Exceptional Symmetry Profile: A Genomic Word Analysis

Vera Afreixo, João M.O.S. Rodrigues, Carlos A.C. Bastos and Raquel M. Silva

Abstract The extension of Chargaff's second rule, also called the DNA symmetry, is pointed as an universal law present in the genomes of species. Previously, a measure of the symmetry above that expected in independence contexts (exceptional symmetry) was proposed to evaluate the phenomenon globally. The global exceptional symmetry was found in several species. However, the analysis of exceptional symmetry by word was not studied in detail. In this work a new exceptional symmetry measure is proposed to evaluate the exceptional symmetry effect by symmetric word pair. We develop a detailed study of the exceptional symmetry by symmetric pairs. We also discuss the exceptional symmetry by symmetric word pair for several organisms: 7 viruses; 5 archaea; 5 bacteria; 14 eukaryotes.

Keywords Chargaff's second parity rule · Exceptional symmetry · Genomic word counts

1 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 [5].

V. Afreixo(✉) · R.M. Silva
iBiMED-Institute of Biomedicine, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal
e-mail: vera@ua.pt

V. Afreixo · J.M.O.S. Rodrigues · C.A.C. Bastos · R.M. Silva
IEETA-Institute of Electronic Engineering and Informatics of Aveiro, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

V. Afreixo
Department of Mathematics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

J.M.O.S. Rodrigues · C.A.C. Bastos
Department of Electronics, Telecommunications and Informatics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

© Springer International Publishing Switzerland 2016

M.S. Mohamad et al. (eds.), *10th International Conference on PACBB*,
Advances in Intelligent Systems and Computing 477,
DOI: 10.1007/978-3-319-40126-3_16

Chargaff's second parity rule states that those quantities are almost equal in a single strand of DNA [8, 10, 11], and this phenomenon holds in almost all living organisms.

The extension to the second parity rule is also known as single strand symmetry phenomenon. The single strand symmetry states that, in each DNA strand, the proportion of an oligonucleotide should be similar to that of its reversed complement [6]. There is no knowledge about why the parity is needed in the DNA sequence and there is no consensual explanation for the occurrence of the single strand phenomenon. There are some attempts to explain the phenomenon related with the species evolution process, for example: stem-loops hypothesis [7]; duplication followed by inversion hypothesis [4]; inversions and inverted transpositions hypothesis [3].

Powdel and others [9] studied the symmetry phenomenon in non-overlapping regions of DNA of a specific size. They analysed 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 tends to be statistically similar. Afreixo et al. [2] 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.

Based on the exceptional symmetry concept, we discuss this effect size measure for each symmetric word pairs, and we also propose a new measure to evaluate the dissimilarity between word occurrences in relation to the word dissimilarities in the corresponding equivalent composition group. We obtain the word symmetry effects for 31 complete genomes. Our results show that the symmetry effect value has the potential to discriminate between species groups. And there are sets of words which present high symmetry effect in all species under study, and we also indicate several species specificities.

2 Methods

In this study, we used the complete DNA sequences of 31 organisms obtained from the National Center for Biotechnology Information (NCBI; ftp://ftp.ncbi.nih.gov/genomes/). The species used in this work are listed in Tab. 1. The study was carried out in representative species of the major taxonomic groups across the tree of life and includes genomes from vertebrates, invertebrates, protozoans, fungi, plants, bacteria (gram-positive and gram-negative), archaea and virus (DNA and RNA viruses). The study could be extended to other organisms to improve the resolution of the species tree, but additional genomes will eventually become redundant.

All genome sequences used under this study were processed to obtain the word counts, considering overlap between successive words. We obtained the word counts for word lengths from 1 to 12 nucleotides.

We proposed in a previous work [2] the exceptional genomic word symmetry for equivalent composition groups (ECG) and globally. Some words are equal to

Table 1 List of organisms whose DNA was used in this study.

Organism	Group	Abbreviation
Homo sapiens[a]	eucarya (animalia)	HSap
Macaca mulatta[a]	eucarya (animalia)	MacM
Pan trogolodytes[a]	eucarya (animalia)	PanT
Mus musculus[a]	eucarya (animalia)	MusM
Rattus norvegicus[a]	eucarya (animalia)	RatN
Danio rerio[a]	eucarya (animalia)	DRer
Apis mellifera[a]	eucarya (animalia)	Apis
Caenorhabditis elegans[b]	eucarya (animalia)	CaeE
Drosophila melanogaster[b]	eucarya (animalia)	DroM
Arabidopsis thaliana[a]	eucarya (plantae)	AraT
Vitis vinifera[a]	eucarya (plantae)	VitV
Saccharomyces cerevisiae[a]	eucarya (fungi)	SacC
Candida albicans[a]	eucarya (fungi)	CanA
Plasmodium falciparum[a]	eucarya (protozoa)	PlaF
Helicobacter pylori[a]	bacteria	HelP
Streptococcus mutans GS[a]	bacteria	StMG
Streptococcus mutans LJ23[a]	bacteria	StML
Streptococcus pneumoniae[a]	bacteria	StPn
Escherichia coli[a]	bacteria	EscC
Aeropyrum camini[a]	archaea	AerC
Aeropyrum pernix[a]	archaea	AerP
Caldisphaera lagunensis[a]	archaea	CalL
Candidatus Korarchaeum[a]	archaea	CanK
Nanoarchaeum equitans[a]	archaea	NanE
NC001341[a]	virus	AbaS
NC001447[a]	virus	AcaT
NC004290[a]	virus	AchD
NC008724[a]	virus	AcPL
NC011646[a]	virus	AcPM
NC011591[b]	virus	SouT
NC012532[b]	virus	ZikV

[a] Downloaded in January 2014. [b] Downloaded in March 2016.

their reversed complement, we denote these as self symmetric words (SSW). We also define a symmetric word pair as the set composed by one word w and the corresponding reversed complement word w', with $(w')' = w$ (for example, CCA and TGG make a symmetric word pair). Let n_w be the total number of occurrences of word w in the sequence, n_m be the total number of occurrences of words in the ECG G_m which contain words composed by m nucleotides A or T.

The measure R was proposed to evaluate and sort words by the intensity of the exceptional symmetry phenomenon [2]. For $w \in G_m = \{w_1, w_2, w_3, ..., w_L\}$,

$$R(w) = \left(\frac{\left(n_w - \frac{n_m}{L}\right)^2}{\frac{n_m}{L}} \right) / \left(\frac{(n_w - n_{w'})^2}{2(n_w + n_{w'})} \right), \tag{1}$$

with L the number of different words in G_m.

One disadvantage of the R measure is related to the unequal evaluation of dissimilarities between symmetric word pairs inside a ECG. If we have two pairs of words in the same ECG with identical dissimilarities between their occurrence frequencies, the R values could present distinct symmetry effects.

Consider for example $G_m = \{w_1, w_2, w_3, w'_1, w'_2, w'_3\}$, with $n_{w_1} = n_{w'_1} + 1 = 20$, $n_{w_2} = n_{w'_2} + 1 = 9$ and $n_{w_3} = n_{w'_3} + 1 = 1$. All symmetric word pairs present identical dissimilarities. The average frequency is $\frac{n_m}{L} = 9.5$ and the R values are

Table 2 Example to elucidate the differences between R and S measures.

words (w)	w_1	w'_1	w_2	w'_2	w_3	w'_3
n_w	20	19	9	8	0	1
R	905.2	741.0	0.9	8.1	19.0	15.2
$\ln(R)$	6.8	6.6	-0.1	2.1	2.9	2.7
S	2.5	2.5	2.5	2.5	2.5	2.5

presented in Tab. 2. The symmetric word pair with occurrences nearest to the average numbers of occurrences in the ECG is considered by the R measure to be less exceptional than the word pair whose number of occurrences is most distant. So, the R measure may be inadequate to sort the genomic words by exceptional symmetry. In order to avoid this disadvantage we introduce, in this study, a new measure S, the symmetric word pair effect,

$$S(w) = \ln \frac{\sqrt{\frac{\sum_{i=1}^{L}\sum_{j=1}^{L}(n_{w_i} - n_{w_j})^2}{L^2 - L}}}{|n_w - n_{w'}|}. \tag{2}$$

We can observe that in the numerator of the S measure we have the global deviation between G_m words. However, the numerator of the R measure is the deviation to the mean of the number of occurrence in G_m. Additionally, in the S measure the effect for both words of one symmetry word pair is the same, while the R measure can produce distinct values for each word of the symmetric pair. Table 2 presents the S values for the example discussed previously, and as expected, all words in this word group show equal exceptional symmetry effect.

We calculate the $S(w)$ value only for non SSW, because the exceptional symmetry of SSW is naturally infinitely high. When $n_w = n'_w$, we obtain $S(w) = \infty$. In our analysis the infinity is replaced, when necessary, by the double of the maximum effect obtained for the other words of the same length in the same species symbolizing a high exceptional symmetry value.

2.1 Control Experiments

In order to evaluate the exceptional symmetry in each word we generate random sequences under second parity rule validity assumption, i.e., using the same composition for complementary nucleotides. We generate the nucleotide sequences assuming nucleotide independence. In this scenario the expected probabilities of the words in each ECG are the same (see details in [1]). We denote these random sequences by *sym*.

2.2 Word Analysis Procedure

To identify a symmetric word pair as exceptional we compare the S values with the critical values obtained by the control experiments. To find words with very exceptional symmetry we use the third quartile has a cutoff threshold.

To compare genomes we use hierarchical clustering. We use a hierarchical bi-clustering procedure to compare both genomes and word S values, simultaneously. Hierarchical clustering was obtained using the UPGMA aggregation criterion with Euclidean distance.

3 Results and Discussion

We consider all genomic words with lengths k ($k \in \{1, ..., 12\}$) and a set of 31 genomes, and we obtain the symmetric word pair effect for each genomic word. As obvious result for $k = 1$, $S(w) = 1$ for all nucleotides.

Table 3 Percentage of words with exceptional symmetry effect ($S > 0$).

k (%)	2	3	4	5	6	7	8	9	10	11	12
HSap	100	100	100	100	100	100	100	100	100	100	100
MacM	100	100	100	100	100	100	100	100	100	100	100
PanT	100	100	100	100	100	100	100	100	100	100	100
MusM	100	100	100	100	100	100	100	100	100	100	100
RatN	100	100	100	100	100	100	100	100	100	100	100
DRer	100	100	100	100	99	99	98	98	99	100	100
Apis	100	100	100	100	100	100	100	51	11	79	99
CaeE	100	100	100	100	100	100	100	100	100	100	100
DroM	100	100	100	100	100	100	100	100	100	100	100
AraT	100	100	100	100	100	100	100	100	100	100	100
VitV	100	100	100	100	100	100	100	100	100	100	100
SacC	100	100	100	100	100	100	100	99	99	100	100
CanA	100	100	100	100	100	100	98	98	99	100	100
PlaF	100	100	100	100	100	100	100	99	100	100	100
HelP	100	100	100	100	100	100	100	99	99	100	100
StMG	100	100	100	100	100	100	99	97	98	100	100
StML	100	100	100	100	100	100	99	96	98	100	100
StPn	100	100	100	100	100	100	99	97	98	100	100
EscC	100	100	100	100	100	100	100	100	100	100	100
AerC	100	100	100	100	100	100	99	98	99	100	100
AerP	100	100	100	100	100	100	99	98	99	100	100
CalL	100	100	100	100	100	100	98	98	99	100	100
CanK	100	100	100	100	100	100	100	99	99	100	100
NanE	100	100	100	100	100	99	95	95	99	100	100
AbaS	100	97	99	98	91	82	80	91	99	100	100
AcaT	100	100	100	99	98	94	88	88	96	100	100
AchD	63	75	81	77	78	73	87	98	100	100	100
AcPL	63	69	78	78	73	74	84	95	99	100	100
AcPM	50	66	70	71	78	80	90	98	100	100	100
SouT	63	63	71	75	76	70	90	99	100	100	100
ZikV	63	78	83	83	81	79	76	97	100	100	100
sym	63	72	75	73	70	69	69	84	97	100	100

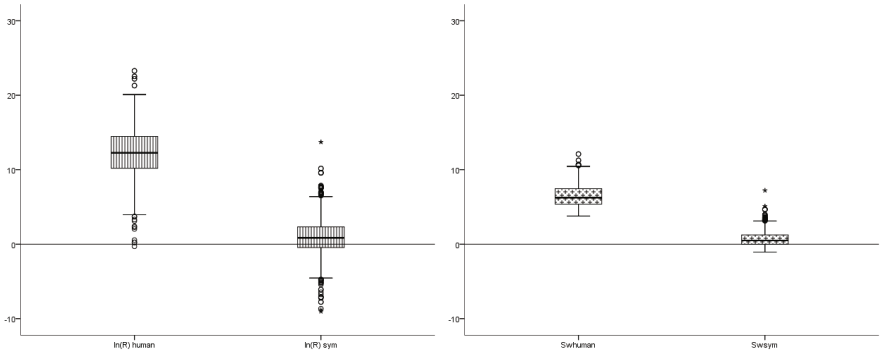

Fig. 1 Boxplots for S (left) and $\log R$ (right) values of human genome and random sequence example (*sym*) for word length 5.

Almost all words in eukaryote genomes show significant exceptional symmetry effect (comparing with the critical values obtained by simulation). Most of the genomic words have some degree of exceptional symmetry $S > 0$. Table 3 shows the percentage of words with $S > 0$ for each species and word length of this study. The viruses present a high percentage of words without exceptional symmetry. However, this result is expectable, since in a previous work, using a global measure of symmetry for all word of a fixed word length [1], it was concluded that some viruses do not have significant exceptional symmetry. Note that, with the increase of the word length some species reveal several symmetric word pairs where both elements have no occurrences. We considered these cases as having exceptional symmetry effect. This option artificially increases the percentage of words with exceptional symmetry in longer words for the species with shorter genomes.

To show the differences between real genomes and the random sequences generated under the second parity rule assumption, Table 3 includes the *sym* row corresponding to one control scenario (sequence with length equal to the length of human genome). Figure 1 shows the S (left) and $\ln R$ (right) values for $k = 5$ in the human genome and boxplots for the corresponding random realization *sym*. The boxplot for the human genome shows high symmetric word pair effects in both measures. The right outliers detected in the human S boxplot are: (GCGTA, TACGC), (ACCGG, CCGGT), (GCCAC, GTGGC), (GCCCA, TGGGC), (CGGGA, TCCCG). And the right outliers detected in the human boxplot of $\ln R(w)$ values are: $TACGC$, $GCGTA$, $TGGGC$, $GCCCA$, $GTGGC$, $GCCAC$. We note that, in this genome, the $R(w)$ outliers are a subset of the S outliers. The S and $\ln R$ measures differ the most in the detection of non symmetric words: S shows exceptional symmetry for every length-5 word, whereas $\ln R$ detects some word pairs with no exceptional symmetry. Due to the R disadvantages discussed in the Methods section we consider S measure results as more reliable.

Figure 2 shows the dendrogram obtained using the UPGMA aggregation criteria with Euclidean distance for $k = 5$. We observe three distinct groups: mammalian

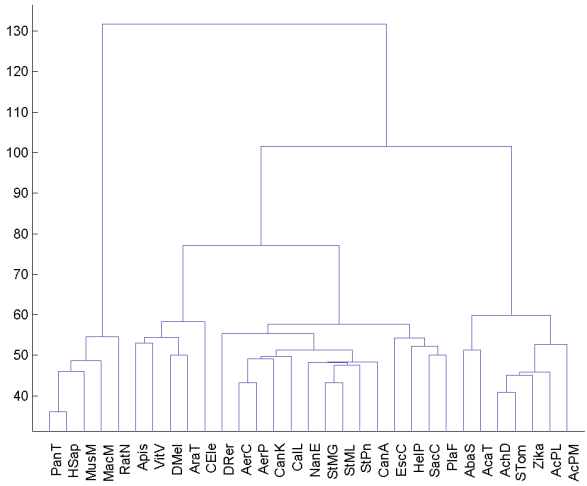

Fig. 2 Dendrogram of S values for all species under study word length 5.

species, viruses and the other species. The dendrograms obtained for other word lengths essentially maintain the same structure (the dendrogram for $k = 3$ is also included in Figure 3). Figure 3 shows the colormap with biclustering organization for $k = 3$. The cluster for the species highlights the viruses and a subgroup of animals. The symmetric word pair effect is stronger on the left side of the colormap and weaker on the right side. The word cluster highlights the group compound by two symmetric word pairs: (CCG, CGG), (GCG, CGC). We analysed the words with

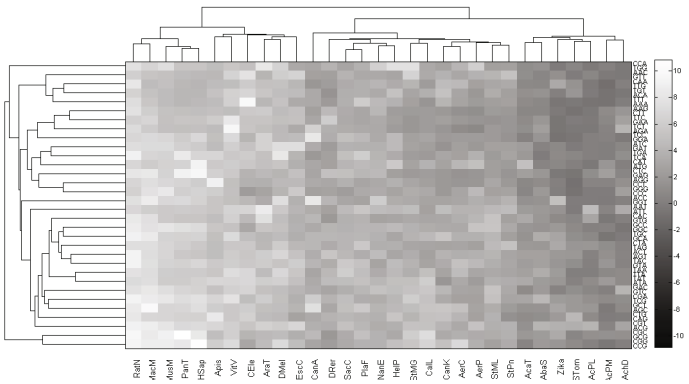

Fig. 3 Colormap with biclustering organization of S values for all species under study and word length 3.

most exceptional symmetry in all species under study, the words with exceptional symmetry effect above the third quartile. Above the third quartile we observe no common word for all the species under analysis. However, we observe some common words in animals group. The strongest symmetric word pair effect is observed in words composed by CpG dinucleotides.

4 Conclusions

We proposed a new measure to evaluate the exceptional symmetry effect. The exceptional symmetry values seem to contain information about the species evolution. The eukaryote group showed the highest exceptional symmetry in this study and the mammalian species has very high exceptional symmetry values distinct from all other species under study.

All cellular organisms under study present high percentages of words with exceptional symmetry effect. We conjecture that exceptional symmetry is an universal law of cellular organisms, but the most exceptional symmetric words are species specific.

We reinforce that some viruses show a behavior opposite to exceptional symmetry in almost all words under study ($S < 0$).

Acknowledgment This work was supported by Portuguese funds through the iBiMED - Institute of Biomedicine, 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: UID/BIM/04501/2013 and PEst-OE/EEI/UI0127/2014.

References

1. Afreixo, V., Rodrigues, J.M.O.S., Bastos, C.A.C.: Exceptional single strand DNA word symmetry: analysis of evolutionary potentialities. Journal of Integrative Bioinformatics **11**(3), 250 (2014)
2. Afreixo, V., Rodrigues, J.M.O.S., Bastos, C.A.C.: Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics **16**(2), 209–221 (2015)
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. Chargaff, E.: Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia **6**(6), 201–209 (1950)
6. Forsdyke, D.R.: Evolutionary Bioinformatics. Springer, New York (2011)
7. Forsdyke, D.R., Bell, S.J.: Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Applied Bioinformatics **3**(1), 3–8 (2004)
8. 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)

9. Powdel, B.R., Satapathy, S.S., Kumar, A., Jha, P.K., Buragohain, A.K., Borah, M., Ray, S.K.: A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (chargaff's second parity rule). DNA Research **16**, 325–343 (2009)
10. 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)
11. 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)