Detection and characterization of local inverted repeats regularities

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

[1] IEETA-Institute of Electronics and Informatics Engineering 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
cbastos@ua.pt, vera@ua.pt, jmr@ua.pt, ap@ua.pt

Abstract. To explore the inverted repeats regularities along the genome sequences, we propose a sliding window method to extract the concentration scores of inverted repeats periodic regularities and the total mass of possible inverted repeats pairs. We apply the method to the human genome and locate the regions with the potential for the formation of large number of hairpin/cruciform structures. The number of found windows with periodic regularities is small and the patterns of occurrence are chromosome specific.

Keywords: cruciform, distance distribution, inverted repeats, periodic regularities

1 Introduction

Hairpin/cruciform structures are a type of non-B DNA structure with importance in biological processes and gene function [1]. DNA motifs that are known to potentially form non-B DNA structures are available at public databases [2, 3]. Hairpins/cruciforms may form dynamically 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 location and mechanism of action are lacking *in vivo* [4, 5].

The stem and loop lengths of hairpin/cruciforms 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 [6, 7, 2]. 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 [4].

The simultaneous occurrence of inverted repeats in a specific region are a required feature of local cruciform structures. However, some regions can greatly enhance the occurrence of hairpin/cruciforms conformations than others.

A DNA word analysis based on the distribution of the distances between adjacent symmetric words of length seven [8] 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 of non-adjacent inverted repeats, since adjacency is not a required condition for cruciform structures to form [9, 10].

The present work focuses on identifying and characterising the local behaviour of inverted repeats. The occurrence of regular peaks and high mass in the cumulative distance distribution of symmetric word pairs will be explored.

2 Methods

This work aims to find, in the human genome, structures with regularity beyond the already well-known repetition structures published in the literature. Thus, we used pre-masked sequences available from the UCSC Genome Browser webpage [11]. These files contain the GRCh38 assembly sequences, with repeats reported by RepeatMasker [12] and Tandem Repeats Finder [13] 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 an inverted repeats pair. For example, (ACT, AGT) is an inverted repeat pair.

In this work we analyse, along the human genome, the regularities in the distance distribution of inverted repeats by dividing the complete genome in successive windows containing 100k nucleotides. Instead of separately analysing the distance distribution for each possible inverted repeat, as done in previous works [9, 10], in the present work we analyse *cumulative* distance distributions of all possible inverted repeats. This keeps the data size manageable.

2.1 Distance between inverted repeats

For all words of length k, we compute the frequency distributions of distances, f, between occurrences of each word and all succeeding reversed complements at distances between k and 4000.

For example, consider the sequence $ACTTTGTACTAAAGTTAAG$. Only four inverted repeats (w, w') of length $k = 3$ occur in this short sequence. The following lines show all occurrences of these inverted repeats, marked by underlines (w) and overlines (w'):

(ACT, AGT): $\underline{ACT}TTGT\underline{ACT}AA\overline{AGT}TAAG$,
(CTT, AAG): $A\underline{CTT}TGTACTA\overline{AAG}TT\overline{AAG}$,
(TTT, AAA): $AC\underline{TTT}GTACT\overline{AAA}GTTAAG$,
(TAA, TTA): $ACTTTGTAC\underline{TAA}AG\overline{TTA}AG$.

The previous sequence includes six distances to all the succeeding reversed complement words (distances: 12, 5, 10, 15, 8, and 5). Thus the cumulative distribution is $f(5) = 2, f(8) = f(10) = f(12) = f(15) = 1$ and $f(i) = 0$ for all other i values.

Motivated by previous work and considering 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 a N symbol is found, the search for w' is stopped, because the length of long stretches of Ns may be artificial.

2.2 Quantifying periodic regularities

In previous work, we detected words with strong periodic regularities in the complete human genome [10]. That work proposed a new measure for quantifying the periodic regularity of distributions, the *concentration score*. The proposed method is also able to find the fundamental period of the regularities.

The concentration score, s, for a given distribution $f(i)$, with $i = 1, 2, \ldots, N$ is computed in several steps [10]:

1. Obtain an auxiliary distribution g by sorting the frequencies in f in descending order.
2. Generate the family of *wrapped distributions* for f,

$$f_n(i) = \sum_{0 \leq j \leq \frac{N-i}{n}} f(i + jn), \text{ for } i = 1, 2, \ldots, n. \tag{1}$$

3. Generate the family of wrapped distributions, g_n, for g, using the previous procedure.
4. Compute the concentration score for each *wrapping period* n by the ratio

$$s(n) = \frac{\max f_n}{g_n(1)}. \tag{2}$$

5. The periodic regularity of distribution f is quantified by $S = \max s(n)$ and $P = \arg\max s(n)$ is its *fundamental period*.

Note that auxiliary distribution g eliminates any periodic regularity from f and that $\max(g_n) = g_n(1)$. Also, note that $s(n)$, S and P are not defined if $f(i) = 0$, for all i.

2.3 Number of possible pairs of inverted repeats

In a sequence, the total mass of the distribution f,

$$M = \sum_{i=k}^{4000} f(i), \tag{3}$$

corresponds to the number of possible pairs of inverted repeats occurring within a range of 4000 nt.

2.4 Windows selection

In order to locate the sequence windows with the highest concentration scores or highest total mass we used quantile 0.999 as the discriminating threshold. This procedure resulted in the isolation of 30 windows with relevant concentration scores and 30 windows with relevant total mass.

3 Results

Table 1 shows the minimum, maximum and quartiles of concentration scores, S, and of total masses, M, over all the windows in each of the chromosomes and in the full human genome. The median of the set of S values over the complete

Table 1. Order statistics of scores, S, and of total masses, M, over all the windows in each of the chromosomes and in the full human genome.

	Concentration Score, S					Total mass, M					
chr	**min**	**Q1**	**med**	**Q3**	**max**	**chr**	**min**	**Q1**	**med**	**Q3**	**max**
1	1.00	1.43	1.50	1.58	4.16	**1**	1	2182	3053	4259	21361
2	1.16	1.42	1.48	1.55	3.56	**2**	15	2613	3517	4732	65495
3	1.21	1.42	1.48	1.56	10.88	**3**	137	2353	3273	4472	38696
4	1.20	1.40	1.47	1.55	6.06	**4**	119	2715	3723	4885	32671
5	1.19	1.42	1.48	1.56	5.69	**5**	42	2531	3586	4913	120365
6	1.21	1.42	1.48	1.55	2.85	**6**	250	2717	3602	4732	112782
7	1.00	1.42	1.48	1.55	4.47	**7**	1	2465	3573	4835	29132
8	1.19	1.42	1.49	1.56	3.00	**8**	3	2268	3259	4487	97380
9	1.19	1.43	1.50	1.57	7.83	**9**	135	2211	3101	4266	19694
10	1.00	1.43	1.48	1.56	3.35	**10**	1	2426	3234	4430	40031
11	1.00	1.43	1.50	1.57	3.79	**11**	1	2193	3030	4123	46349
12	1.00	1.43	1.50	1.56	5.68	**12**	1	2184	3107	4341	36793
13	1.17	1.40	1.46	1.53	5.53	**13**	539	3115	4183	5361	25881
14	1.16	1.43	1.48	1.56	9.43	**14**	135	2430	3374	4678	40828
15	1.00	1.43	1.50	1.56	2.39	**15**	1	2295	3128	4310	35519
16	1.25	1.44	1.50	1.58	4.00	**16**	26	1742	2661	3813	24775
17	1.00	1.43	1.50	1.57	5.92	**17**	1	2180	3114	4557	46507
18	1.00	1.42	1.47	1.54	4.03	**18**	1	2739	3683	4909	21117
19	1.00	1.46	1.55	1.67	7.51	**19**	1	1765	2655	3749	18806
20	1.28	1.44	1.50	1.58	4.79	**20**	2	1801	2507	3729	26786
21	1.00	1.42	1.48	1.55	3.00	**21**	1	2840	3942	5171	18180
22	1.28	1.44	1.50	1.59	4.00	**22**	11	1781	2856	4810	21097
X	1.00	1.47	1.56	1.67	11.70	**X**	1	1595	2412	3513	72594
Y	1.00	1.46	1.55	1.67	11.70	**Y**	1	1604	2727	4229	41404
all	1.00	**1.42**	**1.50**	**1.57**	11.70	**all**	1	2312	3278	4551	120365

genome is 1.50 and the inter-quartile range is 0.15 (Table 1), which shows that the majority of the windows have an S that shows low degree of periodic regularity.

However, the maxima of S reveal that there are genomic regions that show high degree of periodic regularity. The range of fundamental periods found is $30-102$ showing diversity on the period regularities (see Table 2).

The selection criteria defined in Section 2.4 results in a threshold of 4.09 for the concentration score and a threshold of 27519 for the total mass. Using the criteria we observed that the occurrence of windows with high periodic regularity is heterogeneous among the chromosomes. Table 2 shows the windows with the highest concentration scores. Only 14 chromosomes contain regions with high periodic regularity. Figure 1 shows, as an example, the behaviour of the

Table 2. Windows with the 0.1% highest concentration scores and the windows with the 0.1% highest total mass.

Highest S				
chr	**win #**	S	M	P
X	2	11.70	41404	61
Y	2	11.70	41404	61
3	1958	10.88	14845	48
X	3	10.02	18395	61
Y	3	10.02	18395	61
X	1	9.46	27519	61
Y	1	9.46	27519	61
14	1053	9.43	24554	**102**
9	4	7.83	4594	61
X	6	7.70	8879	44
Y	6	7.70	8879	44
19	2Y	7.51	3481	84
X	X	6.21	3786	40
Y	X	6.21	3786	40
4	1865	6.06	5547	28
17	10	5.92	2886	57
17	3	5.69	5182	45
5	4	5.69	14260	62
12	504	5.68	2744	36
5	1781	5.65	6145	43
13	11X	5.53	17242	34
4	1813	5.18	7638	51
13	11Y	4.84	10887	34
20	605	4.79	4145	46
5	1811	4.65	180X	33
12	405	4.48	16585	**30**
7	1547	4.47	18310	44
19	228	4.33	3348	84
19	234	4.32	2670	84
1	2368	4.16	5844	49

Highest M			
chr	**win #**	S	M
5	674	1.39	120365
6	315	1.23	112782
8	1295	1.79	97380
6	1609	1.26	73271
X	531	1.76	72594
8	576	1.57	72053
2	685	1.16	65495
X	91	1.27	60390
X	1157	1.62	50189
17	135	1.48	46507
11	1147	1.88	46349
2	1948	1.97	44469
X	2	11.70	41404
Y	2	11.70	41404
5	1262	2.13	41007
14	859	2.27	40828
10	575	2.80	40031
3	1888	1.81	38696
12	866	2.14	36793
Y	2X	1.78	36129
2	439	1.47	35951
15	231	1.52	35519
2	2265	2.56	34891
6	1703	2.44	32718
4	1763	1.72	32671
3	1119	2.19	31790
8	353	1.24	31471
8	38	1.33	30288
7	581	1.31	29132
15	253	1.58	27826

win #- window number in chromosome
S- concentration score; M- total mass; P- fundamental period;

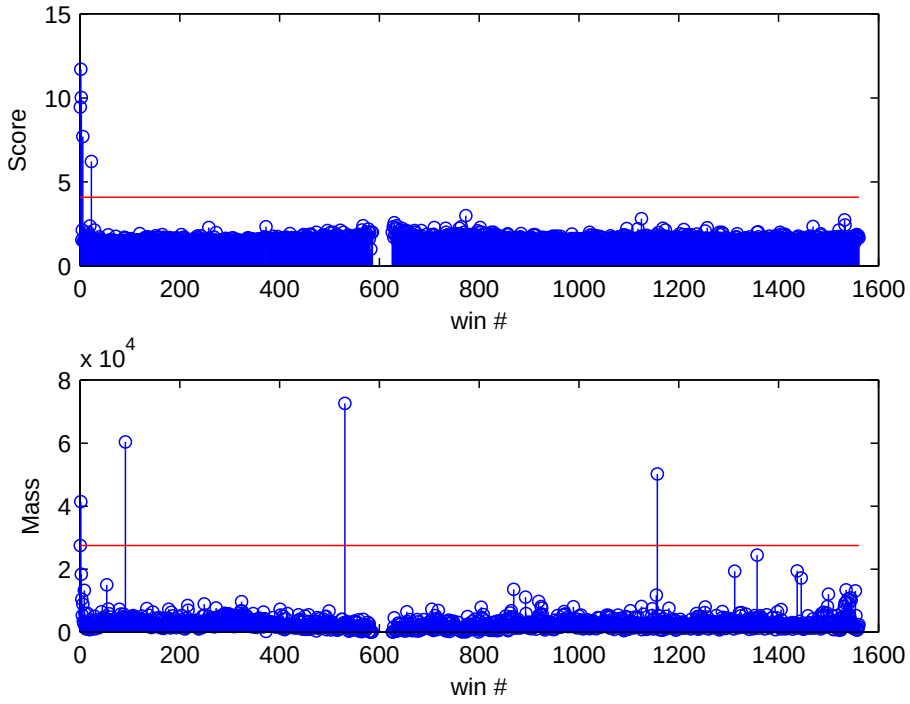

Fig. 1. The concentration score and the total mass of possible inverted repeats pairs, measured in 100 knt windows along chrX. The horizontal red lines correspond to the thresholds obtained with the windows selection criteria. Gaps around window 600 are a consequence of long stretches of the symbol N in the sequence.

concentration score and of the total mass along the X chromosome. There are 5 windows with S above the threshold and 4 windows with M above the threshold, only one of the identified windows is common to both criteria. The scores S and M for the set of all the windows in the human genome are weakly correlated, Pearson's correlation coefficient is 0.29.

Figures 2 and 3 show the cumulative distance distributions of, respectively, the window with the highest concentration score and the window with the highest total mass of possible inverted repeats pairs.

4 Discussion and Conclusion

Motivated by the potential connection between the occurrence of inverted repeats pairs with the possible formation of hairpin/cruciform structures, we explored the behaviour of the inverted repeats pairs in terms of periodic regularity of its occurrence and the total mass.

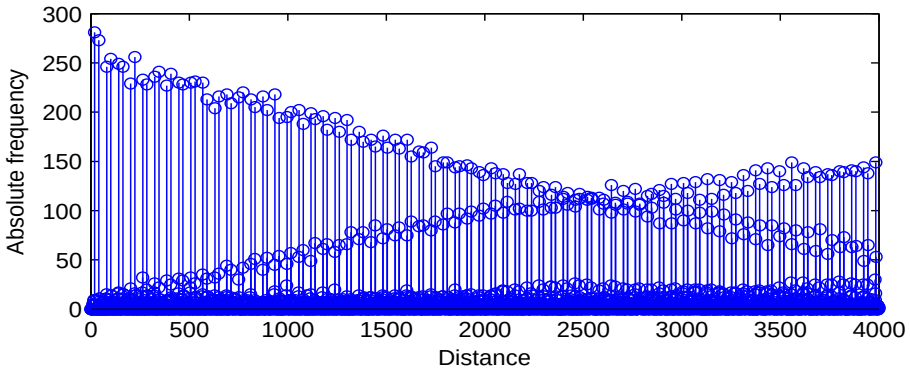

Fig. 2. Cumulative distance distribution of the window with the highest concentration score ($S = 11.70$) in chromosome X (chrX:100001–200000).

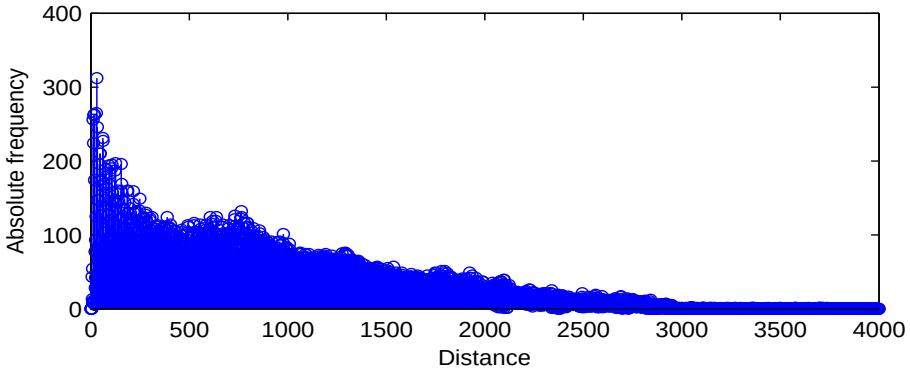

Fig. 3. Cumulative distance distribution of the window with the highest number of possible inverted repeat pairs ($M = 120365$) in chromosome 5 (chr5:67300001–67400000).

We identified genomic regions with atypically high score values indicative of the frequent occurrence of inverted repeats pairs at regular intervals. We also identified regions with a large number of possible inverted repeats pairs.

The patterns of periodic regularities and of total mass seem to be specific for each chromosome. The only exceptions are the patterns for the X and Y chromosomes, which are partially similar, as expected since they share parts of their genomic sequences.

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/04106/2019 to CIDMA (Center for Re-

search and Development in Mathematics and Applications) and UID/CEC/00127/2019 to IEETA (Institute of Electronics and Informatics Engineering of Aveiro).

References

1. Du, Y., Zhou, X.: Targeting non-B-form DNA in living cells. The Chemical Record **13**(4) (2013) 371–384
2. Cer, R.Z., Bruce, K.H., Mudunuri, U.S., Yi, M., Volfovsky, N., Luke, B.T., Bacolla, A., Collins, J.R., Stephens, R.M.: Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. Nucleic acids research **39**(suppl_1) (2010) D383–D391
3. Cer, R.Z., Donohue, D.E., Mudunuri, U.S., Temiz, N.A., Loss, M.A., Starner, N.J., Halusa, G.N., Volfovsky, N., Yi, M., Luke, B.T., et al.: Non-B DB v2. 0: a database of predicted non-B DNA-forming motifs and its associated tools. Nucleic acids research **41**(D1) (2012) D94–D100
4. Bacolla, A., Wells, R.D.: Non-B DNA conformations, genomic rearrangements, and human disease. Journal of Biological Chemistry **279**(46) (2004) 47411–47414
5. Inagaki, H., Kato, T., Tsutsumi, M., Ouchi, Y., Ohye, T., Kurahashi, H.: Palindrome-mediated translocations in humans: A new mechanistic model for gross chromosomal rearrangements. Frontiers in genetics **7** (2016) 125
6. Kolb, J., Chuzhanova, N.A., Högel, J., Vasquez, K.M., Cooper, D.N., Bacolla, A., Kehrer-Sawatzki, H.: Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. Chromosome research **17**(4) (2009) 469–483
7. Wang, Y., Leung, F.C.: Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. FEBS letters **580**(5) (2006) 1277–1284
8. Tavares, A.H., Pinho, A.J., Silva, R.M., Rodrigues, J.M., Bastos, C.A., Ferreira, P.J., Afreixo, V.: DNA word analysis based on the distribution of the distances between symmetric words. Scientific reports **7**(1) (2017) 728
9. Bastos, C.A.C., Afreixo, V., Rodrigues, J.M.O.S., Pinho, A.J.: An analysis of symmetric words in human DNA: adjacent vs non-adjacent word distances. In: PACBB 2018 - 12th International Conference on Practical Applications of Computational Biology & Bioinformatics, Toledo, Spain (June 2018)
10. Bastos, C.A.C., Afreixo, V., Rodrigues, J.M.O.S., Pinho, A.J., Silva, R.: Distribution of distances between symmetric words in the human genome: Analysis of regular peaks. Interdisciplinary Sciences: Computational Life Sciences (2019)
11. Kent, W., Sugnet, C., Furey, T., Roskin, K., Pringle, T., Zahler, A., Haussler, D.: The human genome browser at UCSC. Genome Research **12**(6) (2002) 996–1006
12. Smit, A.F.A., Hubley, R., Green, P.: RepeatMasker Open- 4.0 (2013–2015)
13. Benson, G.: Tandem repeats finder: a program to analyze dna sequences. Nucleic acids research **27**(2) (1999) 573