Journal of Computational Biology
Volume 4, Number 2, 1997
Mary Ann Liebert, Inc.
Pp. 189–192

AN EFFICIENT STATISTIC TO DETECT OVER- AND UNDER-REPRESENTED WORDS IN DNA SEQUENCES

Sophie Schbath[1]

Running head: Detecting Over-/Under-represented DNA Words

Abstract

In this note, we point out a very efficient statistic to detect over- and under-represented words in DNA sequences, when Markov chain models are used to represent the sequences. This statistic is missing from the recent review done on this important problem, and appears to be a better measure of rarity and abundance of words in DNA sequences.

Key words: DNA sequences, word counts, unexpected frequencies, Gaussian approximation, Markov chains.

1 Introduction

This paper is a note on the review Leung *et al.* (1996) made on various measures to test over- and under-representation of words in DNA sequences. They present and compare different statistics that have appeared in the recent literature, based essentially on the comparison of the observed count $N(W)$ of a word W in the sequence, with its expected count under a Markov chain model (z-scores, representation ratios, cross-ratios, ...). Actually, the expected count is replaced by a consistent estimator, denoted by $\widehat{N}(W)$, according to the DNA sequence of interest. For instance, the representation ratio is simply $N(W)/\widehat{N}(W)$ whereas the z-scores are normalizations of the difference $N(W) - \widehat{N}(W)$.

Their main conclusion is that z-scores are more appropriate and more rigorous because the normalization factor, that can be assimilated to a variance, takes into account the self-overlapping structure of the words. They considered three z-scores $z^L(W)$, $z^C(W)$ and $z^M(W)$ from Kleffe and Borodovsky (1992), Prum *et al.* (1995) and Schbath *et al.* (1995), each of them presenting some disadvantage. Indeed, $z^L(W)$ is not asymptotically distributed as a standard Gaussian whereas both of $z^C(W)$ and $z^M(W)$ are. Also, $z^M(W)$ actually tests if the last letter of the word appears with an expected frequency after the first letters of the word, and therefore seems to be of less interest in our framework than $z^L(W)$ and $z^C(W)$. Finally, $z^C(W)$ appears to be the best statistic, but it is not easily and efficiently computable for high-order Markov chains. Their remarks are incontestable and realistic from a practical point of view.

Nevertheless, Leung *et al.* (1996) missed a z-score statistic used in Schbath *et al.* (1995), that fits exactly the problem since it is asymptotically distributed according to the standard Gaussian distribution, is very efficiently computable, and tests whether the word itself appears with an expected frequency. Our aim here is then to point out this adequate statistic, very close to $z^C(W)$, to detect over- and under-represented words in DNA sequences.

[1]I.N.R.A., Unité de Biométrie, 78352 Jouy-en-Josas, France. E-mail: Sophie.Schbath@jouy.inra.fr

2 Efficient way to detect words with unexpected frequencies

Let us consider a stationary Markov chain of order $m \geq 1$ (Model Mm) on the four-letter DNA alphabet $\mathcal{A} = \{A, C, G, T\}$, with transition probabilities $\{\pi(a_1a_2\cdots a_m, a_{m+1}) > 0, a_i \in \mathcal{A}, i = 1, \ldots, m+1\}$, and stationary distribution $\{\mu(a_1a_2\cdots a_m), a_i \in \mathcal{A}, i = 1, \ldots, m\}$. The quantities $\pi(a_1a_2\cdots a_m, a_{m+1})$ and $\mu(a_1a_2\cdots a_m)$ are estimated respectively by their maximum likelihood estimators $\widehat{\pi}(a_1a_2\cdots a_m, a_{m+1}) = N(a_1a_2\cdots a_m a_{m+1})/N(a_1a_2\cdots a_m)$ and $\widehat{\mu}(a_1a_2\cdots a_m) = N(a_1a_2\cdots a_m)/n$, where n is the length of the DNA sequence and $N(\cdot)$ denotes the count of words in the DNA sequence. Thus, the Markov chain looks somehow like the sequence in the sense that it fits to the counts of all the words of length $(m+1)$ in the DNA sequence.

We fix a word $W = w_1w_2\cdots w_h$ of length $h \geq m+2$ on the alphabet $\mathcal{A}$. A z-score is defined to be such a statistic

$$z(W) = \frac{N(W) - \widehat{N}(W)}{\sqrt{\widehat{V}(W)}}$$

where $\widehat{N}(W)$ is a consistent estimator of the expected count of W and $\widehat{V}(W)$ is a consistent estimator of the variance of $N(W) - \widehat{N}(W)$ under the model. Since the expected count of W is given by $(n-h+1)\mu(w_1\cdots w_m)\prod_{j=1}^{h-m}\pi(w_j\cdots w_{j+m-1}, w_{j+m})$ under the model Mm, a natural and consistent estimator of the expected count of W is

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

The central limit theorem for Markov chains ensures the asymptotic normality of $n^{-1/2}(N(W) - \widehat{N}_m(W))$ (Prum *et al.*, 1995), so the associated z-score is asymptotically distributed as a standard Gaussian. The question remaining is to find a suitable normalization $\widehat{V}_m(W)$ easy to compute.

The solution is to come back to the $z^C(W)$ statistic where $\widehat{N}(W)$ is the conditional expectation of $N(W)$ given $\mathcal{S}_m$, the set composed of the m first letters of the DNA sequence and the collection $\{N(a_1a_2\cdots a_{m+1}), a_i \in \mathcal{A}, i = 1, \ldots, m+1\}$, and $\widehat{V}(W)$ is the conditional variance of $N(W)$ given $\mathcal{S}_m$:

$$z^C(W) = \frac{N(W) - \mathbb{E}(N(W) \mid \mathcal{S}_m)}{\mathrm{Var}(N(W) \mid \mathcal{S}_m)}\,. \tag{2}$$

Indeed, the conditional expectation is asymptotically equivalent to $\widehat{N}_m(W)$ given by (1), and the limit of the conditional variance divided by n, as $n \to \infty$, is precisely the asymptotic variance of $n^{-1/2}(N(W) - \widehat{N}_m(W))$ (Prum *et al.*, 1995). The asymptotic variance of $n^{-1/2}(N(W) - \widehat{N}_m(W))$ has been given by Schbath *et al.* (1995) and we can easily derive from it a natural "plug-in" estimator $\widehat{V}_m(W)$:

$$\widehat{V}_m(W) = \widehat{N}_m(W) + 2\sum_{d=1}^{h-m-1}\delta(W;d)\,\widehat{N}_m(W^dW) \tag{3}$$

$$+\widehat{N}_m(W)^2\left(\sum_{a_1,\ldots,a_m}\frac{n(a_1\cdots a_m+)^2}{N(a_1\cdots a_m)} - \sum_{a_1,\ldots,a_{m+1}}\frac{n(a_1\cdots a_{m+1})^2}{N(a_1\cdots a_{m+1})} + \frac{1-2n(w_1\cdots w_m+)}{N(w_1\cdots w_m)}\right),$$

where $\delta(W;d)$ is equal to 1 if the first $h-d$ letters of W are the same as the last $h-d$ letters, and 0 otherwise; W^dW denotes the concatenated word $w_1w_2\cdots w_dw_1w_2\cdots w_h$, small $n(\cdot)$ denotes the count of words inside the word W and $n(a_1\cdots a_m+) = \sum_b n(a_1\cdots a_m b)$.
Finally, we denote by $z^{AC}(W)$ the following statistic that is asymptotically equivalent to $z^C(W)$:

$$z^{AC}(W) = \frac{N(W) - \widehat{N}_m(W)}{\sqrt{\widehat{V}_m(W)}}, \tag{4}$$

where $\widehat{N}_m(W)$ and $\widehat{V}_m(W)$ are given respectively by (1) and (3) and can be calculated efficiently. To give an idea of the computational complexity, the statistics z^{AC} of all the words of length 5 in *M3* are calculated in 4 seconds with the *R'MES* software (Gélis and Schbath, 1996), whereas 56 minutes are required to obtain the statistics for all the words of length 8 in *M6* [these times have been obtained on a SPARC 1000 SUN Station under Solaris 2.4 with a sequence of length about 50kb].

Using the Markov chain model of order $h-2$ to study the words of length h is particularly relevant because it integrates the observed frequencies of all the shorter words of length less or equal to $h-1$. Leung *et al.* (1996) have considered mostly this maximal model. However, they did not notice that, in this particular case, the z-scores $z^M(W)$, $z^C(W)$ and $z^{AC}(W)$ are asymptotically equivalent. Indeed, the asymptotically Gaussian statistic $z^M(W)$ in the model *Mm* is defined by

$$z^M(W) = \frac{N(W) - \widehat{\widehat{N}}_m(W)}{\sqrt{\widehat{\widehat{V}}_m(W)}}, \tag{5}$$

with $\widehat{\widehat{N}}_m(W) = N(w_1\cdots w_{h-1})\widehat{\pi}(w_{h-m}\cdots w_{h-1}, w_h)$ and $\widehat{\widehat{V}}_m(W)$ is any consistent estimator of the limiting variance given in Prum *et al.* (1995) for $m=1$ and in Schbath (1995) in the general case. When $m = h-2$,

$$\widehat{\widehat{N}}_{h-2}(W) = \frac{N(w_1\cdots w_{h-1})N(w_2\cdots w_h)}{N(w_2\cdots w_{h-1})} \tag{6}$$

and a natural plug-in estimator $\widehat{\widehat{V}}_{h-2}(W)$ is

$$\begin{aligned}\widehat{\widehat{V}}_{h-2}(W) \quad &= \quad \frac{N(w_1\cdots w_{h-1})N(w_2\cdots w_h)}{N(w_2\cdots w_{h-1})^3}\Big(N(w_2\cdots w_{h-1}) - N(w_1\cdots w_{h-1})\Big) \\ &\times\Big(N(w_2\cdots w_{h-1}) - N(w_2\cdots w_h)\Big). \end{aligned} \tag{7}$$

We can see that the estimators $\widehat{N}_{h-2}(W)$ and $\widehat{\widehat{N}}_{h-2}(W)$ given respectively by (1) and (6) are identical. This implies the limiting variance in each case is the same, and the statistics $z^M(W)$ and $z^{AC}(W)$ are equivalent, as $n\to\infty$, when $m = h-2$. Moreover, in the maximal model, the z-score $z^M(W)$ really tests whether the word W appears with an expected frequency in the sequence, as do the z-scores $z^C(W)$ and $z^{AC}(W)$; using $z^M(W)$ has the advantage that $\widehat{\widehat{V}}_{h-2}(W)$ is even easier to compute than $\widehat{V}_{h-2}(W)$ (formulae (7)) and (3)). It takes 2 seconds to compute the z-scores z^M for all the words of length 5 in *M3*, and only 28 seconds for all the words of length 8 in *M6*.

3 Conclusion

In conclusion, the statistic $z^{AC}(W)$ given by (1), (3) and (4) provides a very efficient and accurate way to detect over- and under-represented words in DNA sequences, considering a general m-order Markov chain model. When using the maximal model, the statistic $z^M(W)$, given by (5), (6) and (7), is asymptotically equivalent to $z^{AC}(W)$ and is more efficient. The expression of $z^{AC}(W)$ and $z^M(W)$ can also be obtained in 3-periodic Markov chain models adapted for coding DNA sequences (see Schbath *et al.* (1995) and Schbath (1995)).

The computation of the z-scores $z^C(W)$, $z^{AC}(W)$ and $z^M(W)$ is one of the purposes of the *R'MES* software that also proposes several graphical representations allowing, for instance, the scientist to study easily the frequence of a word and all its subwords, to compare the exceptional words of a DNA sequence using several models, or to compare the exceptional words of two different sequences. *R'MES* is available at the following web location
`http://www-bia.inra.fr/J/AB/genome/RMES/welcome.html`.

REFERENCES

Gélis, F. and Schbath, S. (1996). *R'MES*: Recherche de Mots Exceptionnels dans les Séquences d'ADN, Guide de l'utilisateur – Version 1. *INRA, Département de Biométrie et Intelligence Artificielle, F78352 Jouy-en-Josas.*

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.

Leung. M.-Y., Marsh, G. M. and Speed, T. P. (1996). Over- and underrepresentation of short DNA words in Herpesvirus genomes. *J. Comp. Bio.* **3** 345–360.

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

Schbath, S. (1995). 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, France.

Schbath, S., Prum, B. and Turckheim, É. de (1995). Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comp. Biol.* **2** 417–437