Interdiscip Sci Comput Life Sci (2017) 9:14–23
DOI 10.1007/s12539-016-0200-9

ORIGINAL RESEARCH ARTICLE

Exceptional Symmetry by Genomic Word

A Statistical Analysis

Vera Afreixo[1] · João M. O. S. Rodrigues[2] · Carlos A. C. Bastos[2] · Ana H. M. P. Tavares[3] · Raquel M. Silva[4]

Received: 20 July 2016 / Revised: 2 November 2016 / Accepted: 4 November 2016 / Published online: 19 November 2016

Abstract Single-strand DNA symmetry is pointed as a universal law observed in the genomes from all living organisms. It is a somewhat broadly defined concept, which has been refined into some more specific measurable effects. Here we discuss the exceptional symmetry effect. Exceptional symmetry is the symmetry effect beyond that expected in independence contexts, and it can be measured for each word, for each equivalent composition group, or globally, combining the effects of all possible words of a given length. Global exceptional symmetry was found in several species, but there are genomic words with no exceptional symmetry effect, whereas others show a very high exceptional symmetry effect. In this work, we discuss a measure to evaluate the exceptional symmetry effect by symmetric word pair, and compare it with others. We present a detailed study of the exceptional symmetry by symmetric pairs and take the CG content into account. We also introduce and discuss the exceptional symmetry profile for the DNA of each organism, and we perform a multiple comparison for 31 genomes: 7 viruses; 5 archaea; 5 bacteria; 14 eukaryotes.

Keywords Single-strand symmetry · Exceptional symmetry · Multiple organism comparison · Genomic word analysis

✉ Vera Afreixo
vera@ua.pt

[1] iBiMED-Institute of Biomedicine, IEETA-Institute of Electronic Engineering and Informatics of Aveiro, CIDMA-Center for Research and Development in Mathematics and Applications, Department of Mathematics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[2] IEETA-Institute of Electronic Engineering and Informatics of Aveiro, Department of Electronics, Telecommunications and Informatics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[3] iBiMED-Institute of Biomedicine, Department of Mathematics, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

[4] iBiMED-Institute of Biomedicine, IEETA-Institute of Electronic Engineering and Informatics of Aveiro, Department of Medical Sciences, University of Aveiro, Campus Universitário de Santiago, Aveiro, Portugal

1 Introduction

Erwin Chargaff was a biochemist that discovered a set of intriguing rules about the composition of DNA from the analysis of bacterial genomes [1]. The first rule states that the total percentage of complementary nucleotides (A-T and C-G) in double-stranded DNA must be equal. Of course, this is now known to result from the double helix structure of DNA [2]. The second rule sates that the percentage of complementary nucleotides is also identical in each strand [3–5], [6, chap. 4].

A natural extension of Chargaff's second parity rule is that, in each DNA strand, the number of occurrences of a given word (oligonucleotide or *k*-mer) should match that of its reversed complement [6]. The extension to the second parity rule is also known as the single-strand symmetry phenomenon. This symmetry phenomenon refers to the distributions of symmetric pairs, i.e., the distribution of occurrences of all words and the distribution of occurrences of the corresponding reversed complements.

Presently, there is not a generally accepted justification for the need of single-strand parity in DNA sequences, and there is no consensual explanation for the occurrence of the single-strand phenomenon. There are some attempts to explain the phenomenon, which could be classified in two

groups: the conserved patterns model [7–9], and the evolutive models. Evolutive models can further be classified according to several underlying hypothesis, for example: the stem-loops hypothesis [10]; the duplication followed by inversion hypothesis [11]; the inversions and inverted transpositions hypothesis [12, 13]; the non-uniform substitutions hypothesis [14]; and the statistical mechanics equilibrium hypothesis [15].

To characterize the symmetry phenomenon, Powdel and others [16] analyzed the frequency distributions of oligonucleotides in localized windows along a single strand of DNA. They found that the differences between the frequency distributions of reverse complementary oligonucleotides are not statistically significant. Afreixo et al. [17] noted that the frequency of an oligonucleotide is more similar to the frequency of its reversed complement than to the frequencies of other words of equivalent composition (equal-length oligonucleotides with equal CG content). They called this phenomenon exceptional symmetry, defined measures to evaluate it, and identified several word groups with strong exceptional symmetry in the human genome. More recently, a different measure was introduced to overcome a disadvantage of the previous measure of exceptional symmetry by word [18]. This measure evaluates the difference between the number of occurrences of a word and its reversed complement and relates it with the dissimilarities of the number of occurrences in the corresponding equivalent composition group.

Here, we introduce an improved exceptional symmetry measure and use it to obtain the word symmetry effects in 31 complete genomes stratified by equivalent composition group for word lengths up to 14. Results confirm that measures of word exceptional symmetry can be used to form clusters of related species. Also, we identify words that show high symmetry effect across the 31 species, and across the 9 animal species studied.

2 Materials

The genomes analyzed here are available from the website of the National Center for Biotechnology Information (NCBI; ftp://ftp.ncbi.nih.gov/genomes/). The complete list of species is indicated in Table 1. We selected genomes of species representative of the major taxonomic groups across the tree of life. These include vertebrates, invertebrates, protozoans, fungi, plants, bacteria (gram-positive and gram-negative), archaea and viruses (both double-stranded and single-stranded DNA and RNA viruses).

All non-sequenced or ambiguous nucleotides (mostly N symbols in the sequence file) were discarded from the analysis. For genomes composed by several chromosomes, the chromosomes were processed as separate sequences. 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 14 nucleotides.

3 Methods

In a previous work [17], we called equivalent composition group (ECG) to a set of words with length k that contain a given number m of nucleotides a or t [17]. For example, for $k = 2$ there are three ECGs:

$$G_0 = \{cc, cg, gc, gg\};$$
$$G_1 = \{ac, ag, ca, ct, ga, gt, tc, tg\};$$
$$G_2 = \{aa, at, ta, tt\}.$$

The words division created by ECGs is also called a binary partition [19]. Consider the binary classification of nucleotides in two types, $T_1 = \{a, t\}$ and $T_2 = \{c, g\}$, and let G_m^k (or simply, G_m) be the ECG with words of length k where each word has m symbols of type T_1 and $k - m$ symbols of type T_2, with $m \in \{0, 1, ..., k\}$. Taking into account the combinatorial results (permutations with repetition of indistinguishable objects), it can be concluded that G_m has N_m distinct words,

$$N_m = 2^k \times \frac{k!}{m!(k-m)!}.$$

Note that, for k-mers there are $k + 1$ ECGs with a total of 4^k words.

For even values of k, some words are equal to 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).

We proposed in a previous work [17] one exceptional genomic word symmetry measure evaluated for ECGs and globally. Here, we highlight the exceptional genomic symmetry evaluated for each word, discussing the potentialities of the T measure (symmetric word pair effect, Eq. 1), an improvement of the S measure recently proposed in [18].

Let n_w be the total number of occurrences of word w in the sequence, and n_m be the total number of occurrences of words in the ECG G_m, which contains words composed by m nucleotides a or t. The symmetric word pair effect, for $w \in G_m = \{w_1, w_2, w_3, ..., w_{N_m}\}$, was given by,

$$T(w) = T(w') = \ln \frac{\sqrt{\frac{\sum_{i=1}^{N_m}\sum_{j=1}^{N_m}(n_{w_i} - n_{w_j})^2}{N_m^2 - N_m}} + 1}{|n_w - n_{w'}| + 1}. \quad (1)$$

Table 1 List of species whose genomes are analyzed in this work

Species name	Abbreviation	Usable genome size	Taxonomic group
Abalone shriveling syndrome-associated virus	AbaS	34952	dsDNA viruses, no RNA stage
Acanthocystis turfacea Chlorella virus	AcaT	288046	dsDNA viruses, no RNA stage
Acheta domesticus densovirus	AchD	5234	ssDNA viruses
Acholeplasma phage L2	AcPL	11965	dsDNA viruses, no RNA stage
Acholeplasma phage MV-L1	AcPM	4491	ssDNA viruses
Zika virus	ZikV	10794	ssRNA viruses
Southern tomato virus	SouT	3437	dsRNA viruses
Aeropyrum camini SY1	AerC	1595994	Archaea
Aeropyrum pernix K1	AerP	1669696	Archaea
Caldisphaera lagunensis DSM 15908	CalL	1546846	Archaea
Candidatus Korarchaeum cryptofilum OPF8	CanK	1590757	Archaea
Escherichia coli K12 substr DH10B	EscC	4686135	Bacteria
Helicobacter pylori	HelP	1548238	Bacteria
Nanoarchaeum equitans Kin4-M	NanE	490885	Archaea
Streptococcus mutans GS5	StMG	2027088	Bacteria
Streptococcus mutans LJ23	StML	2015626	Bacteria
Streptococcus pneumoniae 670 6B	StPn	2240043	Bacteria
Plasmodium falciparum	PlaF	22853268	Protozoan
Candida albicans	CanA	949626	Fungi
Saccharomyces cerevisiae	SacC	12157105	Fungi
Arabidopsis thaliana	AraT	118960141	Plants
Vitis vinifera	VitV	416169194	Plants
Caenorhabditis elegans	CaeE	100272607	Nematodes
Apis mellifera	Apis	198904823	Insects
Drosophila melanogaster	DroM	137057575	Insects
Danio rerio	DRer	1295489541	Fish
Macaca mulatta	MacM	2646263223	Primates
Pan troglodytes	PanT	2756176116	Primates
Homo sapiens	HSap	2858658094	Primates
Mus musculus	MusM	2647521431	Rodents
Rattus norvegicus	RatN	2442682943	Rodents

Species are identified by name and abbreviations used herein. Usable genome size (excluding *N*s) and taxonomic group are provided. Downloaded in March 2016 from ftp://ftp.ncbi.nih.gov/genomes/

Table 2 Percentage of words (of length k) with exceptional symmetry effect ($T > 0$), measured in the genomes of 31 species and in the random control sequence (*sym*)

k (%)	2	3	4	5	6	7	8	9	10	11	12	13	14
AbaS	100	97	99	98	91	–	–	–	–	–	–	–	–
AcaT	100	100	100	99	98	94	-	–	–	–	–	–	–
AchD	63	75	81	77	–	–	–	–	–	–	–	–	–
AcPL	63	69	78	78	–	–	–	–	–	–	–	–	–
AcPM	50	66	70	–	–	–	–	–	–	–	–	–	–
ZikV	63	78	83	83	–	–	–	–	–	–	–	–	–
SouT	63	63	71	–	–	–	–	–	–	–	–	–	–
AerC	100	100	100	100	100	100	99	97	–	–	–	–	–
AerP	100	100	100	100	100	100	99	97	–	–	–	–	–
CalL	100	100	100	100	100	100	98	95	–	–	–	–	–
CanK	100	100	100	100	100	100	100	98	–	–	–	–	–
EscC	100	100	100	100	100	100	100	95	–	–	–	–	–
HelP	100	100	100	100	100	100	100	98	–	–	–	–	–
NanE	100	100	100	100	100	99	95	–	–	–	–	–	–
StMG	100	100	100	100	100	100	99	94	–	–	–	–	–
StML	100	100	100	100	100	100	99	94	–	–	–	–	–
StPn	100	100	100	100	100	100	99	94	–	–	–	–	–
PlaF	100	100	100	100	100	100	100	99	99	98	–	–	–
CanA	100	100	100	100	100	100	93	–	–	–	–	–	–
SacC	100	100	100	100	100	100	100	99	95	–	–	–	–
AraT	100	100	100	100	100	100	100	100	100	99	98	–	–
VitV	100	100	100	100	100	100	100	100	100	100	100	100	–
CaeE	100	100	100	100	100	100	100	100	100	100	100	–	–
Apis	100	100	100	100	100	100	100	100	100	100	100	–	–
DroM	100	100	100	100	100	100	100	100	100	100	100	–	–
DRer	100	100	100	100	99	99	98	98	97	97	95	98	–
MacM	100	100	100	100	100	100	100	100	100	100	100	100	100
PanT	100	100	100	100	100	100	100	100	100	100	100	100	100
HSap	100	100	100	100	100	100	100	100	100	100	100	100	100
MusM	100	100	100	100	100	100	100	100	100	100	100	100	100
RatN	100	100	100	100	100	100	100	100	100	100	100	100	100
sym	63	72	75	73	70	69	69	68	68	68	68	69	68

The maximum word length under study is given by $\max\{k \in \{1,2,3,...\} : n * 0.25^k > 5\}$, with n the genome size

The $T(w)$ measure may also be expressed as the difference between two terms. The first term assesses the average frequency deviation between any two words in G_m, whereas the second term accounts for the deviation between the frequency of w and that of its reversed complement. Exceptional symmetry, therefore, is revealed by positive values of T.

T differs from the previously defined S measure by a simple correction introduced to avoid indeterminations. Their values are approximately equal for sufficiently large word counts.

3.1 Control Experiments

Small, positive values of T may be obtained for word pairs that are not exceptionally symmetric. In order to establish a magnitude reference for T, we generate random sequences of independent and identically distributed nucleotides, under the assumption of the validity of the second parity rule, that is, by constraining the generator to produce complementary nucleotides with equal probabilities. Under these conditions, all words in each ECG have the same probabilities, hence no exceptional symmetry (see details

in [20]). The label *sym* is used to denote these random sequences in the remainder of the document.

3.2 Word Analysis Procedure

A word is declared as exceptionally symmetrical when its T value surpasses the critical value, which is defined as the 95th percentile of the T values obtained from the control experiments. To complement this analysis, we compute the percentage of words with $T \leq 0$ for each word length.

To identify groups of genomes with similar exceptional symmetry profiles ($T(w)$ values), we use a hierarchical clustering procedure, using the UPGMA aggregation criterion with Euclidean distance. A similar clustering procedure is used to identify words with similar exceptional symmetry profiles across species.

4 Results and Discussion

For the set of 31 genomes, the word counts were obtained for all word lengths between 1 and 14 nucleotides, and the symmetric word pair effect was obtained for each genomic word. However, for given genome, we only consider the genomic words with lengths k ($k \in \{1, ..., k_{\max}\}$), with

$$k_{\max} = \max\{k \in \{1, 2, 3, ...\} : n * 0.25^k > 5\}$$

and n the genome size. This threshold motivation is the count representability and the protection of the T measure to the sensitivity of rare counts occurrences.

Obviously, for $k = 1$, each ECG contains only one symmetric word pair, and so $T(w) = 0$, for all nucleotides. Almost all words in eukaryote genomes show significant exceptional symmetry effect (above the critical values obtained in the control experiments). Table 2 shows the percentage of words with $T > 0$ for each species and word length of this study. A high percentage of words in viruses show no exceptional symmetry. This result agrees with a previous work [20], which used a different measure and procedure.

Table 2 includes the *sym* row corresponding to one control scenario (sequence with length equal to the length of the human genome). This may be used as a reference of non-exceptional symmetry results.

4.1 Human Genome

A word analysis in the context of exceptional symmetry for the human genome was carried out.

Figure 1 shows boxplots of the T values for $k = 5$ in the human genome and in the corresponding random realization *sym*. The boxplot for the human genome shows high

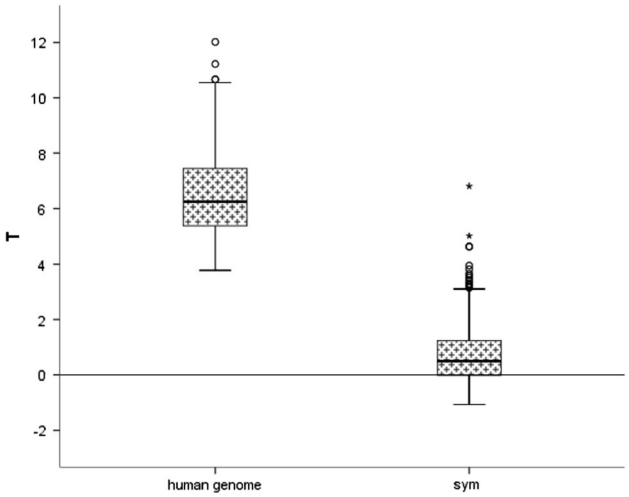

Fig. 1 Boxplots for T values in the human genome and in a random control sequence realization (*sym*) for word length 5

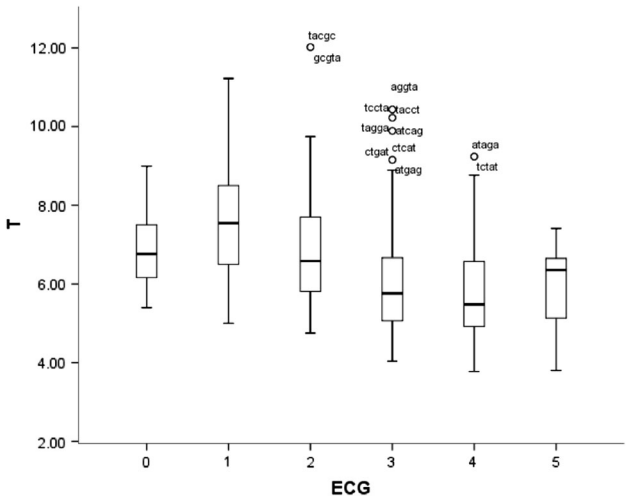

Fig. 2 Boxplots for T values in each ECG for word length 5, in the human genome

and significant symmetric word pair effects. The most exceptionally symmetric word pairs, corresponding to the right outliers, detected in the human T boxplot are: (gcgta, tacgc), (accgg, ccggt), (gccac, gtggc), (gccca, tgggc), (cggga, tcccg).

Figure 2 shows the T values in each ECG for $k = 5$ in the human genome. We observe that as the CG content varies (decreases along the x-axis), the T median values have a non-monotonous behavior. The ECG G_1 has the highest T median value. In general, for the word lengths under study and for the human genome, the T median in ECG G_0 is lower than in G_1, and the T median for G_k is higher than for G_{k-1}. For the control scenario, on the other hand, we observed that the T median values remained essentially constant across all ECGs.

Table 3 The six symmetric word pairs (represented by a single word of the pair) that have the highest (−h) $T(w)$ values, and the six symmetric word pairs that have the lowest (−l) $T(w)$ values for each k, in the human genome

Rank	Word length (k)							
	2	3	4	5	6	7	8	9
1st -h	gg	ccg	cggg	gcgta	taatca	atgttag	atattaac	aaaaatata
	6.9	10.8	9.8	12.0	11.9	12.1	11.7	11.5
2nd -h	tc	ctc	gcga	accgg	acaccg	cagttgg	cgagtggc	atatattta
	6.3	9.7	9.6	11.2	11.9	12.0	11.0	10.2
3rd -h	ct	atg	gggc	tgggc	ggatcg	cttaggc	gagcacgc	aaattaaat
	5.9	9.2	9.2	10.7	11.9	12.0	11.0	10.2
4th -h		gcg	agcc	gtggc	gtcacg	agatcgg	accggcgt	ttaaatata
		8.2	9.2	10.7	11.7	12.0	11.0	10.1
5th -h		tcg	tccg	tcccg	ccgtga	acgaatg	gtcgcgga	aaattttat
		7.6	9.1	10.5	11.5	12.0	11.0	10.1
6th -h		gct	aggg	aggta	ttatct	gcgtacg	cgggtcga	ttattaata
		7.4	9.1	10.4	11.5	11.4	11.0	10.1
6th -l		tct	tctt	gtttt	ttttgt	gtgtgtg	atggaatg	tggaatgga
		4.9	4.4	4.0	3.6	3.3	2.5	1.5
5th -l		ctt	gttt	ttttg	tttttt	aatggaa	tggaatgg	ccaggctgg
		4.8	4.3	3.9	3.6	3.3	2.5	1.5
4th -l		gtt	tttg	tgttt	ttgttt	tttgttt	tgtgtgtg	aatggaatg
		4.4	4.1	3.9	3.6	3.1	2.4	1.4
3rd -l	gt	ttt	tttt	tttgt	tttttg	atggaat	gaatggaa	gtgtgtgtg
	4.8	4.3	4.1	3.8	3.6	3.1	2.3	1.4
2nd -l	tg	tgt	ttgt	ttttt	tttgtt	tgtgtgt	aatggaat	tgtgtgtgt
	4.7	4.2	4.0	3.8	3.5	3.0	2.3	1.3
1st -l	tt	ttg	tgtt	ttgtt	tgtttt	gaatgga	gtgtgtgt	gaatggaat
	4.4	4.2	4.0	3.8	3.5	2.9	2.3	1.2

Table 3 continued

Rank	Word length (k)				
	10	11	12	13	14
1st -h	tttaatataa	atttttataa	aattaaaaatat	atttatttttta	aaatatataaataa
	12.0	11.6	11.1	10.6	10.0
2nd -h	atatattttt	taaaaataatt	ataaaattaaat	taaaatattttaa	aaaaaataatttta
	12.0	10.9	11.1	10.6	10.0
3rd -h	aaatatataa	taaaaatttta	attaaaataaat	aatatatataatt	ataaatatttttaa
	11.6	10.9	11.1	10.6	10.0
4th -h	tacaataaaa	ttaattattaa	aattatatttta	aattaattaaaat	atataaataatata
	11.6	10.9	11.1	10.6	10.0
5th -h	aaattttgta	attaattaatt	aattaaaatata	ataatttaaaata	aattaaaaatatat
	11.1	10.5	11.1	10.6	10.0
6th -h	ttatttagaa	atttaaatttt	aaattaatttaa	aaaatatttatat	atattatttttaaa
	11.1	10.5	11.1	10.6	10.0
6th -l	cgaatggaat	ttgtgatgtgt	tttgtgatgtgt	ctggctaattttt	tatgtccagagttt
	0.8	0.0	−0.7	−1.4	−2.0
5th -l	gaatggaatg	ctggctaattt	tgtccagagttt	ctttgtgatgtgt	tagagcagttttga
	0.7	0.0	−0.8	−1.4	−2.0
4th -l	atggaatgga	tcgaatggaat	ctggctaatttt	tttgtgatgtgtg	agagcagttttgaa
	0.6	−0.2	−0.9	−1.5	−2.1
3rd -l	tgtgtgtgtg	gtgtgtgtgtg	gaatggaatgga	cctggctaatttt	cctggctaattttt
	0.5	−0.2	−0.9	−1.6	−2.3
2nd -l	gtgtgtgtgt	aatggaatgga	tgtgtgtgtgtg	gtgtgtgtgtgtg	tgtgtgtgtgtgtg
	0.5	−0.5	−1.1	−1.8	−2.7
1st -l	aatggaatgg	tgtgtgtgtgt	gtgtgtgtgtgt	tgtgtgtgtgtgt	gtgtgtgtgtgtgt
	0.5	−0.6	−1.1	−2.1	−2.7

In case of a tie, the words are sorted (largest to smallest) by frequency of occurrence. Note that for $k \geq 7$ there are (w, w') pairs where $n_w = n'_w$, exactly

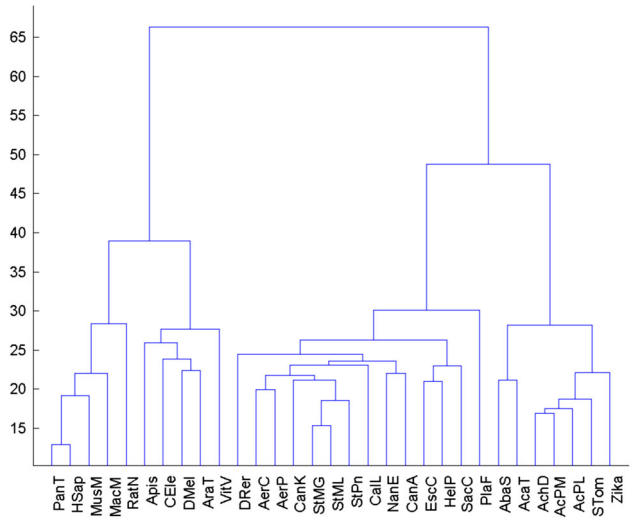

Fig. 3 Dendrogram obtained from the T values for all species under study, word length 4

Table 3 presents, for the word lengths under study, the twelve words with the six highest and the six lowest $T(w)$ values. Some of these extreme words could have some biological interest, e.g., regulatory elements, functional elements, motifs.

Based on the results of the effect size measure, we may conclude that the human genome presents exceptional symmetry. The human genome shows exceptional symmetry for the thirteen different word lengths ($k = 2, ..., 14$) used in this study.

Although the existence of global exceptional symmetry in the human genome was verified, 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, which will be presented in the next section.

It may be also concluded that in the human genome there are ECGs that are more exceptionally symmetric than others. And 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, it was identified that words rich in CG content behave differently from words rich in AT content, in terms of exceptional symmetry.

Table 4 Word pairs with exceptional symmetry effect above the third quartile, which are most common across species, and most common across animal species

k	Word	% of species	Word	% of animals
2	*ca*	42	*cc*	67
3	*acg*	58	*ccg*	78
4	*cgac*	52	*aacg*	67
	cgga		*agcg*	
			cgac	
			cgcc	
			tccg	
5	*attcg*	61	*aacgg*	78
			cgatc	
			gcgcc	
6	*acgcgt*	74	*acggat*	89
			ccgtac	
			gacgta	
7	*tacgtaa*	74	*cgtacga*	100

Each pair is represented by a single word of the pair

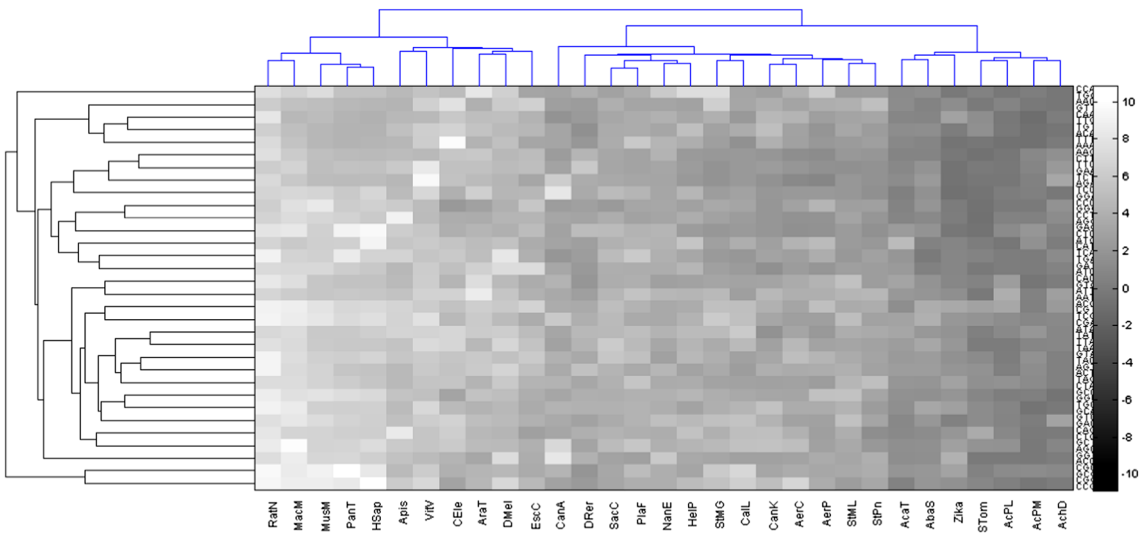

Fig. 4 Heatmap with biclustering organization of the T values for words of length 3 and for all species under study

4.2 Species Comparison

Figure 3 shows the dendrogram obtained with the hierarchical clustering procedure, for $k = 4$. Four distinct groups can be observed in Figure 3: mammalian (on the left); viruses (on the right); a group including the plants and the other animals (except *Danio rerio*); and a group with the unicellular species, plus *Danio rerio*. For other word lengths, the resulting dendrograms essentially maintain the same structure (the dendrogram for $k = 3$ is also included in Figure 4).

Figure 4 shows the heatmap with biclustering organization for trinucleotides. Species are shown on the horizontal axis, and words are shown on the vertical axis. The symmetric word pair effect is stronger on the left side of the heatmap, corresponding to multicellular organisms, and weaker on the right side. The word clustering highlights the group formed by two symmetric word pairs: (ccg, cgg), (gcg, cgc).

We identified the word pairs with high exceptional symmetry (T above the third quartile) in every species under study. From these, we selected the pairs that are highly symmetric across the most species under study, and those that are highly symmetric across the most animal species under study. The results are shown in Table 4. No word pair is considered highly symmetric across all the species under study. However, $T(cgtacga) = T(tcgtacg)$ is above the third quartile in all the animal species under study. The strongest symmetric word pair effect is observed in words composed by CpG dinucleotides.

The results presented in Table 4 are restricted to word lengths between 2 and 7 because for longer word lengths the number of most common symmetric word pair above the third quartile is high. The strongest symmetric word pair effect is observed in words composed by CpG dinucleotides.

5 Conclusions

We evaluated the exceptional symmetry effect in several species, with particular emphasis in the human genome. The word exceptional symmetry values contain information specific to the species and seem to contain information about the species evolution. Taking into account the species in this study, the primates and rodents species have the highest exceptional symmetry values and form a subgroup distinct from all the other species under study. Globally, the eukaryote group showed the highest word exceptional symmetry values, while viruses showed the lowest values. We reinforce that some viruses show a behavior opposite to the exceptional symmetry ($T < 0$) in almost all words under study.

Exceptional symmetry effect was found in a high percentage of words in all cellular organisms under study. Therefore, we conjecture that exceptional symmetry results from some universal law imposed on cellular organisms. Still, the exceptional symmetry profiles are species specific.

Acknowledgements This work was supported by Portuguese funds through the iBiMED-Institute of Biomedicine, IEETA-Institute of Electronics and Informatics Engineering of Aveiro, CIDMA - 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 projects: UID/BIM/04501/2013, PEst-OE/EEI/UI0127/2014 and UID/MAT/04106/2013.

References

1. Chargaff E (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia 6(6):201–209
2. Watson J, Crick F (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738
3. Karkas JD, Rudner R, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands. II. Template functions and composition as determined by transcription with RNA polymerase. Proc Natl Acad Sci USA 60(3):915–920
4. Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands, I. Biological properties. Proc Natl Acad Sci USA 60(2):630–635
5. Rudner R, Karkas JD, Chargaff E (1968) Separation of B. subtilis DNA into complementary strands. III. Direct analysis. Proc Natl Acad Sci USA 60(3):921–922
6. Forsdyke DR (2011) Evolutionary bioinformatics. Springer, New York
7. Sobottka M, Hart AG (2011) A model capturing novel strand symmetries in bacterial DNA. Biochemical and biophysical research communications 410(4):823–828. doi:10.1016/j.bbrc.2011.06.072. http://www.sciencedirect.com/science/article/pii/S0006291X1101045X
8. Zhang SH, Huang YZ (2008) Characteristics of oligonucleotide frequencies across genomes: conservation versus variation, strand symmetry, and evolutionary implications. Nat Precedings:1–28.
9. Zhang SH, Huang YZ (2010) Strand symmetry: characteristics and origins. In: Fourth international conference on bioinformatics and biomedical engineering (iCBBE) 2010. pp. 1–4 (2010). doi:10.1109/ICBBE.2010.5517388
10. Forsdyke DR, Bell SJ (2004) Purine loading, stem-loops and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. Appl Bioinform 3(1):3–8
11. Baisnée PF, Hampson S, Baldi P (2002) Why are complementary DNA strands symmetric? Bioinformatics 18(8):1021–1033
12. Albrecht-Buehler G (2006) Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. Proc Natl Acad Sci USA 103(47):17,828–17,833
13. Albrecht-Buehler G (2007) Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. Genomics 90:297–305
14. Lobry TH (1995) Properties of a general model of DNA evolution under no-strand-bias condition. J Mol Evol 40:326–330

15. Hart A, Martnez S, Olmos F (2012) A gibbs approach to Chargaff's second parity rule. J Stat Phys 146:408–422
16. Powdel B, Satapathy S, Kumar A, Jha P, Buragohain A, Borah M, Ray S (2009) A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (chargaff's second parity rule). DNA Res 16:325–343
17. Afreixo V, Rodrigues JMOS, Bastos CAC (2015) Analysis of single-strand exceptional word symmetry in the human genome: new measures. Biostatistics 16(2):209–221
18. Afreixo V, Rodrigues JMOS, Bastos CAC, Silva RM (2016) Exceptional symmetry profile: A genomic word analysis. In: PACBB
19. Kong SG, Fan WL, Chen HD, Hsu ZT, Zhou N, Zheng B, Lee HC (2009) Inverse symmetry in complete genomes and whole-genome inverse duplication. PLoS ONE 4(11):e7553
20. Afreixo V, Rodrigues JMOS, Bastos CAC (2014) Exceptional single strand DNA word symmetry: analysis of evolutionary potentialities. J Integr Bioinform 11(3):250