one-to-one correspondence is defined when two subwords $w_2 \cdots w_{h-1}$ are deleted from two words W sharing one or no letter, defining the set $\mathcal{S} \setminus W \setminus W$.

The formulae for $\operatorname{Var}(N(W) \mid \mathcal{S})$ and $\operatorname{Cov}(N(W), N(W') \mid \mathcal{S})$ are given in the appendix. Letting n tend to infinity gives the limiting variances and covariances.

Asymptotic conditional variance and covariance in *M1*

$$\begin{aligned}
\sigma_1^2(W) &= \lim_{n \to +\infty} \frac{1}{n} \operatorname{Var}(N(W) \mid \mathcal{S}) \\
&= \mu(W) + 2\sum_{d=1}^{h-2} \delta(W; d)\mu(W^d W) + \mu(W)^2 \left(\sum_a \frac{n(a+)^2}{\mu(a)} - \sum_{a,b} \frac{n(ab)^2}{\mu(ab)} + \frac{1 - 2n(w_1+)}{\mu(w_1)} \right)
\end{aligned}$$

where $n(a+) = \sum_b n(ab)$. In this formula, $\delta(W; d) = 1$ if in W, the first $h - d$ letters are the same as the last $h - d$ letters, and $W^d W$ denotes the concatenated word $w_1 w_2 \cdots w_d w_1 w_2 \cdots w_h$; $\delta(W; d) = 0$, otherwise.

$$\sigma_1(W, W') = \lim_{n \to +\infty} \frac{1}{n} \operatorname{Cov}(N(W), N(W') \mid \mathcal{S}) = \sum_{d=2-h'}^{h-2} \delta(W, W'; d)\, \mu(W^d W')$$

$$+ \mu(W)\, \mu(W') \left(\sum_a \frac{n(a+)\, n'(a+)}{\mu(a)} - \sum_{a,b} \frac{n(ab)\, n'(ab)}{\mu(ab)} - \frac{n(w_1'+)}{\mu(w_1')} - \frac{n'(w_1+)}{\mu(w_1)} + \frac{\mathbb{I}\{w_1 = w_1'\}}{\mu(w_1)} \right) .$$

The notations $n'(ab)$, $\delta(W, W'; d)$ and $W^d W'$ generalize the preceding ones in the natural way.

3.2 Model M1_3

The periodic model with period 3 may be considered as a stationary first-order Markov chain if we consider an extended alphabet $\bar{\mathcal{A}} = \mathcal{A} \times \{1, 2, 3\}$. Each letter in the sequence is then renamed according to its phase. The letter a appearing in position $3j + k$ is denoted (a, k) in $\bar{\mathcal{A}}$. For example, the sequence $AGCCTG$ is written $(A, 1)(G, 2)(C, 3)(C, 1)(T, 2)(G, 3)$. We also denote (W, k) a word with its phase and $N(W, k)$ the number of occurrences of W in phase k. If $W = CGT$, then $(W, 3) = (C, 1)(G, 2)(T, 3)$, $(W, 2) = (C, 3)(G, 1)(T, 2)$ and $(W, 1) = (C, 2)(G, 3)(T, 1)$. The transition matrix with this 12-letter alphabet is $\bar{\pi}$

$$\bar{\pi}((a, k), (b, \ell)) = \pi_\ell(a, b) \quad \text{if} \quad \ell = k + 1 \quad \text{or} \quad \ell = k - 2$$

$$\bar{\pi}((a, k), (b, \ell)) = 0 \quad \text{otherwise}$$

and $\pi_k(a, b)$ is defined in (3). The cardinal of the set $\bar{\mathcal{S}}$ of the set of sequences with fixed X_1 and $N(ab, k)$'s is a direct application of Whittle's formula. Considering a word W in a given position i, it is possible to delete its inner letters without changing the phase of the letters after w_h only if $h = 3p + 2$. In that case, the same formulae are used to calculate $\mathbf{E}(N(W, k) \mid \bar{\mathcal{S}})$ and $\mathrm{Var}(N(W, k) \mid \bar{\mathcal{S}})$. Considering words of $3p+1$ letters as words of $3p+2$ letters with an unspecified first letter gives the following decomposition

$$N(W, k) = \sum_{a \in \mathcal{A}} N(aW, k) + \varepsilon$$

where $\varepsilon = 1$ if the sequence starts with a word (W, k), $\varepsilon = 0$ otherwise. The conditional expectation of $N(W, k)$ is then the sum of five terms, where the last one is negligible, and the conditional variance is the sum of the corresponding variances and covariances (Schbath, 1995b). Words with $3p$ letters will be treated with the same method

$$N(W, k) = \sum_{a, b \in \mathcal{A}} N(abW, k) + \varepsilon + \sum_a \varepsilon_a \ ,$$

where $\varepsilon_a = 1$ if the sequence starts with a word (aW, k), $\varepsilon_a = 0$ otherwise. In the appendix, we give the formulae for the asymptotic variances and covariances in this model.

3.3 Model MC

The conditional model associated with MC is the set $\mathcal{C}$ of sequences with equal probabilities which have fixed codon counts $N(abc,3)$ and fixed 2-letter word counts $N(ab)$. We suppose that the sequence length is a multiple of 3. The phased 2-letter word counts are also fixed because

$$N(ab,2) = \sum_c N(abc,3)$$

$$N(bc,3) = \sum_a N(abc,3)$$

$$N(ca,1) = N(ca) - N(ca,2) - N(ca,3) \ ,$$

as are the counts of spaced words $(a\cdot c,3)$

$$N(a\cdot c,3) = \sum_b N(abc,3) \ .$$

Altschul and Erickson (1985) propose a method to simulate the sequences in $\mathcal{C}$. They use the fact that once the letters in phases 1 and 3 are known, it is easy to fill up phase 2 by a random uniform choice of each X_{3j+2} in the table of codons having X_{3j+1} and X_{3j+3} as first and last letters. We define $\mathcal{S}'$ as a set of sequences with no letters in phase 2 and with the fixed 2-letter word counts $N(ca,1)$ and $N(a\cdot c,3)$. The preceding remark implies that the cardinal of $\mathcal{C}$ is the product of the cardinal of $\mathcal{S}'$ and the number of distinct permutations of the codons with given first and last letters

$$\mathrm{card}(\mathcal{C}) = \mathrm{card}(\mathcal{S}') \prod_{a,c\in\mathcal{A}} \frac{N(a\cdot c,3)!}{\prod_{b\in\mathcal{A}} N(abc,3)!} \ .$$

We define $(\widetilde{W},k)$ as the word (W,k) without its letters in phase 2. Given that $(\widetilde{W},k)$ appears at position i, the conditional probability of observing (W,k) at position i does not depend on i and is denoted $p_k(W)$. Therefore

$$\mathbf{E}(N(W,k)\mid\mathcal{C}) = p_k(W)\ \mathbf{E}(N(\widetilde{W},k)\mid\mathcal{S}') \ .$$

The conditional expectation $\mathbf{E}(N(\widetilde{W},k)\mid\mathcal{S}')$ is obtained with the method for first-order Markov chain with periodic transition presented in section 3.2 changing the period to 2 letters. No de-phasing problem occurs when (W,k) is of type a) or b), where

a) $(W,k) = C_1\cdots C_\ell$

b) $(W,k) = aC_1\cdots C_\ell b$,

where the C_i's are codons, that is 3-letter words in phase 3.

For other words, the same trick as for $(3p+1)$ and $(3p)$-words in *M1_3* may be used.

Conditional variances and covariances are calculated with the same method but two words in $\mathcal{S}'$ are considered and conditional probabilities $p_k(W,W)$ and $p_{k,k'}(W,W')$ used to fill in the words in $\mathcal{S}'$ with letters in phase 2. Formulae for words of type a) or b) are in the appendix.

3.4 Model Mm

We first consider the second-order case. As a second-order Markov chain with states in $\mathcal{A}$ is also a first-order Markov chain with states in $\mathcal{A}^2$, the method to calculate the conditional expectation and variance of $N(W)$ given X_1, X_2 and the $N(abc)$'s seems straightforward. The corresponding sequence in the alphabet $\mathcal{A}^2$ is written $Y_1, \ldots, Y_{n-1}$ with $Y_i = (X_i, X_{i+1})$. The conditioning set $\mathcal{S}_2$ of all sequences $X_1, X_2, \ldots X_n$ with fixed X_1, X_2 and N_{abc}'s is in a one-to-one correspondence with the set $\mathbb{S}$ of sequences $Y_1, Y_2, \ldots Y_{n-1}$ with fixed Y_1 and $N(AB)$'s where $A = (a_1, a_2)$, $B = (b_1, b_2)$, $N(AB) = N(a_1 a_2 b_2)$ if $a_2 = b_1$, $N(AB) = 0$, otherwise. The word W in the alphabet $\mathcal{A}$ is associated with a word $\mathbb{W}$ in the alphabet $\mathcal{A}^2$, where $\mathbb{W} = W_1 W_2 \cdots W_{h-1}$ and $W_i = w_i w_{i+1}$. The modified set $\mathbb{S} \setminus \mathbb{W}$ has a two-letter word $W_1 W_{h-1}$ that is not associated with a 3-letter word in $\mathcal{A}$ when $w_2 \neq w_{h-1}$. Despite that difference from the case of section 3.1, the same formulae are valid for the conditional expectation and variance of $N(\mathbb{W})$.

The asymptotic developments must be adapted from the case of section 3.1 to that case to take into account that some of the $N(AB)$'s do not tend to infinity (Schbath 1995b). Nevertheless, the conditional expectation and variance are such that

$$\mathbf{E}(N(W) \mid \mathcal{S}_2) = \mathbf{E}(N(\mathbb{W}) \mid \mathbb{S}) = n\hat{\mu}_2(W) + O(1) ,$$

where

$$\hat{\mu}_2(W) = \frac{1}{n} \frac{\prod_{j=1}^{h-2} N(w_j w_{j+1} w_{j+2})}{\prod_{j=2}^{h-2} N(w_j w_{j+1})} ,$$

and the conditional variance is a consistent estimator of the limiting variance of $T_2(W) = n^{-1/2}(N(W) - n\widehat{\mu}_2(W))$.

This method is obviously extended to the case of an m-order Markov chain. The set $\mathcal{S}_m$ of sequences in the alphabet $\mathcal{A}$ with fixed $X_1, X_2, \ldots, X_m$ and fixed $N(a_1 \cdots a_{m+1})$ is in a one-to-one correspondence with the set $\mathbb{S}_m$ of sequences in the alphabet $\mathcal{A}^m$ with given first letter and 2-letter word counts. This remark solves the problem of the asymptotic variance in *Mm* of $T_m(W) = n^{-1/2}(N(W) - n\widehat{\mu}_m(W))$ with $\widehat{\mu}_m(W)$ defined in (2) and we get

$$\sigma_m^2(W) = \lim_{n \to +\infty} \frac{1}{n} \mathrm{Var}(N(W) \mid \mathcal{S}_m) = \mu(W) + 2 \sum_{d=1}^{h-m-1} \delta(W; d)\, \mu(W^d W)$$

$$+\mu(W)^2\left(\sum_{a_1,\dots,a_m}\frac{n(a_1\cdots a_m+)^2}{\mu(a_1\cdots a_m)}-\sum_{a_1,\dots,a_{m+1}}\frac{n(a_1\cdots a_{m+1})^2}{\mu(a_1\cdots a_{m+1})}+\frac{1-2n(w_1\cdots w_m+)}{\mu(w_1\cdots w_m)}\right).$$

These results can be extended to *Mm_3* just as the formulae for *M1* are extended to *M1_3*.

4 Applications

4.1 Phased three- and four-words in *E. coli*

To illustrate how exceptional motifs can give insight on DNA sequences, we fitted models *M1_3* and *M2_3* to 184 genes of *Escherichia coli* (181083 bp) extracted from four sequences (L10328, M87049, L19201 and U00006 in Genbank), located on the same strand and on the same side of the origin of replication. Coding sequences translated from the other strand are not taken in this set.

3-words under *M1_3*

Fitting the first-order model and computing the phased 3-words statistics shows that exceptional words are mainly in phase 3 (codons). For the different phases, the statistics ranges are :

$$\begin{array}{rcccl} -12 & < & U_{1-3}(abc,1) & < & 9\,, \\ -13 & < & U_{1-3}(abc,2) & < & 19\,, \\ -30 & < & U_{1-3}(abc,3) & < & 48\,. \end{array}$$

Therefore, most of the discrepancy between the phased 3-words counts and the model is explained by codons. This fact is also clear from Figure 1 where the total statistics $U_{1-3}(abc)$ are represented versus the statistics on phase 3, $U_{1-3}(abc,3)$. Very few points are far from the line $x=y$.

Table 1 gives the list of the words abc with a statistic $|U_{1-3}(abc,3)|>19$. Some of them are also exceptional on phase 2 with $|U_{1-3}(abc,2)|>10$ and TAG also appears on phase 1.

Considering the codons of Table 1 and listing the statistics of their synonymous codons in Table 2, we see that the single codon TGG which specifies the amino acid tryptophan is over-represented. This is also true for the two codons of tyrosine, TAT and TAC. The other codons of Table 1 have synonymous codons with large statistics of the other sign. Therefore, for most of the exceptional 3-words, the lack of fit of *M1_3* is not explained by the amino acid composition of the encoded proteins but by the codon usage.

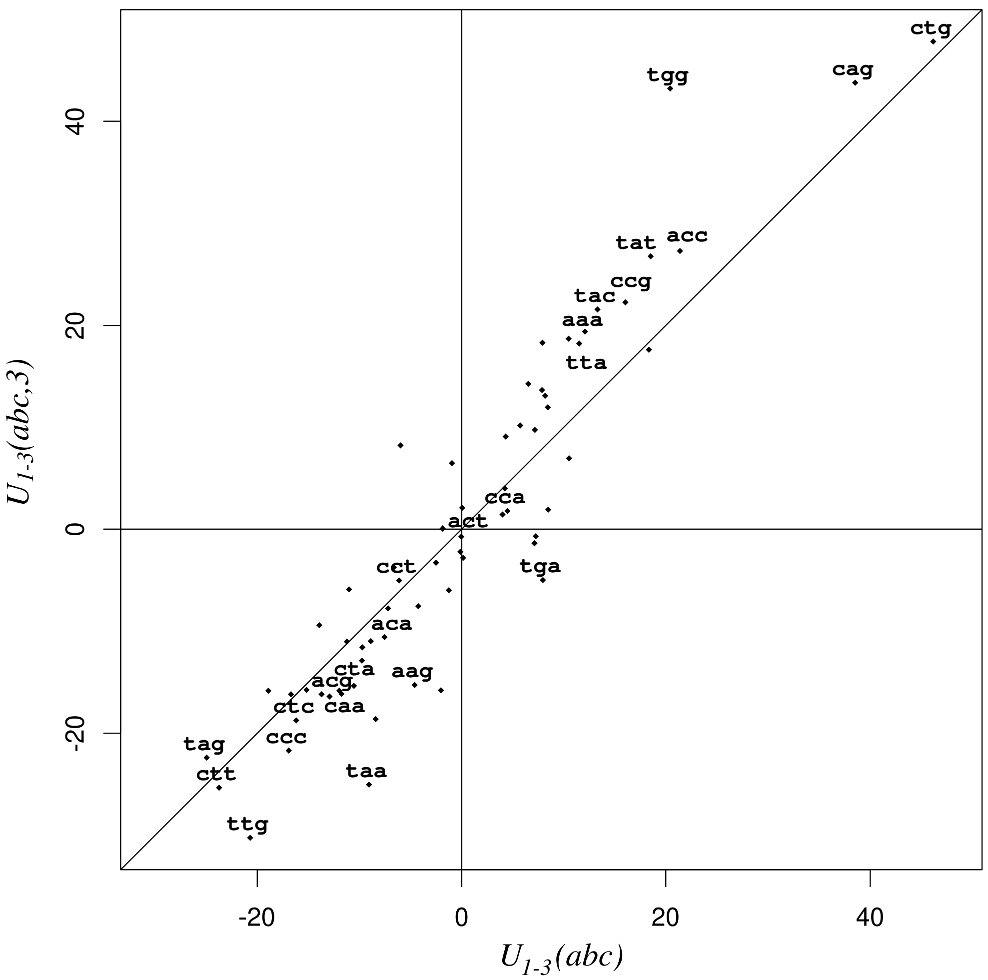

Figure 1: Exceptional 3-words in *E. coli*.

Each point is represented with $(U_{1-3}(abc), U_{1-3}(abc,3))$ *as coordinates. As* $U_{1-3}(abc)$ *is derived from the difference between the global count* $N(abc)$ *and its expected value, the* x*-axis coordinate summarizes the deviation of a word-count on the three phases (three reading frames) whereas the* y*-axis shows the deviation on the codon phase (codon reading frame). Words of Table 2 are named.*

Table 1: Exceptional 3-words in *E. coli*.

abc	$U_{1-3}(abc,3)$	$U_{1-3}(abc,2)$	abc	$U_{1-3}(abc,3)$	$U_{1-3}(abc,2)$	$U_{1-3}(abc,1)$
<u>CTG</u>	48	16	<u>TTG</u>	-30	-11	
<u>CAG</u>	44	19	<u>CTT</u>	-25	-13	
<u>TGG</u>	43		TAA	-25		
<u>ACC</u>	27		<u>TAG</u>	-22	-10	-10
<u>TAT</u>	27		<u>CCC</u>	-22		
<u>CCG</u>	22					
TAC	22					
AAA	19					

Words with a statistic $|U_{1-3}(abc,3))| > 19$ are shown. Their statistics on other phases are noted if $|U_{1-3}(abc,k))| > 10$ for $k = 1, 2$. Underlined words are also found in Brendel et al. *(1986).*

Table 2: Exceptional codons and their synonymous codons in *E. coli*.

abc ($U_{1-3}(abc,3)$)	amino acid
CTG (48) TTA (18) CTA (-16) CTC (-19) CTT (-25) TTG (-30)	leucine
CAG (44) CAA (-16)	glutamine
TGG (43)	tryptophan
ACC (27) ACT (-1) ACA (-11) ACG (-16)	threonine
TAT (27) TAC (22)	tyrosine
CCG (22) CCA (2) CCT (-5) CCC (-22)	proline
AAA (19) AAG (-15)	lysine
TAA (-25) TAG (-22) TGA (-5)	stop

Each codon of Table 1 and its synonymous codons are shown. The most under-represented codons all appear as synonymous codons of an over-represented codon, except for the three stop codons.

4-words under *M1_3*

Exceptional 4-words under *M1_3* appear mainly in phases 1 and 3, as shown by the statistics ranges,

$$\begin{array}{rcccl} -20 & < & U_{1-3}(abcd,1) & < & 34\,, \\ -9 & < & U_{1-3}(abcd,2) & < & 15\,, \\ -22 & < & U_{1-3}(abcd,3) & < & 43\,. \end{array}$$

This is easily explained by the following two facts :

- the 4-words in phases 1 and 3 have a codon as subword, whereas 4-words in phase 2 do not,

- 4-words are exceptional because at least one of their subwords is exceptional.

The second fact is shown in Table 3 where the exceptional words with a statistic $|U_{1-3}(abcd,k)| > 15$ are listed. Nearly all 4-words of this list have an exceptional subword appearing in brackets for codons or in parenthesis for 3-words in phase 2.

Table 3: EXCEPTIONAL 4-WORDS IN *E. coli* UNDER M1_3.

Phase 1		Phase 2		Phase 3	
abcd	$U_{1-3}(abcd,1)$	*abcd*	$U_{1-3}(abcd,2)$	*abcd*	$U_{1-3}(abcd,3)$
$*$[CTG]G	34	C(CAG)	15	$*$C[TGG]	44
[CAG]G	30			G[CTG]	30
[TGG]C	24			(CAG)C	25
$\triangleright$[TAT]C	22			T[CAG]	24
[TGG]T	20			C[CAG]	21
[TGG]G	18			G[TGG]	20
[AAA]G	17			G[CAG]	19
[CTG]A	17			$\triangleright$TATC	17
[CAG]C	17			C[ACC]	17
GAAG	16			C[TAC]	16
				T[TAT]	16
				GGCG	15
[TTG]G	-20			C[TTG]	-22
CAAG	-16				
[TAA]G	-15				

The 4-words with statistics $|U_{1-3}(abcd,k)| > 15$ are listed here. Most words contain a subword from Table 1. Subwords that are exceptional codons are in brackets and exceptional subwords in phase 2 are in parentheses. Except CCG, all over-represented codons from Table 1 have at least one exceptional 4-letter extension which is exceptional. There are very few under-

represented words. Two words appear on both phases and are marked with a sign $*$ or $\triangleright$.

Nevertheless, there is new information in the list of exceptional 4-words that is not found in the 3-words, since only some extensions of exceptional 3-words are exceptional 4-words.

4-words under *M2_3*

Other words are expected under *M2_3* since the 3-word counts are taken into account in this model. Table 4 shows that

- some words are exceptional under both *M1_3* and *M2_3*,
- if one letter is changed in the following over-represented words CAGG,TTGC,CTGG,GGCG,CGCC, we get the under-represented words CAAG,TTGG,TTGG,GGCC,CGCG.

Table 4: Exceptional 4-words in *E. coli* under M2_3.

Phase 1		Phase 3	
abcd	$U_{2-3}(abcd,1)$	*abcd*	$U_{2-3}(abcd,3)$
<u>**CAGG**</u>	15	<u>GGCG</u>	11
TTGC	10	CTTC	10
ACGC	10	TGCC	10
ATTC	10	AGAG	9
<u>**CTGG**</u>	9	<u>CGCC</u>	9
<u>TTCC</u>	9	**TATC**	9
<u>CAAC</u>	8	<u>GGTG</u>	8
		GCAA	8
		CGAC	8
<u>**CAAG**</u>	-16	$*$<u>GGCC</u>	-17
$\bullet$<u>**TTGG**</u>	-12	<u>CCAA</u>	-10
GAGG	-11	**CTTG**	-10
$*$<u>GGCC</u>	-10	$\bullet$<u>TTGG</u>	-10
<u>**TAAG**</u>	-8	CGTG	-8
		CGCG	-8

Words with statistics $|U_{2-3}(abcd,k)| > 8$ are listed here. Bold-faced words are exceptional in the same phase under both M1_3 and M2_3. Words appearing in both phases are marked with a common sign ($*$, $\bullet$). Underlined words are also found in Brendel et al. *(1986).*

Genes of the other strand

The same analysis applied to the genes of the other strand shows the same words with very similar statistics (not shown). This means that the information borne by 3- and 4-words is mainly related to the coding function of the sequence and not to the linear structure of a strand.

Brendel *et al.* (1986) gave a list of exceptional words in four sequences of genes of *E. coli*. They use *M1* for 3-words, *M2* for 4-words and approximate the variance by $n\widehat{\mu}_M(W)$. Despite these differences, many words are common to both studies (they are underlined in Tables 1 and 4) which can be explained by different facts :

- exceptional 3-words under *M1_3* and *M1* are the same, in the same order, in *E. coli* (not shown),

- exceptional 3- and 4-words are the same for genes translated from the other strand.

Nevertheless, some differences should be explained. Brendel *et al.* (1986) find the following words that we do not have : GAT(+), GGT(+), TGA(+), GAG(-), GTC(-), GGG(-) and AATA(+), CTAC(+), GAAG(+), GATG(+), TGCT(+), TTCC(+), CTAG(-), GAAT(-), GGCT(-), GGTT(-), GTCC(-),
and we find also TAC(+), AAA(+), AGC(+), GTA(+), TAA(-), ATA(-), and TTGC(+), ACGC(+), ATTC(+), CTTC(+), TGCC(+), AGAG(+), TATC(+), GCAA(+), CGAC(+), GAGG(-), CTTG(-), CGTG(-), CGCG(-).
Differences of the same kind appear between our analysis and Phillips *et al.* (1987a). For overlapping words, the differences may be explained by the overlapping occurrences of the words which have to be taken into account in the estimation of the variance. For instance, for AAA we estimate the expectation by 1562 and the variance by 771, and for GGG we have 786 and 383.
The new fact in this paper is that we show that the 3-words are exceptional on one phase, the codon phase, and the 4-words are mainly exceptional in one phase, phase 1 or phase 3, that is words starting or ending with a codon.

4.2 Phased three- and four-words in *B. Subtilis*

In the same way, we fitted model *M1_3* and *M2_3* to 159 genes (134559 bp) extracted from two sequences of *Bacillus Subtilis* (X73124 in Genbank and A. Sorokine, personal communication). As before, the genes are located on the same strand and on the same side of the origin of replication.

3-words under *M1_3*

Again, most exceptional 3-words are codons, but there are also exceptional 3-words in phase 2. Ranges for the statistics $U_{1-3}(abc,k)$ are

$$\begin{array}{rcccl} -10 & < & U_{1-3}(abc,1) & < & 12\,, \\ -18 & < & U_{1-3}(abc,2) & < & 26\,, \\ -27 & < & U_{1-3}(abc,3) & < & 34\,. \end{array}$$

Table 5 shows the words with a statistic $|U_{1-3}(abc,k))| > 14$. Four of them appear on two phases. Notice the case of TGG : the over-representation of TGG on phase 3 (tryptophan) is compensated by its under-representation on phase 2. As in *E. coli*, the over-representation of TGG is related to the frequency of tryptophan, that of TAT and TAC to tyrosine. The other exceptional codons show an important bias in codon usage.

Table 5: EXCEPTIONAL 3-WORDS IN *B. Subtilis*.

Phase 3		Phase 2	
abc	$U_{1-3}(abc,3)$	abc	$U_{1-3}(abc,2)$
TAT	34	CGG	26
→**TGG**	23	∗**CAG**	16
CCG	19	GCT	15
AAA	18		
∗**CAG**	17		
TAC	16		
TTA	15		
ACA	15		
CTG	15		
TAA	-30	→TGG	-18
○**TAG**	-21	CGA	-16
•AAT	-17	•AAT	-15
TGA	-14	○**TAG**	-14
CCA	-14		

Words with a statistic $|U_{1-3}(abc,k))| > 14$ appear here. Bold-faced words are exceptional in the same phase in both B. Subtilis and E. coli. Words appearing in both phases are marked with a common sign (→, ∗, ○, •).

If we compare the 3-words in *E. coli* and *B. Subtilis*, many exceptional codons are common : they are bold-faced in Table 5. Two codons are specific of *E. coli*, ACC(+27) and CTT(-25), and three are specific of *B. Subtilis*, ACA(+15), TGA (-14) and CCA(-14). On phase 2, CTG(+16) and CTT(-13) are specific words of *E. coli*, and CGG(+26), GCT(+15), TGG(-18) and CGA(-16) are specific of *B. Subtilis*. We can summarize the comparison of the two sequences if we represent each word with coordinates equal to the statistics $U_{1-3}(abc)$ associated with *E. coli* (x-axis) and *B. Subtilis* (y-axis). In Figure 2, common and specific exceptional words are represented with an index showing the phase on which they are significant. Only TGG has a position that is due to the combination of two very different positions in phases 2 and 3 (shown with arrows).

4-words under *M2_3*

As for *E. coli*, exceptional 4-words under *M1_3* are mainly explained by exceptional subwords (not shown). We now consider 4-words under *M2_3*.
Table 6 shows exceptional words of *B. Subtilis* with a statistic $|U_{2-3}(abcd,k)| > 7$. Only TAAA is common to both *M1_3* and *M2_3*.

Table 6: Exceptional 4-words under M2_3 for *B. Subtilis*.

Phase 1		Phase 3	
$abcd$	$U_{2-3}(abcd,1)$	$abcd$	$U_{2-3}(abcd,3)$
TTTA	7	GTTT	9
CTTA	-8	GTTG	-9
TAAA	-7	AGAT	-8
CAAC	-7	GGCC	-7
		TTTT	-7

Words with a statistic $|U_{2-3}(abcd,k)| > 7$ *are shown.*

Comparing the two bacteria reveals that very few exceptional words are common. This is also shown on Figure 3 where the clouds of points for both phases 2 and 3 have a fairly round shape. Unlike codons, exceptional 4-words in *M2_3* are very much different for the two bacteria.

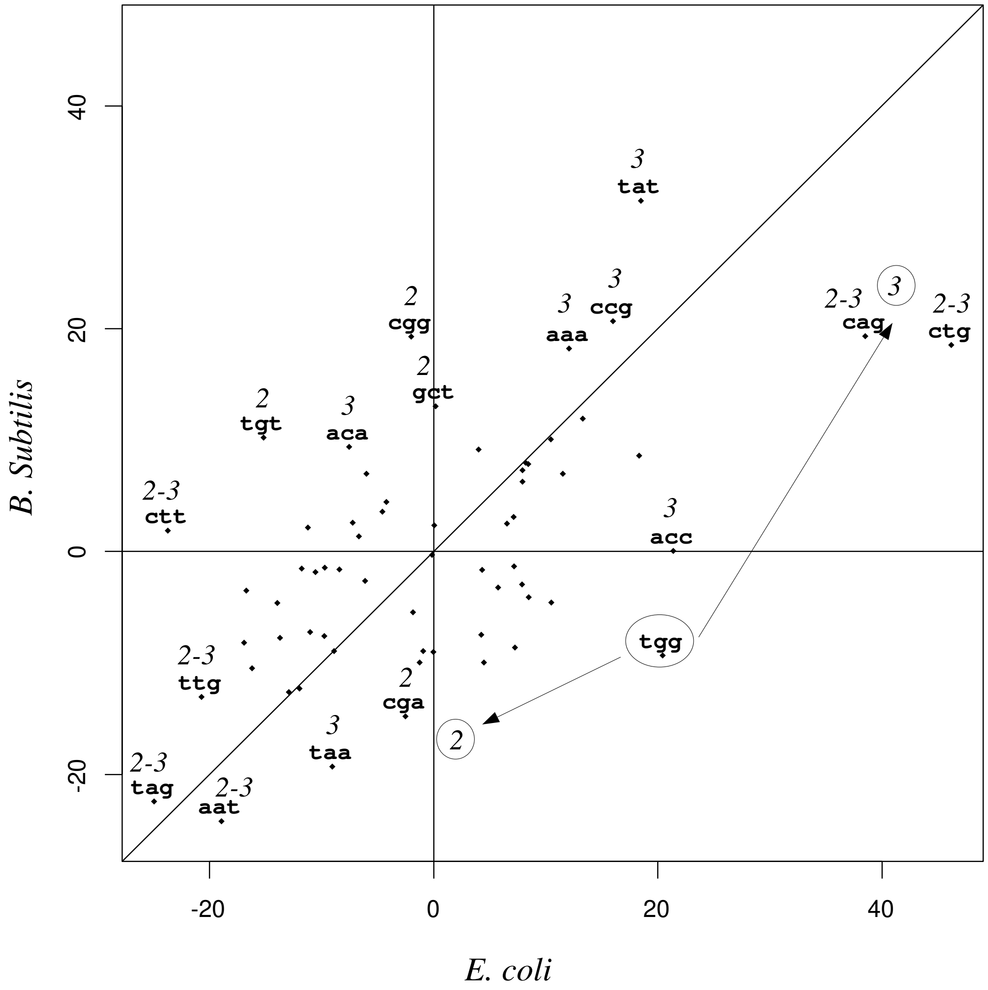

Figure 2: Comparison of *E. coli* and *B. Subtilis* through exceptional 3-words.

Words are represented with their global statistics $U_{1-3}(abc)$ in E. coli versus B. Subtilis. As the statistic $U_{1-3}(abc)$ is derived from the global count $N(abc) = \sum_k N(abc, k)$, this figure summarizes the information from the three phases. For all words of the edge of the cloud, the location in this graph corresponds to a very similar location on one or two given phases (marked above the word name) and a rather central location on the other phases. Only TGG has a very different location in this global representation and in the phase representations. The location of TGG in phases 2 and 3 are shown with arrows.

a)

b)

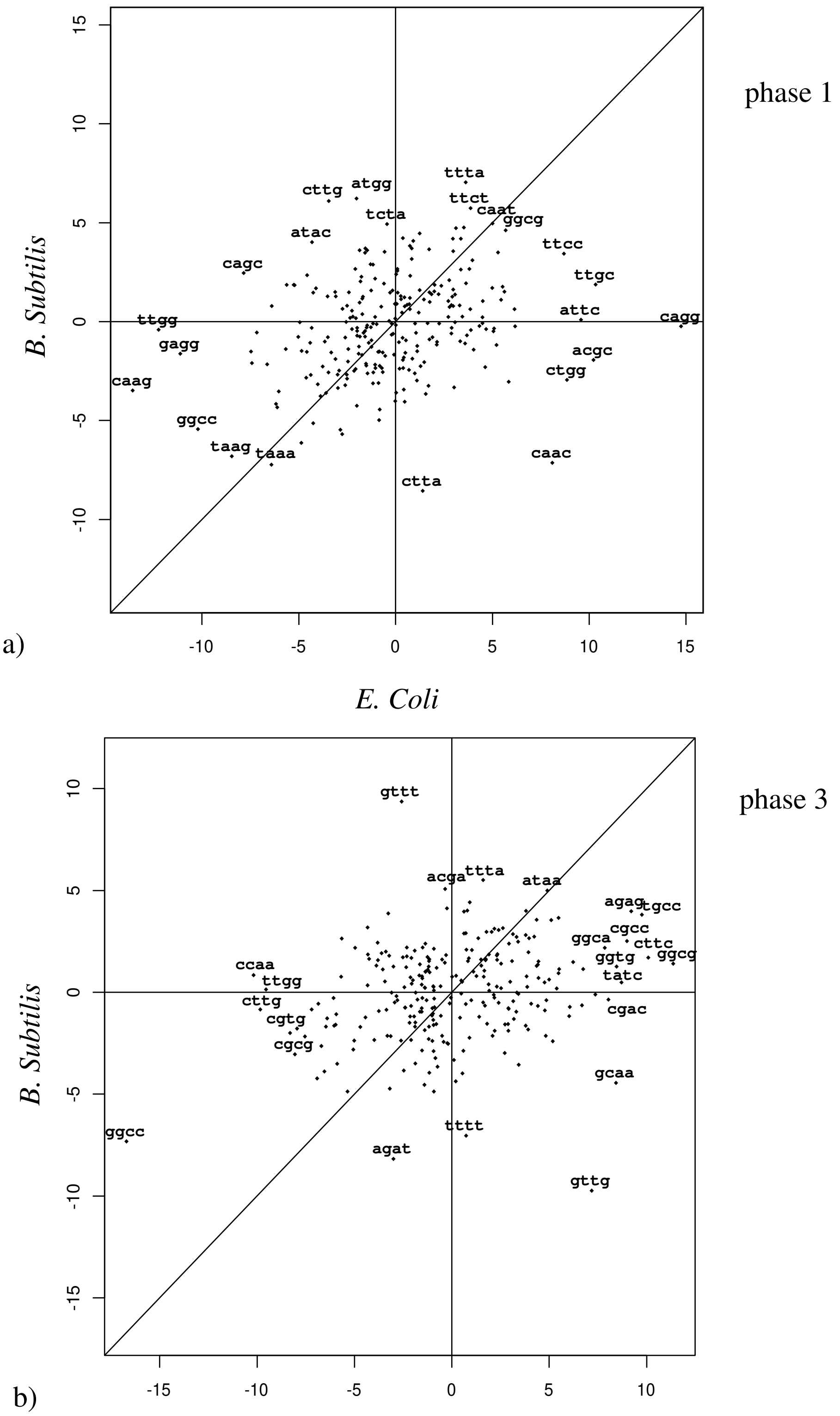

Figure 3: Comparison of *E. coli* and *B. Subtilis* through exceptional 4-words under M2_3.

Words are represented with their statistics $U_{2-3}(abcd, 1)$ in a) and $U_{2-3}(abcd, 3)$ in b), in E. coli versus B. Subtilis. Words on the edges of the cloud are named. Other words are only represented by a point. Except a few words as GGCC in phase 3 and TTTA in phase 1, most exceptional words of this analysis are specific to one of the two bacteria.

4.3 Palindromes in *E. coli*

Palindrome counts are known to be significantly low in many phages, in *E. coli* and in *B. Subtilis* sequences. This has been interpreted as a means to avoid target sites of restriction enzymes. In a study of palindrome counts in a broad range of organisms, Karlin *et al.* (1992) show the relation between these counts and the existence of restriction systems with palindrome targets. Different mechanisms can be suggested as a means to avoid these words. We give here more insight on the under-representation of a word considering its statistics under different models. The sequence is taken from four sequences (L10328, M87049, L19201 and U00006 in Genbank) and consisted of 384243 bp on the right side of the origin of replication.

4-palindromes under *M1* and *M2*

Figure 4 shows the 4-words statistics under *M1* versus *M2*, where palindromes are identified. Most palindromes are under-represented at least under one model. CCGG is the only over-represented palindrome. CTAG is the most avoided with respect to the prediction based on the dinucleotide frequencies and GGCC is the most avoided with respect to the trinucleotide frequencies. This means that a different mechanism is used to avoid these two words. CTAG is avoided mainly because CTA and TAG are avoided. Also, AATT is exceptional only because AAT is under-represented. If trinucleotide frequencies are taken into account, AATT is no more exceptional and CTAG has rank 8 only. On the contrary, GGCC is rare because G is avoided before GCC or C avoided after GGC while both GCC and GGC are equal represented themselves. GATC is avoided despite the over-representation of ATC and GAT. There is a new fact in the paucity of GATC and GGCC that is not compatible with trinucleotide frequencies.

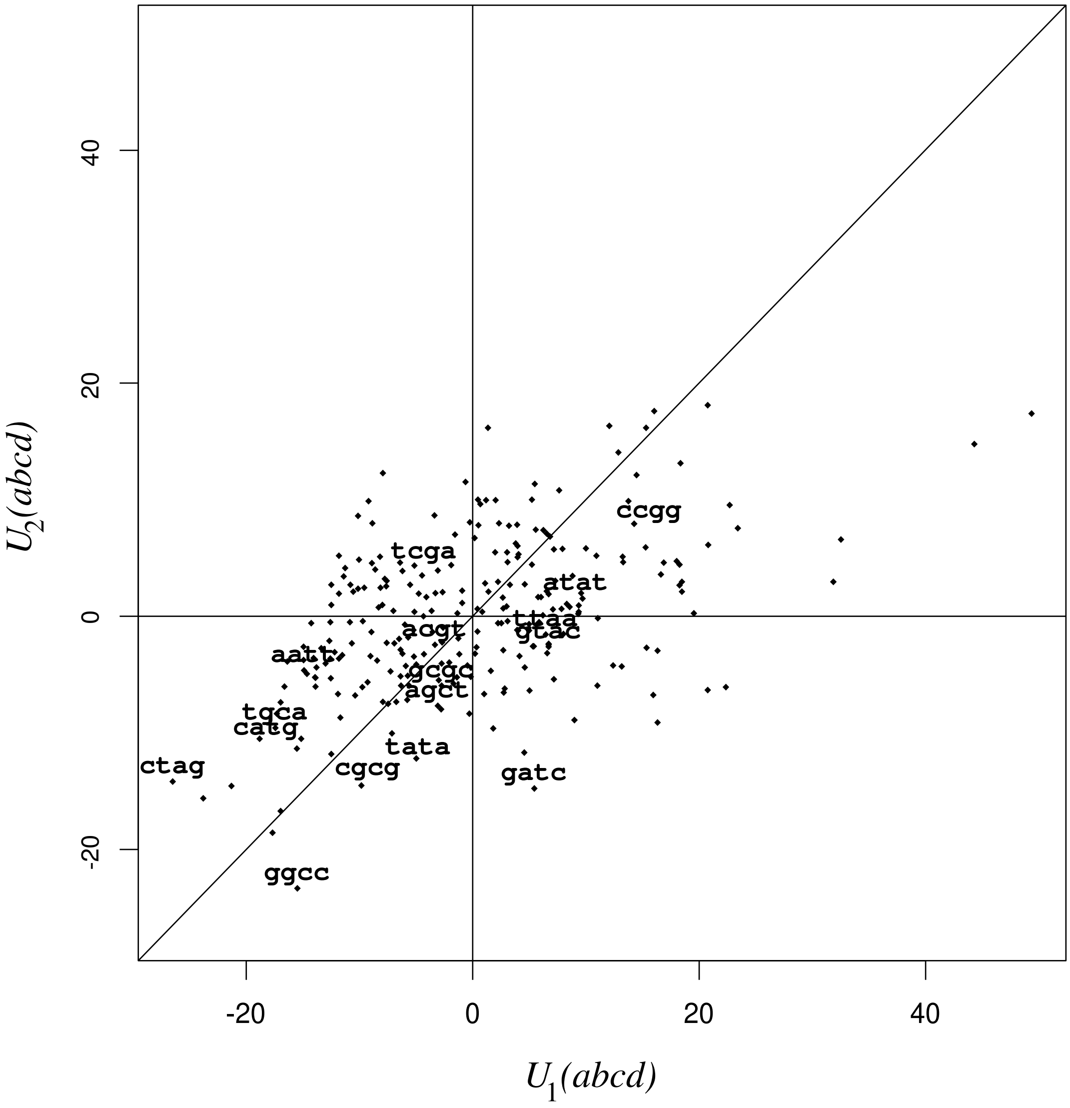

Figure 4: Exceptional 4-words under M1 versus M2.

Each word abcd is represented with coordinates $(U_1(abcd), U_2(abcd))$. *All palindromes are named in the figure, other words are just represented with a point. Most palindromes are avoided in both models.*

6-palindromes under *M1* to *M4*

6-words can be analyzed in four models. Figure 5 shows the statistics of the palindromes compared to the statistics of the whole set of 6-words.

Most palindromes are under-represented in the four models and the most under-represented words are palindromes. This is more visible in higher order models as shown in Table 7 where the number of palindromes among the 23 most under-represented words increases with the order of the model.

Table 7: UNDER-REPRESENTED 6-WORDS UNDER EACH MODEL IN *E. coli*.

M1	*M2*	*M3*	*M4*	Ranks
• GGC (-13)	• GCC (-14)	• GGC (-19)	• GGC (-17)	1
• GCC (-12)	• GGC (-14)	• GCC (-16)	• GCC (-14)	2
∘ GCA (-11)	• CGG (-12)	∘ AGC (-8)	⋆ CTG (-11)	3
• CGG (-10)	⋆ CTG (-10)	⋆ CTG (-8)	∘ AGC (-10)	4
⋆ ttggaa (-9)	∘ TGG (-10)	GGT (-8)	TCC (-7)	5
∘ gcttgg (-9)	⋆ CCG (-10)	• CGG (-8)	• CGG (-7)	6
CAC (-9)	tggccg (-9)	∘ CAG (-8)	GAG (-7)	7
TTG (-9)	⋆ ttggaa (-9)	⋆ CCG (-8)	⋆ CCG (-7)	8
cgcatg (-9)	∘ gcttgg (-8)	⋆ ttggaa(-7)	∘ CAG (-7)	9
cttgga (-9)	tttgga (-8)	∘ TGG (-7)	∘ GCA (-6)	10
–	CAC (-8)	CAC (-7)	–	11
–	–	–	CCC (-6)	12
–	–	–	GTC (-6)	13
–	CAG (-8)	–	–	14
GGG (-8)	–	–	CGT (-6)	15
–	–	–	AAA (-6)	16
–	–	–	GGT (-5)	17
–	–	–	–	18
–	–	–	ATC (-5)	19
–	–	–	CAC (-5)	20
–	GCA (-7)	–	–	21
–	–	–	AAG (-5)	22
ATG (-8)	tccaaa (-7)	ttccga (-5)	GGA (-5)	23
8	9	10	19	# of palindromes

The most under-represented 6-words in E. coli are listed here with their statistics under each model in parentheses. 6-palindromes are denoted by their 3 first letters. Non identified words (–) are not palindromes. In the *first 10 lines of the table, some words appear under more than one model : three words are exceptional under the 4 models (marked with $\bullet$), three words are exceptional under 3 models ($\star$) and five are exceptional under 2 models ($\circ$).*

Many palindromes are under-represented in more than one model, which says that the "paucity" is obtained in both using avoided subwords and extending them with a carefully chosen letter. Moreover, the fourth-order model maybe understood itself as a complex code on 6-words which avoids a list of palindromes and of a few other words and favors the occurrence of one-letter mutations of these palindromes (Table 8).

Table 8: OVER-REPRESENTED 6-WORDS UNDER M4 AND THEIR UNDER-REPRESENTED ONE-LETTER MUTATIONS IN *E. coli.*

Over-represented words [rank of U_4]	Under-represented one-letter mutations [rank of $-U_4$]	
TCCGGC [1]	TCCGGa [5]	gCCGGC [2]
AGCGCC [2]	AGCGCt [4]	gGCGCC [1]
GCCGGA [3]	GCCGGc [2]	tCCGGA [5]
GGCGCT [4]	GGCGCc [1]	aGCGCT [4]
GAGCTG [5]	GAGCTc [7]	cAGCTG [9]
TTGCAG [6]	–	cTGCAG [3]
GCCGGG [7]	GCCGGc [2]	cCCGGG [12]
CTAGTA [8]	–	–
TTGAAG [9]	–	cTGAAG[a] [11]
CAGCTT [10]	CAGCTg [9]	aAGCTT [4]
GGCGCG [11]	GGCGCc [1]	–

[a] *not a palindrome but in the list of most under-represented words under M4.*
A lowercase letter indicates the mutated base. Most of the 12 most under-represented words are a one-letter mutation of an over-represented word ; only the words of rank 6, 8, 10 of Table 7 (M4) do not appear.

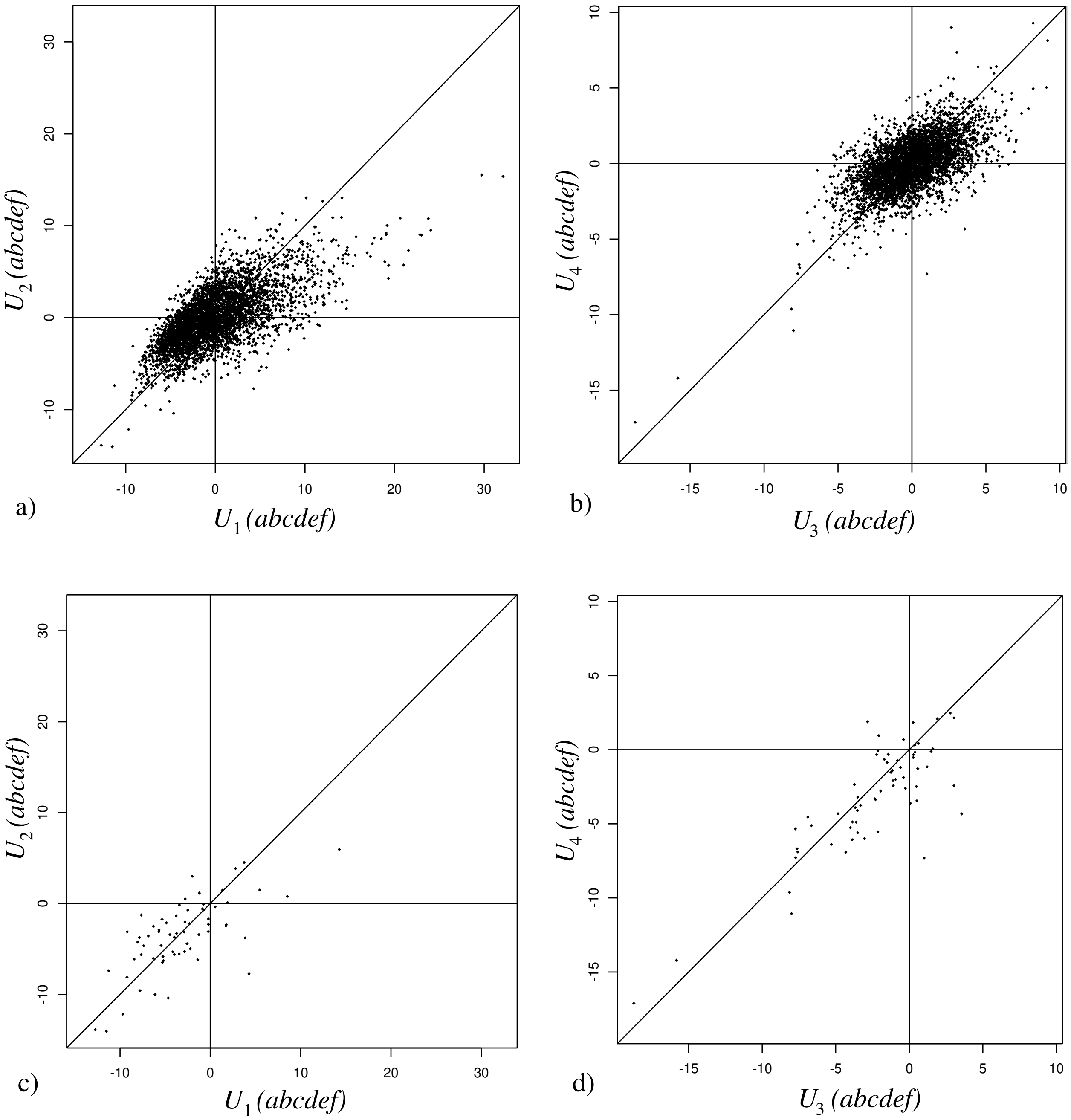

Figure 5: Exceptional 6-words and palindromes under M1 to M4.

6-words a) and 6-palindromes c) are represented under M1 versus M2 ; 6-words b) and 6-palindromes d) are represented under M3 versus M4.

This balance between avoided and prefered words which differ by only one letter is partially true under the other models as shown in Table 9.

Table 9: OVER- AND UNDER-REPRESENTED WORDS DIFFERING BY ONE LETTER IN *E. coli.*

Model	Over-represented word [rank of U]	Under-represented word [rank of $-U$]
M1	GCCAGC [3]	GCCgGC [2]
	GCTGGC [7]	GCcGGC [2]
M2	TCGCCA [4]	TgGCCA [5]
	TGGCGA [7]	TGGCcA [5]
M3	GCCGGA [1]	GCCGGc [2]
	TCCGGC [4]	gCCGGC [2]
	TCGCCA [7]	TgGCCA [10]
	AGGCCG [10]	cGGCCG [6]

The lowercase letter indicates the mutated base.

8-palindromes under *M1* to *M6*

Considering the distribution of the palindrome statistics in the set of all word statistics shows that many 8-palindromes are avoided under *M1* and *M2*. This is less true when the order of the model increases. Under *M5* and *M6*, palindrome statistics are scattered over the whole range of word statistics (not shown). In Table 10, the ranks (in brackets) of the under-represented palindromes show that the palindromes are not a majority among avoided words.

Finally, most of these palindromes are an extension of a 6- or 4-palindrome under-represented in the same model (in parentheses in Table 10). Only two 8-palindromes are avoided after 6 or 7 letters have been chosen. This means that these two palindromes CAGATCTG and CAGGCCTG may have a function determined by the whole word and not by a part of it. For the other 8-palindromes, on the contrary, the under-representation of, say, CGGCGCCG and TGGCGCCA corresponds to avoiding GGCGCC and avoiding to add C and G, or T and A, to the ends of it.

Table 10: UNDER-REPRESENTED 8-PALINDROMES IN *E. coli*.

M1	*M2*	*M3*	*M4*	*M5*	*M6*
C(GGC) [3]	C(GGC) [2]	C(GCC) [3]	T(GGC) [3]	CAGA [21]	CAGG [24]
T(GCA) [6]	C(GCC) [8]	T(GGC) [5]	C(GCC) [6]	CAGG [36]	
C(GCC) [14]	T(GGC) [13]	C(GGC) [7]	C(GGC) [10]		
C(GCA) [24]	CA(GA) [32]	T(GCC) [41]	T(GCC) [16]		
G(CAC) [28]	G(CGG) [35]	C(AGC) [71]	CA(GG) [18]		
TT(GG) [78]	G(CCG) [57]	A(GGC) [73]	C(AGC) [62]		
T(GGC) [89]	CT(GA) [75]	C(TGG) [105]			
	G(GCC) [91]				
	T(CAG) [99]				
	G(CTG) [101]				

8-palindromes with a statistic $U_m < -3$ *are listed here for* $m = 1, \cdots, 6$ *with their ranks in brackets, where the most under-represented word has rank 1. They are represented by their first 4 letters. Under-represented 6-palindromes of Table 7 and under-represented 4-palindromes appearing as subwords are in parentheses. Two underlined 8-palindromes are the repetition of an under-represented 4-palindrome.*

The statistical characterisation of under-represented 8-palindromes is therefore quite different from that of 6-palindromes, showing a specific function of 6-palindromes in *E. coli*.

5 Discussion

Many related problems have not been considered in this paper. Two of them, kindly pointed out by the referees, may be partly solved as follows:

– the question of groups of similar words, corresponding to ambiguous letters for instance, is solved using the covariances given in this paper. If $\mathcal{W}$ is a set of words, for instance the words associated to the intron splice site $YYRA$ (Avery 1987):

$$\mathcal{W} = YYRA = \{abcA,\ a = C \text{ or } T,\ b = C \text{ or } T,\ c = A \text{ or } G\},$$

we get the variance of $T_M(\mathcal{W}) = \sum_{W\in\mathcal{W}} T_M(W)$ with

$$\mathrm{Var}\,(T_M(\mathcal{W})) = \sum_{W\in\mathcal{W}} \mathrm{Var}(T_M(W)) + \sum_{W\in\mathcal{W}, W'\in\mathcal{W}, W\neq W'} \mathrm{Cov}(T_M(W), T_M(W')).$$

Sets of words of different lengths as those corresponding to indels can be analysed with the same techniques.

– The question of heterogeneity of the sequence can be treated if breakpoints are known between which a stationary model is acceptable. For instance the usual G+C variation along the chromosome suggests cutting the sequence into subsequences of two kinds, where the model would be $M^{(1)}$ and $M^{(2)}$ alternatively. If a few letters around the breakpoints are ignored, as well as words overlapping two subsequences, the two statistics $T^{(1)}(W)$ and $T^{(2)}(W)$ corresponding to the two sets of subsequences are independent and the standard statistic associated with $T^{(1)}(W) + T^{(2)}(W)$ is straightforward. But, if the number of breakpoints increases with n, words overlapping two subsequences have to be considered. Also the question of estimating the breakpoints is serious. It can be approached in the frame of Hidden Markov Chains Models as in Churchill (1992). Practical and theoretical work remains to be done to know how the set of exceptionnal words would be modified in such a context.

Finally, the two main results of this paper are the following:

1. The exact asymtotic variance of $T_M(W) = n^{-1/2}(N(W) - n\widehat{\mu}_M(W))$ is now available for the Markov chain models *Mm* and *Mm_3*, and also for the intermediate model MC where dinucleotide counts and codon counts are fixed. The same ideas could be used for the other intermediate models of higher order. Obviously, if we use this exact variance instead of the previous approximations found in the literature, we get sets of exceptional words that have many common words with the previous set. It is however rather difficult to measure the improvement of the biological interpretation since examination of each word separately would need a lot of precise biological knowledge.

2. However, there is a source of new information in comparing different models. The question is not to find the best model to fit to a given sequence but to describe more precisely the difference of fit between two nested models. It is interesting to

know, for instance, that the main information in phased 3-words when phased 2-words frequencies are known comes from the codons for both *E. coli* and *B. Subtilis*. This observation indicates how the codon usage is an appropriate summary of a sequence to be used in comparison between sequences or subsequences. The new information carried by 4-words is also summarized by the few words which are exceptional under *M2_3* and the method can be used further for longer words.
The palindrome example shows that the new information between 5- and 6-words in *E. coli* is partially described through a set of underrepresented 6-palindromes. This means that the underrepresentation of these words is obtained not only because subwords are avoided but also because biological constraint is present at the highest order corresponding to 6-words themselves.

These two examples are just old ones revisited but it would be useful to analyze many other genomes with this method.

Automatic methods are now needed to perform routine analyses of this kind on a larger scale. Infographics and other software developments could be appropriated for further analysis of DNA sequences, based on exceptional words under a series of nested models.

6 Appendix : Formulae

6.1 Formulae for M1

Whittle's formula

$$\mathrm{card}(\mathcal{S}) = \prod_a \frac{N(a+)!}{\prod_b N(ab)!} H(X_n),$$

where $N(a+) = \sum_b N(ab)$, $H(r)$ is the cofactor of any element in line r of the matrix $I - \hat{\pi}$ and $\hat{\pi}(a,b) = N(ab)/N(a+)$.

To avoid subscripts in the following formulae, we now denote u and v the first and last letters of W, $w_1 = u$, $w_h = v$, $w'_1 = u'$ and $w_{h'} = v'$.
We recall that $\mathrm{card}(\mathcal{S} \setminus W)$ is calculated as $\mathrm{card}(\mathcal{S})$ but replacing $N(ab)$ and $N(a+)$ with the corrected counts taking account of the deletion of the inner letters of W. With the same method, $\mathcal{S} \setminus W \setminus W$ or $\mathcal{S} \setminus W \setminus W'$ correspond to the deletion of the inner letters of two words successively, W and W, or W and W'.

Conditional variance

$$\begin{aligned}\mathrm{Var}\,(N(W) \mid \mathcal{S}) &= \mathbf{E}(N(W) \mid \mathcal{S}) + 2\sum_{d=1}^{h-2} \delta(W;d)\mathbf{E}(N(W^dW) \mid \mathcal{S}) \\ &+ (N(uv) - n(uv) + 1)(N(uv) - n(uv)) \frac{\mathrm{card}(\mathcal{S} \setminus W \setminus W)}{\mathrm{card}(\mathcal{S})} \\ &- \mathbf{E}(N(W) \mid \mathcal{S})^2 .\end{aligned}$$

Conditional covariance

$$\mathrm{Cov}(N(W), N(W') \mid \mathcal{S}) = \sum_{d=2-h'}^{h-2} \delta(W,W';d)\mathbf{E}(N(W^dW') \mid \mathcal{S})$$

$$+\widetilde{N}(uv)\left(\widetilde{N}(u'v') + \mathbb{1}\{uv = u'v'\}\right) \frac{\mathrm{card}(\mathcal{S} \backslash W \backslash W')}{\mathrm{card}(\mathcal{S})} - \mathbf{E}(N(W) \mid \mathcal{S})\,\mathbf{E}(N(W') \mid \mathcal{S}),$$

where $\widetilde{N}(uv) = N(uv) - n(uv) - n'(uv) + 1$, $\delta(W,W';d) = 1$ if the word W starting at position 1 and word W' starting at position $1+d$ are compatible, and in this case W^dW' denotes the concatenated word obtained by overlapping. Note that $1 - h' \leq d \leq h - 1$. If overlapping is impossible, $\delta(W,W';d)$ is zero.

6.2 Formulae for M1_3

Asymptotic conditional variance and covariance for words of length $3p+2$

$$\sigma^2_{1\_3}(W,k) = \lim_{n\to+\infty}\frac{1}{n}\mathrm{V}ar(N(W,k)\mid\bar{\mathcal{S}}) = \mu(W,k) + 2\sum_{d=1}^{h-2}\delta((W,k);d)\mu((W,k)^d(W,k))$$
$$+ \quad \mu(W,k)^2\left(\sum_{a,\ell}\frac{n(a+,\ell+1)^2}{\mu(a,\ell)} - \sum_{a,b,\ell}\frac{n(ab,\ell)^2}{\mu(ab,\ell)} + \frac{1-2n(u+,k)}{\mu(u,k-1)}\right),$$

$$\sigma_{1\_3}((W,k),(W',k')) = \lim_{n\to+\infty}\frac{1}{n}\mathrm{Cov}(N(W,k),N(W',k')\mid\bar{\mathcal{S}}) =$$
$$\sum_{d=2-h}^{h-2}\delta((W,k),(W',k');d)\,\mu((W,k)^d(W',k'))$$
$$+\,\mu(W,k)\,\mu(W',k')\left(\sum_{a,\ell}\frac{n(a+,\ell+1)\,n'(a+,\ell+1)}{\mu(a,\ell)} - \sum_{a,b,\ell}\frac{n_(ab,\ell)\,n'(ab,\ell)}{\mu(ab,\ell)}\right.$$
$$\left. - \frac{n(u'+,k')}{\mu(u',k'-1)} - \frac{n'(u+,k)}{\mu(u,k-1)} + \frac{\mathbb{I}\{u=u',k=k'\}}{\mu(u,k-1)}\right),$$

where $n(a+,\ell) = \sum_b n(ab,\ell)$, $n'(a+,\ell) = \sum_b n'(ab,\ell)$.

For words W of $3p+1$ letters, we use

$$\lim_{n\to+\infty}\frac{1}{n}\mathrm{Var}(N(W,k)\mid\bar{\mathcal{S}}) = \sum_{a,b}\lim_{n\to+\infty}\frac{1}{n}\mathrm{Cov}(N(aW,k)N(bW,k)\mid\bar{\mathcal{S}}),$$

and for words W of $3p$ letters, we use

$$\lim_{n\to+\infty}\frac{1}{n}\mathrm{Var}(N(W,k)\mid\bar{\mathcal{S}}) = \sum_{a,b,c,d}\lim_{n\to+\infty}\frac{1}{n}\mathrm{Cov}(N(abW,k),N(cdW,k)\mid\bar{\mathcal{S}}).$$

6.3 Formulae for MC

We denote $n(abc,3)$ the number of codons abc in the word (W,k) and $n(a\cdot c,3) = \sum_b n(abc,3)$. The probability in $\mathcal{C}$ of observing (W,k) at position i given that $(\widetilde{W},k)$ is at position i does not depend on k for words of type a) or b), and is

$$p_k(W) = \prod_{a,c}\frac{\prod_b N(abc,3)(N(abc,3)-1)\cdots(N(abc,3)-n(abc,3)+1)}{N(a\cdot c,3)(N(a\cdot c,3)-1)\cdots(N(a.c,3)-n(a.c,3)+1)},$$

$$p_k(W,W) = \prod_{a,c} \frac{\prod_b N(abc,3)(N(abc,3)-1)\cdots(N(abc,3)-n(abc,3)+1)}{N(a\cdot c,3)(N(a\cdot c,3)-1)\cdots(N(a.c,3)-2n(a.c,3)+1)},$$

$$p_{k,k'}(W,W') = \prod_{a,c} \frac{\prod_b N(abc,3)(N(abc,3)-1)\cdots(N(abc,3)-n(abc,3)-n'(abc,3)+1)}{N(a\cdot c,3)(N(a\cdot c,3)-1)\cdots(N(a\cdot c,3)-n(a\cdot c,3)-n'(a\cdot c,3)+1)},$$

$$\mathbf{E}(N(W,k)^2 \mid \mathcal{C}) = \mathbf{E}(N(W,k) \mid \mathcal{C}) + 2\sum_{d=1}^{h-1} \delta((W,k);d)\mathbf{E}(N((W,k)^d(W,k)) \mid \mathcal{C})$$
$$+p_k(W,W)(N(uv,k)-n(uv,k)+1)(N(uv,k)-n(uv,k))\frac{\mathrm{card}(\mathcal{S}' \setminus (\widetilde{W},k) \setminus (\widetilde{W},k))}{\mathrm{card}(\mathcal{S}')}.$$

Acknowledgments The problem considered in this paper was defined through discussions with S.D. Ehrlich from the Bacterial Genetics Department of INRA. Françoise Gélis provided an efficient toolkit of C++ and S+ programs to get and display the results of the methods presented here. They are available by contacting her by E-mail (Francoise.Gelis@jouy.inra.fr). This research has been partially supported by the Groupement d'Études et de Recherches sur le Génome (grant 61) of the French Ministère de l'Enseignement Supérieur et de la Recherche.

References

Altschul, S. F. and Erickson, B. W. (1985). Significance of nucleotide sequence alignments : A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.* **2** 526–538.

Arquès, D. G. and Michel, C. J. (1990a). A model of DNA sequence evolution. *Bull. Math. Biol.* **52** 741–772.

Arquès, D. G. and Michel, C. J. (1990b). Periodicities in coding and noncoding regions of the genes. *J. Theor. Biol.* **143** 307–318.

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

Beckmann, J. S., Brendel, V. and Trifonov, E. N. (1986). Intervening sequences exhibit distinct vocabulary. *J. Biomol. Struct. Dynamics.* **4** 391–400.

Blaisdell, B. E. (1985). Markov chain analysis finds a significant influence of neighboring bases on the occurrence of a base in eucariotic nuclear DNA sequences both protein-coding and noncoding. *J. Mol. Evol.* **21** 278–288.

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.

Burge, C., Campbell, A. and Karlin, S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc. Natl. Acad. Sci. USA.* **89** 1358–1362.

Chryssaphinou, O. and Papastavridis, S. (1988). A limit theorem for the number of non-overlapping occurrences of a pattern in a sequence of independent trials. *J. Appl. Prob.* **25** 428–431.

Churchill, G. A. (1992). Hidden Markov chains and the analysis of genome structure. *Computers Chem.* **16** 107–115.

Cowan, R. (1991). Expected frequencies of DNA patterns using Whittle's formula. *J. Appl. Prob.* **28** 886–892.

Gelfand, M. S., Kozhukhin, C. G. and Pevzner, P. A. (1992). Extendable words in nucleotide sequences. *Comp. Applic. Biosci.* **8** 129–135.

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

Karlin, S., Burge, C. and Campbell, A. M. (1992). Statistical analyses of counts and distributions of restriction sites in DNA sequences. *Nucl. Acids Res.* **20** 1363–1370.

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.

Lagunez-Otero, J. and Trifonov, E. N. (1992). mRNA periodical infrastructure complementary to the proof-reading site in the ribosome. *J. Biomol. Struct. Dynamics.* **10** 455–464.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A. and Danchin, A. (1991). Evidence for horizontal gene transfer in Escherichia coli speciation. *J. Mol. Biol.* **222** 851–856.

NUSSINOV, R. (1981). The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice. *J. Mol. Evol.* **17** 237–244.

PEVZNER, P. A. (1992). Nucleotide sequences versus Markov models. *Computers Chem.* **16** 103–106.

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. (1987a). 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.

PHILLIPS, G. J., ARNOLD, J. and IVARIE, R. (1987b). Mono- through hexanucleotide composition of the *escherichia coli* genome : a Markov chain analysis. *Nucl. Acids Res.* **15** 2611–2626.

PIETROKOVSKI, S. and TRIFONOV, E. N. (1992). Imported sequences in the mitochondrial yeast genome identified by nucleotide linguistics. *Gene.* **122** 129–137.

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

RAFTERY, A. and TAVARÉ, S. (1994). Estimation and modelling repeated patterns in high order Markov chains with the Mixture Transition Distribution Model. *Appl. Statist.* **43** 179–199.

RAFTERY, A. E. (1985). A model for high-order Markov chains. *J. R. Statist. Soc. B.* **47** 528–539.

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.

SHEPHERD, J. C. W. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proc. Natl. Acad. Sci. USA.* **78** 1596–1600.

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

TRIFONOV, E. N. (1987). Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *J. Mol. Biol.* **194** 643–652.

TRIFONOV, E. N. (1989). The multiple codes of nucleotide sequences. *Bull. Math. Biol.* **51** 417–432.

WHITTLE, P. (1955). Some distribution and moment formulae for the Markov chain. *J. R. Statist. Soc. B.* **17** 235–242.

Address reprint requests to:
Élisabeth de Turckheim
INRA
Laboratoire de Biométrie
F78352 Jouy-en-Josas Cedex, France
et@jouy.inra.fr

Received for publication November 4, 1994; accepted as revised March 31, 1995.