Statistical tests to compare motif count exceptionalities

Stephane Robin, Sophie Schbath, Vincent Vandewalle

BMC Bioinformatics

Methodology article

Statistical tests to compare motif count exceptionalities

Stéphane Robin*[1], Sophie Schbath*[2] and Vincent Vandewalle[1]

Address: [1]INA PG/ENGREF/INRA, UMR518 Unité Mathématiques et Informatique Appliquées, 75005 Paris, France and [2]INRA, UR1077 Unité Mathématique, Informatique et Génome, 78350 Jouy-en-Josas, France

Email: Stéphane Robin* - robin@inapg.inra.fr; Sophie Schbath* - sophie.schbath@jouy.inra.fr; Vincent Vandewalle - vincent_vandewalle@yahoo.fr

* Corresponding authors

Published: 8 March 2007

BMC Bioinformatics 2007, **8**:84 doi:10.1186/1471-2105-8-84

Received: 15 September 2006
Accepted: 8 March 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/84

Abstract

Background: Finding over- or under-represented motifs in biological sequences is now a common task in genomics. Thanks to *p*-value calculation for motif counts, exceptional motifs are identified and represent candidate functional motifs. The present work addresses the related question of comparing the exceptionality of one motif in two different sequences. Just comparing the motif count *p*-values in each sequence is indeed not sufficient to decide if this motif is significantly more exceptional in one sequence compared to the other one. A statistical test is required.

Results: We develop and analyze two statistical tests, an exact binomial one and an asymptotic likelihood ratio test, to decide whether the exceptionality of a given motif is equivalent or significantly different in two sequences of interest. For that purpose, motif occurrences are modeled by Poisson processes, with a special care for overlapping motifs. Both tests can take the sequence compositions into account. As an illustration, we compare the octamer exceptionalities in the *Escherichia coli* K-12 backbone versus variable strain-specific loops.

Conclusion: The exact binomial test is particularly adapted for small counts. For large counts, we advise to use the likelihood ratio test which is asymptotic but strongly correlated with the exact binomial test and very simple to use.

Background

Detecting motifs with a significantly unexpected frequency in DNA sequences has become a very common task in genome analysis. It is generally addressed to propose candidate functional motifs based on their statistical properties [1-3]. Lots of statistical methods have been developed to that purpose (see the recent surveys by [4] or [5] and references therein) and satisfactory solutions exist now to find exceptional motifs thanks to *p*-value calculations.

More recently, a new related question has arisen in the literature concerning the comparison of motif exceptionalities in two sequences. One wants for instance to compare particular sets of genes [6], upstream regions of CDSs versus whole chromosome [7], structural domains [8], CDSs versus intergenic regions, conserved regions versus strain-specific regions of bacterial genomes [9], or chromosomes from the same species [10]. Chromosomes from different species can also be compared from a comparative genomics point of view. In all these works, one would like to know if a given motif is significantly more exceptional in one sequence compared to another one. This criterion is

usually used to identify motifs which are specific from some regions or expected to be more frequent in some particular parts of the genome. Transcription factor binding sites, for instance, are expected to be more frequent in upstream regions than along the whole genome.

Surprisingly, no rigorous statistical method has been proposed yet to decide if a given motif, exact or not, is significantly more exceptional in one sequence compared to a second one. Of course, two *p*-values can be calculated separately on each sequence to know if the motif is exceptional in these sequences but the difficult point is how to compare these two *p*-values from a statistical point of view. It is indeed not sufficient to make the difference or the ratio to know if the two *p*-values are significantly different; One needs a statistical test.

In this paper, we propose two statistical tests to compare the motif count exceptionalities in two independent sequences. In the Results Section, we first present the underlying model for motif occurrences and the null hypothesis to test, namely the motif is similarly exceptional in both sequences. Then we derive an exact binomial test and an asymptotic likelihood ratio test adapted for frequent motifs. Usage conditions and power of both tests are described in the Discussion Section, together with a more refined model for occurrences of overlapping words and the associated tests. An illustration of the method is finally given; We compare the octamer exceptionalities in two sets of regions (backbone/loops) from the *Escherichia coli* K12 leading strands. These two sets correspond to the mosaic structure of *E. coli*'s genome when comparing the two strains K12 and O157:H7: the backbone represents the common regions whereas the loops are specific to the K12 strain. As a toy example all along this paper, we will treat in detail the case of the palindromic octamer cagcgctg which occurs respectively 30 times in the loops (758434 bps long) and 113 times in the backbone (3 882 513 bps long).

Results

Poisson model

In sequence i, the motif count N_i is supposed to have a Poisson distribution with mean (and variance) λ_i. This distribution has been shown to fit correctly theoretical (in Markovian sequences, for example) as well as observed count distributions of non-overlapping words [11]; A non-overlapping word is a word such that two occurrences of itself can not overlap in a sequence.

The mean count λ_i in sequence i must account for three parameters: (*i*) the length ℓ_i of the sequence, (*ii*) the composition of the sequence, (*iii*) the possible exceptionality of the motif in the sequence.

Expected intensity

The composition of the sequence can be accounted for via the probability μ_i for the motif to occur at any position in the sequence under a simple model. The most popular models are Markov chain models which can fit the frequencies in mono-, di-, tri-nucleotides, etc. Indeed, the Markov chain model of order m (denoted by Mm) takes the $(m + 1)$-mer composition into account. Under such models, the occurrence probability μ_i of a h-letter motif $\mathbf{w} = w_1 w_2 \ldots w_h$ on the {a, c, g, t} alphabet can be expressed in terms of counts of its subwords of length m and $m + 1$ [5]. For instance, here are the expression of μ_i in models M0, M1 and M(h - 2) which fit respectively the composition in bases, in dinucleotides and in oligonucleotides of length h - 1:

$$\text{M0}: \quad \mu_i = \frac{\prod_{j=1}^{h} N_i(w_j)}{\ell_i^h},$$

$$\text{M1}: \quad \mu_i = \frac{\prod_{j=1}^{h-1} N_i(w_j w_{j+1})}{\ell_i \prod_{j=2}^{h-1} N_i(w_j)},$$

$$\text{M}(h-2): \quad \mu_i = \frac{N_i(w_1 \cdots w_{h-1}) N_i(w_2 \cdots w_h)}{(\ell_i - h + 3) N_i(w_2 \cdots w_{h-1})},$$

where $N_i(\cdot)$ denotes the count in sequence i.

If one does not want to account for the sequence composition (this case will be referred to as model M00), then μ_i simply depends on the motif, hence $\mu_1 = \mu_2 = (1/4)^h$.

The choice of the Markov chain model depends on the sequence composition one wants to fit. For instance, model M2 is often used for coding DNA sequences to take the codon bias into account. Higher the model order, better the fit, but usually the model order is bounded either by h - 2 or because the sequence is too small (the number of parameters to be estimated increases exponentially with the order).

Table 1 gives the expected counts $\ell_i \mu_i$ for the motif cagcgctg in the *E. coli* loops/backbone sequences. Since N_1 = 30 and N_2 = 113, we see that this motif is highly over-represented in both sequences under models M00, M0 and M1. However, under the richest possible model (M6), it is over-represented in sequence 1 (loops) but under-represented in sequence 2 (backbone).

Exceptionality coefficient

When the motif is not exceptional with respect to the considered model, the mean count λ_i is simply $\ell_i \mu_i$. For exceptional motifs, *i.e.* motifs with an observed count N_i far from its expectation $\ell_i \mu_i$, under a given model, the mean count λ_i should reflect this exceptionality.

Table 1: Expected count for cagcgctg in the loops (1) and in the backbone (2) of *E. coli* leading strands under different models.

Model	M00	M0	M1	M6	Count
$\ell_1 \mu_1$	11.6	9.4	13.9	24.8	$n_1 = 30$
$\ell_2 \mu_2$	59.2	66.0	106.2	126.1	$n_2 = 113$

We therefore introduce an exceptionality coefficient k_i which allows λ_i to be greater (or smaller) than the expected value:

$$\lambda_i := k_i \ell_i \mu_i.$$

In the following, parameters ℓ_i and μ_i will be supposed to be known *a priori*: they can be considered as two correction terms. The inference will only be made on k_i.

Hypothesis testing

Comparing the (potential) exceptionality of a motif in two sequences is equivalent to test the null hypothesis $\mathbf{H}_0 = \{k_1 = k_2\}$.

We emphasize that the respective values of k_1 and k_2 can be larger than one (unexpectedly frequent motif), smaller than one (unexpectedly rare motif) or close to one (motif with expected count). These values do not matter: our only concern is to know if they are significantly different or not.

Exact binomial test

We first propose an exact test based on a general property of the Poisson distribution. If N_1 and N_2 are two independent Poisson counts with respective means λ_1 and λ_2, the distribution of N_1 given their sum $N_+ := N_1 + N_2$ is binomial [12]: $N_1 \sim \mathcal{B}(N_+, \pi)$ with

$$\pi = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{(k_1/k_2)\ell_1\mu_1}{(k_1/k_2)\ell_1\mu_1 + \ell_2\mu_2}.$$

Under $\mathbf{H}_0$, we have $\pi = \pi_0$ with

$$\pi_0 = \frac{\ell_1\mu_1}{\ell_1\mu_1 + \ell_2\mu_2} \qquad (1)$$

because $k_1 = k_2$. In absence of correction (M00 model) for the sequence composition (*i.e.* $\mu_1 = \mu_2$), we have $\pi_0 = \ell_1/(\ell_1 + \ell_2)$. If furthermore the two sequences have the same length, we get $\pi_0 = 1/2$.

Moreover, the proportion π and then the expectation of N_1, increases as the ratio k_1/k_2 increases. Therefore, the *p*-value for the one-sided alternative $\mathbf{H}_1 = \{k_1 > k_2\}$ is $p_B = \Pr\{\mathcal{B}(n_+, \pi_0) \geq n_1\}$, i.e.

$$p_B = 1 - \sum_{d=0}^{n_1-1} \binom{n_+}{d} \pi_0^d (1-\pi_0)^{n_+-d}$$

where n_+ and n_1 are the observed values of N_+ and N_1.

Table 2 gives the probability π_0 and the *p*-value p_B for the motif cagcgctg in *E. coli*. At level 5%, the null hypothesis is accepted under models M00 and M6 meaning that the motif is similarly exceptional in both sequences with respect to their length and/or 7-mer composition. However, $\{k_1 = k_2\}$ is rejected at level 5% against $\{k_1 > k_2\}$ under models M0 and M1; since cagcgctg is over-represented in both sequences, it means that it is significantly more exceptionally over-represented in sequence 1 (loops) with respect to the base and/or dinucleotide compositions of both sequences.

Likelihood ratio test (LRT)

Another test statistic based on the comparison of the likelihood of the data under the $\mathbf{H}_0$ and the alternative hypothesis $\mathbf{H}_1 = \{k_1 \neq k_2\}$ can be derived. This statistic is known as the Likelihood Ratio Test (see [13], vol. IV). In our model (see the Methods Section), it is defined as

$$LRT = 2\left[N_1 \ln\left(\frac{N_1/N_+}{\pi_0}\right) + N_2 \ln\left(\frac{N_2/N_+}{1-\pi_0}\right)\right]$$

where π_0 is defined in (1). Under the null hypothesis, its asymptotic distribution is a chi-square distribution with one degree of freedom.

This test is two-sided, because, under $\mathbf{H}_1$, parameters k_1 and k_2 are estimated independently (in particular, without the constraint $k_1 > k_2$). The exact distribution of *LRT* could be calculated via permutation techniques but the computation time would be tremendeous for large counts. We will then calculate the following asymptotic *p*-value:

$$p_L = \Pr\left\{\chi^2 \geq 2\left[n_1 \ln\left(\frac{n_1/n_+}{\pi_0}\right) + n_2 \ln\left(\frac{n_2/n_+}{1-\pi_0}\right)\right]\right\},$$

where n_2 is the observed value of N_2 and $\chi^2 \sim \chi^2(1)$.

Table 2: Probability π_0 and *p*-value p_B under different models for cagcgctg in the *E. coli* loops/backbone comparison.

Model	M00	M0	M1	M6
π_0 (%)	16.3	12.4	11.6	16.4
p_B	$8.6\ 10^{-2}$	$2.7\ 10^{-3}$	$9.1\ 10^{-4}$	$9.1\ 10^{-2}$

Table 3 gives the LRT statistic and the associated *p*-value for the motif cagcgctg in *E. coli*. Remember that the *LRT* is two-sided, so p_L have to be divided by two when compared to the one-sided binomial *p*-value p_B. We see that the significances obtained with the *LRT* are different from the ones obtained with the exact binomial test, but the qualitative conclusions are the same.

Chi-square test

Another standard asymptotic test is the chi-square test where the counts N_i are compared to their expected values $\widehat{N}_i$ under $\mathbf{H}_0$ given the total count N_+:

$$X^2 = \sum_{i=1}^{2} \frac{(N_i - \widehat{N}_i)^2}{\widehat{N}_i}$$

where $\widehat{N}_1 = \pi_0 N_+$ and $\widehat{N}_2 = (1 - \pi_0)N_+$. Under the null hypothesis, X^2 has also an asymptotic chi-square distribution with one degree of freedom. It is also an intrinsically two-sided test. Further analyzes (including simulations) not presented here (see [14]) show that this test performs very similarly to the LRT in every situations. Note that the chi-square test is the same as the score test [13].

Discussion

LRT distribution

The chi-square distribution of the LRT statistic is only asymptotic, so the actual level may be different from the nominal one (typically α = 5%). To measure this difference, we have calculated this actual level for different values of π_0 and N_+. Since *LRT* is a function of N_1, the actual level can be derived from the exact distribution of N_1 given N_+ which is binomial (see Results Section).

Figure 1 compares both levels (actual and nominal). Since the counts are discrete, the actual level can never be exactly α leading to oscillations in the plot. We see that the nominal level is only reached with $N_+ \simeq 1000$ for π_0 = 0.5 and even later for π_0 = 0.95 (or π_0 = 0.05). It means that the chi-square approximation of the LRT statistics is only valid for motifs with many total occurrences.

Regarding the motif cagcgctg, because π_0 is about 15% (cf. Table 2), the picture is close to the right plot of Figure 1; In fact, with a total count of 143, the actual level is respectively 0.095%, 1.1%, 5.1% and 12.5% for a nominal level α equal to 0.1%, 1%, 5% and 10%.

LRT as a contrast measure

The LRT statistic can still be used as a contrast measure, i.e. a measure of the difference, between the two exceptionalities. For large values of N_+ it is much faster and easier to compute than the binomial *p*-value. We will see in the illustration below that the two quantities are strongly correlated.

Decidability limits for the binomial test

Because the binomial test is exact, the actual and nominal levels are equal. The significance can then always be determined. It would be maximal when $N_1 = N_+$ (*i.e.* $N_2 = 0$) and the corresponding *p*-value p_B would be equal to $\pi_0^{N_+}$. Therefore, if this minimal *p*-value is greater than the desired level α (typically 5%), no significance conclusion can be made. This happens when $\pi_0^{N_+} \geq \alpha$, *i.e.* when $N_+ \geq \ln(\alpha)/\ln(\pi_0)$.

Figure 2 gives this critical value of N_+ for various values of π_0 and α. We see, for instance, that for π_0 = 0.7 and N_+ = 10, one may get significant results at a level greater than 5% but not at a level smaller than 1%.

Power

An important property for a statistical test is its ability to detect departure from the null hypothesis. This ability is measured by the power of the test which is the probability

Table 3: *LRT* statistic and associated *p*-value p_L under different models for cagcgctg in the *E. coli* loops/backbone comparison.

Model	M00	M0	M1	M6
LRT	2.1	8.2	10.2	2.0
p_L	$1.5\ 10^{-1}$	$4.2\ 10^{-3}$	$1.4\ 10^{-3}$	$1.6\ 10^{-1}$

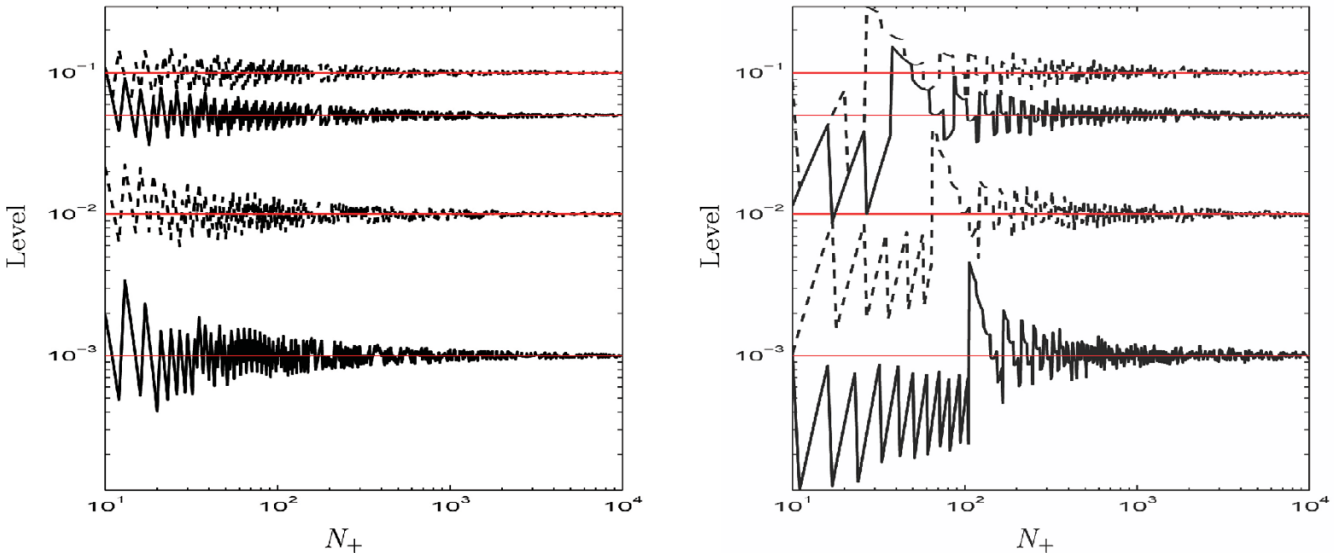

Figure 1
Actual level (log scale) of the LRT as a function of N_+ (log scale) for a nominal level α = 0.1%, 1%, 5% or 10% and probability π_0 = 0.5 (left) and 0.95 (right). (Since the LRT test is two-sided, the right plot also holds for π_0 = 0.05).

to exceed the significance threshold (defined under $\mathbf{H}_0$) when the true parameter satisfies $\mathbf{H}_1$. In our case, the parameter of interest is

$$\pi = \frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{(k_1/k_2)\ell_1\mu_1}{(k_1/k_2)\ell_1\mu_1 + \ell_2\mu_2}$$

which is equal to π_0 when $k_1 = k_2$. So the departure from $\mathbf{H}_0$ will be measured by the ratio k_1/k_2 when it differs from 1.

Exact binomial

Figure 3 presents the power of the exact binomial test when k_1/k_2 increases. As expected, the power increases with N_+. Moreover, it decreases when π_0 increases i.e. when the expected ratio $\ell_1\,\mu_1/(\ell_2\,\mu_2)$ increases. It means that, when the motif is already expected to be much more frequent in sequence 1 than in sequence 2, it is more difficult to detect that its exceptionality in the first sequence is also higher.

The motif cagcgctg occurs N_+ = 143 times in the whole genome. In the different models considered in Table 2, probability π_0 is between 11.6% and 16.4%. The power of the binomial test in this case can therefore be read in Figure 3, in the two top plots between the black and red solid lines. We see that a ratio k_1/k_2 = 2 can be detected with probability greater than 90%, while a ratio of 1.5 will be detected with a bit more than 50% probability.

LRT

The same analysis can be made for the LRT tests. However, this only makes sense for sufficiently large N_+, to guaranty the validity of the chi-square distribution.

Case of overlapping words

Compound Poisson model

The distribution of overlapping word occurrences is better modeled by a compound Poisson process (see [15]) in the following way:

• The word occurs in clumps distributed according to a Poisson process. The number of clumps C_i in sequence i is hence a random Poisson variable with mean denoted by $\tilde{\lambda}_i$.

• The size V_{ic} of the c-th clump (in sequence i) is random with geometric distribution:

$$\Pr\{V_{ic} = v\} = a_i^{v-1}\ (1 - a_i).$$

The clump sizes are supposed to be independent. Parameter a_i is the *overlapping probability* of the motif and can be calculated under various Markovian models (see [5]).

In this setting, the count N_i is hence the sum of the sizes of C_i clumps and has the Polya-Aeppli (or *geometric Pois-*

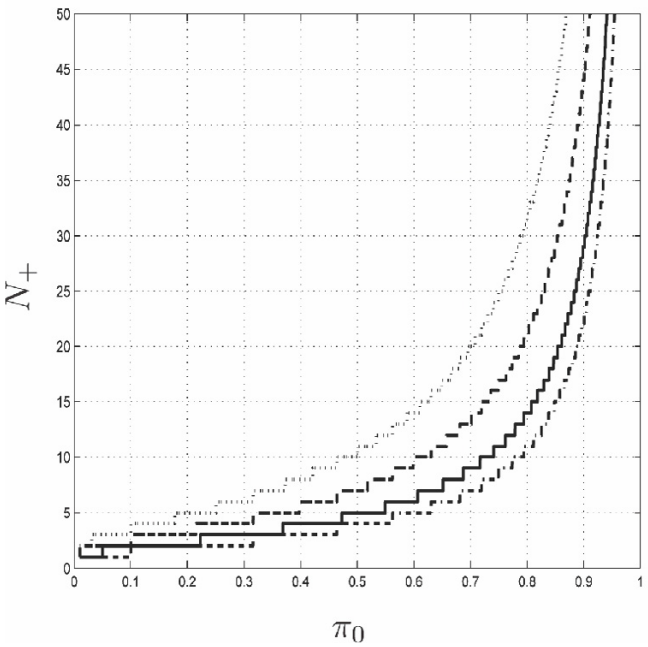

Figure 2
Minimal count N_+ to get a significant result with the binomial test as a function of the probability π_0. Curves correspond to different levels α = 0.1%, 1%, 5% or 10% (from top to bottom).

son) distribution (see [12]). We have (see [5]) $\tilde{\lambda}_i = (1 - a_i)\lambda_i$. In the case of a non-overlapping word, we have $C_i = N_i$, $a_i = 0$ and $\lambda_i = \lambda_i$. For overlapping words, the mean clump size is equal to $1/(1 - a_i)$ and increases with a_i.

Tests

An overlapping word can occur with an exceptional frequency (*i*) because of an exceptional number of clumps or (*ii*) because of exceptional sizes of clumps. Then comparing the exceptionalities of an overlapping word in two sequences leads to compare the number of clumps C_1 with C_2, and/or the sizes V_{1c}'s with V_{2c}'s.

Comparison of the number of clumps

In this compound Poisson model, the number of clumps in each sequence is Poisson distributed. The comparison of the counts C_1 and C_2 is then exactly equivalent to the comparison of the counts N_1 and N_2 studied in the Results Section, replacing λ_i by $\tilde{\lambda}_i$ and μ_i by $\tilde{\mu}_i := (1 - a_i)\mu_i$.

Exact test for the overlapping probability under M00

The question is now to test the null hypothesis $\mathbf{H}_0 = \{a_1 = a_2\}$. This comparison is made conditionally to the observed counts N_1 and N_2. It only makes sense if the motif occurs at least once in each sequence, *i.e.* if N_1, N_2, C_1 and C_2 are all larger than (or equal to) 1. In this case, the first occurrence necessarily corresponds to the first clump and the C_i - 1 last clumps have to be chosen among the other N_i - 1 motif occurrences. Since a motif occurrence (except the first one) corresponds to a clump occurrence with probability 1 - a_i, the number of clumps (except the first one) has a binomial distribution:

$$C_i - 1 \sim \mathcal{B}\ (N_i - 1, 1 - a_i) \quad (2)$$

which means that the expected number of clumps decreases when the overlapping probability increases.

Following the same strategy as for the non-overlapping case, we base our test on the distribution of C_1 given the total clump count $C_+ = C_1 + C_2$. Under $\mathbf{H}_0$, $(C_1 - 1)$ has an hyper-geometric distribution $\mathcal{H}\ (N_+ - 2, N_1 - 1, C_+ - 2)$ (see [12], Eq. (3.23)):

$$\Pr\{C_1 = c_1 \mid N_1, N_2, C_+\} = \frac{\binom{N_1 - 1}{c_1 - 1}\binom{N_2 - 1}{C_+ - c_1 - 1}}{\binom{N_+ - 2}{C_+ - 2}}$$

The overlapping probability a_1 is then significantly greater than a_2 if the probability $\Pr\{C_1 \le c_1 | N_1, N_2, C_+\}$ is smaller than a given level α.

Exact test in the general case

The previous test does not account for the composition of the sequences. The overlapping probabilities a_1 and a_2 can be expected to be different, according to some null model. In this case, the true overlapping probability in sequence i is $b_i = h_i a_i$, where h_i is an exceptionality coefficient (analogous to k_i for the mean count). The problem is then to test $\mathbf{H}_0 = \{h_1 = h_2\}$. Such a test is proposed in Appendix: it involves the generalized negative hyper-geometric distribution.

Asymptotic tests

As for the counts *N* and *C*, asymptotic tests such as likelihood ratio, chi-square or score tests can be derived to compare exceptionalities in terms of overlaps. These tests are not presented here to avoid further statistical developments but also because the small overlapping probabilities generally observed make them rarely relevant.

Illustration

Materials

Comparing complete genomes of strains of single bacterial species allows to determine highly conserved regions (so-called *backbone*) and numerous strain-specific DNA

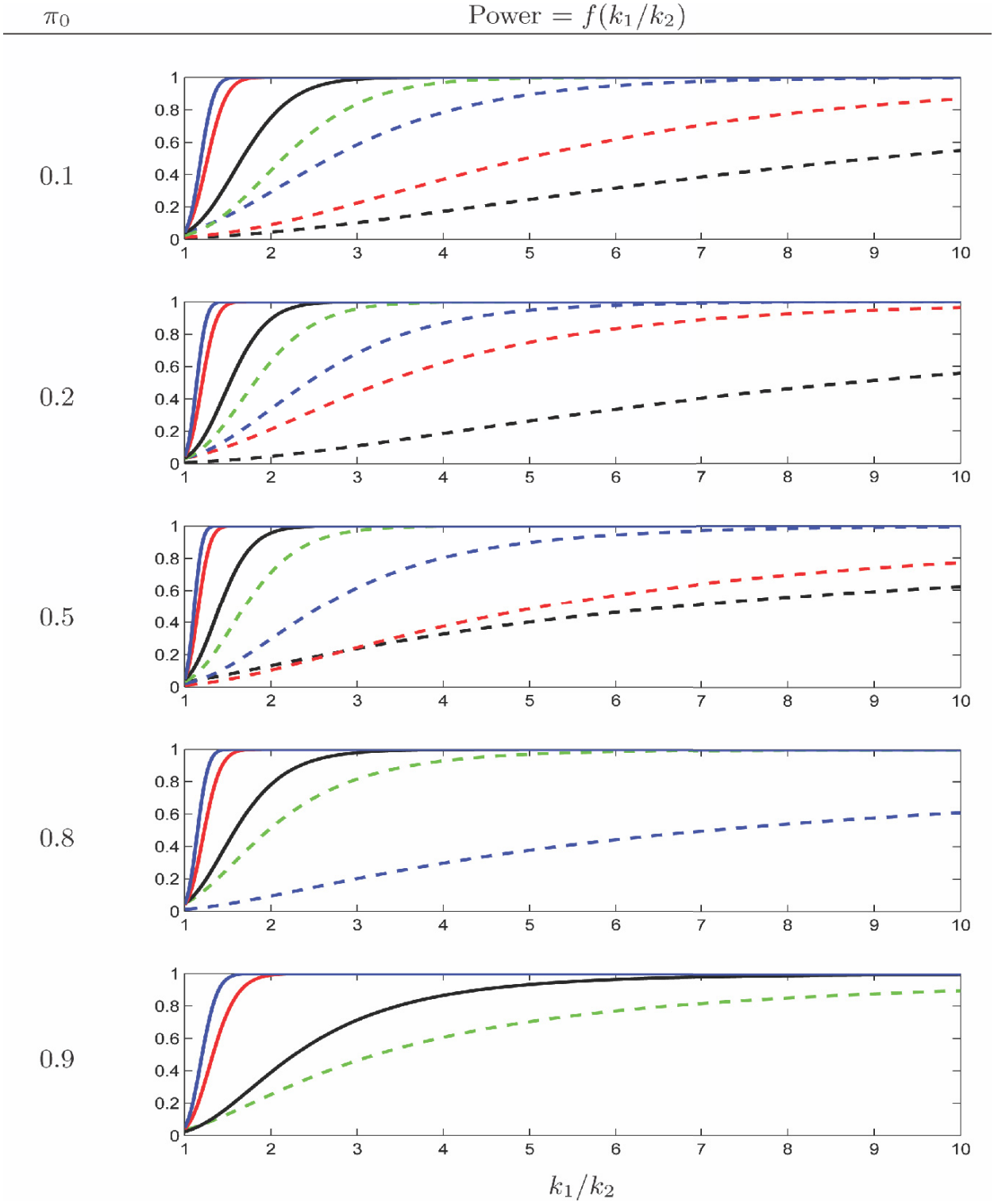

Figure 3
Power of the exact binomial test with level α = 5% as a function of k_1/k_2 (*x*-axis) for different values of π_0. Curves correspond to different values of the total count N_+ = 5 (dashed black), 10 (dashed red), 20 (dashed blue), 50 (dashed green), 100 (solid black), 500 (solid red) and 1000 (solid blue). Missing curves correspond to the values of N_+ for which no significant results at level α = 5% can be obtained (cf. the Discussion Section).

segments (so-called *loops*) for each strain. These mosaic structures help to understand the evolution of bacterial genomes. Indeed, the backbone probably corresponds to the common ancestral strain and is under vertical pressure whereas the loops may be associated with mobile elements or strain-specific pathogenicity. Such backbone/loops segmentation has been systematically performed [9] and store in the public MOSAIC database [16]. We have extracted from this database the *E. coli* K-12 specific loops (sequence 1) and the backbone (sequence 2) obtained from the pairwise alignment of the complete genomes of *E. coli* K-12 laboratory strain and the enterohemorrhagic *E. coli* O157:H7 strain. As an illustration, we have compared the exceptionalities of all the 65536 octamers in the backbone versus in the loops. Such comparison will point out octamers which do not have the same constraint, with respect to their frequency, on the loops versus on the backbone.

Exact binomial test

Figure 4 presents the significance of the binomial test for all octamers in the backbone/loops comparison. The limits between the different significance levels are clear under M00 because the probability π_0 is the same for all octamers, while they are fuzzy under M1 because π_0 depends on the octamer composition. In this case, same counts (N_1, N_2) may result in different p_B values. The distribution of the *p*-value p_B is summarized in Table 4. The 10 motifs with smallest *p*-values, i.e. with an exceptionality coefficient significantly higher in the loops than in the backbone, are listed in the top of Table 5. Multiple testing problems arise when we compare the exceptionalities of the 65 536 octamers simultaneously. Table 6 gives the number of significant octamers and the corresponding threshold when adjusting for a False Discovery Rate (FDR, [17]) of 1%. For example, under model M1 only 154 octamers are significantly more exceptional in the loops. These octamers have all a *p*-value p_B smaller than 2.2 10^{-5}.

Symmetrically, to find the motifs with an exceptionality coefficient significantly higher in the backbone than in the loops, we have to test H0 versus $\mathbf{H}_1' = \{k2 > k1\}$ using the p-value p_B' defined as $p_B' = \Pr\{\mathcal{B}(n+, \pi 0) \leq n1\}$. The 10 most significant motifs for this test are given at the bottom of Table 5. When adjusting for a False Discovery Rate of 1%, only 14 octamers under model M1 are significantly more exceptional in the backbone than in the loops. These octamers have all a p-value p_B' smaller than 1.8 10-6. Note that under model M6, no octamer is significant after multiple testing adjustment.

According to the p_B list, the motif cagcgctg has rank 1 115 among 65 536 under the M1 model. Note that the well known Chi motif (gctggtgg) which is the most overrepresented octamer in *E. coli* genome has a p_B' value of 5.1 10^{-5} (rank 1 281) under the same model; It means that k_{backbone} is significantly higher than k_{loops} but due to multiple testing Chi is not among the significant octamers.

LRT versus binomial

We now compare the results provided by the two tests: binomial and LRT. Because the former is one-sided and the latter is two-sided, we use a signed version LRT^s of the LRT statistic which is equal to *LRT* when N_1 is greater than expected ($N_1 \geq \pi_0 N_+$) and to – *LRT* otherwise ($N_1 < \pi_0 N_+$). To make the graph more readable, we also transform the *p*-value p_B into a Gaussian score $S_B \in \mathbb{R}$:

$$S_B = \Phi^{-1}(1 - p_B)$$

where Φ is the cumulative distribution function of the standard Gaussian distribution. High positive values of S_B correspond to motifs with an exceptionality coefficient in sequence 1 significantly higher than in sequence 2, while high negative values of S_B correspond to motifs having an exceptionality coefficient in sequence 1 significantly lower than in sequence 2.

We see in Figure 5 that the two statistics give very similar results for all the octamers in the backbone/loops comparison. Table 7 gives the Spearman and Kendall correlation coefficients between the two statistics for different models. Recall that Spearman's coefficient is the correlation between the ranks, while Kendall's one is the proportion of concordant pairs between the two rankings. This confirms that the LRT statistics is a reliable exceptionality comparison score, although the associated *p*-value is questionable for small counts.

Note that the naive comparison between the two *p*-values simply associated with the exceptionality of each motif in each sequence does not provide the same sets of significant octamers (see Figure 6). Such *p*-values have been calculated using the Poisson approximation of the number of clumps.

Test for overlaps

Very few motifs have significant differences in their clumps sizes. Table 8 presents the results for the 4 motifs having a *p*-value smaller than 10%. For all of them, no overlap is observed in the backbone ($C_2 = N_2$ means that all clumps are of size 1 while few are observed in the loops ($C_1 < N_1$). The probability *a* is the overlapping probability under model M00.

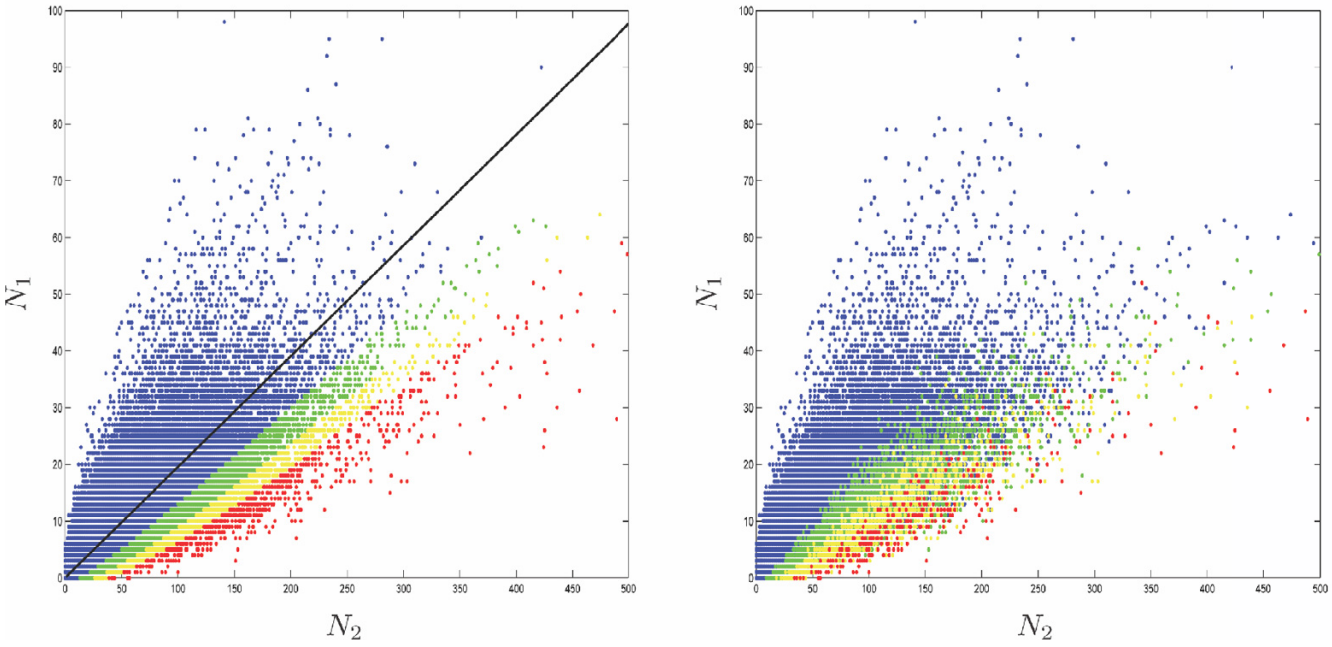

Figure 4
Counts N_2 (*x*-axis: backbone) and N_1 (*y*-axis: loops) for all the octamers under the M00 (left) and M1 (right) models. The color indicates the significance of the binomial test in the M00 model: blue = '$p_B > 0.01$', green = '$p_B < 0.01$', yellow = '$p_B < 0.001$', red = '$p_B < 0.0001$'. The solid black line on the left indicates the expected ratio $N_1/N_2 = \pi_0/(1 - \pi_0)$.

Conclusion

We have proposed two complementary statistical tests to compare the exceptionalities of motif counts in two sequences. The binomial test is exact and particularly of interest for small counts (from a computational point of view). For large counts, we advise to use the likelihood ratio test which is asymptotic but strongly correlated with the exact binomial test. The LRT statistics is simple to calculate and can be directly interpreted as a contrast measure between the exceptionalities; its *p*-value can be derived from the chi-square distribution. Both tests will be implemented in the R'MES software already devoted to exceptional motifs [18].

The likelihood ratio test can be generalized to more than two sequences. Suppose we want to compare I sequences $S_1, S_2, \ldots, S_I$. In each of them, we assume that the count N_i has a Poisson distribution with parameter $\lambda_i = k_i \ell_i \mu_i$ and we want to test $\mathbf{H}_0 = \{k_1 = k_2 = \cup = k_I\}$ versus $\mathbf{H}_1 = \{$At least one k_i differs from the others$\}$. The *LRT* statistics is

$$LRT = 2\sum_i N_i \ln\left(\frac{N_i}{N_+}\frac{\sum_j \mu_j \ell_j}{\mu_i \ell_i}\right).$$

Under $\mathbf{H}_0$, *LRT* has an asymptotic chi-square distribution with $(I - 1)$ degrees of freedom. The Chi-square test can be generalized as well.

Under the Poisson model, both tests can be easily used for degenerated motifs or more generally for sets of motifs. Let denote by $\mathcal{W}$ a set of motifs; The count N_i (respectively the occurrence probability μ_i) will be the sum of the counts (resp. occurrence probability) of **w** for all motifs **w** from $\mathcal{W}$. However, the generalization is much more

Table 4: Number of significantly unbalanced octamers under different models and for different thresholds.

		Model:	M00	M0	M1	M6
	p_B	$< 10^{-4}$	277	126	83	37
$10^{-4} \leq$	p_B	$< 10^{-3}$	519	303	247	4
$10^{-3} \leq$	p_B	$< 10^{-2}$	1758	1330	1143	104
$10^{-2} \leq$	p_B		62982	63777	64063	65391

Table 5: Top: 10 motifs with smallest *p*-value p_B ($k_{loops} > k_{backbone}$) for model M00, M0, M1 and M6. * indicates overlapping words. Bottom: 10 motifs with smallest *p*-value p'_B ($k_{backbone} > k_{loops}$).

M00		M0		M1		M6	
cggataag	1.2 10^{-19}	cggataag	3.9 10^{-20}	cggataag	2.7 10^{-18}	gggataaa	2.4 10^{-4}
ggataagg*	8.6 10^{-16}	ccgcatcc*	2.0 10^{-16}	taaggcgt*	9.1 10^{-15}	tcgaccaa	3.0 10^{-4}
taaggcgt*	4.6 10^{-15}	ggataagg*	3.0 10^{-16}	ccgcatcc*	4.0 10^{-14}	agttttta*	4.5 10^{-4}
gataaggc	1.2 10^{-14}	tgtaggcc	1.1 10^{-15}	acgccgca*	4.0 10^{-14}	aagtgata*	5.3 10^{-4}
taataaaa	1.9 10^{-14}	tcaggcct*	2.9 10^{-15}	ataaggcg	3.2 10^{-13}	gatagcgc	8.1 10^{-4}
ataaggcg	5.6 10^{-14}	taaggcgt*	2.9 10^{-15}	gccgcatc	1.0 10^{-12}	gggtcagg*	1.5 10^{-3}
ctgataag	1.2 10^{-13}	gataaggc	4.9 10^{-15}	gataaggc	2.2 10^{-12}	agccgaga*	1.7 10^{-3}
tgtaggcc	4.0 10^{-13}	ggcctaca	1.1 10^{-14}	gttccccg*	4.0 10^{-12}	gaggttac	1.7 10^{-3}
cttatccg	5.5 10^{-13}	ccggccta	1.2 10^{-14}	cgcatccg*	4.4 10^{-12}	cagagtcc*	1.8 10^{-3}
ccttatcc*	6.0 10^{-13}	aggcctac	1.4 10^{-14}	tgtaggcc	4.7 10^{-12}	ccctggcc*	2.0 10^{-3}
ggcgctgg*	< 10^{-20}	ctggaaga	6.8 10^{-10}	ctggaaga	1.2 10^{-10}	tcggttac	4.9 10^{-4}
gcgctgga	2.5 10^{-14}	cgatgaag	2.9 10^{-9}	atctggtg	3.3 10^{-8}	ggttgatg*	5.4 10^{-4}
cggcgctg	3.0 10^{-13}	gaagtgct	7.2 10^{-9}	gaagtgct	4.6 10^{-8}	gcgcatcc	6.8 10^{-4}
tggcgctg*	5.8 10^{-12}	tgaaactg*	4.0 10^{-8}	ggcgctgg*	5.2 10^{-8}	taggccgc	8.5 10^{-4}
gcgctggt	7.2 10^{-12}	atctggtg	4.9 10^{-8}	cgatgaag	6.6 10^{-8}	aagcttcg	1.1 10^{-3}
cgctggtg	8.9 10^{-12}	gcgctgga	8.0 10^{-8}	tatctggt*	1.1 10^{-7}	cgatgaag	1.1 10^{-3}
cgcgctgg	1.0 10^{-10}	cggtaaag	1.1 10^{-7}	cggtaaag	1.4 10^{-7}	cggataaa	1.2 10^{-3}
gctggcga	1.3 10^{-10}	ggttgatg*	1.4 10^{-7}	ggttgatg*	2.0 10^{-7}	gggggac	1.4 10^{-3}
tggcgcag	1.7 10^{-10}	gtgctgga	1.6 10^{-7}	gtgctgga	2.5 10^{-7}	caggcgtt	1.6 10^{-3}
ctggaaga	3.1 10^{-10}	aattgtcg	2.1 10^{-7}	tgggcttc	5.6 10^{-7}	acgccttc	1.8 10^{-3}

involved for the compound Poisson model because of the possible overlaps between motifs from the set; In particular, the overlapping probability a_i becomes a matrix [19].

We emphasize that these tests are valid only for independent sequences. They can not be used to detect skewed oligomers because the leading strand is not independent from the lagging strand [20]. This particular question requires the development of another rigorous statistical method; this is an ongoing work.

Finally, note that the exceptionality comparison of word counts in sequences is actually equivalent to the differential analysis of SAGE expression data [21]. Indeed, in the SAGE technology, the expression level of a given gene is measured by a number of associated tags and the problem is to compare the number of tags between two conditions. In such problem, no correction has to be done except for the total number of tags and our test statistics under model M00 are adapted.

Methods

Likelihood ratio test

The model presented in the Results Section can be rephrased as two Poisson processes with respective intensity $k_i u_i$ (i = 1,2). To calculate the likelihood, we need to estimate the exceptionality coefficients k_1 and k_2. Under the alternative hypothesis, their respective maximum likelihood estimates (MLE) are $\hat{k}_1 = N_1/(\ell_1 \mu_1)$ and $\hat{k}_2 = N_2/$

Table 6: Top: numbers of octamers significantly more exceptional in the loops when adjusting for a False Discovery Rate of 1% and associated thresholds for the p-value p_B for different models. Bottom: idem for octamers significantly more exceptional in the backbone.

Model	M00	M0	M1	M6
Nb. of significant octamers	677	257	154	0
Threshold for p_B	1.0 10^{-4}	3.9 10^{-5}	2.2 10^{-5}	_
Nb. of significant octamers	159	23	14	0
Threshold for p'_B	2.4 10^{-5}	3.4 10^{-6}	1.8 10^{-6}	-

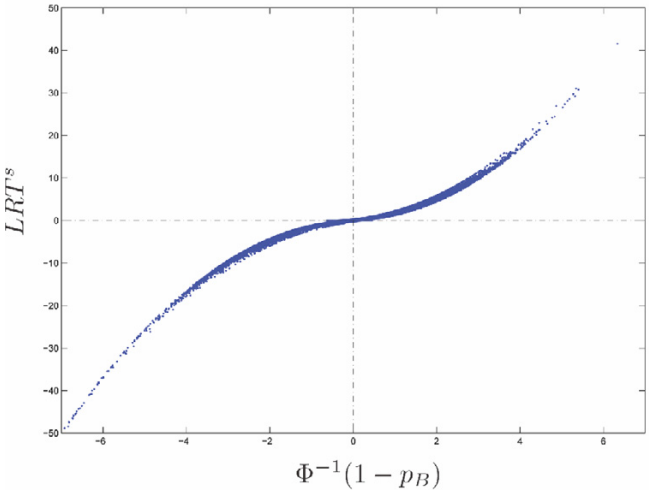

Figure 5
Signed LRT statistic LRT^s (*y*-axis) versus transformed binomial *p*-value Φ^{-1} (1 - p_B) (*x*-axis) under model M1 for all octamers in the *E. coli* backbone/loops comparison.

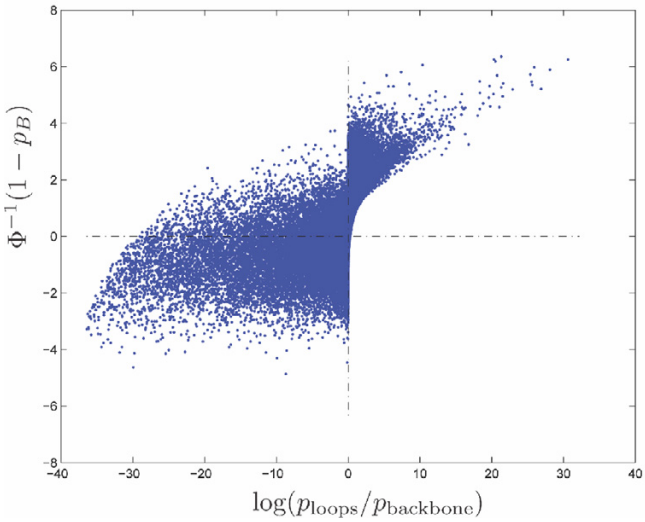

Figure 6
Transformed binomial *p*-value Φ^{-1} (1 - p_B) (*y*-axis) versus log ratio between the two *p*-values associated with the exceptionality of the motif in each sequence (*x*-axis) under model M1 for all octamers in the *E. coli* backbone/loops comparison.

Table 7: Spearman and Kendall correlation coefficients between LRT^s and S_B for different models.

Model	M00	M0	M1	M6
Spearman (%)	99.7	99.7	99.7	99.3
Kendall (%)	96.0	95.6	95.5	93.3

Table 8: Octamers with significant differences in terms of overlaps in the backbone/loops comparison.

Motif	Loops		backbone		*p*	a
	C_1	N_1	C_2	N_2	(%)	(%)
accactac	7	9	44	44	2.20	0.02
tattatta	38	41	69	69	4.83	1.56
tcggggtc	2	3	24	24	8.00	0.02
cgcgccgc	27	28	246	246	9.93	0.10

($\ell_2\ \mu_2$). Assuming that the two sequences are independent, the log-likelihood of the two processes is

$$\mathcal{L}_1 = \sum_{i=1}^{2}\Big[\, N_i \ln(\hat{k}_i \mu_i) - \hat{k}_i \mu_i \ell_i \,\Big] = \sum_{i=1}^{2}[N_i \ln(N_i / \ell_i) - N_i].$$

Under the null hypothesis, the common MLE of k_1 and k_2 is $\hat{k} = (N_1 + N_2)/(\ell_1 \mu_1 + \ell_2 \mu_2)$ and the log-likelihood is

$$\mathcal{L}_0 = \sum_{i=1}^{2}\Big[\, N_i \ln(\hat{k}\mu_i) - \hat{k}\mu_i \ell_i \,\Big] = \sum_{i=1}^{2}\Big[\, N_i \ln(\hat{k}\mu_i) - N_i \,\Big].$$

The LRT is defined as twice the difference between $\mathcal{L}_1$ and $\mathcal{L}_0$: $LRT = 2(\mathcal{L}_1 - \mathcal{L}_0)$. The result follows after standard algebraic manipulations.

Appendix

Exact hyper-geometric test

Conditional distribution of the number of clumps

The conditional distribution of C_i - 1 given in (2) can be modified as

$$N_i - C_i \sim \mathcal{B}\ (N_i - 1, b_i)$$

where $b_i = h_i a_i$ is the true overlapping probability. This version is preferable, since the exceptionality coefficient h_i directly appears here as a multiplicative constant. The conditional distribution of the difference N_i - C_i given the clump counts C_1 and C_2 and the total count N_+ is a generalized negative hyper-geometric distribution (see [12] p. 264 for the classical version and p. 270 for the generalization):

$$\Pr\{N_1 = n_1 \mid C_1, C_2, N_+\} = A^{-1}\frac{\binom{n_1 - 1}{C_1 - 1}\binom{N_+ - n_1 - 1}{C_2 - 1}}{\binom{N_+ - 2}{C_+ - 2}}\left(\frac{b_1}{b_2}\right)^{n_1 - C_1}$$

where A is the constant such that the sum over all n_1 between C_1 and N_+ is equal to one.

Test

Under $\mathbf{H}_0 = \{h_1 = h_2\}$, the term b_1/b_2 can be replaced by a_1/a_2. The overlapping probability b_1 is significantly greater than b_2 if N_1 is significantly large, *i.e.* if $\Pr\{N_1 \geq n_1 | C_1, C_2, N_+\}$ is small. The power of this test can also be studied: under $\mathbf{H}_0$, b_1/b_2 equals a_1/a_2, while under the alternative hypothesis, it is equal to $(h_1/h_2)\ (a_1/a_2)$. The power of the test is then a function of h_1/h_2.

Authors' contributions

SR and SS developed the statistical methodology, analyzed the examples and wrote the paper. VV studied the usage conditions. All authors read and approved the final manuscript.

Acknowledgements

We thank Meriem El Karoui and Marie-Agnès Petit for helpful discussions. We also thank the referees for their remarks. This work has been supported by the French Action Concertée Incitative IMPBio.

References

1. van Helden J, André B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281:**827-842.
2. El Karoui M, Biaudet V, Schbath S, Gruss A: **Characteristics of Chi distribution on several bacterial genomes.** *Research in Microbiology* 1999, **150:**579-587.
3. Bigot S, Saleh O, Lesterlin C, Pages C, El Karoui M, Dennis C, Grigoriev M, Allemand JF, Barre FX, Cornet F: **KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase.** *EMBO J* 2005, **24:**3770-3780.
4. Lothaire M: *Applied Combinatorics on Words, Volume 105 of Encyclopedia of Mathematics and its Applications* Cambridge University Press; 2005.
5. Robin S, Rodolphe F, Schbath S: *DNA Words and Models* Cambridge University Press; 2005. [English version of *ADN, mots et modéles*, BELIN 2003].
6. Davidsen T, Rodland E, Lagesen K, Seeberg E, Rognes T, Tonjum T: **Biased distribution of DNA uptake sequences towards genome maintenance genes.** *Nucleic Acids Research* 2004, **32:**1050-1058.
7. Touzain F, Schbath S, Debled-Rennesson I, Aigle B, Leblond , Kucherov G: **SIGffRid: Searching for transcription factor binding sites in bacterial genomes using comparative approach and biologically driven statistics.** 2006. [Preprint. Preliminary version in *JOBIM 2005 proceedings*, 417-426].
8. Valens M, Penaud S, Rossignol M, Cornet F, Boccard F: **Macrodomain organization of the *Escherichia coli* chromosome.** *EMBO J* 2004, **23:**4330-4341.
9. Chiapello H, Bourgait I, Sourivong F, Heuclin G, Jacquemard A, Petit MA, El Karoui M: **Systematic determination of the MOSAIC structure – backbone versus strain specific loops – of bacterial genomes.** *BMC Bioinformatics* 2005, **6:**171.
10. McNeil J, Smith K, Hall L, Lawrence J: **Word frequency analysis reveals enrichment of dinucleotide repeats on the human X chromosome and [GATA]n in the X escape region.** *Genome Research* 2006, **16:**477-484.
11. Schbath S: **Compound Poisson approximation of word counts in DNA sequences.** *ESAIM: Probability and Statistics* 1995, **1:**1-16.
12. Johnson NL, Kotz S, Kemp AW: *Univariate Discrete Distributions* Wiley: New-York; 1992.
13. Armitage P, Colton T, (Eds): *Encyclopedia of Biostatistics* Wiley; 1998.
14. Vandewalle V: **Etude de motifs dans les séquences d'ADN : comparaison d'exceptionnalités.** In *Master's thesis* Institut National Agronomique Paris-Grignon; 2005.
15. Robin S: **A compound Poisson model for words occurrences in DNA sequences.** *J R Statist Soc C* 2002, **51:**437-451.
16. MOSAIC: [http://genome.jouy.inra.fr/mosaic/].
17. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerfull approach to multiple testing.** *JRSS B* 1995, **57:**289-300.
18. Hoebeke M, Schbath S: **R'MES: Finding Exceptional Motifs.** *User guide, version 3* 2006 [http://genome.jouy.inra.fr/ssb/rmes/].
19. Roquain E, Schbath S: **Improved compound Poisson approximation for the number of occurrences of multiple words in a stationary Markov chain.** *Adv Appl Prob* 2007, **39:**.
20. Salzberg S, Salzberg A, Kerlavage A, Tomb JF: **Skewed Oligomers and Origins of Replication.** *Gene* 1998, **217:**57-67.
21. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Research* 1997, **7:**986-995.