Atelier INSERM, La Londes les Maures, April 27-28, 2006

Statistics of motifs

Sophie SCHBATH
Institut National de la Recherche Agronomique
Unité Mathématique, Informatique & Génome, Jouy-en-Josas, France

1 Introduction

In this lecture we will essentially focus on the statistical analysis of the number of overlapping occurrences (*count*) of a given oligonucleotide (*word*), or a given degenerated oligonucleotide (*motif or word family*), in a DNA sequence. Of course, there is no restriction to sequences on a 4 letter alphabet. Related topics will be just mentioned at the end, with appropriate references. Moreover, note that this lecture is part of a more complete presentation published in the book *DNA, Words and Models* (Robin *et al.*, 2003, 2005) that contains much more references.

The question we would like to address is "does this word occur in this sequence with an expected frequency?" In other words, can we observe it so many times, or so few times, just by chance? Usually, when the answer is no, such word is candidate to get a particular biological meaning; only a candidate: statistical significance is not equivalent to biological significance.

As a guiding example, we will look at the occurrences of the octamer `gctggtgg` in the complete genome of *Escherichia coli* (leading strands). This word is known as the Chi motif of the bacterium; it is very frequent, with 762 occurrences on the leading strands and it is necessary for the stability of the chromosome. Let us do the following simple calculation: "if all the 4^8 octamers would have the same occurrence probability in a sequence of length 4638858, then one expects to see each of them $4638851/4^8 \simeq 70$ times in the sequence. At this point, the Chi motif seems very over-represented in *E. coli* because we compare 762 occurrences with 70 occurrences.

The key idea is indeed to compare the observed count with the one we could expect given some knowledge on the sequence. To decide if a word count is expected or not, we need to know what to expect. This will be defined by a probabilistic model, i.e. by the description of what is "random". After choosing the appropriate model (Section 2),

one needs to evaluate the significance of the difference between observed and expected count (Section 3). In fact, one will calculate the p-value which is the probability, under our model, to observed as much (or as few) occurrences of our word of interest (Section 4). As we will see in Section 5, the p-value intrinsically depends on the chosen model: a word can be exceptionally frequent in one model but expected in another one which, for instance, takes more information on the sequence composition into account. Therefore, when claiming that an observation is statistically significant, do not forget to mention your a priori, your reference, your model.

2 A model as reference

As said in the introduction, to decide if a count is significantly too high or too low, one needs to know its expected count. In genome analysis, we usually have a single observation for the observed count of a particular word. There is no way to get independent and identically distributed copies of this count: words are not independent and genomes are "unique". Therefore, the expected count has to be evaluated thanks to "random sequences" that look like, in some sense, the genome of interest.

Markov chain models are widely used in genome analysis for two reasons. First of all, when their parameters are correctly estimated according to the analyzed genome, such models fit the composition of the genome in words of length 1 up to $(m+1)$, where m is the chosen order for the Markov chain. It means in particular that the observed sequence is compared to sequences that have the same composition in short words. It is also possible to take the 3-periodicity of coding sequences into account (phased Markov chains) and some heterogeneities (hidden Markov chains). Second, many theoretical results exist and analytical probability calculations can often be performed, avoiding heavy simulations.

In the remainder, we will denote by Mm the Markov model of order m. From a theoretical point of view, letters in a Markov chain of order m depend on the m previous letters and they are generated thanks to transition probabilities. The Bernoulli model which assumes independent letters is a particular case of the model M1.

3 Observed and expected count

Let consider the following notations:

- n is the genome length,
- $\mathcal{A}$ is the four letter DNA alphabet,

- $\mathbf{X} = X_1 X_2 \cdots X_n$ is a random sequence of letters from $\mathcal{A}$ (model Mm),
- $\mathbf{w}$ is a word of length h on $\mathcal{A}$,
- $N(\cdot)$ denotes the count,
- $Y_i(\mathbf{w})$ is 1 if $\mathbf{w}$ occurs at position i in $\mathbf{X}$, and 0 otherwise.

The number of occurrences of $\mathbf{w}$ in $\mathbf{X}$ can be written like

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

and its expectation is simply $\mathbb{E}N(\mathbf{w}) = (n-h+1)\mathbb{P}(\mathbf{w} \text{ at } i)$. The probability $\mathbb{P}(\mathbf{w} \text{ at } i)$ can be easily expressed with respect to the transition probabilities. Usually, these transition probabilities are estimated according to the observed genome, for instance in model M1, the probability that `t` is followed by `a` is estimated by $N^{\text{obs}}(\texttt{ta})/N^{\text{obs}}(\texttt{t})$[1] where the exponent $^{\text{obs}}$ indicates that it is the observed count in the genome.

We will then compare the observed count $N^{\text{obs}}(\mathbf{w})$ of $\mathbf{w}$ with the following natural estimator $\widehat{N}_m(\mathbf{w})$ of the expected count under model Mm. Here are some examples for the 5-letter word `atcga` under models M0 (Bernoulli model) to M3:

Model Mm	Fit	Estimated expected count
M0	bases	$\widehat{N}_0(\mathbf{w}) = \dfrac{N^{\text{obs}}(\texttt{a})N^{\text{obs}}(\texttt{t})N^{\text{obs}}(\texttt{c})N^{\text{obs}}(\texttt{g})N^{\text{obs}}(\texttt{a})}{n^4}$
M1	dinucl.	$\widehat{N}_1(\mathbf{w}) = \dfrac{N^{\text{obs}}(\texttt{at})N^{\text{obs}}(\texttt{tc})N^{\text{obs}}(\texttt{cg})N^{\text{obs}}(\texttt{ga})}{N^{\text{obs}}(\texttt{t})N^{\text{obs}}(\texttt{c})N^{\text{obs}}(\texttt{g})}$
M2	trinucl.	$\widehat{N}_2(\mathbf{w}) = \dfrac{N^{\text{obs}}(\texttt{atc})N^{\text{obs}}(\texttt{tcg})N^{\text{obs}}(\texttt{cga})}{N^{\text{obs}}(\texttt{tc})N^{\text{obs}}(\texttt{cg})}$
M3	tetranucl.	$\widehat{N}_3(\mathbf{w}) = \dfrac{N^{\text{obs}}(\texttt{atcg})N^{\text{obs}}(\texttt{tcga})}{N^{\text{obs}}(\texttt{tcg})}$

We clearly see that **if we choose model Mm then, the estimated expected count only depends on the composition of the genome in words of length** $(m+1)$ and m. It means that our count of reference only takes the composition in $(m+1)$-letter words (and shorter) into account. This is a key point as regard to the choice of the order m in practice (see Section 5).

[1] To be completely rigorous one should divide by $N(\texttt{t}+)$, the number of `t`'s followed by a letter ...

Moreover, note that M3 is the maximal model to analyze the exceptionality of a 5-letter word because, in M4 and higher models, the estimated expected count would be the observed count itself. More generally the maximal model will be of order $h-2$.

Table 1 gives the estimated expected count under various models of the Chi motif, together with two other octamers, in both leading strands of *E. coli*. Clearly these counts of reference change with the model: one can see for instance that the three octamers are over-represented in all the models, despite its 70 occurrences `ccggccta` "seems" exceptionally frequent as we take more and more information on *E. coli*'s composition whereas the 828 occurrences of `ggcgctgg` "seems" expected given the heptamers of the genome. Only the p-values will tell us if the observed counts are significantly different from the estimated expected counts under each model.

		`gctggtgg`	`ggcgctgg`	`ccggccta`
	Fit	762 occurrences	828 occurrences	71 occurrences
M0	bases	85.944	85.524	70.445
M1	dinucl.	84.943	125.919	48.173
M2	trinucl.	206.791	255.638	35.830
M3	tetranucl.	355.508	441.226	14.697
M4	pentanucl.	355.312	392.252	15.341
M5	hexanucl.	420.867	633.453	27.761
M6	heptanucl.	610.114	812.339	25.777

Table 1: Estimated expected count of 3 octamers in both leading strands of *E. coli* under models M0 to M6.

4 Scores and p-value

The first score of exceptionality that have been used in the literature was the ratio observed count over (estimated)[2] expected count. The problem with this crude score is that one does not know its variability around 1 and one cannot give a significant threshold.

z-score asymptotically Gaussian Then, people thought to normalize the difference between observed count and (estimated) expected count and to assume that this so-called

[2]I put some parenthesis because most of the time, people forget that the parameters of the model have been estimated and then depend on the observed sequence; an estimator is then usually a random variable.

z-score is distributed, at least asymptotically, according to the $\mathcal{N}(0,1)$. Before that the variance of the count was provided, the normalizing factor used was the square root of the (estimated) expected count as if the count would follow a Poisson distribution. As we will see, this was not a so bad idea. In 1992, the formula for the variance came out (Kleffe and Borodovsky (1992)) solving half of the problem. The z-score is indeed asymptotically distributed according to the $\mathcal{N}(0,1)$ distribution, but the limiting variance 1 is correct only if we assume that the parameters are the true ones. If they are estimated according to the observed sequence, the square root of the estimated variance is no more the appropriate normalizing factor. The good normalizing factor was finally proposed by Prum *et al.* (1995) under M1, and generalized later in models Mm. Like the variance of the count, the normalizing factor explicitly depends on the periods of the word; an integer $p < h$ is a period of the word $\mathbf{w}$ if and only if two occurrences of $\mathbf{w}$ may occur at a distance p apart.

p-value The p-value $\mathbb{P}(N(\mathbf{w}) \geq N^{\text{obs}}(\mathbf{w}))$ can then be approximated by the probability that a $\mathcal{N}(0,1)$ random variable is greater that the observed value of the z-score. If the p-value is close to zero, then the word is significantly frequent; if it is close to 1, it means that $\mathbb{P}(N(\mathbf{w}) < N^{\text{obs}}(\mathbf{w}))$ is close to zero and the word is significantly rare.

Compound Poisson approximation Because a word count is positive, a Gaussian distribution is not really appropriate to approximate the distribution of the count of an expectedly rare word (small estimated expected count). Poisson approximations are known to be better for the count of rare events. In fact, a Poisson approximation is satisfactory for the count of a non-overlapping word, but it is not for overlapping words. Indeed, occurrences of an overlapping word produce *clumps* of overlapping occurrences. The number of clumps can be correctly approximated by a Poisson variable but we need to deal with the number of occurrences of the word in each clump. Since this clump size is geometrically distributed, it leads to a compound Poisson approximation for the count $N(\mathbf{w})$ (Schbath (1995)). The p-value will then be approximated by the tail distribution of the limiting compound Poisson distribution $\mathbb{P}(\mathcal{CP} \geq N^{\text{obs}}(\mathbf{w}))$.

Exact distribution Later, the exact distribution of the word count in a Markovian sequence was provided either via a recursive formula (Robin and Daudin (1999), Robin *et al.* (2005)) or its generating function (Régnier (2000)). In practice this exact distribution is not really used, except for short sequences (<10kb), because numerical instabilities happen with the recursive formula and symbolic calculation are required to get the

Taylor expansion of the generating function. However, the exact distribution allows to measure the quality of the approximations (Gaussian and compound Poisson) for medium sequences.

Comparison Numerical comparisons performed in Robin and Schbath (2001) indicate that the Gaussian distribution is well adapted when estimated expected counts are far from 100, but should not be used when the expected count is less than 10. The compound Poisson distribution performs very well, but in practice numerical instabilities may arise to calculate the tail distribution (i.e. the approximate p-value) if the expected count is large, say more than 100 (however works are in progress in this direction).

Large deviation A third method to approximate the p-value consists in using the theory of large deviation (Nuel (2001)). It is particularly of interest to get an accurate approximation of the p-value for very exceptional words.

Software Let just mention two softwares dedicated to the detection of exceptional words: *R'MES* (`http://genome.jouy.inra.fr/ssb/rmes/`) and
SPatt (`http://stat.genopole.cnrs.fr/spatt/`).

5 Choice and influence of the model

The most frequent question about exceptional motifs is "how to choose the order m of the Markov model?". In fact there is no a unique answer. Here are some elements that have to be kept in mind when we are interested by the statistical significance of a word count.

- Choosing model Mm means that the composition of the genome in oligonucleotides of length (m+1) and shorter will be taken into account to get the estimated expected count and the p-value.
- Higher the order m, better the fit and fewer unexpected events.
- The number of parameters to be estimated in model Mm is 3×4^m; the sequence should be long enough to have accurate estimates (ideally more than 1000 times the number of parameters).

Model	Fit	Expected	z-score	p-value	Rank
M0	bases	85.944	72.9	$< 10^{-323}$	3
M1	dinucl.	84.943	73.5	$< 10^{-323}$	1
M2	trinucl.	206.791	38.8	$< 10^{-323}$	1
M3	tetranucl.	355.508	22.0	$1.4\,10^{-107}$	5
M4	pentanucl.	355.312	22.9	$2.3\,10^{-116}$	2
M5	hexanucl.	420.867	19.7	$1.0\,10^{-86}$	1
M6	heptanucl.	610.114	10.6	$1.5\,10^{-26}$	3

Table 2: Estimated expected counts, z-scores and p-values (Gaussian approximation) for the Chi motif in both strands of *E. coli* genome under models M0 to M6. The rank corresponds to the rank of Chi when the 65536 octamers are sorted with respect to their scores in decreasing order. Results obtained with the *R'MES* software.

As regard to these remarks, the maximal model (order $m = h - 2$) is of real interest for rather short words because it allows to identify h-letter words having an exceptional frequency which cannot be explained by the composition of the genome in shorter words.

When we are interested by some particular words, it should be fruitful to use all the models. Either the word is exceptional in all models, meaning that biological investigations should be done to understand such constraint on the genome, like for the Chi motif in *E. coli* (see table 2). Or it looses its exceptionality as we increase the order of the model, meaning that its frequency can be explained by the frequency of its subwords (advantage of a pyramidal display, see Robin *et al.* (2005)), or it is exceptional in the maximal model, meaning that it represents a real bias in the genome composition.

Table 3 is just to illustrate the fact that exceptionally frequent (resp. rare) words are not necessarily the ones with a high (resp. low) count. The analysis has been made on a DNA sequence of 111 402 bps from *E. coli* genome. It shows for instance that `ggcct` occurs 91 times which is few under models M0, M1 and M2, but as soon as we take into account the tetranucleotide composition of the sequence, 91 becomes significantly too high; `ggcct` is the most exceptionally frequent 5-mer in the sequence. We have the opposite situation with `cctgg` which occurs 150 times and is the most under-represented 5-mer under M3 in the sequence. The explanation is simply that `ggcct` is composed of an exceptionally rare tetranucleotide (`ggcc`) and `cctgg` is composed of an exceptionally frequent tetranucleotide (`cctg`) and only M3 knows these information.

	ggcct			cctgg		
	obs	exp	z-score	obs	exp	score
M0	91	127	-3.2	150	127	2.0
M1	91	107	-1.6	150	96	5.6
M2	91	105	-1.5	150	158	-0.7
M3	91	55	5.7	150	205	-5.4
	the most over-represented			the most under-represented		
	given the tetranucleotide composition					

Table 3: Estimated expected counts and z-scores for `ggcct` and `cctgg` in a sequence of 111 402 bps from *E. coli* genome under models M0 to M3. Results obtained with the *R'MES* software.

6 Related topics

Results exist for the number of occurrences of non-overlapping occurrences. For the number of clumps, the exact distribution can be obtained via its generating function (Stefanov *et al.* (2006)), and a Poisson approximation has been proposed (Schbath (1995)). For the number of renewals, see Reinert *et al.* (2005) and references therein (exact distribution, Gaussian and Poisson approximations).

Results exist to decide if distances between successive occurrences, or cumulative distances, are significantly too high or too low. Two kind of models are considered to determine the reference: a Markov model on the sequence (Robin and Daudin (1999)) or a compound Poisson process for the word occurrences (Robin (2002)). The advantage of the later model is that it takes the word frequency into account. Another approach has been proposed by Gusto and Schbath (2005) to study statistically favored or avoided distances between two motifs. Here the null hypothesis is that both motifs occur independently, and we look at the correlation profiles that capture the departure from the null hypothesis.

Finally, let mention that statistics of structured motifs (two words, called boxes, separated by a variable but bounded distance) is much more complicated than for simple motifs because we cannot describe the complete overlapping structure of the structured motifs. Some works have been done (Robin *et al.* (2002), Stefanov *et al.* (2006)) but there is still room for improvements regarding generalizations to more than two boxes or to degenerated boxes.

References

Gusto, G. and Schbath, S. (2005). FADO: a statistical method to detect favored or avoided distances between motif occurrences using the hawkes' model. *Statistical Applications in Genetics and Molecular Biology.* **4** 0. Article 24.

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.

Nuel, G. (2001). *Grandes déviations et chaînes de Markov pour l'étude des mots exceptionnels dans les séquences biologiques.* PhD thesis, Université d'Evry Val d'Essonne.

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

Reinert, G., Schbath, S. and Waterman, M. (2005). Statistics on Words with Applications to Biological Sequences, chapter in *Applied Combinatorics on Words*, volume 105 of *Encyclopedia of Mathematics and its Applications*, Cambridge University Press.

Régnier, M. (2000). A unified approach to word occurrence probabilities. *Discrete Applied Mathematics.* **104** 259–280.

Robin, S. (2002). A compound Poisson model for words occurrences in DNA sequences. *J. R. Statist. Soc. C.* **51** 437–451.

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

Robin, S., Daudin, J.-J., Richard, H., Sagot, M.-F. and Schbath, S. (2002). Occurrence probability of structured motifs in random sequences. *J. Comp. Biol.* **9** 761–773.

Robin, S., Rodolphe, F. and Schbath, S. (2003). *ADN, mots et modèles.* BELIN.

Robin, S., Rodolphe, F. and Schbath, S. (2005). *DNA, Words and Models.* Cambridge University Press (english version of *ADN, mots et modèles*, BELIN 2003).

Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.* **8** 349–359.

Schbath, S. (1995). Compound Poisson approximation of word counts in DNA sequences. *ESAIM: Probability and Statistics*. **1** 1–16.

Stefanov, V., Robin, S. and Schbath, S. (2006). Waiting times for clumps of patterns and for structured motifs in random sequences. *Discrete Applied Mathematics*. To appear.