An analysis of symmetric words in human DNA: adjacent vs non-adjacent word distances

Carlos A. C. Bastos[1,2], Vera Afreixo[2,3,4], João M.O.S. Rodrigues[1,2], and Armando J. Pinho[1,2]

[1] IEETA-Institute of Electronic Engineering and Informatics of Aveiro
[2] Department of Electronics, Telecommunications and Informatics, University of Aveiro
[3] CIDMA-Center for Research and Development in Mathematics and Applications
[4] Department of Mathematics, University of Aveiro

Abstract. It is important to develop methods for finding DNA sites with high potencial for the formation of hairpin/cruciform structures. In a previous work, we studied the distances between adjacent reversed complement words (symmetric words), and we observed that for some words some distances were favored. In the work presented here, we extended the study to the distance between non-adjacent reversed complement words and we observed strong periodicity in the distance distribution of some words. This may be an indication of potential for the formation of hairpin/cruciform structures.

Keywords: cruciform, distance distribution, genomic word, reversed complement

1 Introduction

Several genomic studies have focused on the analysis of word counts and word distances. Namely, phylogeny studies [8], alignment-free methods [1, 4], CpG detection [6], coding detection [2] and DNA structure analysis [10].

A DNA word analysis based on the distribution of the distances between adjacent symmetric words of length seven was performed [10], and the distributions showed a strong overrepresention of distances up to 350, a feature that may be associated with the occurrence of hairpin/cruciform structures. However, the cruciform structure can occur between reversed complements that are not necessarily adjacent. The stem and loop lengths of cruciform structures seem to vary over a wide range. According to different authors, the stem length varies between 6 and 100 nucleotides, while loop lengths may range from 0 to 2000 nucleotides [7, 11, 5]. The aim of this work is to analyse the occurrence of adjacent and non-adjacent symmetric words along the sequence.

2 Methods

The human genome is the subject of this study and the main purpose is to explore the human DNA structure. Specifically, we want to explore structures beyond the

well-known repetition structures. Thus, we used pre-masked sequences available from the UCSC Genome Browser (http://genome.ucsc.edu) downloads page. These files contain the GRCh38 assembly sequences, with repeats reported by RepeatMasker [9] and Tandem Repeats Finder [3] masked with Ns.

Consider the alphabet $\mathcal{A} = \{A, C, G, T\}$ and let w be a symbolic sequence (word) defined in $\mathcal{A}^k$, where k is the length of w. The pair composed by 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 the adjacent reversed complement word $f_{w,w'}$. We also compute the frequency distributions of distances between occurrences of each word and all succeeding reversed complements, $f_{w,w'\ldots w'}$. We compare both distance distributions using the well-known Kullback-Leibler divergence (KL).

For example, consider the following sequence:

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

For word $w = ACT$ we have only two valid distances between adjacent reversed complement words (7 and 6), but we have five distances between adjacent and non adjacent reversed complement words (7, 14, 30, 13, 6).

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

In order to analyse abundant words in the DNA sequence, we exclude symmetric word pairs with relative occurrence frequency lower than $1/4^7$.

3 Results and Discussion

Table 1 shows the 10 words with the greatest divergence between $f_{w,w'}$ and $f_{w,w'\ldots w'}$. Note that there are no CG pairs in the composition of the words and that there is no obvious pattern in words composition.

By visual comparison of the distance distributions of several words, we found distinct patterns of divergence between adjacent and non-adjacent words. Figures 1 and 2 show two of those divergence patterns. For word $w = TGTCACC$, the distribution of $f_{w,w'}$ shows a single strong peak at distance 28 (see Fig. 1 top), whereas the $f_{w,w'\ldots w'}$ distribution displays a periodic pattern of peaks at distances of $28 + 48i$, with $i = 0, 1, 2, \ldots$ (see Fig. 1 bottom). For word $w = TGCATGC$, the divergence between $f_{w,w'}$ and $f_{w,w'\ldots w'}$ is also high. The $f_{w,w'}$ distribution has a single strong peak at distance 324, which is weakened in $f_{w,w'\ldots w'}$, but the extra peaks in the $f_{w,w'\ldots w'}$ distribution do not introduce additional regularity (see Fig. 2).

With our exploratory analysis by visual inspection, we found an interesting regular pattern in words with high divergence between $f_{w,w'}$ and $f_{w,w'\ldots w'}$: words with a few peaks in $f_{w,w'}$ and with several nearly periodic peaks in $f_{w,w'\ldots w'}$

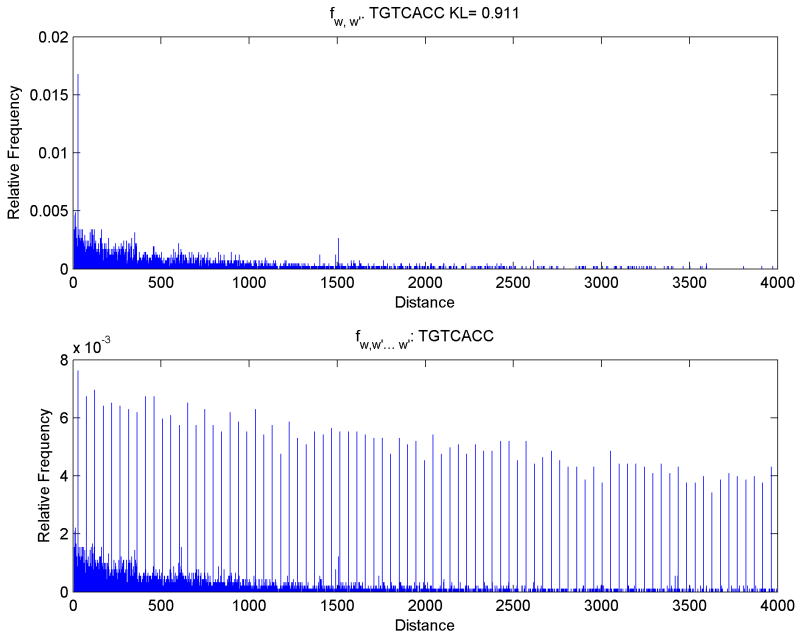

Fig. 1. $f_{w,w'}$ and $f_{w,w'\ldots w'}$ $TGTCACC$.

(see for example Fig. 1). To find words that have this divergence pattern, we implemented the following algorithm:

- compute
  - $m = \max\{f_{w,w'}\}$
  - $n_1 = \#\{d \in 1, \ldots, 4000 : f_{w,w'}(d) > c\,m\}$
  - $n_2 = \#\{d \in 1, \ldots, 4000 : f_{w,w'\ldots w'}(d) > c\,m\}$, with $c \in]0; 1[$
- select the words ($w \in \mathcal{A}^k$) for which $n_1 \leq 2$ and $n_2 - n_1 > 2$

For $c = 0.4$, the algorithm selected 34 words. Table 2 shows the 10 words with exactly periodic peaks in $f_{w,w'\ldots w'}$. It is interesting to note that the peak period of 6 out of 10 distributions has a value of 84. The most frequent spacing between peaks of $f_{w,w'\ldots w'}$, for each word, is not longer than 102.

In order to characterize the chromosomal distribution of the peak distances (d_p) we searched, in each chromosome, the positions where the word w appears at a distance d_p before w' and we counted the number of occurrences in each chromosome. Table 2 contains the chromosome with the highest percentage of occurrence of the first peak distance. It is evident that the pattern of distance distributions studied here reflects a local behaviour that occurs mainly in a single

Table 1. The ten words with greatest divergence between $f_{w,w'}$ and $f_{w,w'\ldots w'}$. For each word, the Kullback-Leibler divergence and the nucleotide composition are shown.

Word	KL	n_A	n_C	n_G	n_A
$TGTGCAC$	1.175	1	2	2	2
$TGGGCCC$	1.141	0	3	3	1
$GGAGCTC$	1.054	1	2	3	1
$TGGGTAA$	1.053	2	0	3	2
$TTACCCA$	1.033	2	3	0	2
$TGTCCAC$	0.961	1	3	1	2
$GGGCCCA$	0.941	1	3	3	0
$AGAATTC$	0.918	3	1	1	2
$TGTCACC$	0.911	1	3	1	2
$TGCATGC$	0.905	1	2	2	2

Table 2. Words with a peak in $f_{w,w'}$ and regular peaks in $f_{w,w'\ldots w'}$. For each word, the peak period, the first peak distance and the nucleotide composition are shown. The column Chr* contains the number of the chromosome with the highest occurrence of the first peak distance and the last column contains the percentage of occurrences at the Chr* chromosome.

Word	peak period	first peak distance	n_A	n_C	n_G	n_T	Chr*	%
AAGCTTT	84	83	2	1	1	3	19	70
AGGCCTT	84	83	1	2	2	2	19	76
AGTGTGG	84	52	1	0	4	2	19	34
ATTCATA	84	21	3	1	0	3	19	54
CCACACT	84	32	2	4	0	1	19	43
TATGAAT	84	63	3	0	1	3	19	55
TCACCAT	44	36	2	3	0	2	7	91
TGGGTAA	42	13	2	0	3	2	11	91
TGTCACC	48	28	1	3	1	2	3	86
TTACCCA	42	29	2	3	0	2	11	91

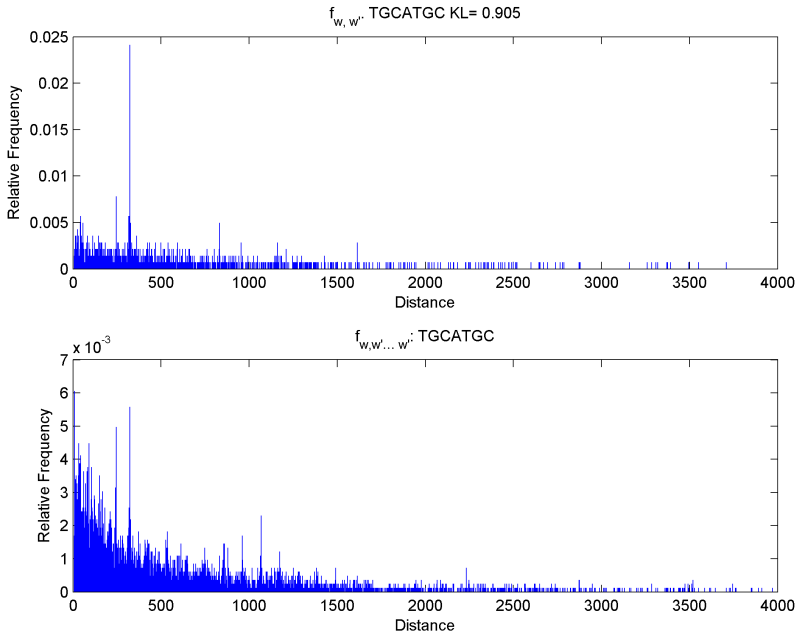

Fig. 2. $f_{w,w'}$ and $f_{w,w'...w'}$ $TGCATGC$.

chromosome. Moreover, all of the words of Table 2 with the same peak period occurred in the same chromosome.

Figures 3 and 4 show examples of the distance distributions found by the previous algorithm for words $w = GTGGCCA$ and $w = TCACCAT$. Both words have the expected regular pattern of peaks in $f_{w,w'...w'}$. Figure 3 presents an interesting pattern with a very strong peak at a distance of 980. Figure 4 also shows a regular pattern of peaks in $f_{w,w'...w'}$, with a heavier tail, resembling a mixture of two distributions.

Figure 5 shows the positions, in chromosome 19, of the first peak distance of the words in Table 2 with peak period of 84. These positions seem to form three groups of words: group 1 - $AGGCCTT$, $ATTCATA$ and $TATGAAT$; group 2 - $AGTGTGG$ and $CCACACT$; and group 3 - $AAGCTTT$. Both group 1 and group 2 contain each other a reverse complement pair which might indicate that the words w and w' form a regular pattern of occurrence in the regions of chromosome 19 shown in the figure.

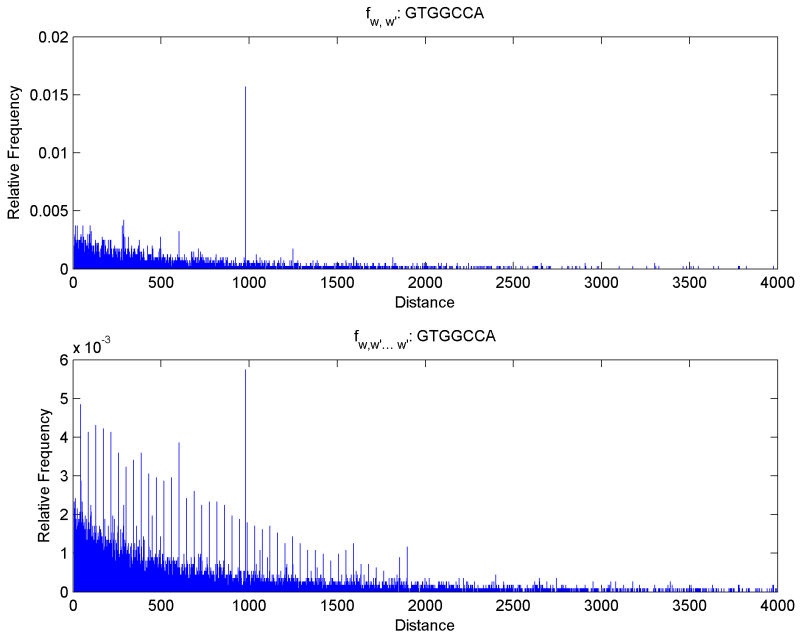

Fig. 3. $f_{w,w'}$ and $f_{w,w'...w'}$ $GTGGCCA$.

4 Conclusions

We believe that strong overrepresention of some distances between symmetric words is a feature that may be associated with the occurrence of cruciform structures.

Our analysis identified a set of words with unusual distribution of distances to the corresponding reversed complements. Since we use masked sequences, the observed regularities are not due to the known repeated structures.

The regular periodic pattern of the distance between reversed complements occurs mostly at some regions of a single chromosome.

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.

Acknowledgment

This work was supported by FEDER ("Programa Operacional Fatores de Competitividade" COMPETE) and 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) and

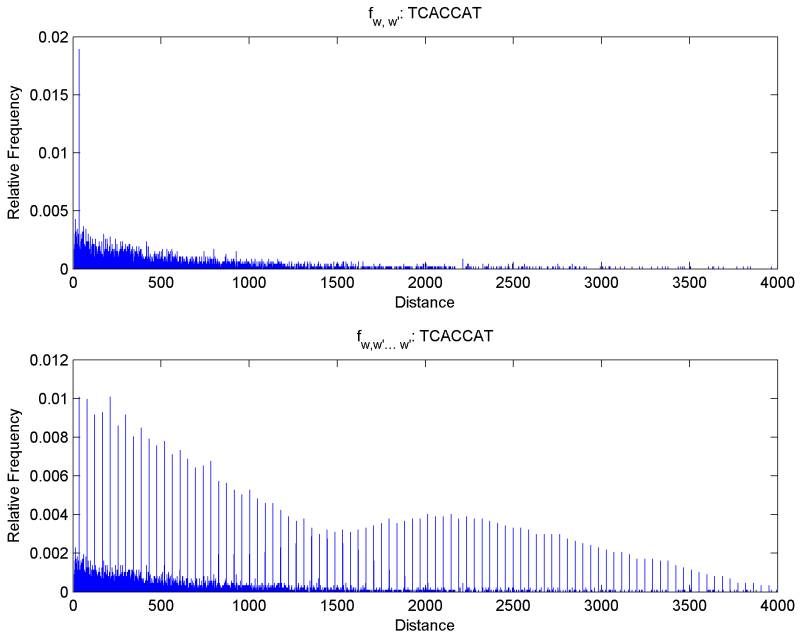

Fig. 4. $f_{w,w'}$ and $f_{w,w'...w'}$ $TCACCAT$.

UID/CEC/00127/2013 to IEETA (Institute of Electronics and Informatics Engineering of Aveiro).

References

1. Vera Afreixo, Carlos A. C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, December 2009.
2. Carlos AC Bastos, Vera Afreixo, Sara P Garcia, and Armando J Pinho. Inter-stop symbol distances for the identification of coding regions. *Journal of integrative bioinformatics*, 10(3):31–39, 2013.
3. Gary Benson. Tandem repeats finder: a program to analyze dna sequences. *Nucleic acids research*, 27(2):573, 1999.
4. Guillaume Bernard, Cheong Xin Chan, Yao-ban Chan, Xin-Yi Chua, Yingnan Cong, James M Hogan, Stefan R Maetschke, and Mark A Ragan. Alignment-free inference of hierarchical and reticulate phylogenomic relationships. *Briefings in bioinformatics*, 2017.
5. Regina Z Cer, Kevin H Bruce, Uma S Mudunuri, Ming Yi, Natalia Volfovsky, Brian T Luke, Albino Bacolla, Jack R Collins, and Robert M Stephens. Non-b db: a database of predicted non-b dna-forming motifs in mammalian genomes. *Nucleic acids research*, 39(suppl_1):D383–D391, 2010.

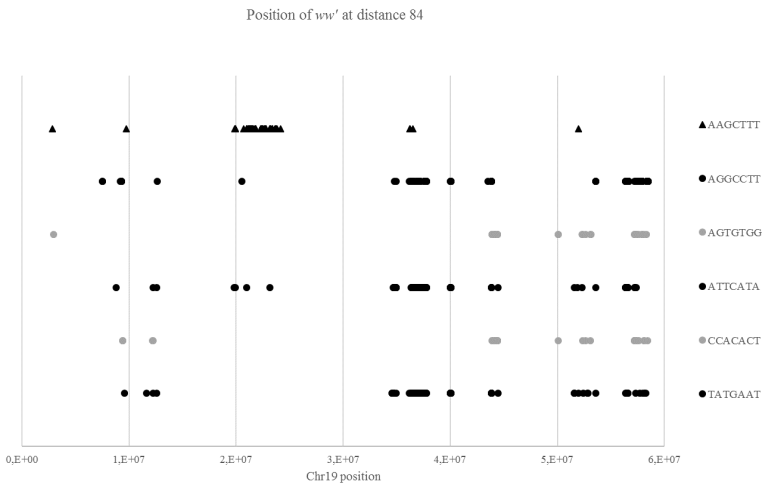

Fig. 5. Positions of the first peak of ww' distance for six words with peak period of 84 (in chromosome 19).

6. Michael Hackenberg, Christopher Previti, Pedro Luis Luque-Escamilla, Pedro Carpena, José Martínez-Aroza, and José L Oliver. Cpgcluster: a distance-based algorithm for cpg-island detection. *BMC bioinformatics*, 7(1):446, 2006.
7. Jessica Kolb, Nadia A Chuzhanova, Josef Högel, Karen M Vasquez, David N Cooper, Albino Bacolla, and Hildegard Kehrer-Sawatzki. Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. *Chromosome research*, 17(4):469–483, 2009.
8. Gregory E Sims and Sung-Hou Kim. Whole-genome phylogeny of escherichia coli/shigella group by feature frequency profiles (ffps). *Proceedings of the National Academy of Sciences*, 108(20):8329–8334, 2011.
9. Arian FA Smit, Robert Hubley, and P Green. Repeatmasker, 1996.
10. Ana HMP Tavares, Armando J Pinho, Raquel M Silva, João MOS Rodrigues, Carlos AC Bastos, Paulo JSG Ferreira, and Vera Afreixo. Dna word analysis based on the distribution of the distances between symmetric words. *Scientific reports*, 7(1):728, 2017.
11. Yong Wang and Frederick CC Leung. Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. *FEBS letters*, 580(5):1277–1284, 2006.