Analysis of Word Symmetries in Human Genomes Using Next-Generation Sequencing Data

Vera Afreixo[2], João M.O.S. Rodrigues[1], and Sara P. Garcia[1]

[1] Signal Processing Lab, IEETA and Department of Electronics Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal
{spgarcia,jmr}@ua.pt

[2] CIDMA - Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal
vera@ua.pt

Abstract. We investigate Chargaff's second parity rule and its extensions in the human genome, and evaluate its statistical significance. This phenomenon has been previously investigated in the reference human genome, but this sequence does not represent a proper sampling of the human population. With the 1000 genomes project, we have data from next-generation sequencing of different human individuals, constituting a sample of 1092 individuals. We explore and analyze this new type of data to evaluate the phenomenon of symmetry globally and for pairs of symmetric words.

Our methodology is based on measurements, traditional statistical tests and equivalence statistical tests using different parameters (e.g. mean, correlation coefficient).

We find that the global symmetries phenomenon is significant for word lengths smaller than 8. However, even when the global symmetry is significant, some symmetric word pairs do not present a significant positive correlation but a small or non positive correlation.

1 Introduction

Chargaff's second parity rule asserts that the percentage of complementary nucleotides should be similar in each of the two strands of a DNA sequence [11] [5] [12]. Different authors suggest and describe that there are similarities between the frequencies of words and of their inverted complements (which we call symmetry phenomenon), even for longer word lengths (e.g [10] [4] [3] [7] [14]). No previous work used genomes of several individuals from the same species to characterize the significance of the symmetry phenomenon in the species. In this work, we explore and characterize the significance of the symmetry phenomenon in the human genome, using data from multiple genomes made available by the 1,000 Genomes Project [2].

The contribution of this work is to present novel methodologies to explore the similarities between symmetric words using sequencing data obtained with next-generation methodologies. We explore the symmetry phenomenon in word lengths between 1 and 12 nucleotides.

M.S. Mohamad et al. (Eds.): *7th International Conference on PACBB*, AISC 222, pp. 7–13.
DOI: 10.1007/978-3-319-00578-2_2

2 Methods

We evaluate the symmetry phenomenon using word frequency counts. Words are interchangeably called k-mers. We study word lengths $k \in \{1,2,...,12\}$. Our sample has $n = 1092$ human genomes. For each individual, all words of lengths k were counted, and for each word length, the word (w) and its corresponding symmetric word (w') counts are paired to obtain symmetric pair counts $(N_w, N_{w'})$.

Note that, the number of distinct k-mers is 4^k. For $i \in \{1,2,...,n\}$, N_w^i is the number of times the word w appears in the genome sequence of individual i and

$$\sum_w N_w^i = \sum_{w'} N_{w'}^i = S^i.$$

The corresponding relative frequencies are represented by $f_w^i = N_w^i/S^i$ and $f_{w'}^i = N_{w'}^i/S^i$.

2.1 Statistical Hypothesis Testing

Traditional statistical hypothesis testing may be used to assess differences. However, it is well known that when traditional hypothesis tests are applied to large data sets, any small effect is always deemed significant [9] [6] [8]. Furthermore, we want to evaluate if there are similarities, not differences, between the occurrence of symmetric words. To overcome this drawback, we use equivalence tests for accepting/rejecting the equivalence between symmetric words.

We studied the equivalence between pairs of symmetric words (w, w') using the ratio of the frequency of the symmetric pair R_w and a practical tolerance δ (> 1) , and concluding the equivalence when $1/\delta < R_w < \delta$. Let μ_{R_w} denote the (population) mean of the ratio of the w word frequency and its corresponding reversed complement word frequency (ratio of the frequency of the symmetric pair). Let $\overline{R_w}$ denote the corresponding sample mean and for each individual i the ratio is given by $R_w^i = f_w^i/f_{w'}^i$.

The statistical hypotheses for the equivalence test are:

$$H_{0_w} : \mu_{R_w} \geq \delta \text{ or } \mu_{R_w} \leq 1/\delta \text{ vs } H_{1_w} : 1/\delta < \mu_{R_w} < \delta.$$

The ratio between two frequencies, $r_w^i = f_w^i/f_{w'}^i$, is an effect size measure. As in many studies, e.g. [13], we consider the effect to be weak when it assumes values between 1.1 and 1.3 and we explore these lower effect size values as a tolerance to conclude practical equivalence. When the sample size is high, by the central limit theorem, we use the z interval for the unknown true value of μ_{R_w}, which is,

$$\left(\overline{R_w} \mp z \times SE(R_w)\right)$$

where $SE(R_w) = S_{R_w}/\sqrt{n}$.

In this case, the equivalence tests procedure consists of obtaining the confidence interval for the parameter and checking if it is contained in the interval $]1/\delta, \delta[$. If so, H_{0_w} is rejected and for the (w, w') pair, the equivalence can be assumed.

For each word length k, we construct 4^k equivalence tests. When we reject all of the 4^k null hypotheses, we consider that the symmetry phenomenon is present, as all symmetric pairs are equivalent in a global way. Since we conduct simultaneous tests, we apply the Bonferroni correction.

2.2 Correlation

We use the Pearson's correlation to measure the global symmetry effect in each individual. In particular, we use the coefficient as a score of global symmetry in each individual

$$SS_i = \frac{\sum_w \left[(N_w^i - \overline{N^i})(N_{w'}^i - \overline{N^i}) \right]}{\sum_w \left[(N_w^i - \overline{N^i})^2 \right] \sum_w \left[(N_{w'}^i - \overline{N^i})^2 \right]} \tag{1}$$

$i \in \{1,2,...,n\}$.

To evaluate the significance of correlation between a pair of symmetric words, we apply the one tailed Pearson correlation test. Considering ρ the Pearson correlation parameter, the tests hypothesis are:

$$H_0 : \rho = 0 \text{ vs } H_1 : \rho > 0.$$

with $T = c\sqrt{\frac{n-2}{1-c^2}} \underset{under\ H_0}{\sim} t_{n-2}$ and c the sample Pearson correlation coefficient.

2.3 Generating 1,092 Individual Human Genomes

We use the GRCh37.1 reference human genome assembly [2] and version 3 (March 16, 2012) of a Phase 1 integrated variant call set based on both low coverage and exome whole genome sequencing data from 1,092 individuals [1]. The VCF files contain the alterations necessary to incorporate in the reference human genome in order to obtain a different, individual human genome. We developed a package of custom-made C programs to generate alternate FASTA genomes from population sequencing VCF data, and to count occurrences of words from these individual genomes.

3 Results

We apply traditional statistical tests to compare the means of the occurrences of symmetric words. As expected, globally, there are significant differences for all word lengths. Also, for each word length, almost all pairs have significant differences.

As discussed in the methods section, we use equivalence testing in this analysis. Table 1 displays the percentage of equivalent pairs (in the sense of what has been described in subsection 2.1) for each k-mer length and each tolerance value (δ). We verify equivalence between symmetric pairs for $k \leq 7$, for both tolerance values $\delta = 1.1$ and $\delta = 1.3$.

Figure 1 displays an error bar plot of the global scores of symmetry (SS, equation 1). We observe high score values (close to 1) for all word lengths $k \in \{1,2,...,12\}$. Note that, though all global scores of symmetry have high values, these might be attributable to the contribution of a few outliers. In this figure, we observe a high association between k and the scores (approximately parabolic behavior, concavity down with

Table 1. Percentage of equivalent tests that reject the null hypothesis

k	$\delta = 1.1$	$\delta = 1.3$
1	100	100
2	100	100
3	100	100
4	100	100
5	100	100
6	100	100
7	100	100
8	99.67	100
9	96.60	99.82
10	85.17	97.98
11	64.84	89.50
12	40.20	71.52

inflection point in $k = 4$). Hence, the global symmetry score has high values in the human genome. However, this score has a tendency to decrease as the word length increases.

For each word length k, there are 4^k pairs of symmetric words. The correlation coefficient and the corresponding statistical test p-value are obtained for each symmetric word pair based on a sample of 1092 individuals. Table 2 displays the frequency table of the correlation coefficients, highlighting the corresponding conclusion of t correlation tests. We observe that, for $k \leq 4$, the result is in accordance with the previous conclusion. However, for $k > 4$, the correlation values are very low. We also observe some not significantly positive correlations.

Table 2. Percentage of correlation coefficients in each class of effect size. *the p-value of one tailed Pearson correlation test is <0.05.

Correlation	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8	k=9	k=10	k=11	k=12
$[-1;0[$	0	0	0	0	0.2	2.0	8.9	21.3	37.8	46.1	48.7	49.5
$[0;0.05[$	0	0	0	0	0	3.1	7.3	18.3	26.4	27.1	27.3	28.3
$[0.05;0.10[*$	0	0	0	0	0.4	2.7	6.6	18.6	18.2	15.5	14.5	13.8
$[0.10;0.30[*$	0	0	0	0	2.5	10.0	31.4	35.1	16.2	10.7	9.2	8.3
$[0.30;0.50[*$	0	0	0	0	5.5	8.8	31.1	4.8	1.0	0.3	0.2	0.1
$[0.50;1]*$	100	100	100	100	91.4	73.4	14.6	2.0	0.4	0.2	0.0	0.0

To clarify the obtained correlation values/tests for each symmetric pair, we compute, for each k, the mean, minimum, maximum and standard deviation of the correlation. Figure 2 displays these statistics for all words of length k. We observe a curious tendency that as k increases, the mean of the correlation tends to zero. Hence, the previous high values of the global symmetry score may be a consequence of only a few pairs of very frequent symmetric words.

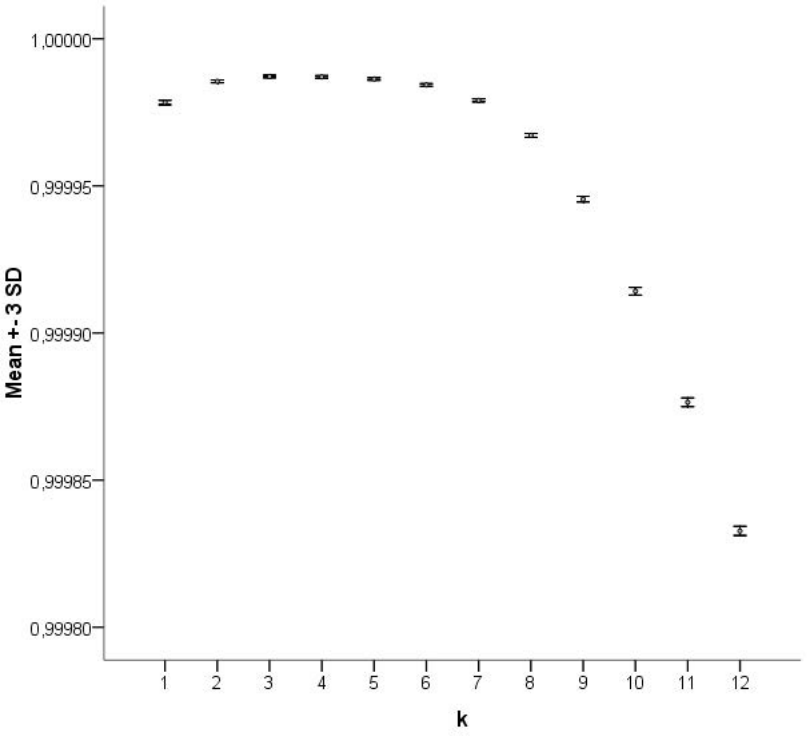

Fig. 1. Error bar of the scores of symmetry (*SS*) in 1092 human genomes

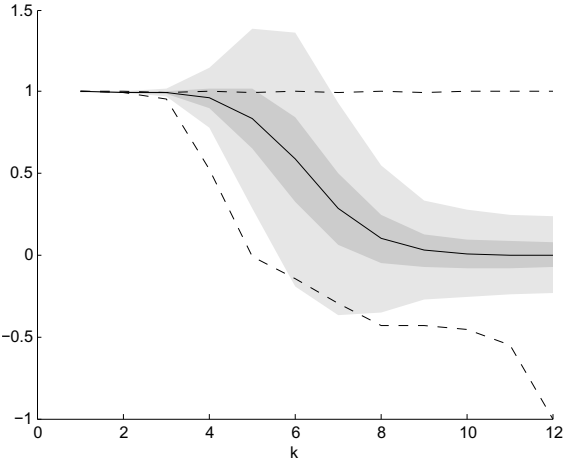

Fig. 2. Summary of statistics of the correlation coefficients between pairs of symmetry words in 1092 human genomes: mean (continuous line), minimum (dashed), maximum (dashed). The shaded region represents the standard deviation around the mean (mean $\mp$ standard deviation in the darker gray region and mean $\mp$ 3 standard deviations in the lighter gray region.

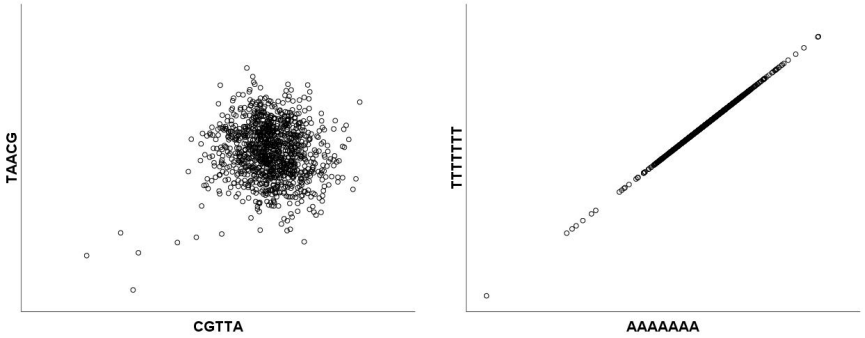

Fig. 3. Left: Scatter plot of the frequencies of the (CGTTA, TAACG) symmetric pair, with $r = -0.008$ and p value 0.790. Right: Scatter plot of the frequencies of the(AAAAAAA, TTTTTTTT) symmetric pair, with $r = 0,870$ and p value < 0.001.

For k=5, the symmetric pair (CGTTA, TAACG) is the single pair responsible for the hypothesis test not rejecting the null hypothesis. Moreover, there are 8.6% pairs where the correlation is not strong (Table 2). For $k > 5$, there are many more pairs responsible for the non rejection of the null hypothesis. The percentage of not strongly correlated pairs also increases. The left panel of Figure 3 shows a scatter plot for the symmetric word pair (CGTTA, TAACG), which is the single pair, in the set of words of length 5, that does not present significant positive correlation.

However, for all $k \in \{1,2,...,12\}$, there are several pairs of symmetric words where the correlation is significantly positive and strong. An example is displayed in the right panel of Figure 3, representing the word symmetric pair (AAAAAAA, TTTTTTT).

4 Conclusion and Future Work

Here, we studied the word symmetry phenomenon, characterized through word frequencies, in 1092 human genomes. We confirmed the global tendency of the symmetry phenomenon using equivalence tests and a global score of symmetry.

We identified an interval of word lengths where the global symmetry phenomenon tendency starts to be non significant. Whereas the global score of symmetry has high values for all word lengths investigated, the equivalence tests show a breakdown of symmetry for $k > 7$.

When we looked into symmetric word pairs, we identified several indicators of lack of similarity between the occurrences of symmetric words even when the global symmetry phenomenon tendency was present. This is surprising, as no symmetry breakdown is expected for random sequences with equal frequencies of complementary nucleotides (i.e. $N_A \sim N_T$ and $N_C \sim N_G$). We conclude that the symmetry phenomenon is less prevalent in human genomes than previously thought. It will be interesting to investigate this symmetry phenomenon for selected genomic regions.

5 Funding

This work was supported in part by *FEDER* funds through *COMPETE*– Operational Programme Factors of Competitiveness ("Programa Operacional Factores de Competitividade") and by Portuguese funds through the *Center for Research and Development in Mathematics and Applications* and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-C/MAT/UI4106/2011 with COMPETE number FCOMP-01-0124-FEDER- 022690. SPG acknowledges funding from the European Social Fund and the Portuguese Ministry of Education and Science, and from project FCOMP-01-0124-FEDER-010095 of Portuguese Science Foundation.

References

1. The 1000 genomes project data release: Integrated variant call set for phase 1, version 3
2. Grch37 Reference human genome assembly
3. Albrecht-Buehler, G.: Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. Genomics 90, 297–305 (2007)
4. Baisnée, P.-F., Hampson, S., Baldi, P.: Why are complementary DNA strands symmetric? Bioinformatics 18(8), 1021–1033 (2002)
5. Karkas, J.D., Rudner, R., Chargaff, E.: Separation of B. subtilis DNA into complementary strands. II. template functions and composition as determined by transcription with RNA polymerase. Proceedings of the National Academy of Sciences of the United States of America 60(3), 915–920 (1968)
6. Kline, R.B.: Beyond Significance testing: Reforming Data Analysis Methods in Behavioral Research. American Psychological Association (2004)
7. Kong, S.-G., Fan, W.-L., Chen, H.-D., Hsu, Z.-T., Zhou, N., Zheng, B., Lee, H.-C.: Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS One 4(11), 7553 (2009)
8. Migliorati, S., Ongaro, A.: Adjusting p-values when n is large in the presence of nuisance parameters. In: Statistics for Industry and Technology, Vienna, pp. 305–318 (September 2010)
9. Moore, D.S.: Statistics: Concepts and Controversies, 4th edn. Freeman (1997)
10. Qi, D., Jamie Cuticchia, A.: Compositional symmetries in complete genomes. Bioinformatics 17(6), 557–559 (2001)
11. Rudner, R., Karkas, J.D., Chargaff, E.: Separation of B. subtilis DNA into complementary strands, I. biological properties. Proceedings of the National Academy of Sciences of the United States of America 60(2), 630–635 (1968)
12. Rudner, R., Karkas, J.D., Chargaff, E.: Separation of B. subtilis DNA into complementary strands. III. direct analysis. Proceedings of the National Academy of Sciences of the United States of America 60(3), 921–922 (1968)
13. Thanassoulis, G., Vasan, R.S.: Genetic cardiovascular risk prediction — Will we get there? Circulation 122(22), 2323–2334 (2010)
14. Zhang, S.-H., Huang, Y.-Z.: Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. Bioinformatics 26(4), 478–485 (2010)