Interdisciplinary Sciences: Computational Life Sciences (2019) 11:367–372
https://doi.org/10.1007/s12539-019-00326-x

ORIGINAL RESEARCH ARTICLE

Distribution of Distances Between Symmetric Words in the Human Genome: Analysis of Regular Peaks

Carlos A. C. Bastos[1] · Vera Afreixo[2] · João M. O. S. Rodrigues[1] · Armando J. Pinho[1] · Raquel M. Silva[3]

Received: 24 August 2018 / Revised: 24 January 2019 / Accepted: 27 February 2019 / Published online: 25 March 2019

Abstract
Finding DNA sites with high potential for the formation of hairpin/cruciform structures is an important task. Previous works studied the distances between adjacent reversed complement words (symmetric word pairs) and also for non-adjacent words. It was observed that for some words a few distances were favoured (peaks) and that in some distributions there was strong peak regularity. The present work extends previous studies, by improving the detection and characterization of peak regularities in the symmetric word pairs distance distributions of the human genome. This work also analyzes the location of the sequences that originate the observed strong peak periodicity in the distance distribution. The results obtained in this work may indicate genomic sites with potential for the formation of hairpin/cruciform structures.

Keywords Cruciform · Distance distribution · Genomic word · Reversed complements · Inverted repeats · Regular peaks

1 Introduction

Several genomic studies have focused on the analysis of word counts and word distances, namely, phylogeny studies [1], alignment-free methods [2, 3], CpG detection [4], coding regions detection [5] and other DNA structure analysis [6, 7].

In the context of DNA structure analysis, non-B conformations have been shown to play an important role in DNA damage and repair, genetic instability, gene regulation, and chromatin architecture [8]. In particular, hairpin/cruciforms structures are important regulators for biological processes and gene function [9].

Inverted repeats are a required feature of cruciform structures, but not all inverted repeats will form cruciforms. Cruciforms are dynamic structures that may occur when certain conditions are met, such as the coiling state of DNA, but are less stable than the normal B-DNA conformation. Although their properties and relevance in several biological processes are acknowledged, evidence of their genomic localization and mechanism of action are lacking in vivo [10, 11].

The stem and loop lengths of cruciform structures seem to vary over a wide range. According to different authors, the stem lengths vary between 6 and 100 nucleotides, while loop lengths may range from 0 to 2000 nucleotides [12–14]. Shorter distances could favour the occurrence of these structures, but long distances have also been reported, such as the translocation breakpoints associated with human developmental diseases or infertility [10].

Computational techniques have been used to identify DNA motifs that are known to potentially form non-B DNA structures [6, 14]. A DNA word analysis based on the distribution of the distances between adjacent symmetric words of length seven [7] showed a strong over-representation of distances up to 350, a feature that the authors considered might be associated with the potential for the occurrence of cruciform structures. Recently, the same research group extended their analysis to include distance distributions

✉ Carlos A. C. Bastos
cbastos@ua.pt

[1] Department of Electronics, Telecommunications and Informatics, IEETA-Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[2] Department of Mathematics, IEETA-Institute of Electronics and Informatics Engineering of Aveiro, CIDMA-Center for Research and Development in Mathematics and Applications, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[3] Department of Medical Sciences, iBiMED, IEETA-Institute of Electronics and Informatics Engineering of Aveiro, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

of non-adjacent inverted repeats, since adjacency is not a required condition for cruciform structures to form [15].

The present work focuses on identifying and characterizing a particular type of motif, the inverted repeat, whose distance distribution contains some atypical frequencies (peaks) at regular intervals (the occurrence of regular peaks in the distance distribution of some symmetric word pairs was first reported in [15]).

2 Methods

We want to find, in the human genome, structures beyond the already well-known repetition structures published in the literature. Thus, we used pre-masked sequences available from the UCSC Genome Browser (http://genome.ucsc.edu) webpage. These files contain the GRCh38 assembly sequences, with repeats reported by RepeatMasker [16] and Tandem Repeats Finder [17] masked with Ns.

2.1 Distance Between Symmetric Word Pairs

Consider the alphabet $\mathscr{A} = \{\mathrm{A, C, G, T}\}$ and let w be a symbolic sequence (word) defined in $\mathscr{A}^k$, where k is the length of w. The pair composed of one word, w, and the corresponding reversed complement word, w', is called a symmetric word pair. For example, (ACT, AGT) is a symmetric word pair.

For a given word length k, we compute the frequency distributions of distances between occurrences of each word and all succeeding reversed complements, $f_{w,w' \dots w'}$ up to a maximum distance (4000 in this work).

For example, consider the following sequence:

$$\underline{\mathrm{ACT}}\mathrm{GGAA}\overline{\mathrm{AGT}}\mathrm{AAGA}\overline{\mathrm{AGT}}\underline{\mathrm{ACT}}\mathrm{TTGT}\underline{\mathrm{ACT}}\mathrm{GGG}\overline{\mathrm{AGT}}\mathrm{TTGT}$$

For word $w = \mathrm{ACT}$, we have, in the previous sequence, five distances to all the succeeding reversed complement words (distances 7, 14, 30, 13, and 6).

Motivated by previous work and the stem length of possible cruciform structures and considering computational limitations , we study words of length $k = 7$. For each word w, we analyse distances up to 4000 nucleotides, but, if an N symbol is found, the search for w' is stopped. To avoid the direct dependencies associated with the nucleotide composition of some words, we exclude distances shorter than k from the analysis.

2.2 Detecting Symmetric Word Pairs with Atypical Distance Distributions

To find the symmetric word pairs with atypical (high) frequencies in the distance distribution, we developed and used a simple algorithm based on finding outliers within the distance distribution. The algorithm is the following:

- For each symmetric pair distance distribution, compute the distances that are outliers.
    - use MATLAB function `isoutlier` to find the distances whose frequencies are more than six local scaled MAD (median absolute deviation) from the local median computed over a window with length 101;
    - verify if the mass of the distances candidates to be outliers is greater than ten occurrences.
- Select as atypical the symmetric word pairs whose distance distribution contains more than ten outliers as computed above.

The application of the above algorithm to all the 4^7 words resulted in the identification of 247 distance distributions that were considered to have atypical (high) frequencies.

The visualization of the distance distributions with atypical frequencies revealed many words with regular peaks and with different periods of repetition of the peaks. Those observations lead to the development of a method to detect periodic regularities in the distance distributions (see Sect. 2.3).

Figure 1 shows, as an example, the distance distribution of $w = \mathrm{CCAGCTG}$ with regular peaks spaced by 102 positions. The distances with atypical frequencies are marked with red dots.

2.3 Detecting Periodic Regularities

Consider a frequency distribution $f(i)$ of an integer variable defined for $i = 1, 2, \dots, N$. We define a family of distributions, derived by "wrapping" f around itself, modulo n:

$$f_n(i) = \sum_{j=0}^{J(i)} f(i + jn), \tag{1}$$

for $i = 1, 2, \dots, n$. The upper bound of the summation is $J(i) = \left\lfloor \frac{N-i}{n} \right\rfloor$.

If f contains a periodic pattern of peaks at positions $i = a + jn$, with $j \in \{0, 1, \dots\}$, where n is the period and a is the position of the initial peak, then those peaks will be superimposed at the single position $i \equiv a \pmod{n}$, with $i \in \{1, 2, \dots, n\}$, on the n-wrapped distribution f_n. On the contrary, if f also contains peaks spaced with a distinct period $m \neq n$, then those peaks will be spread over several positions in f_n. Therefore, any component peaks with period n in f will be relatively amplified and stand out against the other components in the n-wrapped distribution f_n. This is the rationale for using this analysis tool.

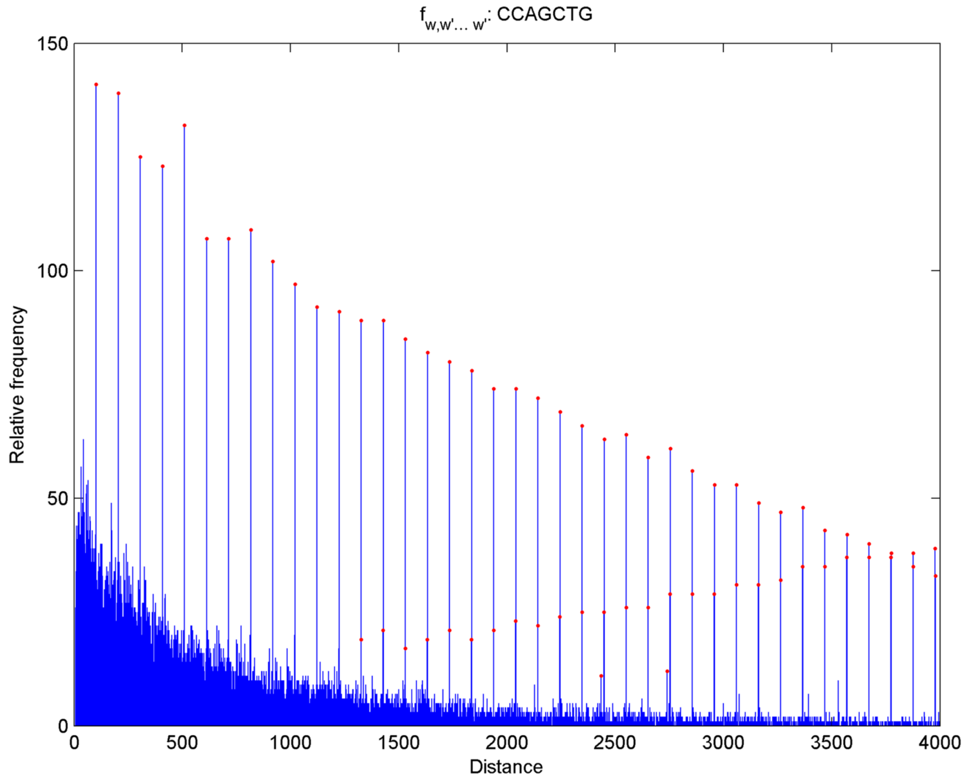

Fig. 1 The distance distribution for word CCAGCTG showing also the distances considered atypical (red dots)

Figure 2 shows an example of a distribution *f*(*i*) defined for $i \in \{1, 2, \dots, 100\}$. This distribution has a total mass of 1000 and half of this mass is concentrated in positions $i = 3, 13, 23, \dots, 93$.

Figure 3 shows four wrapped distributions obtained from the *f* distribution in Fig. 2. The ten-wrapped distribution (Fig. 3b) displays a distinct concentration of mass in $f_{10}(3)$, which allows the correct identification of the period and initial position of the pattern of peaks in *f*. The concentration of mass vanishes with just a minimal change in the wrapping period *n*, as shown in Fig. 3a, c. Some concentration of mass is expected when *n* is a multiple of the period, as demonstrated by Fig. 3d.

2.3.1 Finding the Fundamental Period

To find a periodic component in a distribution *f*, we can generate the family of *n*-wrapped distributions f_n for $n = 1, 2, \dots$ and select the one with the most concentration of mass in a few positions. We use the maximum frequency, $M_f(n) = \max f_n$,

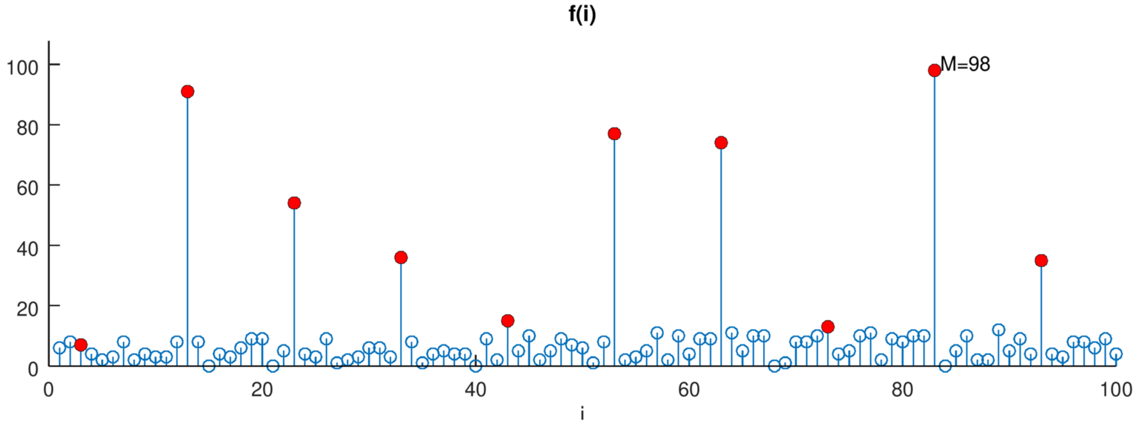

Fig. 2 An example distribution *f* with a periodic pattern of peaks. The peaks at positions $i \equiv 3 \pmod{10}$ are highlighted. These contain 50% of the total mass in the distribution. The maximum frequency *M* is shown in position $i = 83$

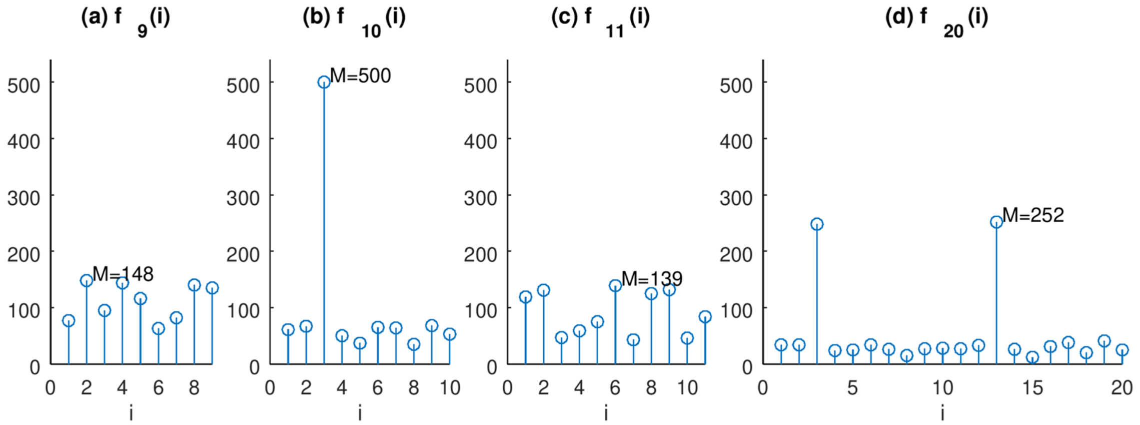

Fig. 3 Wrapped distributions f_n of the distribution in Fig. 2. From left to right: f_9, f_{10}, f_{11} and f_{20} are shown. The maximum is identified in each distribution

as an indicator of the concentration of mass in each n-wrapped distribution. However, we expect the maxima of n-wrapped distributions to grow with decreasing n even if the original distribution has no periodic pattern of peaks. Therefore, the maxima for different periods n are not directly comparable. This is quite evident in Fig. 4 (top), which shows the maxima $M_f(n)$ derived from the example distribution of Fig. 2. The maximum $M_f(10) = 500$ stands out, as expected, since $n = 10$ is the period of the pattern of peaks in f. But that is surpassed by $M_f(5)$ and $M_f(2)$, unsurprisingly since 5 and 2 are divisors of the period.

To make the true period stand out even against its divisors, we define a *concentration score* by the ratio

$$s(n) = \frac{\max f_n}{\max g_n}, \qquad (2)$$

where g_n is the n-wrapped distribution of a distribution g obtained by sorting the frequencies in f in descending order. This score effectively normalizes the maxima of the n-wrapped distributions of f against those of a derived distribution from which all regularities have been removed.

The maxima $\max g_n$ are shown as a thin line in Fig. 4 (top) for the example f distribution. Figure 4 (bottom) displays the corresponding concentration scores.

3 Results

The application of the method to detect regularities on the distance distribution of the 247 previously identified symmetric word pairs produced the following results.

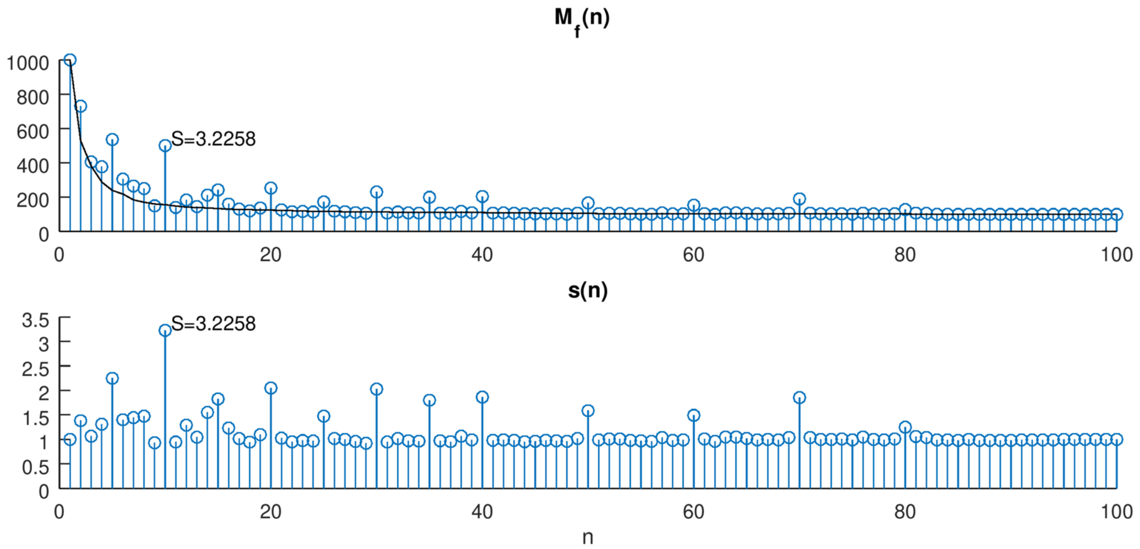

Fig. 4 Maxima $M_f(n)$ of the n-wrapped distributions of f (top) and the corresponding concentration scores for the same distributions (bottom)

Table 1 The ten words with the highest concentration scores for the peak period (T)

Word	Peak period (T)	$s(T)$
GCAGACT	61	17.0
AGTCTGC	61	16.7
TGTCACC	48	15.2
GGTGACA	48	14.9
CTAGGTC	61	13.9
CCTAGGT	61	11.4
GGAGCTC	44	11.2
GACCTAG	61	10.8
GAGCGCT	46	10.1
AGCGCTC	46	10.0

For each word, the peak period and the concentration scores are shown

Table 2 The most frequent periods (# of words ≥ 10) and the corresponding mean and median concentration scores

T	# of words with period T	Mean ($s(T)$)	Median ($s(T)$)
61	21	7.2	5.2
24	16	2.9	2.7
84	15	3.2	3.4
34	13	2.5	2.2
48	11	4.5	2.4
32	10	2.5	2.2
44	10	8.0	7.4

Table 3 Words with a peak period of 84 and a significant number of occurrences

Word	1st peak distance	Chr*	%
AAACCTT	27	Chr19	56.9
AAAGCTT	85	Chr19	70.1
AAGCTTT	83	Chr19	70.7
AAGGCCT	85	Chr19	73.1
AAGGCTT	85	Chr19	62.5
AGGCCTT	83	Chr19	75.9
AGTGTGG	52	Chr19	33.6
ATTCATA	21	Chr19	54.0
CACACTG	30	Chr19	66.3
CAGTGTG	54	Chr19	70.9
CCACACT	32	Chr19	43.4
TATGAAT	63	Chr19	59.3

For each word, the first peak distance, the number of the chromosome (Chr*) with the highest occurrence of the first peak distance and the percentage of occurrences at the Chr* chromosome

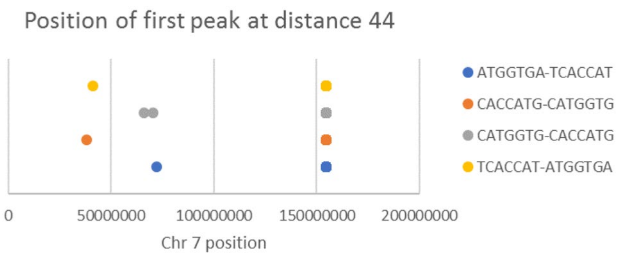

Fig. 5 Positions of the first peak for four words with peak period of 44 (in chromosome 7)

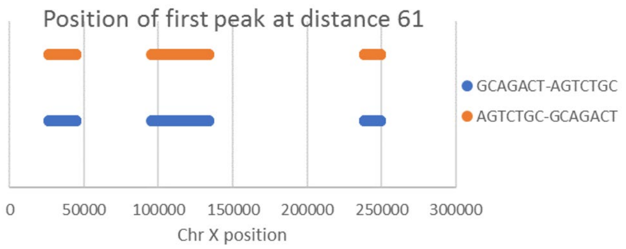

Fig. 6 Positions of the first peak for two words with peak period of 61 (in chromosome X)

Table 1 shows the 10 words with the highest concentration scores and Table 2 lists the most frequent periods (T) found in the set of the 247 words. It may be observed in Table 1 that the symmetric word pairs have, in general, similar scores.

The analysis of the data in Table 2 shows that there are several distinct symmetric word pairs whose distance distributions contain peak regularities with the same period.

At least two questions may be asked: are the sequences that lead to these periodic peaks spread over the entire genome? or are they localized in specific chromosomes?

The periods 44, 61 and 84 were selected (highest median($s(T)$)) to carry out a genomic local analysis to find the positions of the sequences that originate the regular peaks in the distance distribution.

From the analysis of the local distribution of sequences that originate the regular peaks, it was found that, from the set of words with period 44, only four words (ATGGTGA, CACCATG, CATGGTG and TCACCAT) had significant number of occurrences and that those occurred mainly in chromosome 7. For the words with period 61, only two (GCAGACT and AGTCTGC) had significant number of occurrences and mainly in the X and Y chromosomes. For the words with period 84, only 12 words were considered relevant and occurrence mainly in chromosome 19 (see Table 3).

Figures 5, 6 and 7 show the positions in the relevant chromosomes of the selected words for periods 44, 61 and 84.

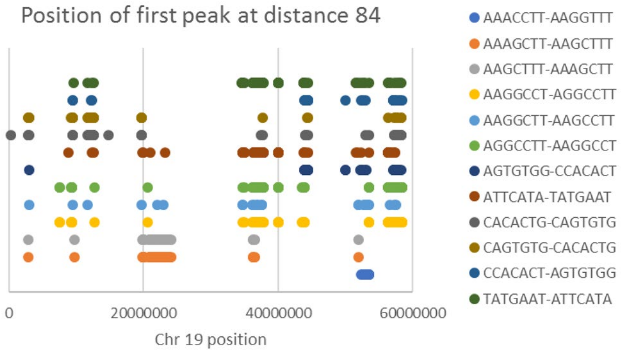

Fig. 7 Positions of the first peak for 12 words with peak period of 84 (in chromosome 19)

4 Conclusion

We studied the occurrence or regular peaks (over-representation) of some distances between symmetric words.

The results of this work revealed sets of words with unusual distribution of distances to the corresponding reversed complements and also with distinct periods of peak regularity.

Since we use masked sequences, the observed regularities are, to the authors knowledge, not due to the known repetitive structures in the human genome and may indicate possible sites for the occurrence of cruciform structures.

We developed a method for detecting periodic regularities in distance distributions that is also able to find the fundamental period of the regularities.

A local analysis was carried out for some symmetric word pairs and it was found that the regular periodic pattern of peaks of the distance distribution occurs mostly at some regions of a single chromosome. Moreover, the symmetric word pairs with the same period of the regular peaks tend to occur in the same chromosome(s).

We expect that this analysis contributes to clarify the possible association between the features of distances between symmetric words and the occurrence of cruciform structures.

To the authors knowledge, the regularly spaced inverted repeats found in this work are a novel genomic feature. We believe that this new feature may be associated with the potential of occurrence of more complex non-B conformations with medium or long length.

Acknowledgements This work was supported by FEDER ("Programa Operacional Fatores de Competitividade"—COMPETE) and by National Funds through FCT ("Fundação para a Ciência e a Tecnologia"), within the projects UID/MAT/04106/2013 to CIDMA (Center for Research and Development in Mathematics and Applications), UID/CEC/00127/2013 to IEETA (Institute of Electronics and Informatics Engineering of Aveiro) , UID/BIM/04501/2013 and POCI-01-0145-FEDER-007628 to iBiMED, and the Grant SFRH/BPD/111148/2015 to RMS.

References

1. Sims GE, Kim SH (2011) Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). Proc Natl Acad Sci 108(20):8329–8334
2. Afreixo V, Bastos CAC, Pinho AJ, Garcia SP, Ferreira PJSG (2009) Genome analysis with inter-nucleotide distances. Bioinformatics 25(23):3064–3070
3. Bernard G, Chan CX, Chan Yb, Chua XY, Cong Y, Hogan JM, Maetschke SR, Ragan MA (2017) Alignment-free inference of hierarchical and reticulate phylogenomic relationships. Brief Bioinform. https://doi.org/10.1093/bib/bbx067
4. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL (2006) CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinform 7(1):446
5. Bastos CAC, Afreixo V, Garcia SP, Pinho AJ (2013) Inter-STOP symbol distances for the identification of coding regions. J Integr Bioinform 10(3):31–39
6. Cer RZ, Donohue DE, Mudunuri US, Temiz NA, Loss MA, Starner NJ, Halusa GN, Volfovsky N, Yi M, Luke BT et al (2012) Non-B DB v2. 0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic Acids Res 41(D1):D94–D100
7. Tavares AH, Pinho AJ, Silva RM, Rodrigues JMOS, Bastos CAC, Ferreira PJSG, Afreixo V (2017) DNA word analysis based on the distribution of the distances between symmetric words. Sci Rep 7(1):728
8. Wang G, Vasquez KM (2014) Impact of alternative DNA structures on DNA damage, DNA repair, and genetic instability. DNA Repair 19:143–151
9. Du Y, Zhou X (2013) Targeting non-B-form DNA in living cells. Chem Rec 13(4):371–384
10. Bacolla A, Wells RD (2004) Non-B DNA conformations, genomic rearrangements, and human disease. J Biol Chem 279(46):47411–47414
11. Inagaki H, Kato T, Tsutsumi M, Ouchi Y, Ohye T, Kurahashi H (2016) Palindrome-mediated translocations in humans: a new mechanistic model for gross chromosomal rearrangements. Front Genet 7:125
12. Kolb J, Chuzhanova NA, Högel J, Vasquez KM, Cooper DN, Bacolla A, Kehrer-Sawatzki H (2009) Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. Chromosom Res 17(4):469–483
13. Wang Y, Leung FC (2006) Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. FEBS Lett 580(5):1277–1284
14. Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, Bacolla A, Collins JR, Stephens RM (2010) Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. Nucleic Acids Res 39(suppl–1):D383–D391
15. Bastos CAC, Afreixo V, Rodrigues JMOS, Pinho AJ (2018) An analysis of symmetric words in human DNA: adjacent vs non-adjacent word distances. In: 12th international conference on practical applications of computational biology and bioinformatics, PACBB 2018, Toledo, Spain
16. Smit AF, Hubley R, Green P (2013–2015) RepeatMasker Open-4.0. http://www.repeatmasker.org
17. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573