Austin Journal of Proteomics, Bioinformatics & Genomics

Open Access

Austin Publishing Group

Research Article

Exploring the Exceptional Genomic Word Symmetry among Regions

Afreixo V[1,2,3,4]*, Rodrigues JMOS[4,5] and Bastos CAC[4,5]

[1]Department of Mathematics, University of Aveiro, Portugal

[2]Institute for Research in Biomedicine, University of Aveiro, Portugal

[3]Center for Research and Development in Mathematics and Applications, University of Aveiro, Portugal

[4]Institute of Electronics and Telematics Engineering of Aveiro, University of Aveiro, Portugal

[5]Department of Electronics Telecommunications and Informatics, University of Aveiro, Portugal

***Corresponding author:** Vera Afreixo, Department of Mathematics, University of Aveiro, Aveiro, Portugal

Received: February 13, 2015; **Accepted:** October 05, 2015; **Published:** October 10, 2015

Abstract

The second Chargaff's parity rule and its extensions are reported as a universal phenomenon in DNA sequences [1-3]. However, parity of the frequencies of reverse complementary oligonucleotides could be a mere consequence of the single nucleotide parity rule, if nucleotide independence is assumed. Exceptional symmetry has been proposed as a meaningful measure of the extension of the second parity rule to oligonucleotides (symmetry above that expected in independence contexts) [4]. For short genomic word length (lower than 13) the global exceptional symmetry was detected in long and short organism genomes [5]. But there are some issues that have not been explored: Is the local symmetry behavior distinct from the symmetry in the full organism? What is the variation of the exceptional symmetry along the sequence? To explore the exceptional genomic word symmetry along the genome sequences, we propose a sliding window method to extract the values of exceptional symmetry (for all words or by word groups). We compare the exceptional symmetry effect distribution in real genomes against control scenarios, testing the differences and performing a residual analysis. We compare and sort the word groups taking into account the exceptional symmetry variation along the sequence and the exceptional symmetry effect distribution by word group.

Introduction

Chargaff's first parity rule states that, in any sequence of double-stranded DNA molecules, the total number of complementary nucleotides is exactly equal. Chargaff's second parity rule states that those quantities are almost equal in a single strand of DNA.

%**A** ≅ %**T** and %**C** ≅ %**G**

The extensions to the second parity rule state that, in each DNA strand, the proportion of an oligonucleotide should be similar to that of its reversed complement [1].

%ACTGG ≅ %CCAGT

w = A C T G G
| | | | |
T G A C C → Reversed complement of *w*
CCAGT

There is no single accepted reason that justifies this single strand symmetry. However, the relative ubiquity of this phenomenon suggests a relationship with genomic evolution.

Powdel and others [6] studied the symmetry phenomenon, by defining and analysing the frequency distributions of the local abundance of oligonucleotides along a single strand of DNA, and found that the frequency distributions of reverse complementary oligonucleotides tend to be statistically similar. Afreixo *et al.* [4] introduced a new symmetry measure, which emphasizes that the frequency of an oligonucleotide is more similar to the frequency of its reversed complement than to the frequencies of other equivalent composition oligonucleotides. They also identified several word groups with a strong exceptional symmetry (e.g. G_0 for word length equal to 4).

Materials

We analyzed the whole human genome, reference assembly build 37.3, available from the website of the National Center for Biotechnology Information, discarding all non-ACGT symbols. In our data processing, the chromosomes were processed as separate sequences, words were counted with overlap and non-ACGT symbols were considered as sequence separators. As a control experiment, we generated random sequences assuming independence between nucleotides and using the human chromosomes nucleotide composition.

Exceptional genomic word symmetry

Exceptional genomic word symmetry was proposed for equivalent composition groups (ECG) and globally [4].

G_m denotes a set of words with equivalent composition, containing the same number of As + Ts (m). For k=2, G_0 = {**CC, CG, GC, GG**}; G_1 = {**AC, AG, CA, GA, CT, GT, TC, TG**} and G_2 = {**AA, AT, TA, TT**}.

The proposed exceptional symmetry measure for G_m is given by

$$\mathrm{VR}(G_m)=\sqrt{\frac{\left(X_u^2(G_m)+\varepsilon\right)/\mathrm{df}_u(G_m)}{\left(X_s^2(G_m)+\varepsilon\right)/\mathrm{df}_s(G_m)}}$$

where $X_s^2(G_m)=\sum_{w\in G_m}\frac{\left(n_w-\frac{n_w+n_{w'}}{2}\right)^2}{\frac{n_w+n_{w'}}{2}}=\frac{1}{2}\sum_{w\in G_m}\frac{(n_w-n_{w'})^2}{n_w+n_{w'}}$,

$X_u^2(G_m)=\sum_{w\in G_m}\frac{\left(n_w-\bar{n}_{G_m}\right)^2}{\bar{n}_{G_m}}=\sum_{w\in G_m}\frac{n_w^2}{\bar{n}_{G_m}}-n_m$ and df are the degrees of freedom.

Citation: Afreixo V, Rodrigues JMOS and Bastos CAC. Exploring the Exceptional Genomic Word Symmetry among Regions. Austin J Proteomics Bioinform & Genomics. 2015; 2(1): 1011.

G_1	AC GT AG CT CA TG GA TC	Avg	$X_S^2(G_1)$	$X_U^2(G_1)$	$df_s(G_1)$	$df_u(G_1)$	VR (G_1)
Example 1(n_w)	31 30 31 30 29 30 29 30	30	0,1	0,2	3	6	1,2
Example 2 (n_w)	41 40 21 20 11 10 51 50	30	0,1	66,8	3	6	18,8

If $\mathbf{VR}(\boldsymbol{G}_m)\approx\mathbf{1}$, there is no exceptional symmetry and if $\mathbf{VR}(\boldsymbol{G}_m)>>\mathbf{1}$, there is exceptional symmetry.

To measure the global exceptional symmetry we proposed [4],

$$VR=\sqrt{\frac{\left(X_u^2+\varepsilon\right)/df_u}{\left(X_s^2+\varepsilon\right)/df_s}}$$

where $X_s^2=\sum X_s^2(G_m)$ and $X_u^2=\sum X_u^2(G_m)$

The exceptional genomic word symmetry values were determined in all non-overlapping sub-chromosomal regions of specific size (10 000, 50 000 and 100 000).

Results and Conclusion

In this study, we analysed local exceptional word symmetry in the complete human genome. In particular, we analyzed words of lengths up to 12 in each human chromosomes and globally. We show results only for word length 5.

The local exceptional symmetry in the human genome is clearly higher than in the random scenario.

The human chromosomes appear to have a synchronized behavior. As G_0 and G_k are the sets with fewer elements under independence assumption, higher variability in VR results is expected. However, for human chromosomes (word length=5) G_1 and G_k present the highest variability.

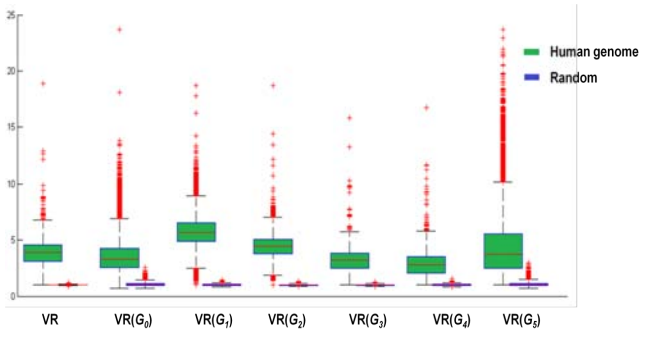

Figure 1: VR results (window size = 100 000 and word length=5.

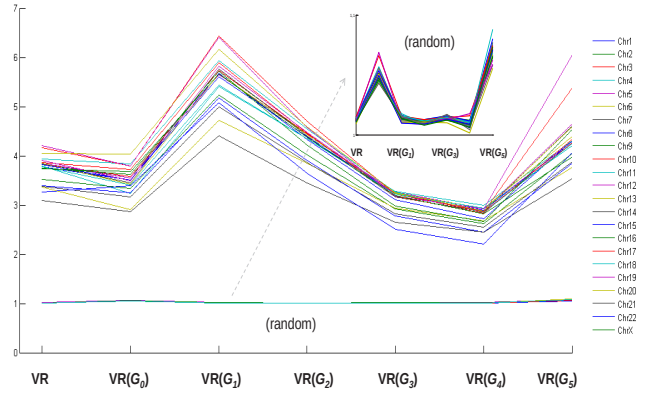

Figure 2: VR mean values organized by chromosome (window size = 100 000 and word length=5).

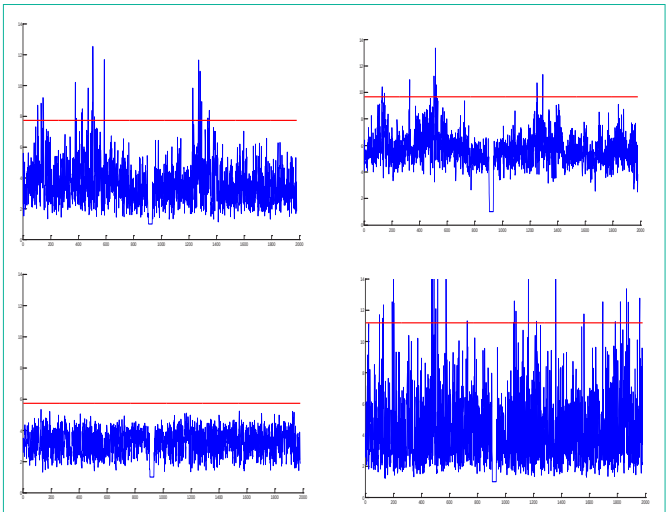

Figure 3: VR values for chromosome 3: VR(G_0), VR(G_1), VR(G_3) and VR(G_5) (window size=100 000, word length=5). Red line y=3*std + mean.

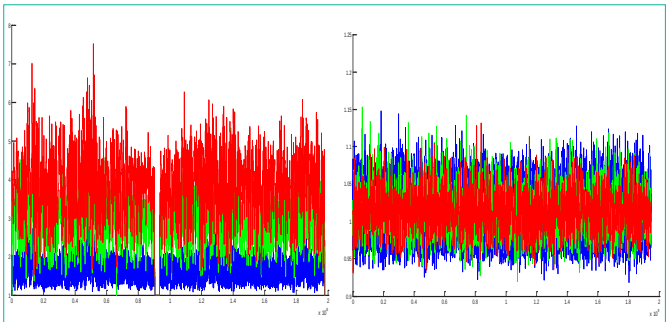

Figure 4: VR values for chromosome 3 and random, word length=5: multi-scale analysis for window sizes=10 000 (blue), 50 000 (green) and 100 000 (red).

In the human genome, there are several chromosome regions with local behavior distinct from the global.

Unlike the random sequence, the human genome exhibits increasing exceptional symmetry for larger window sizes.

References

1. Forsdyke DR. Evolutionary Bioinformatics. Springer, Berlin. 2010.
2. Qi D, Cuticchia AJ. Compositional symmetries in complete genomes. Bioinformatics. 2001; 17: 557–559.
3. Zhang S-H, Huang Y-Z. Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. Bioinformatics. 2010; 26: 478–485.
4. Afreixo V, Rodrigues JMOS, Bastos CAC. Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics. 2014; 16: 209-221.
5. Afreixo V, Rodrigues JMOS, Bastos CAC. Exceptional single strand DNA word symmetry: analysis of evolutionary potentialities. J Integrat Bioinfo. 2014; 11: 250.
6. Powdel BR, Satapathy SS, Kumar A, Jha PK, Buragohain AK, Borah M, et al. A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). DNA Research. 2009; 16: 325-343.