Short interrupted palindromes on the extragenic DNA of Escherichia coli *K-12,* Haemophilus influenzae *and* Neisseria meningitidis

Ana T. Vasconcelos[1,4]*, Marco A. Grivet Mattoso Maia*[2] *and Darcy F. de Almeida*[3]

[1]Laboratório Nacional de Computação Científica, MCT, Petrópolis, RJ, 25651-070, Brazil, [2]Centro de Estudos em Telecomunicações, PUC/RJ, Rio de Janeiro, RJ, 22453-900, Brazil and [3]Instituto de Biofísica Carlos Chagas Filho, UFRJ, Rio de Janeiro, RJ, 21941-590, Brazil

Received on March 18, 2000; revised and accepted on June 6, 2000

Abstract

Motivation: *The importance of the various kinds of repetitive nucleotide sequences for the workings of bacterial DNA has been widely recognized. This work is concerned with the distribution of a particular group of repetitive sequences, the short-sequenced interrupted extragenic palindromes, on the genetic maps of Escherichia coli K-12, Haemophilus influenzae Rd and Neisseria meningitidis Z2491 and MC58. A tool has been developed based upon a statistical hypothesis test taking into account the markovian structure of random sequences in order to determine the non-random character of extragenic palindromes.*

Results: *Totals of 7631, 12 904, 4722 and 5477 non-random short interrupted palindromes have been found on the E.coli, H.influenzae, and N.meningitidis serogroup A and serogroup B genomes, respectively. Their distribution patterns on the respective genomes vary according to the bacterial species considered. Based on their position on the genome, palindromes could be distinguished as those which integrate longer, repetitive sequences; those which stand in isolation, and still others are associated to specific genome sites.*

Availability: *The complete list of the observed palindromes is available at the site http://www/lncc.br/∼atrv.*

Contact: *atrv@lncc.br*

Introduction

There are numerous reports dealing with the occurrence of repeat sequences in non-coding DNA regions on prokaryotic genomes (for recent reviews see Bachellier *et al.*, 1996, 1999; Rudd, 1999). Although their functions are not completely elucidated, they seem to be involved in physiological processes as diverse as the expression of both upstream and downstream genes, transcription termination and the structural organization of the bacterial nucleoid, among others (Bachellier *et al.*, 1996, 1999). Palindromes are a peculiar kind of repeat often found associated with those sequences. Besides, due to their structural organization palindromes stand out from the surrounding sequences as a highly suitable structure for recognition by specific proteins (Calladine and Drew, 1997; Pinder *et al.*, 1998).

Efforts to investigate a possible functional role for non-coding (extragenic) sequences, including palindromes, have often resorted to statistical methods. In search of over- and under-represented sequences, Schbath *et al.* (1995) used Markov chain models to assess statistical properties of motifs in DNA sequences. More recently, Gelfand and Koonin (1997) demonstrated that short palindromic sequences recognized by restriction enzymes are avoided at a statistical level in the genome of several bacteria. In a somewhat distinct and more general context, linguistic text analysis techniques have shown that non-coding sequences exhibit statistical properties similar to both natural and artificial languages, which denote their information-containing nature (Mantegna *et al.*, 1994). Using statistical analysis on lines different from those in the above-mentioned works, we confirm here the non-random character of extragenic short sequence (4–13 bp or ‘words’) interrupted (by up to 30 bp) palindromes present on the *Escherichia coli* K-12 genome (Blattner *et al.*, 1997), implying an organization compatible with the presence of biological information. Similar findings have also been observed on the genome of two bacterial pathogens, *Haemophilus influenzae* Rd (Fleischmann *et al.*, 1995) and *Neisseria meningitidis* serogroup A strain Z2491 (Parkhill *et al.*, 2000) and serogroup B strain

[4]To whom correspondence should be addressed at Laboratório Nacional de Computação Científica, Rua Getúlio Vargas 333, Quitandinha, Petrópolis.

MC58 (Tettelin *et al.*, 2000). The results obtained indicate that the distribution pattern of short palindromes on the bacterial genomes is species-dependent.

Systems and methods

Problem statement

In the quest for evidence of under- or over-represented palindromic structures in the DNA, Vasconcelos (1995) and Vasconcelos *et al.* (1996) have produced a series of computer programs for searching sequences of general structure w-g-w' (w is a sequence and w' is its inverted complementary repeat; g is a gap sequence) limited to the extragenic region, protein- and stable RNAs-coding regions being excluded.

A question that naturally arises is whether the observed frequencies of these structures depart from those obtained when DNA sequences are randomly generated. The extent of such departure may be evidence of the presence of biological information. The present work addresses this question by means of a hypothesis test, taking advantage of the fact that truly random sequences correspond to independent and equally probable symbols. Moreover, a simple regression model is generated that allows to extrapolate the behavior of random sequences for cases where w length is greater than 10 bp.

Problem description

Let $S_L = \{s_0, s_1, \ldots, s_{L-1}\}$ be a sequence of L symbols generated by some probabilistic scheme over an alphabet $F = \{f_1, f_2, f_3, f_4\} = \{A, C, T, G\}$ of size $Q = 4$.

Let $S_T(i) = \{s_i, s_{i+1}, \ldots, s_{i+T-1}\}$ $i \in \{0, 1, \ldots, L - T\}$ be the subsequence of S_L made by the T symbols starting at position i. We use a star-notation to represent the inverted complementary repeat (i.c.r. for short) of a sequence or an isolated symbol. Therefore the sequence $S^*_T(i)$ is the i.c.r. of $S_T(i)$ while s^*_n stands for the i.c.r. of the symbol s_n. A sequence S is called a palindrome when $S = S^*$.

In general terms, when the previously mentioned structure w-g-w' is found in the sequence S_L, and w is located at position i, the length of the gap sequence g is referred as the i.c.r. distance $d_i(T)$.

The graphic of these distances occurrences can be formally introduced by means of the sequence $\{d_i(T)\}$ produced by the following algorithm:

```
• For each $i$ ranging from 0 to $K = L - 2.T$ do:
    • Search for the smallest $j$ in the range $\{i + T, \ldots, L - T\}$ such that:
        1. $S_T(j) = S^*_T(i)$
        2. if $i > 0$ and $j < L - T$ then $s_{i-1} \neq s^*_{j+T}$ (constraint C1)
        3. if $j \geqslant i + T + 2$ then $s_{i+T} \neq s^*_{j-1}$ (constraint C2)
    • If such $j$ does not exist, make $d_i(T) = -1$. Otherwise:
    • Search for the smallest $k$ in the range $\{i + 1, \ldots, j - T\}$ such that $S_T(k) = S_T(i)$
          If such $k$ exists,   make $d_i(T) = -1$
          Otherwise             make $d_i(T) = j - i - T$
• Next $i$.
```

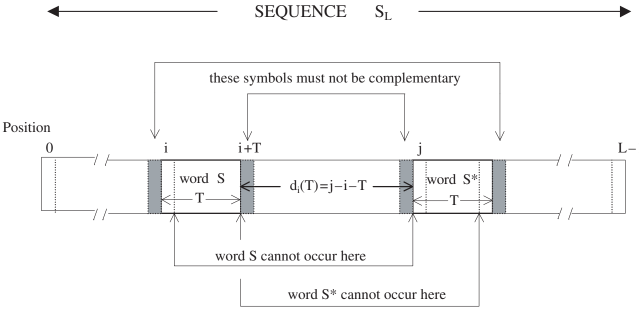

Fig. 1. Schematic representation of the process for the detection of a perfect short-sequence interrupted palindrome embedded in a sequence (S_L), using the algorithm described in this work.

The variables $d_i(T)$ can assume values in the set $\Omega_i = \{-1, 0, 1, \ldots, K - i\}$. When this distance is out of range, or there are nested subsequences, this fact is indicated by forcing this variable to assume the value -1. Constraints C1 and C2 are included in order to avoid the i.c.r. found to be part of a longer one.

The algorithm just described seeks in the sequence S_L for symbol arrangements as illustrated in Figure 1.

We also denote by $\{N_j(T) j \in \{-1, 0, 1, \ldots, K\}\}$ the set of random variables (r.v. for short) that represent the number of $d_i(T)$ elements with value j. Formally speaking, the graphic of i.c.r. distances occurrences is one realization of such random variables.

From a formal point of view, we wish to use the observed sequence $\{N_j(T)\}$ as the decision statistics associated to the following binary hypothesis test:

$$\begin{aligned} H_0 :& S_L \text{ is made of statistically independent} \\ & \text{and equiprobable symbols.} \\ H_1 :& S_L \text{ is not made of statistically independent} \\ & \text{and equiprobable symbols.} \end{aligned}$$

If we are able to assess the expected value m_j and the standard-deviation σ_j of the r.v. $N_j(T)$ under the H_0-assumption, then the knowledge of the observed values of these r.v.′s for real data will allow to accept or reject H_0 for this specific value of j.

In order to make this matter absolutely clear, the rejection of H_0 for a particular value of j, may be an indication of the presence of some biological information and its acceptance by no means should be interpreted as lack of this information. In this way, H_0 rejection should be seen as an indicator to the researcher of what i.c.r. distances are more likely to convey biological meaning.

The expected value of the r.v. $N_j(T)$ under H_0, for j in the range $\{0, 1, \ldots, K\}$, is given by:

$$\begin{aligned} &E\{N_j(T)\} \\ &\quad = \left\{2 + (K - j - 1) \cdot \left(1 - \frac{1}{Q}\right)\right\} \cdot P\{d_0(T) = j\} \\ &\quad = \vartheta_K(j) \cdot P\{d_0(T) = j\}. \end{aligned}$$

Evaluation of $P\{d_0(T) = j\}$ under H_0 assumption

For the sake of clarity, let us introduce some useful notations and definitions:

(a) Let X be the set of all sequences of size T over the alphabet $\{A, C, G, T\}$ whose generic element is denoted by S.

(b) the events $A_i(S) = \{S_T(i) = S\}$ where $S \in X$.

(c) the event $R_j = \{s_T \neq s^*_{T-1+j}\}$.

We can formally state that the concerned probabilities can be expressed as:

$$P\{d_0(T) = j\} = \left(\frac{1}{Q}\right)^T \sum_{S \in X} P_j(S)$$

where $P_j(S)$ is the probability of $\{d_0(T) = j\}$ conditionally to $\{s_T(0) = S\}$.

It is easy to show that, for the cases $j = 0$ and $j = 1$, this probability does not depend on S and is given by:

$$P_j(S) = \left(\frac{1}{Q}\right)^T \cdot \left[1 - 2 \cdot \left(\frac{1}{Q}\right)^T\right]^j \qquad j \in \{0, 1\}.$$

For $j \geqslant 2$, and representing $\overline{A}$ as the complementary event of A, then:

$$P_j(S) = P\left\{\bigcap_{i=1}^{j} \overline{A}_i(S);\ \bigcap_{i=T}^{T+j-i} \overline{A}_i(S^*);\ A_{T+j}(S^*);\ R_j\right\} \qquad j \geqslant 2.$$

No doubt that the assessment of these probabilities is a formidable task. Instead, we propose to evaluate them by means of a finite state machine and the associated homogeneous Markov chain having S_L as the input sequence. The purpose of this machine is to detect the first occurrence of the word S^* without previously detecting the word S, except at the very beginning of the input sequence and assuring the occurrence of constraint C2.

We can define the associated Markov chain as $\{W_n : n \in \{0, 1, \ldots\}\}$ representing the *state* of the machine after receiving symbol s_n as input; for the sake of simplicity we assume that this happens ‘at instant n’.

The machine state is here defined as a triple (x, y, z) where:

(i) x and y respectively represents, at instant n, the maximum number of trailing symbols of the input subsequence $\{s_0, s_1, \ldots, s_n\}$ which agrees with the same size prefix of the word S^* and word S. Therefore x and y belong to the set $\{0, 1, \ldots, T\}$.

(ii) z represents, at instant n, the input symbol located at position n- $\max(x, y)$ if such value is non-negative. In this case, when word S^* is detected, z is the input symbol that precedes it. If n- $\max(x, y)$ is negative (which corresponds to the case where S_L begins with word S^*), we indicate this fact by forcing z to be the symbol Z. Hence z belongs to the set $\{Z, A, C, T, G\}$.

Due to the fact that all event probabilities we are willing to assess are conditioned to the event $\{s_T(0) = S\}$, the state W_{T-1} must be of the form (x, T, Z). Therefore all subsequent states have a z-value different of Z.

The requirement of first occurrence of word S^* can be easily met by forcing the chain to jump from states of the form $(T, y, z \in F)$ to an special absorbing state denoted by $(T + 1, T + 1, Z)$ irrespective of the input symbol.

For the cases where S is not a palindrome, it is desired to detect word S^* without previously detecting the word S. States of the form $(x, T, z \in F)$ should be made absorbent in order to prevent reaching states like $(T, y, ?)$ (where ? stands for ‘don’t care’) at later instants. Unfortunately there are two singular situations where this may prove incorrect. Nevertheless they can be circumvented by the addition of dummy states.

Table 1 presents the next-state transition table for the case where the sequence S is ACT. The probability $P_j(S)$ can be evaluated by assessing the probability the chain reaches a state of the form $(3, 0, z)$ from state $(0, 3, Z)$ after j steps. Notice that all transitions occur with probability $1/Q$.

It is convenient at this point to discuss the number of states of this chain. It is possible to show that the associated Markov chain is characterized by a state transition probability matrix $\mathbf{P}$ with dimension M not greater than $(Q + 1) \cdot (2T + 1)$, thus exhibiting a linear growth with variable T.

Table 1. Next state transition matrix for sequence $S = ACT$

Current state	Input symbol			
	A	C	G	T
$03Z$	$11T$	$00C$	$00G$	$00T$
$00z$	$11z$	$00C$	$00G$	$00T$
$11z$	$11A$	$02z$	$20z$	$00T$
$02z$	$11C$	$00C$	$00G$	$03z$
$20z$	$11G$	$00C$	$00G$	$30z$
$03z$	$03z$	$03z$	$03z$	$03z$
$30z$	$44Z$	$44Z$	$44Z$	$44Z$
$44Z$	$44Z$	$44Z$	$44Z$	$44Z$

z belongs to the set $\{A, C, G, T\}$

Table 2. Number of total and distinct words of length T

Value of T	Number of words	Number of representatives	Reduction factor
4	256	36	0.1406
6	4.096	528	0.1289
8	65.536	8.256	0.1259
10	1048.576	131.328	0.1252

Note that the probability under scrutiny can now be rewritten as:

$$P_k(S) = \frac{1}{Q} \cdot \sum_{u=1}^{Q} \sum_{\substack{v=1 \\ fv \neq f_u^*}}^{Q} P\{W_{k+2T-1} = (T, y, f_u)/W_T = E(f_v)\}$$

where $E(f_v)$ is the state reached from (x, T, Z) when the input symbol is f_v.

The evaluation of $P_k(S)$ must be done for all Q^T possible words S. This calculation effort can be somewhat reduced if we observe that there are distinct words with the same associated matrix $\mathbf{P}$ and when this occurs, we say the words are *equivalent*. In order to establish the formal rules that define this equivalency, let the words be $S_1 = \{s_{11}, s_{12}, \ldots, s_{1T}\}$ and $S_2 = \{s_{21}, s_{22}, \ldots, s_{2T}\}$ and consider the function Φ that associates s_{1k} to s_{2k} for k ranging from 1 to T. If Φ is a 1-1 function and $(\Phi(\alpha) = \beta \Rightarrow \Phi(\alpha^*) = \beta^*)$ then words S_1 and S_2 are equivalent for the purposes of Markov chain.

This equivalence relation induces a partition in the set of all words of size T and a representative word for each member of this partition can be established. It is quite easy to observe that the above partition substantially reduces the calculation, as shown in Table 2.

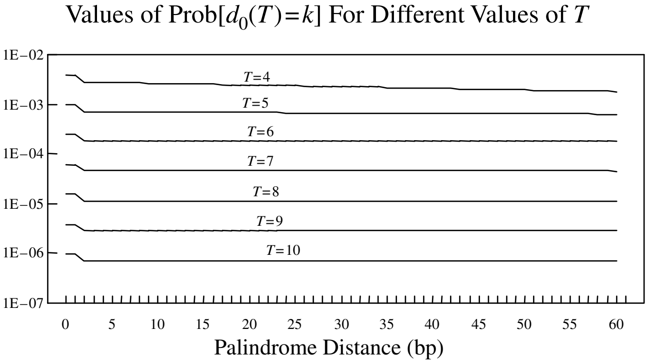

Fig. 2. Probability distribution of the distance $d_0(T)$ for T varying from 4–10.

An upper bound for the variance of $N_j(T)$ under H_0

Straightforward but cumbersome calculations, together with the fact that sequence S_L is stationary, allow to state that the variance of $N_j(T)$ can be upper-bounded by:

$$\mathrm{Var}\{N_j(T)\} \leqslant V_0 \cdot \chi_{00}(j) + V_1 \cdot \chi_{11}(j)$$

where:

$$V_0 = 1 + 2 \cdot \left(1 + \frac{1}{Q}\right) \cdot \min(k - j, 2 \cdot T + j - 1)$$
$$V_1 = (1 + 2 \cdot \rho) \cdot (k - j) - \rho \cdot (\rho - 1)$$
$$\rho = \min(K - j - 1, 2 \cdot T + j - 1)$$
$$\chi_{oo}(j) = P(d_0(T) = j) \cdot [1 - P(d_0(T) = j)]$$
$$\chi_{11}(j) = \left(1 - \frac{1}{Q}\right) \cdot P(d_0(T) = j) \times \left[1 - \left(1 - \frac{1}{Q}\right) P(d_0(T) = j)\right].$$

This result was obtained by taking into account that $N_j(T)$ is a sum of Bernouilli distributed random variables and in this case, we have $E[x \cdot y] \leqslant E[x]$.

Needless to say that this bound can be a coarse exaggeration for the above variance, but since we are particularly interested in the regions for k where assumption H_0 is rejected, the use of the upper-bound for the cases of H_0 failure makes us extremely confident that we are far from the random situation.

A regression model for $P\{d_0(T) = k\}$

The computation just described was implemented as a program in a microcomputer platform and the value of $P\{d_0(T) = k\}$ was computed for words of size T in the range 4–10 and palindrome distances (variable k) in the range 0–60. Their logarithmic plots can be seen in Figure 2.

Table 3. Parameters $a(T)$ and $b(T)$ for the model $P\{d_0(T) = k\} = a(T) \cdot \exp\{-b(T) \cdot k\}$

T	$a(T)$	$b(T)$	R square
4	2.844 788E−03	7.569 6519E−03	9.999 979E−01
5	7.186 220E−04	1.918 8781E−03	9.999 991E−01
6	1.815 364E−04	4.829 2903E−04	9.999 968E−01
7	4.554 371E−05	1.213 4177E−04	9.999 910E−01
8	1.141 839E−05	3.041 2735E−05	9.999 977E−01
9	2.857 318E−06	7.616 9201E−06	9.999 986E−01
10	7.148 496E−07	1.905 6144E−06	9.999 973E−01

Table 4. Parameters U_x and V_x for the model $a(T) = U_a \cdot V_a^T \quad b(T) = U_b \cdot V_b^T$

Parameter	U	V	R squared
a	7.225 9014E−01	2.510 1301E−01	0.999 998
b	1.951 5888E+00	2.505 9080E−01	0.999 993

Since no closed form was obtained for these probabilities and due to the fact that an exponential behavior can be observed on most of the equations, one may wonder if an exponential model can be fitted into these probabilities. A regression analysis was made for the previously mentioned values of T and k and the results are summarized in Table 3.

The extreme closeness to one of the R squared coefficients does reveal that an exponential model is a fair choice. Nevertheless we still have the job of investigating the relationship of the parameters a and b in connection with T. Again, exponential models seem quite adequate and Table 4 summarizes the results concerning the regression analysis for them.

These results suggest as a general model for the referred probability the following expression:

$$\text{Prob}\{d_0(T) = k\} = U_a \cdot V_a^T \cdot \exp\{-U_b \cdot V_b^T \cdot k\}.$$

Results and discussion

Some general considerations on the observed results which apply (unless otherwise stated) to the four examined genomes are aligned below.

As expected, the total number of palindromes decreases with the increase in the sequence length (Figures 3, 5 and 6). Palindromes with $w = 3$ bp are not distinguishable from random distribution. The search had been initially planned to cover g values in the range 0–60 bp; however, the results have shown that frequencies for $g > 20$ bp, with a few exceptions, could not be distinguished from random occurrences; therefore, the cut-off score has been set at 30 bp. Perfect palindromes with $w \geqslant 13$ bp are so rare as to be considered isolate cases (Figure 3, panel H, and results not shown). Taken together, these results suggest that rather than perfect palindromes, the occurrence of some degree of mismatching between w and w' seems to be the rule, particularly for larger values of w.

The maximal frequencies of occurrence in each of the examined palindrome length categories were found at low g values (Figures 3, 5 and 6). Exceptions were noted for both *N.meningitidis* genomes, where frequency peaks at g values of 18 bp and 23 bp ($w = 6$ bp), and 27 bp ($w = 9$ and 10 bp) have been identified (Figure 6; see details below), and in the case of *E.coli*, notably at g values of 11, 16–19 bp (all with $w = 7$ bp; Figure 3, panel D and discussion below). Some short perfect palindromes may eventually be part of a larger (usually repetitive) structure or an operon. Under this condition, palindromes may serve to localize specific regions on the genome map.

It is quite common to find a palindrome embedded in another palindrome, but it is not known whether this arrangement serves any functional purpose. Besides, since our searches were exclusively directed to the finding of perfect palindromes, there follows that frequencies reported in this paper are the minimal figures attributable to the various short palindrome sizes found on the bacterial genomes. Obviously a perfect palindrome and a sequence differing from it by a single or just a few mismatches usually belong to the same homology group.

E.coli K-12. The frequencies of occurrence of variously sized (13 bp $\geqslant w \geqslant$ 4 bp) interrupted (30 bp $\geqslant g \geqslant$ 0 bp) palindromes in the extragenic DNA of the *E.coli* K-12 genome are shown in Figure 3. As described above, the non-random character is based on the numerical occurrence of each length group. A total of 7631 palindromes were found. Their complete listing and localization on the *E.coli* genome are available for download from http://www.lncc.br/∼atrv. This total includes palindromes which are part of repetitive sequences of various types, as described by Bachellier *et al.* (1996, 1999) and by Rudd (1999), whose findings are here confirmed.

We have noticed that perfect palindromes are distributed all round the *E.coli* K-12 genome, but in a few regions their concentrations are particularly high, as is the case around 6, 35, 44, and 93 min on the genetic map (Figure 4). Two examples will serve to illustrate this point. First, in the 44 min region, downstream of gene *asnW*, there are 10 short perfect palindromes of the types 6-6-6; 5-5-5 (2); 4-4-4 (3); 5-4-5; 7-17-7; 7-11-7, and 10-10-10, within a sequence 453 nt long (from nt 2 055 596 to 2 056 049 on the *E.coli* K-12 genetic map). Second, between ORFs f338 and o139 there are 23 perfect palindromes over a 1287 nt long region (from nt 2 065 343 to 2 066 630): 4-4-4 (6); 5-4-5 (5); 4-7-4 (5); 4-5-4 (3); 5-5-5 (2); 6-4-6, and 6-10-6. On the other

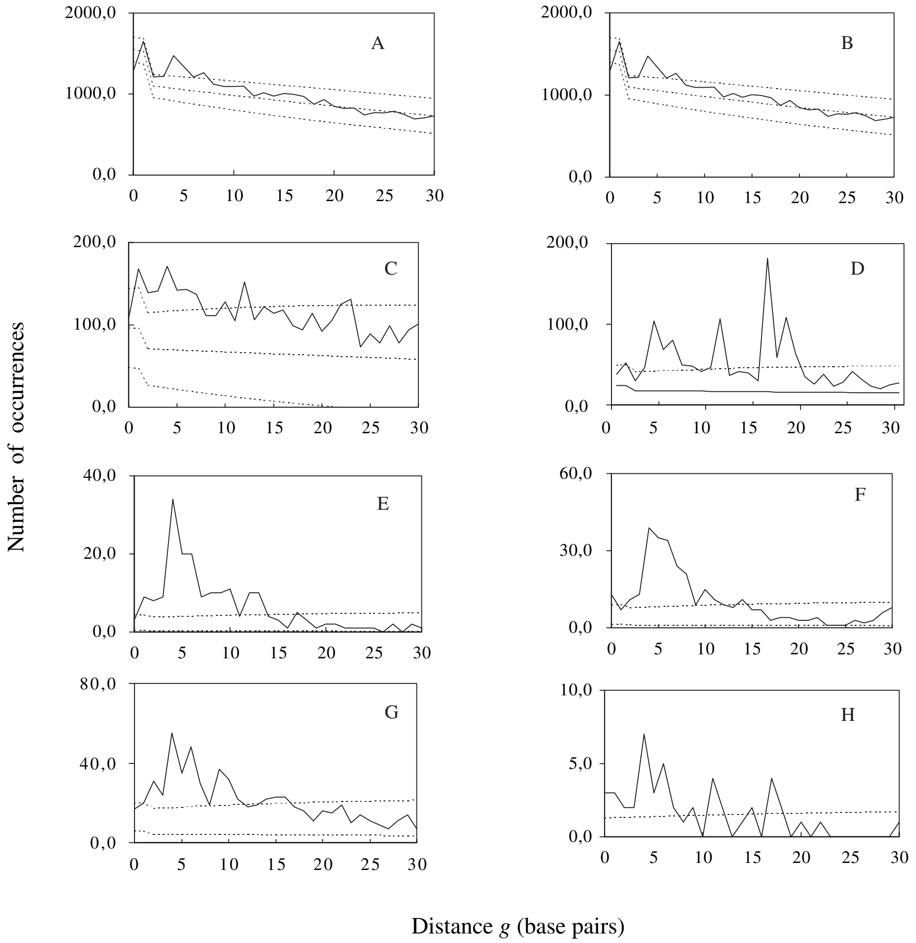

Fig. 3. Comparative distribution of randomic (calculated) and observed palindromes in the *E.coli* K-12 DNA extragenic regions. - - - - - -, randomic (average ± standard deviation); ——, observed (from data base). w size (in bp): from A–G, 4–10, respectively; H, 13. g is the gap size between w and its inverted complementary sequence w'.

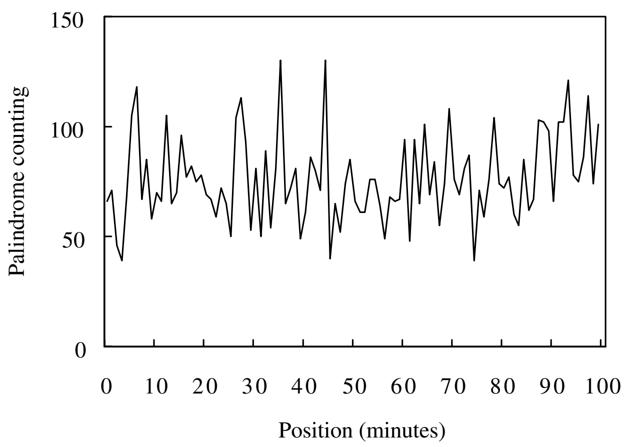

Fig. 4. Distribution of nonrandom occurrences of all extragenic short sequence (4–10 bp) interrupted perfect palindromes on the genetic map of the *E.coli* K-12 chromosome.

hand, they are relatively scarce around 50–60 min on the genetic map (Figure 4). Preliminary analysis of our results indicates that some particular palindromic sequences are preferentially associated with certain g lengths; a more detailed description of such findings is forthcoming (Vasconcelos and de Almeida, to be published).

The detection of short perfect palindromes previously known to be present on the *E.coli* K-12 genome was an essential preliminary requisite to validate the presently proposed tool. That this was the case has been demonstrated by focusing on palindrome-containing repetitive sequences such as the palindrome units (PU; Gilson *et al.*, 1987) or the bacterial interspersed mosaic elements (BIME; Bachellier *et al.*, 1997). All such previously described structures (Bachellier *et al.*, 1999) have been recognized by our method. Thus, on Figure 3, panel D, peaks corresponding to $g = 11$, 16–19 bp are essentially BIMEs constituted by the Z^1, Y, and

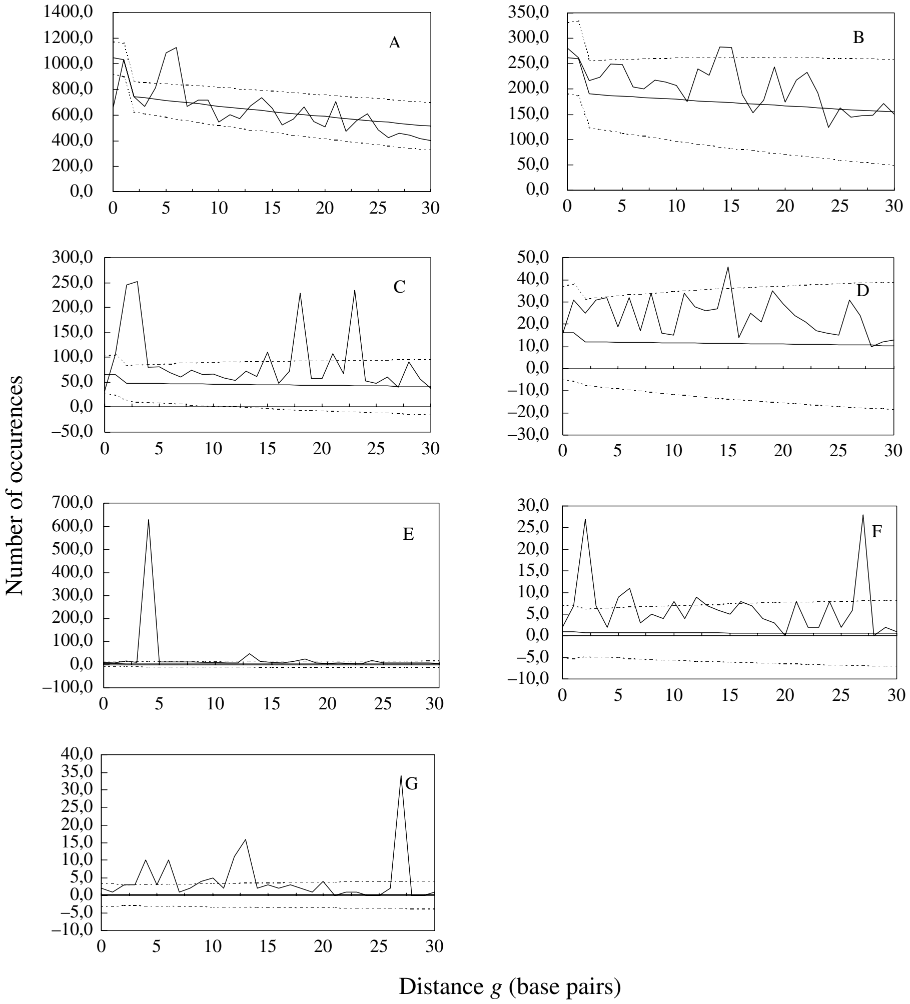

Fig. 5. Comparative distribution of randomic (calculated) and observed palindromes in the *N.meningitidis* Z2491 DNA extragenic regions. - - - - - -, randomic (average ± standard deviation); ——, observed (from data base). w size (in bp): from A–G, 4–10, respectively. g is the gap size between w and its inverted complementary sequence w'.

Z^2 categories of PUs, as described by Bachellier *et al.* (1996, 1999). It is worth noting that interrupted palindromes with $g = 12$, 13, 14, or 15 bp (that is, intermediate between the significant 11 and 16 bp distances mentioned above) could not be differentiated from DNA sequences randomly generated (Figure 3, panel D), a result that adds up to the reliability of the method.

H.influenzae Rd and N.meningitidis strains Z2491 and MC58.

To extend the results obtained with *E.coli* K-12, and to determine whether specific types of repeats found in other bacterial species would be identified using the approach here described, the complete genomes of three pathogenic bacteria (*H.influenzae* and *N.meningitidis* serogroup A strain Z2491 and serogroup B strain MC58) have been submitted to a similar study. Methods were analogous to those described above for *E.coli*, except that for each strain the respective individual parameters have been used. The frequencies of occurrence of short palindromes on those genomes as a function of their w size are shown in Figures 5 and 6. The total figures are 12 904 and 4722, for *H.influenzae* and *N.meningitidis* Z2491, respectively. Their complete listing is available at http://www.lncc.br/~atrv. The finding that the distribution pattern of short palindromes on the genomes of *N.meningitidis* serogroups A and B strains are quite similar was not surprising, considering that they are similar by more than 90% (Tettelin *et al.*, 2000); the present search has been centered on the serogroup A strain, but for our purposes serogroups A and B strains have been considered as a single entity.

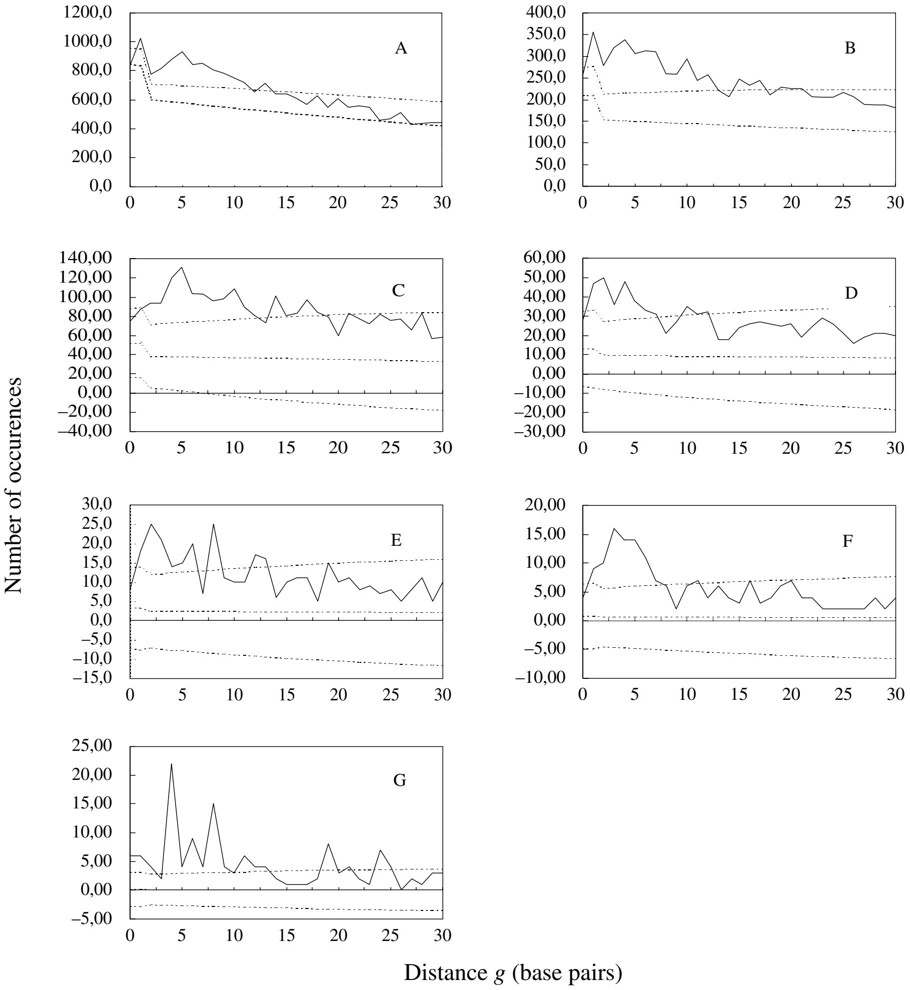

Fig. 6. Comparative distribution of randomic (calculated) and observed palindromes in the *H.influenzae* DNA extragenic regions. ------, randomic (average ± standard deviation); ——, observed (from data base). w size (in bp): from A–G, 4–10, respectively. g is the gap size between w and its inverted complementary sequence w'.

The palindromes on the *H.influenzae* genome have practically always a g size below 10 bp, an exception being a peak at $g = 24$ bp for $w = 10$ bp. The distribution is quite distinct for the palindromes on the *N.meningitidis* genomes, which are mostly concentrated at g sizes $\leqslant 5$ bp; the conspicuous peaks at higher g sizes are accounted for by palindromes found within repetitive sequences.

A few instances are worth reporting, such as that of the Correia elements which, on the *N.meningitidis* genomes, differ somewhat from the original sequence reported on the *N.gonorrhoeae* genome (Correia *et al.*, 1988). We found that the sequence between the inverted complementary repeats contains several short, generally overlapping palindromes, in the succession 4-6-4 (w = GGCG); 6-18-6 (w = CCTTAG); 6-23-6 (w = CGATTC), and 4-5-4 (w = GTAC), plus a 6-3-6 type (w = AGAGAA) embedded in the 6-18-6 and a 6-2-6 type (w = GGTGCT) embedded in the 6-23-6 sequence. Everyone of these palindrome types is represented by a high frequency of occurrence in Figure 6, panels A and C.

The dRS3 repeat sequence, reported as a 6-8-6 palindrome (Parkhill *et al.*, 2000), has been identified as a 8-4-8 type (Figure 6, panel E), since two additional inner nucleotides, at the ATTCCC end, are also inverted, in line with the observation of Van der Ende *et al.* (1999). The complementary GGGAAT end extends farther as an internal 10-27-20 palindrome (48 occurrences) which corresponds to the peak at Figure 6, panel 7. These examples indicate that the searching approach here developed, besides the identification of species-specific sequences, provides further insight into their molecular composition.

Several possibilities may be raised to account for the widespread distribution of the short interrupted palindromes on bacterial genomes. Whether they represent a leftover from ancestral species or still play some structural (adjuvant of DNA coiling, for example) or regulatory role (as described for other repeat sequences), is not yet clear. However, the present work reveals that the most notable feature of the examined genomes is the presence of very prominent peaks at $g = 4$ and/or 5 bp, irrespective of w size. Since they are not essentially repetitive in nature, we are inclined to consider that we are dealing here with repetitive structures rather than repetitive sequences. Taking into account that such structures stand out from the surrounding sequences as highly suitable for recognition by protein motifs such as α-helix, β-sheet, and leucine zipper (Calladine and Drew, 1997), and also that palindrome-containing repetitive structures do interact with different proteins for physiological purposes (Stern *et al.*, 1988; Engelhorn *et al.*, 1995), it is tempting to advance the suggestion that the repetitive palindromic structures may act as extragenic sites for interaction with specific proteins.

Acknowledgements

This work was supported by grants # 62.0113/97.9, from PADCT III / CNPq SBIO, #76971046.00 from PRONEX, and # E-26/170.782/96-APQ1, from FAPERJ.

References

Bachellier,S., Clement,J.M. and Hofnung,M. (1999) Short palindromic repetitive DNA elements in enterobacteria: a survey. *Res. Microbiol.*, **150**, 627–639.

Bachellier,S., Clement,J.M., Hofnung,M. and Gilson,E. (1997) Bacterial interspersed mosaic elements (BIMEs) are a major source of sequence polymorphism in *Escherichia coli* intergenic regions including specific associations with a new insertion sequence. *Genetics*, **145**, 551–562.

Bachellier,S., Gilson,E., Hofnung,M. and Hill,C.W. (1996) Repeated sequences. In Neidhardt,F.C., Curtiss III,R., Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik,B., Reznikoff,W.S., Riley,M., Schaechter,M. and Umbarger,H.E. (eds), *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* ASM, Washington, DC, pp. 2012–2040.

Blattner,F.R., Plunkett,G.,III, Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Calladine,C.R. and Drew,H.R. (1997) *Understanding DNA*. Academic Press, San Diego.

Correia,F.F., Inouye,S. and Inouye,M. (1988) A family of small repeated elements with some transposon-like properties in the genome of *Neisseria gonorrhoeae*. *J. Biol. Chem*, **263**, 12 194–12 198.

Engelhorn,M., Boccard,F., Murtin,C., Prentki,P. and Geiselmann,J. (1995) In vivo interaction of the *Escherichia coli* integration factor with its specific binding sites. *Nucleic Acids Res.*, **23**, 2959–2965.

Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J., Dougherty,B.A., Merrick,J.M., McKenney,K., Sutton,G.G., FitzHugh,W., Fields,C.A., Gocayne,J.D., Scott,J.D., Shirley,R., Liu,L.I., Glodek,A., Kelley,J.M., Weidman,J.F., Phillips,C.A., Spriggs,T., Hedblom,E., Cotton,M.D., Utterback,T., Hanna,M.C., Nguyen,D.T., Saudek,D.M., Brandon,R.C., Fine,L.D., Fritchman,J.L., Fuhrmann,J.L., Geoghagen,N.S., Gnehm,C.L., McDonald,L.A., Small,K.V., Fraser,C.M., Smith,H.O. and Venter,J.C. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.

Gelfand,M.S. and Koonin,E.V. (1997) Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.*, **25**, 2430–2439.

Gilson,E., Clément,J.-M., Perrin,D. and Hofnung,M. (1987) Palindromic units: a case of highly repetitive DNA sequences in bacteria. *Trends Genet.*, **3**, 226–230.

Mantegna,R.N., Buldyrev,S.V., Goldberger,A.L., Havlin,S., Peng,C.-K., Simons,M. and Stanley,H.E. (1994) Linguistic features of noncoding DNA sequences. *Phys. Rev. Lett.*, **73**, 3169–3172.

Parkhill,J., Achtman,M., James,K.D., Bentley,S.D., Churcher,C., Klee,S.R., Morelli,G., Basham,D., Brown,D., Chillingworth,T., Davies,R.M., Davis,P., Devlin,K., Feltwell,T., Hamlin,N., Holroyd,S., Jagels,K., Leather,S., Moule,S., Mungall,K., Quail,M.A., Rajandream,M.A., Rutherford,K.M., Simmonds,M., Skelton,J., Whitehead,S., Spratt,B.G. and Barrell,B.G. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria menigitidis* Z2491. *Nature*, **404**, 502–506.

Pinder,D.J., Blake,C.E., Lindsey,J.C. and Leach,D.R.F. (1998) Replication strand preferences for deletions associated with DNA palindromes. *Mol. Microbiol*, **28**, 719–727.

Rudd,K.E. (1999) Novel intergenic repeats of *Escherichia coli* K-12. *Res. Microbiol.*, **150**, 653–664.

Schbath,S., Prum,B. and Turckheim,E. (1995) Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. *J. Comput. Biol.*, **3**, 417–437.

Stern,M.J., Prossnitz,E. and Ferro-Luzzi Ames,G. (1988) Role of the intercistronic region in post-transcriptional control of gene expression in the histidine operon of *Salmonella typhimurium*: involvement of REP sequences. *Mol. Microbiol.*, **2**, 141–152.

Tettelin,H., Saunders,N.J., Heidelberg,J., Jeffries,A.C., Nelson,K.E., Eisen,J.A., Ketchum,K.A., Hood,D.W., Peden,J.F., Dodson,R.J., Nelson,W.C., Gwinn,M.L., DeBoy,R., Peterson,J.D., Hickey,E.K., Haft,D.H., Salzberg,S.L., White,O., Fleischmann,R.D., Dougherty,B.A., Mason,T., Ciecko,A., Parksey,D.S., Blair,E., Cittone,H., Clark,E.B., Cotton,M.D., Utterback,T.R., Khouri,H., Qin,H., Vamathevan,J., Gill,J., Scarlato,V., Masignani,V., Pizza,M., Grandi,G., Sun,L., Smith,H.O., Fraser,C.M., Moxon,E.R., Rappuoli,R. and Venter,J.C. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**, 1809–1815.

van der Ende,A., Hopman,C.T.P. and Dankert,J. (1999) Deletion of *porA* by recombination between clusters of repetitive extragenic palindromic sequences in *Neisseria meningitidis*. *Inf. Imm.*, **67**, 2928–2934.

Vasconcelos,A.T. (1995) Análise de seqüências de nucleotídeos no ADN extragênico de procariotos: estudo de palíndromos interrompidos, com aplicação ao sistema SOS (in Portuguese). *MSc Thesis*, Instituto de Biofísica Carlos Chagas Filho, Universidade Federal do Rio de Janeiro.

Vasconcelos,A.T., Machado,R.S. and Almeida,D.F. (1996) Analysis of enteric bacteria SOS operator sequences and description of potential DNA damage-inducible genes. *Brazil. J. Genet.*, **19**, 189–195.