Analysis of single-strand exceptional word symmetry in the human genome: new measures

VERA AFREIXO*

Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
CIDMA, University of Aveiro, 3810-193 Aveiro, Portugal
IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
vera@ua.pt

JOÃO M. O. S. RODRIGUES, CARLOS A. C. BASTOS

Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal
IEETA, University of Aveiro, 3810-193 Aveiro, Portugal

SUMMARY

Some previous studies suggest the extension of Chargaff’s second rule (the phenomenon of symmetry in a single DNA strand) to long DNA words. However, in random sequences generated under an independent symbol model where complementary nucleotides have equal occurrence probabilities, we expect the phenomenon of symmetry to hold for any word length. In this work, we develop new statistical methods to measure the exceptional symmetry. Exceptional symmetry is a refinement of Chargaff’s second parity rule that highlights the words whose frequency of occurrence is similar to that of its reversed complement but dissimilar to the frequencies of occurrence of other words which contain the same number of nucleotides A or T. We analyze words of lengths up to 12 in the complete human genome and in each chromosome separately. We assess exceptional symmetry globally, by word group, and by word. We conclude that the global symmetry present in the human genome is clearly exceptional and significant. The chromosomes present distinct exceptional symmetry profiles. There are several exceptional word groups and exceptional words with a strong exceptional symmetry.

Keywords: Effect size measure; Exceptional symmetry; Hypothesis testing; Single-strand symmetry.

1. INTRODUCTION

The discovery of the double helix structure of DNA (Watson and Crick, 1953) made evident that the total percentage of complementary nucleotides (A-T and C-G) in a double-stranded molecule should be equal. This property had been previously reported by Chargaff and it is accordingly known as Chargaff’s first parity rule (Chargaff, 1950).

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
The detailed analysis of some bacterial genomes led to the formulation of Chargaff’s second parity rule, which asserts that the percentage of complementary nucleotides should also be similar in each of the two strands (Rudner and others, 1968a, 1968b; Karkas and others, 1968; Forsdyke, 2010, Chapter 4).

A natural extension of Chargaff’s second parity rule is that, in each DNA strand, the proportion of a given word (oligonucleotide) should be similar to that of its reversed complement. Chargaff’s rule and its extension have been extensively confirmed in bacterial and eukaryotic genomes, including some recent results, e.g. Qi and Cuticchia (2001), Baisnée and others (2002), Albrecht-Buehler (2007), Okamura and others (2007), Kong and others (2009), Zhang and Huang (2010), Forsdyke (2010, Chapter 4), Afreixo, Bastos and others (2013), Afreixo, Garcia and others (2013), Mascher and others (2013). However, the universality of Chargaff’s second parity rule has been questioned for organellar DNA and some viral genomes (Mitchell and Bridge, 2006). Powdel and others (2009) studied the symmetry phenomenon from an interesting new perspective, by defining and analyzing the frequency distributions of the local abundance of mono/oligonucleotides along a single strand of DNA. They found that the frequency distributions of reverse complementary mono/oligonucleotides tend to be statistically similar. This intrastrand frequency distribution parity (ISFDP) was verified in several chromosomes of bacteria, archaea, and eukaryotes, but parity violations were identified in a few strains of bacteria/archaea and in chromosomes of an eukaryote. ISFDP may be considered a refinement of Chargaff’s second parity rule, since violation of the latter implies violation of the former, but not the reverse.

If we constrain a random generator to respect Chargaff’s second parity rule (%A = %T ∧ %C = %G), then extensions of the rule to longer words (e.g. %ACC = %GGT) are expected to hold. Under this model, however, not only reversed complements will have the same prevalence. In fact, every word comprising a given number of As or Ts (e.g. AA, AT, TA, and TT) will also be equally prevalent. We say that words that satisfy such condition have equivalent composition.

A genomic word, or oligonucleotide, of length k is here interchangeably denoted as a k-mer. Moreover, we refer to the pair constituted by one word and the corresponding reversed complement as a symmetric pair.

We will analyze not only the symmetry phenomenon (similarity between symmetric pairs frequencies) but also the exceptional symmetry phenomenon. This exceptional symmetry will be evaluated by some new measures that compare the similarity between symmetric words against the similarity between equivalent composition words. We focus our study on the human genome.

2. Materials and Methods

2.1 Materials

We analyze the human genome and use the reference assembly build 37.1 available from the website of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/), discarding all ambiguous or non-sequenced nucleotides (non ACGT symbols) from the analysis. All chromosomes of the human genome were processed as separate sequences and the words were counted with overlapping (but non ACGT symbols were considered sequences separators).

2.2 Methods

Chargaff’s second parity rule states that in a single strand of DNA complementary nucleotides occur with similar frequencies. Let \( A \) be the set \{A, C, G, T\} and let \( \pi_S \) denote the occurrence probability of nucleotide or symbol \( S \in A \). Then, assuming Chargaff’s second parity rule states the equality between the occurrence of complementary nucleotides, we have:

\[
\pi_A = \pi_T \land \pi_C = \pi_G.
\]

(2.1)
We define by symmetric word pair, the set composed by one word \( w \) and the corresponding reversed complement word \( w' \), with \( (w')' = w \). Extensions of Chargaff’s second parity rule state that all symmetric words have similar occurrence frequency.

If genomic sequences were generated from independent symbols subject only to restriction (2.1), it would be expected that all symmetric words have similar occurrence frequencies. However, it would also be expected that other different words present similar occurrence frequencies (e.g. \( AACT, AGTT, AAAG \), etc.). We call equivalent composition group (ECG) to a set of words with length \( k \) which contain a given number \( m \) of nucleotides \( A \) or \( T \). For example, the EGCs for \( k = 2 \) are: \( G_0 = \{ CC, CG, GC, GG \} \); \( G_1 = \{ AC, AG, CA, CT, GA, GT, TC, TG \} \); and \( G_2 = \{ AA, AT, TA, TT \} \). The words division created by ECGs is also called binary partition (Kong and others, 2009). When all words in an ECG have similar frequencies we have a particular single-strand symmetry phenomenon that we call uniform symmetry.

For example, \( \chi^2 \) statistic denoted by \( \chi^2_{\pi} \)

\[
\chi^2_{\pi}(G_m) = \sum_{w \in G_m} \frac{(n_w - \hat{e}_w)^2}{\hat{e}_w}
\]

with \( n_w \) the total number of occurrences of word \( w \) in the sequence, \( \hat{e}_w \) is the estimated expected value for the occurrence of word \( w \) in the symmetry context (\( \pi_w = \pi_{w'} \)).

Considering the hypothesis of equality of probabilities of occurrence of symmetric words \( \pi_w = \pi_{w'} \), we estimate the expected value by \( \hat{e}_w = (n_w + n_{w'})/2 \).

Under the symmetry assumption, the statistic \( \chi^2_{\pi}(G_m) \) follows a \( \chi^2 \) distribution with d.f.,

\[
V_{\pi}(G_m) = \sqrt{\frac{\chi^2_{\pi}(G_m)}{n_m \times \text{d.f.}(G_m)}},
\]

Table 1 shows an example of two different scenarios of symmetry behavior using an ECG of dinucleotides (\( G_1^2 \)): the first scenario presents uniform symmetry, whereas the second presents exceptional symmetry.

To measure the lack of symmetry in each ECG, we can use a \( \chi^2 \) statistic denoted by \( \chi^2_{\pi} \)

\[
\chi^2_{\pi}(G_m) = \sum_{w \in G_m} \frac{(n_w - \hat{e}_w)^2}{\hat{e}_w}
\]

with \( n_w \) the total number of occurrences of word \( w \) in the sequence, \( \hat{e}_w \) is the estimated expected value for the occurrence of word \( w \) in the symmetry context (\( \pi_w = \pi_{w'} \)).

Considering the hypothesis of equality of probabilities of occurrence of symmetric words \( \pi_w = \pi_{w'} \), we estimate the expected value by \( \hat{e}_w = (n_w + n_{w'})/2 \).

Under the symmetry assumption, the statistic \( \chi^2_{\pi}(G_m) \) follows a \( \chi^2 \) distribution with d.f.,

\[
V_{\pi}(G_m) = \sqrt{\frac{\chi^2_{\pi}(G_m)}{n_m \times \text{d.f.}(G_m)}},
\]

Thus, there is no exceptional symmetry in this type of sequences.

Table 1 shows an example of two different scenarios of symmetry behavior using an ECG of dinucleotides (\( G_1^2 \)): the first scenario presents uniform symmetry, whereas the second presents exceptional symmetry.

To measure the lack of symmetry in each ECG, we can use a \( \chi^2 \) statistic denoted by \( \chi^2_{\pi} \)

\[
\chi^2_{\pi}(G_m) = \sum_{w \in G_m} \frac{(n_w - \hat{e}_w)^2}{\hat{e}_w}
\]

with \( n_w \) the total number of occurrences of word \( w \) in the sequence, \( \hat{e}_w \) is the estimated expected value for the occurrence of word \( w \) in the symmetry context (\( \pi_w = \pi_{w'} \)).

Considering the hypothesis of equality of probabilities of occurrence of symmetric words \( \pi_w = \pi_{w'} \), we estimate the expected value by \( \hat{e}_w = (n_w + n_{w'})/2 \).

Under the symmetry assumption, the statistic \( \chi^2_{\pi}(G_m) \) follows a \( \chi^2 \) distribution with d.f.,

\[
V_{\pi}(G_m) = \sqrt{\frac{\chi^2_{\pi}(G_m)}{n_m \times \text{d.f.}(G_m)}},
\]

Thus, there is no exceptional symmetry in this type of sequences.
Table 1. Example of two different scenarios of symmetry behavior using $G^2_1$: the first presents uniform symmetry and the second exceptional symmetry ($\tau = 10^{-10}$)

<table>
<thead>
<tr>
<th>Word (w) from $G_1$</th>
<th>Word absolute frequency ($n_w$)</th>
<th>Scenario 1</th>
<th>Scenario 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC</td>
<td>30</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>GT</td>
<td>30</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td>30</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>30</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>CA</td>
<td>30</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>TG</td>
<td>30</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>GA</td>
<td>30</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>TC</td>
<td>30</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>$\chi^2_s$</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$\chi^2_u$</td>
<td>0</td>
<td>66.7</td>
<td></td>
</tr>
<tr>
<td>VR</td>
<td>1</td>
<td>1154700.5</td>
<td></td>
</tr>
</tbody>
</table>

where $n_m$ is total number of occurrences of words in $G_m$. $V_s(G_m) = 0$ indicates exact symmetry. As $V_s(G_m)$ grows, the lack of symmetry increases.

In the given example (Table 1), we identify exact symmetry phenomenon for both scenarios. However, in the first scenario there is uniformity, whereas in the second there is not.

To measure the lack of uniformity in different groups, we propose another $\chi^2$ statistic

$$\chi^2_u(G_m) = \sum_{w \in G_m} \frac{(n_w - \hat{e}'_{G_m})^2}{\hat{e}'_{G_m}}$$

If we assume the hypothesis of uniformity of occurrence inside an ECG, the expected values $e'$ can be estimated by the mean occurrence frequency over all words of the $G_m$ group: $\hat{e}'_{G_m} = n_m / N_m$. In this context, the measure of lack of uniformity is given by

$$\chi^2_u(G_m) = \sum_{w \in G_m} \frac{(n_w - n_m / N_m)^2}{n_m / N_m}.$$  

Under the word independence assumption, the $\chi^2_u(G_m)$ statistic asymptotically follows a $\chi^2$ distribution with $\text{d.f.}_u(G_m) = N_m - 2$ degrees of freedom.

To evaluate the effect size of the uniformity phenomenon, we can also use Cramér’s $V$ coefficient

$$V_u(G_m) = \sqrt{\frac{\chi^2_u(G_m)}{n_m \times \text{d.f.}_u(G_m)}}.$$  

Considering the example in Table 1, the first scenario presents absolute uniformity and the second lack of uniformity.

To measure the exceptional symmetry in ECG groups, we propose the following ratio

$$R_s(G_m) = \frac{\chi^2_u(G_m) + \tau}{\chi^2_s(G_m) + \tau}.$$  

where $\tau > 0$ is a residual value to avoid an indeterminate ratio in the presence of exact uniform symmetry.
When \( R_s \gg 1 \), the ECG under study has exceptional symmetry.

In order to obtain an effect size measure able to compare the symmetry effect of all \( k \)-mers, we propose the VR ratio coefficients, based on the Cramér’s \( V \) coefficients:

\[
VR(G_m) = \frac{V_u(G_m)}{V_s(G_m)} = \sqrt{\frac{\text{d.f.}_s(G_m) \, R_s(G_m)}{\text{d.f.}_u(G_m)}} = \sqrt{\frac{R_s(G_m)}{2}}. \tag{2.3}
\]

The effect size measure \( VR(G_m) \) takes values \( \geq 0 \). If the \( VR(G_m) \) measure takes values under 1, then the observed effect is lower than what is expected when we have uniform distribution in the ECG and we cannot identify the word symmetry behavior. If \( VR(G_m) \) values are close to 1, we identify no special word symmetry behavior. Finally, if \( VR(G_m) \) takes values \( \gg 1 \), we can conclude that there is special word symmetry behavior in the \( G_m \) group.

When we measure the exceptional symmetry in the example of Table 1, the two scenarios present different results; Scenario 2 presents exceptional symmetry and Scenario 1 does not.

2.2.2 Global measures. For a global analysis of symmetry in a genomic DNA sequence, we propose the following measures:

\[
X_s^2 = \sum_{m=0}^{k} \chi_s^2(G_m).
\]

Under word independence and symmetry hypothesis, \( \chi_s^2 \) is asymptotically \( \chi^2 \) distributed with \( \text{d.f.}_s = (4^k/2) - 1 \) degrees of freedom.

To evaluate the effect size of global symmetry phenomenon, we also use Cramér’s \( V \) coefficient

\[
V_s = \sqrt{\frac{\chi_s^2}{n \times \text{d.f.}_s}} \tag{2.4}
\]

with \( n \) the total number of words of length \( k \) in the sequence.

For a global analysis of independence/uniformity, we propose the following measures:

\[
X_u^2 = \sum_{m=0}^{k} \chi_u^2(G_m).
\]

Under uniform assumption, \( \chi_u^2 \) has \( \text{d.f.}_u = \sum_{m=0}^{k} (N_m - 1) - 1 = 4^k - (k + 1) - 1 \) degrees of freedom.

To measure the symmetry of interest in a global way, we propose the following ratio

\[
R_s = \frac{\chi_u^2 + \tau}{\chi_s^2 + \tau}. \tag{2.5}
\]

We observe that \( R_s \) statistic does not depend on the sample size dimension, but depends on the degrees of freedom of \( \chi_u^2 \) and \( \chi_s^2 \). As a consequence, this measure depends indirectly on the word length.
Table 2. Simulation scenarios for the probabilities of the nucleotides. The values $\hat{\pi}_A$, $\hat{\pi}_C$, $\hat{\pi}_G$, and $\hat{\pi}_T$ correspond to the nucleotide composition in the human genome.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>$\pi_A$</th>
<th>$\pi_C$</th>
<th>$\pi_G$</th>
<th>$\pi_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.26</td>
<td>0.24</td>
<td>0.24</td>
<td>0.26</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.30</td>
<td>0.20</td>
<td>0.20</td>
<td>0.30</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.35</td>
<td>0.15</td>
<td>0.15</td>
<td>0.35</td>
</tr>
<tr>
<td>$S_5$</td>
<td>$(\hat{\pi}_A + \hat{\pi}_T)/2$</td>
<td>$(\hat{\pi}_C + \hat{\pi}_G)/2$</td>
<td>$(\hat{\pi}_C + \hat{\pi}_G)/2$</td>
<td>$(\hat{\pi}_A + \hat{\pi}_T)/2$</td>
</tr>
<tr>
<td>$S_c$</td>
<td>$\hat{\pi}_A \approx 0.2953$</td>
<td>$\hat{\pi}_C \approx 0.2044$</td>
<td>$\hat{\pi}_G \approx 0.2045$</td>
<td>$\hat{\pi}_T \approx 0.2957$</td>
</tr>
</tbody>
</table>

In order to obtain an effect size measure able to compare the symmetry effect of all $k$-mers, we create the following $V$ ratio measure

$$VR = \sqrt{\frac{d.f._s}{d.f._w}} R_s.$$  \hspace{1cm} (2.6)

In accordance with $VR$ measures for ECG, if $VR$ assumes values $\gg 1$, we conclude exceptional symmetry.

Note that, the $VR$ measure can be described as a ratio of two Cramér’s coefficients and this ratio of two effect size measures can be considered an effect size measure.

2.2.3 Word measure. To evaluate and sort words by the intensity of the exceptional symmetry phenomenon, we also create the following measure

$$R(w) = \frac{(n_w - (n_m/N_m))^2/(n_m/N_m)}{(n_w - \hat{e}_w)^2/\hat{e}_w}. \hspace{1cm} (2.7)$$

We only calculate the $R(w)$ value for non-SSW because the exceptional symmetry of SSW is naturally infinitely high.

2.2.4 Testing the statistical significance. In this subsection, we describe a Monte-Carlo technique for testing the statistical significance of the exceptional symmetry in the human genome.

In this work, we use $R_s$ as the test statistic and since the analytical expression of its distribution is not known we use a Monte-Carlo method to estimate it. Note that in this context, word independence is not verified.

Several scenarios were simulated assuming independent nucleotides with different distributions (see Table 2). In each scenario, 100 sequences with the same length of the human genome ($\sim 2.8 \times 10^9$) were generated. For scenarios $S_1$ through $S_5$, we forced $\pi_A = \pi_T$ and $\pi_C = \pi_G$ to ensure uniform symmetry.

We reject uniform symmetry, concluding exceptional symmetry, when the observed value of the $R_s$ statistic is higher than a certain critical value. The critical values, obtained using the described simulated data and considering a significance level $\alpha = 0.05$, are presented in Table 3. Note that the critical values for each $k$ vary only slightly between different scenarios.

2.2.5 Control scenario. As a control experiment, we also generated 100 random sequences assuming independence and using the human genome nucleotide composition as input, $S_c$ (see Tables 2 and 3).
Table 3. Critical values to the test based on $R_s$ for $\alpha = 0.05 : c_{0.95}$

<table>
<thead>
<tr>
<th>$k$</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
<th>$S_5$</th>
<th>$S_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.89</td>
<td>4.86</td>
<td>4.75</td>
<td>5.92</td>
<td>5.73</td>
<td>1.01</td>
</tr>
<tr>
<td>3</td>
<td>2.95</td>
<td>2.88</td>
<td>3.38</td>
<td>3.05</td>
<td>3.23</td>
<td>1.02</td>
</tr>
<tr>
<td>4</td>
<td>2.70</td>
<td>2.60</td>
<td>2.68</td>
<td>2.46</td>
<td>2.68</td>
<td>1.04</td>
</tr>
<tr>
<td>5</td>
<td>2.26</td>
<td>2.26</td>
<td>2.25</td>
<td>2.22</td>
<td>2.28</td>
<td>1.11</td>
</tr>
<tr>
<td>6</td>
<td>2.14</td>
<td>2.16</td>
<td>2.13</td>
<td>2.14</td>
<td>2.14</td>
<td>1.29</td>
</tr>
<tr>
<td>7</td>
<td>2.07</td>
<td>2.07</td>
<td>2.06</td>
<td>2.07</td>
<td>2.06</td>
<td>1.58</td>
</tr>
<tr>
<td>8</td>
<td>2.03</td>
<td>2.04</td>
<td>2.03</td>
<td>2.04</td>
<td>2.03</td>
<td>1.84</td>
</tr>
<tr>
<td>9</td>
<td>2.02</td>
<td>2.02</td>
<td>2.02</td>
<td>2.02</td>
<td>2.01</td>
<td>1.95</td>
</tr>
<tr>
<td>10</td>
<td>2.01</td>
<td>2.01</td>
<td>2.01</td>
<td>2.01</td>
<td>2.01</td>
<td>1.99</td>
</tr>
<tr>
<td>11</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
</tr>
<tr>
<td>12</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
</tr>
</tbody>
</table>

Fig. 1. Plots of $R_s$ mean values vs word length from simulation values using independence model ($S_1, S_2, S_3, S_4, S_5$, and $S_c$).

Figure 1 presents the $R_s$ mean values for the six scenarios of nucleotide probabilities, all the values are low which describes non-exceptional symmetry. The control case $S_c$ has a different $R_s$ mean profile, which is acceptable due to the differences between the frequencies of occurrence of complementary nucleotides. Moreover, in this scenario the expected probabilities of the reversed complements are not equal but there are words in an ECG (e.g. AT, TA) with equal expected probabilities.

3. Results and discussion

Although for some words ($k > 7$) the lack of symmetry is significant (Afreixo, Bastos and others, 2013), for all studied word lengths the global effect of lack of symmetry can be negligible. Figure 2 presents
Cramér’s $V_s$ coefficient for the complete human genome, $S_5$ and $S_c$ scenarios, and for all $k$-mers we obtain values <0.1 which can be considered a negligible effect (Rea and Parker, 1992). Furthermore, Figure 2 shows that random sequences generated using the uniformity hypothesis ($S_5$) have stronger symmetry effect than the real human genome. In the human genome the symmetry effect is similar to the mean effect measured in scenario $S_c$. Based on the Figure 2 results, we can conclude that the symmetry effect is high in all presented data, but the effect in the human genome is not higher than what is expected by some independent symbol models. However, the $V_s$ statistic does not allow us to conclude how exceptional the phenomenon is.

Since for $k = 1$ each ECG is composed only by a symmetric word pair, the $R_s$ statistic cannot measure exceptional symmetry. Thus, henceforth we present only the results for $k > 1$. For the other word lengths ($k \in \{2, \ldots, 12\}$), the exceptional symmetry is always significant. We can confirm in Table 4 that the observed $R_s$ values surpass all the critical values presented in Table 3.

The strength of the effect is studied with the effect size measure $VR$. Figure 3 presents the effect sizes ($VR_s$) for the complete human genome and for the 24 human chromosomes. We can observe that the highest effect size value in the human genome is obtained for $k = 5$.

Figure 3 also presents the exceptional symmetry profile, $(VR_s(k))_{k\in\{2,3,\ldots,12\}}$, for human chromosomes. And we observe that the chromosomes profiles show differences between them and also to the complete human genome profile. The maximum VR for chromosomes 3, 4, 6, 7, 8, 9, 11, 12, 18, and 19 is obtained for word length 2; the maximum VR for chromosomes 1, 2, 10, 13, 15, 17, 21, 22, and $Y$ is obtained for word length 3; the maximum VR for chromosomes 5, 16, and $X$ is obtained for word length 4; and the
maximum VR for chromosomes 14 and 20 is obtained for word length 5. These results suggest that the exceptional symmetry profiles may be used as a chromosome signature.

As we can see in Figure 3, the complete human genome profile does not follow, in general, the chromosomes profiles. In general, the maximum VR values for chromosomes are lower than the maximum VR for the complete genome. This phenomenon can result from a certain offsetting between chromosomes that leads to increased exceptional symmetry in the complete genome. For example, the words CCG and CGG in individual chromosomes have higher absolute differences of relative frequency on average (weighted average) \(5.18 \times 10^{-6}\) than the corresponding difference in the complete genome \(3.24 \times 10^{-7}\). However, there are some chromosomes (chromosomes 1, 3, 4, 6, 8, 9, 10, 11, and 12) that show higher exceptional symmetry values, for some \(k \leq 4\), than the complete human genome.

We also explore ECGs using the VR\((G_m)\) measure. Figure 4 shows 12 comparative box plots of all VR\((G_m)\) coefficients for each word length \(k\) in the human genome. We can observe that for shorter word lengths the dispersion is higher than for longer word lengths. We identify one outlier for lengths \(k \geq 8\): the ECG composed exclusively by nucleotides of type \(T_1\). For \(k = 4\), we also identify one outlier: the \(G_0\) group.

To find which ECG groups present higher exceptional symmetry, we sorted all VR\((G_m)\) values. Table 5 presents the 16 ECGs with highest VR\((G_m)\).

Table 6 shows, for each word length, the ECGs with the maximum and minimum VR\((G_m)\) values. We observe that the most exceptional ECGs for \(k \leq 6\) are composed exclusively or mostly by nucleotides of type \(T_2\) while for \(k > 6\) they are composed exclusively by nucleotides of type \(T_1\). The nucleotide composition groups with the lowest VR\((G_m)\) values do not show a clear rule.
Fig. 4. Box plots of $\text{VR}(G_m)$ coefficients organized by word length in the human genome. All outliers are represented by circles, and extreme outliers are represented by stars.

Table 5. Top 16 ECGs with the highest $\text{VR}(G_m)$ values of all ECGs under study for the complete the human genome

<table>
<thead>
<tr>
<th>ECG</th>
<th>$k$</th>
<th>VR($G_m$) highest values</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G_0$</td>
<td>4</td>
<td>2685.9</td>
</tr>
<tr>
<td>$G_0$</td>
<td>2</td>
<td>1469.8</td>
</tr>
<tr>
<td>$G_0$</td>
<td>3</td>
<td>1382.5</td>
</tr>
<tr>
<td>$G_1$</td>
<td>5</td>
<td>698.1</td>
</tr>
<tr>
<td>$G_1$</td>
<td>4</td>
<td>661.3</td>
</tr>
<tr>
<td>$G_0$</td>
<td>5</td>
<td>585.0</td>
</tr>
<tr>
<td>$G_{11}$</td>
<td>11</td>
<td>530.0</td>
</tr>
<tr>
<td>$G_1$</td>
<td>6</td>
<td>525.1</td>
</tr>
<tr>
<td>$G_{12}$</td>
<td>12</td>
<td>521.3</td>
</tr>
<tr>
<td>$G_{10}$</td>
<td>10</td>
<td>516.7</td>
</tr>
<tr>
<td>$G_9$</td>
<td>9</td>
<td>462.8</td>
</tr>
<tr>
<td>$G_1$</td>
<td>3</td>
<td>459.3</td>
</tr>
<tr>
<td>$G_2$</td>
<td>5</td>
<td>386.6</td>
</tr>
<tr>
<td>$G_8$</td>
<td>8</td>
<td>380.0</td>
</tr>
<tr>
<td>$G_2$</td>
<td>6</td>
<td>378.6</td>
</tr>
<tr>
<td>$G_7$</td>
<td>7</td>
<td>306.6</td>
</tr>
</tbody>
</table>
Table 6. ECGs, for each word length, with the maximum (and the minimum) \( V R(G_m) \) values in the human genome

<table>
<thead>
<tr>
<th>( k )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>( G_0 )</td>
<td>( G_0 )</td>
<td>( G_0 )</td>
<td>( G_1 )</td>
<td>( G_1 )</td>
<td>( G_7 )</td>
<td>( G_8 )</td>
<td>( G_9 )</td>
<td>( G_{10} )</td>
<td>( G_{11} )</td>
<td>( G_{12} )</td>
</tr>
<tr>
<td>Min</td>
<td>( G_2 )</td>
<td>( G_2 )</td>
<td>( G_3 )</td>
<td>( G_4 )</td>
<td>( G_5 )</td>
<td>( G_6 )</td>
<td>( G_6 )</td>
<td>( G_7 )</td>
<td>( G_7 )</td>
<td>( G_7 )</td>
<td>( G_0 )</td>
</tr>
</tbody>
</table>

Table 7. Percentage of words with \( R(w) \) value \( \leq 1 \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>%( R(w) \leq 1 )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.2</td>
<td>0.3</td>
</tr>
<tr>
<td>Max ( R(w) )</td>
<td>1.2E+06</td>
<td>1.3E+09</td>
<td>6.6E+08</td>
<td>1.3E+10</td>
<td>3.2E+10</td>
<td>3.8E+11</td>
<td>4.6E+10</td>
</tr>
<tr>
<td>Min ( R(w) )</td>
<td>1.0E+04</td>
<td>2.1E+03</td>
<td>3.1E+01</td>
<td>7.5E-01</td>
<td>1.3E-05</td>
<td>2.9E-05</td>
<td>3.1E-04</td>
</tr>
<tr>
<td>Highest ( R(w) )</td>
<td>GG</td>
<td>CGG</td>
<td>GGGC</td>
<td>TACGC</td>
<td>TGATTA</td>
<td>CTGTCTC</td>
<td>CCTCCTCA</td>
</tr>
<tr>
<td>2nd highest</td>
<td>CC</td>
<td>CCG</td>
<td>GCCC</td>
<td>GGCTA</td>
<td>TAATCA</td>
<td>GAGCAG</td>
<td>TGAGGAGG</td>
</tr>
<tr>
<td>3rd highest</td>
<td>CT</td>
<td>CTC</td>
<td>GAGG</td>
<td>TGGGC</td>
<td>CGGTGT</td>
<td>AAGGCA</td>
<td>TGGTGTGT</td>
</tr>
<tr>
<td>4th highest</td>
<td>AG</td>
<td>GAG</td>
<td>CTCC</td>
<td>GCCCA</td>
<td>ACACCG</td>
<td>TGTCCTT</td>
<td>ACACCCA</td>
</tr>
<tr>
<td>5th highest</td>
<td>GA</td>
<td>ATG</td>
<td>AGGC</td>
<td>GTGGC</td>
<td>CGATCC</td>
<td>GAGGTA</td>
<td>GTCCTCA</td>
</tr>
<tr>
<td>6th highest</td>
<td>TC</td>
<td>CAT</td>
<td>GGCT</td>
<td>GCCAC</td>
<td>GGATCG</td>
<td>TCACCTC</td>
<td>GTGAGAAC</td>
</tr>
</tbody>
</table>

The maximum and minimum \( R(w) \) values for each \( k \). The six words which present the highest \( R(w) \) values and the six words which present the lowest \( R(w) \) values for each \( k \).
We also carried out word analysis in the context of exceptional symmetry. The results are summarized in Table 7. Using the \( R(w) \) measure, we identify for \( k \geq 5 \) words without any exceptional symmetry (\( R(w) \leq 1 \)). For \( k = 2 \), we present all non-SSW. For the other word lengths, we present only 12 words with the 6 highest and the 6 lowest \( R(w) \) values. Note that for the high values of \( R(w) \), in Table 7, we find \( w \) and \( w' \) in consecutive positions, which means they show similar values of exceptional symmetry.

We suspect that some of these extreme words have biological interest (e.g. regulatory elements, functional elements, and motifs).

4. Conclusion

In this work, we contribute with a new method to evaluate the DNA symmetry phenomenon: the exceptional symmetry. For each word length, we propose measures of symmetry for global, ECG and word analysis.

We applied our method over the complete human genome and in the human chromosomes. We conclude that the human genome presents exceptional symmetry both through an effect size measure and through statistical testing. We showed the exceptional symmetry profiles using 12 different word lengths.

Although we verified the existence of global exceptional symmetry in the human genome, there are distinct profiles for each chromosome. Consequently, the exceptional symmetry profile may be used as a signature of each chromosome. Preliminary results also suggest that exceptional symmetry profiles are distinct between species, but this will be explored in future work.

We also studied the exceptional symmetry in each ECG, and in each word separately. We conclude that there are ECGs which are more exceptionally symmetric than others. We conclude also that a large percentage of the genomic words present some exceptional symmetry. However, for longer word lengths (\( k \geq 5 \)), there are some words without any exceptional symmetry. With this analysis, we identify that words rich in CG content and AT content behave differently in terms of exceptional symmetry. At the end of our analysis, we list a set of words with extreme behaviors, which we believe might have relevant association with biological phenomena.

5. Software

Software in the form of C and Matlab code, together with a sample input dataset and complete documentation is available on request from the corresponding author.

Acknowledgments

Conflict of Interest: None declared.

Funding

This work was supported by Portuguese funds through the CIDMA—Center for Research and Development in Mathematics and Applications, IEETA—Institute of Electronics and Telematics Engineering of Aveiro, and the Portuguese Foundation for Science and Technology (FCT—Fundação para a Ciência e a Tecnologia), within projects PEst-OE/MAT/UI4106/2014, PEst-OE/EEI/UI0127/2014 and EXPL/MAT-STA/1674/2013.

References

Exceptional DNA word symmetry


[Received April 8, 2014; revised June 21, 2014; accepted for publication July 21, 2014]