Exceptional single strand DNA word symmetry: analysis of evolutionary potentialities

Vera Afreixo[1,2,4,*], João M. O. S. Rodrigues[2,3,4] and Carlos A. C. Bastos[2,3,4]

[1]CIDMA – Center for Research and Development in Mathematics and Applications, Department of Mathematics

[2]IEETA – Institute of Electronics and Telematics Engineering of Aveiro

[3]Department of Electronics Telecommunications and Informatics

[4]University of Aveiro, 3810-193 Aveiro, Portugal

Summary

Some previous studies point to the extension of Chargaff's second rule (the phenomenon of symmetry) to words of large length. However, in random sequences generated by an independent symbol model where the probability of occurrence of complementary nucleotides is the same, we expect that the phenomenon of symmetry holds for all word lengths. In this work, we measure the symmetry above that expected in independence contexts (exceptional symmetry), for several organisms: viruses; archaea; bacteria; eukaryotes. We also create 27 control scenarios with the same length of each genome under study. The results for each organism were compared to those obtained in control scenarios. We created a new organism genomic signature consisting of a vector of the measures of exceptional symmetry for words of lengths 1 through 12. We show that the proposed signature is able to capture essential relationships between organisms.

1 Introduction

The detailed analysis of some bacterial genomes led to the formulation of Chargaff's second parity rule, which asserts that complementary nucleotides occur with similar frequencies in each of the two DNA strands [15, 9, 16, 7]. Extensions of this rule state that the frequencies of inverted complementary words (such as AAC and GTT) should also be similar. Chargaff's second parity rule and its extensions have been extensively confirmed in bacterial and eukaryotic genomes, including recent results (e. g. [13, 14, 6, 5, 12, 10, 17, 7, 11, 3, 4]).

The views about the origins and biological significance of Chargaff's second parity rule and its extensions are conflicting [18]. For example, Forsdyke and Bell (2004) argue that this symmetry results from DNA stem-loop secondary structures [8]. However, Albrecht-Buehler (2007) argues that the presence of Chargaff's second parity rule and its extensions are due to the existence of certain mechanisms of inversions, transpositions, and inverted transpositions [5].

*To whom correspondence should be addressed. Email: vera@ua.pt

If a sequence is randomly generated using an independent symbol model that assigns equal probability to complementary nucleotides, then it is expected that the extensions of Chargaff's second parity rule will hold. In this case, however, words other than inverted complements (e.g. AAG, TTC, CAA) will also be equally likely. In real sequences, we found that the similarity between the frequency of each word and that of the corresponding inverted complement is stronger than the similarity to the frequency of any other word [2].

We will analyze not only the symmetry phenomenon (similarity between the frequencies of symmetric pairs) but also the exceptional symmetry phenomenon. This exceptional symmetry will be evaluated by a relative measure that corresponds to the ratio of the goodness of fit of the symmetry hypothesis and the goodness of fit of the uniformity in equivalent composition group hypothesis. An equivalent composition group is composed by words with equal expected frequencies under independence and Chargaff's second parity rule hypothesis. We focus our study on the analysis of several species and compare the results of the exceptional symmetry with those obtained in the control scenarios.

Based on the exceptional symmetry measure, we created a kind of organism genomic signature by using a vector with the first 12 measures of exceptional symmetry (word lengths from 1 to 12). Our results show that the genomic signature has the potential to discriminate between species groups.

2 Methods

Let $\mathcal{A}$ be the set $\{A, C, G, T\}$ and let π_S denote the occurrence probability of symbol $S \in \mathcal{A}$. Chargaff's second parity rule states the equality between the occurrence of complementary nucleotides: $\pi_A = \pi_T$ and $\pi_C = \pi_G$.

We define a symmetric word pair as the set composed by one word w and the corresponding inverted complement word w', with $w'' = w$. The extensions of Chargaff's second parity rule state that all symmetric word pairs have similar occurrence frequencies.

We call equivalent composition groups (ECG) to the sets of words with length k which contain the same number of nucleotides A or T. Every symmetric word pair is a subset of an ECG, which contains several distinct symmetric word pairs. Note that, for k-mers (word of length k) we have $k + 1$ ECGs and we denote the i-th ECG by G_i, $0 \leq i \leq k$ where i represents the number of nucleotides A or T.

When all words in each ECG have similar frequency, we have a particular single strand symmetry phenomenon that we call uniform symmetry. A random sequence generated under independence with $\pi_A = \pi_T$ and $\pi_C = \pi_G$ is obviously expected to exhibit uniform symmetry. We expect that, in a natural DNA sequence, the frequency of a word is generally more similar to the frequency of its inverted complement than to the frequencies of other words in the same ECG. We call this exceptional symmetry.

We use uniform symmetry as the reference to evaluate the possible exceptional symmetry (non uniform symmetry) of a DNA sequence.

2.1 Exceptional symmetry measure

To measure exceptional symmetry in a global way we propose the following ratio

$$R_s = \frac{X_u^2 + \tau}{X_s^2 + \tau}. \tag{1}$$

where $X_u^2 = \sum_{i=0}^{k} X_u^2(G_i)$ with $X_u^2(G_i)$ being the chi-square statistic used to evaluate the uniformity in ECG G_i, and $X_s^2 = \sum_{i=0}^{k} X_s^2(G_i)$ with $X_s^2(G_i)$ being the chi-square statistic to evaluate the symmetry in ECG G_i. $\tau > 0$ is a residual value to avoid an indeterminate ratio in the presence of exact uniform symmetry.

We observe that R_s statistic does not depend on the sample size dimension (n), but depends on the degrees of freedom of X_u^2 and X_s^2. This measure depends, in an indirect way, on the word length. X_u^2 has $df_u = 4^k - (k+1) - 1$ degrees of freedom and X_s^2 has $df_s = 4^k/2 - 1$ degrees of freedom. If some symmetric word pairs are not present in one genome, we have to correct the X_s^2 degrees of freedom to $df_s = (4^k - n_m)/2 - 1$, where n_m is the number of words in missing symmetric pairs. Likewise, if some ECGs have no occurrences, df_u must also be corrected to $df_u = 4^k - \sum \sharp ECG_m - (k - n_m^{ECG} + 1) - 1$, where n_m^{ECG} is the number of ECGs with no occurrences and $\sharp ECG_m$ the number of elements in one ECG with no occurrences.

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

$$VR = \sqrt{\frac{df_s}{df_u} R_s}. \tag{2}$$

The equivalence between symmetric word pairs may be evaluated using Cramér's coefficient. If a DNA sequence reveals symmetry, it is worthwhile to evaluate the existence of exceptional symmetry. If VR takes on values $\gg 1$ and there is equivalence between symmetric word pairs, we conclude that there is exceptional word symmetry in the sequence being analyzed.

Note that, VR measure can be described as the ratio of two Cramér's coefficients which may be considered an effect size measure.

2.2 DNA sequences

In this study, we used the complete DNA sequences of 27 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 and were downloaded in January 2014.

All genome sequences were processed to obtain the word counts. The words were counted considering overlapping. We use the word lengths from 1 to 12.

Table 1: List of organisms whose DNA was used in this study. Genome length n and measured nucleotide composition are shown.

Organism	Group	Abbreviation	n	f_A	f_C	f_G	f_T
Homo sapiens	eucarya (animalia)	H sapiens	2.86E+9	0.295	0.204	0.205	0.296
Macaca mulatta	eucarya (animalia)	M mulatta	2.65E+9	0.296	0.204	0.204	0.296
Pan trogolodytes	eucarya (animalia)	P troglod	2.76E+9	0.296	0.204	0.204	0.296
Mus musculus	eucarya (animalia)	M musculu	2.65E+9	0.292	0.208	0.208	0.292
Rattus norvegicus	eucarya (animalia)	R norvegi	2.44E+9	0.290	0.210	0.210	0.290
Danio rerio	eucarya (animalia)	D rerio	1.30E+9	0.317	0.183	0.183	0.317
Apis mellifera	eucarya (animalia)	A mellife	1.99E+8	0.330	0.170	0.170	0.330
Arabidopsis thaliana	eucarya (plantae)	A thalian	1.19E+8	0.320	0.180	0.180	0.320
Vitis vinifera	eucarya (plantae)	V vinifer	4.16E+8	0.328	0.172	0.172	0.328
Saccharomyces cerevisiae	eucarya (fungi)	S cervis	1.22E+7	0.310	0.191	0.191	0.309
Candida albicans	eucarya (fungi)	C albican	9.50E+5	0.331	0.167	0.168	0.334
Plasmodium falciparum	eucarya (protozoa)	P falcipa	2.29E+7	0.403	0.097	0.097	0.403
Helicobacter pylori	bacteria	H pylori	1.55E+6	0.303	0.197	0.196	0.304
Streptococcus mutans GS	bacteria	S mutansG	2.03E+6	0.314	0.185	0.183	0.318
Streptococcus mutans LJ23	bacteria	S mutansL	2.02E+6	0.316	0.184	0.187	0.313
Streptococcus pneumoniae	bacteria	S pneumon	2.24E+6	0.300	0.199	0.197	0.304
Escherichia coli	bacteria	E coli	4.69E+6	0.246	0.254	0.254	0.246
Aeropyrum camini	archaea	A camini	1.60E+6	0.214	0.286	0.281	0.218
Aeropyrum pernix	archaea	A pernix	1.67E+6	0.216	0.284	0.280	0.221
Caldisphaera lagunensis	archaea	C lagunen	1.55E+6	0.351	0.149	0.152	0.349
Candidatus Korarchaeum	archaea	C Korarch	1.59E+6	0.257	0.241	0.249	0.253
Nanoarchaeum equitans	archaea	N equitan	4.91E+5	0.342	0.158	0.158	0.342
NC001341	virus	NC001341	4.49E+3	0.399	0.176	0.157	0.268
NC001447	virus	NC001447	1.20E+4	0.357	0.149	0.171	0.324
NC004290	virus	NC004290	5.23E+3	0.288	0.187	0.211	0.313
NC008724	virus	NC008724	2.88E+5	0.246	0.247	0.247	0.260
NC011646	virus	NC011646	3.50E+4	0.307	0.204	0.192	0.298

2.3 Simulated data

To serve as control data, we generated random sequences designed to emulate some of the statistical characteristics of the studied genomes. For each genome, of length n and composition (f_A, f_C, f_G, f_T), we generated 100 random sequences of the same length n, under the independent symbol assumption, with nucleotide probabilities set to $\widehat{\pi_A} = \widehat{\pi_T} = (f_A + f_T)/2$ and $\widehat{\pi_C} = \widehat{\pi_G} = (f_C + f_G)/2$.

Software in the form of C and Matlab code, together with a sample input data set and complete documentation is available on request from the corresponding author (vera@ua.pt).

3 Results

Table 2 shows the number of words in missing symmetric pairs and the number of missing ECGs in the studied genomes. These are totals for the 12 word lengths. Unsurprisingly, shorter genomes have more missing symmetric word pairs. Note also that only viruses had ECGs with no occurrences.

The VR values computed for the 27×100 random sequences show virtually no correlation ($|r| < 0.1$) with the sequence lengths. This is shown (for $k = 12$) in figure 1 and summarized in table 3. Since there is no association between sequence length and the VR values in the random sequences, we propose as a control scenario the mean of all random sequences (2700 distinct random sequences).

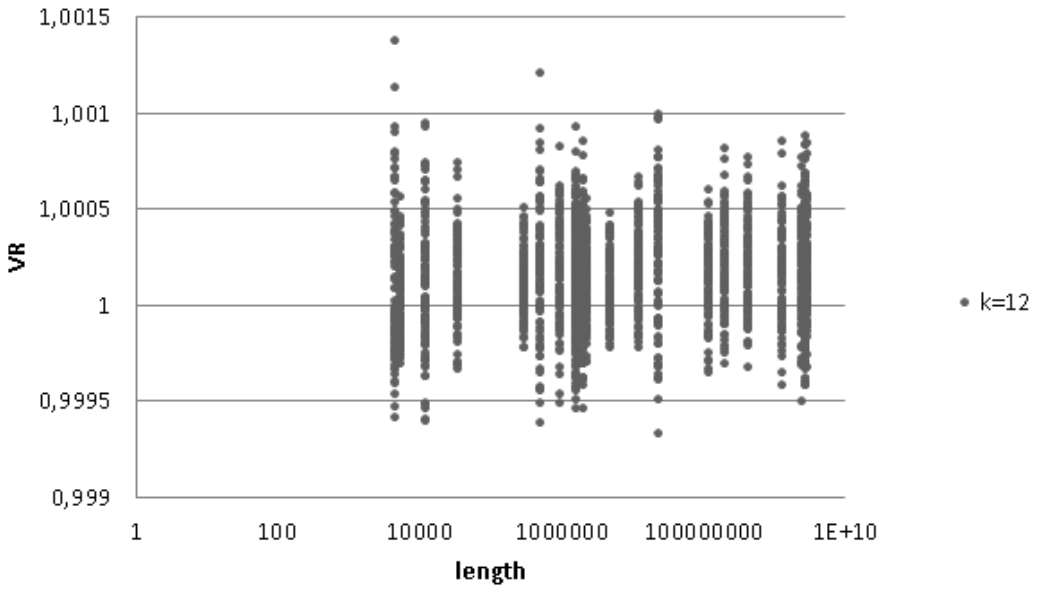

Figure 1: VR **values measured in random sequences for k=12 vs sequence length.**

Figure 2 presents the mean exceptional symmetry profiles for the four types of organisms and for the random scenario (mean of all random sequences). We observe that for all the organism profiles and for all word lengths there is a global exceptional symmetry tendency: all groups

Table 2: Missing symmetric pairs and missing ECGs in the studied genomes. The minimum word length with missing elements is shown in parentheses.

organism	length	total number of missing symmetric words	total number of missing ECGs
H sapiens	2858658094	45335 (11)	0
M mulatta	2646263223	43926 (11)	0
P troglod	2756176116	45630 (11)	0
M musculu	2647521431	53492 (11)	0
R norvegi	2442682943	28143 (11)	0
D rerio	1295489541	18096 (11)	0
A mellife	198904823	371067 (11)	0
A thalian	118960141	18335091 (10)	0
V vinifer	416169194	17084237 (11)	0
S cerevis	12157105	16865183 (9)	0
C albican	949626	17092221 (7)	0
P falcipa	22853268	20385612 (8)	0
H pylori	1548238	17994569 (6)	0
S mutansG	2027088	16572925 (7)	0
S mutansL	2015626	16564420 (7)	0
S pneumon	2240043	15938290 (8)	0
E coli	4686135	12504174 (8)	0
A camini	1595994	18969581 (8)	0
A pernix	1669696	7698754 (8)	0
C lagunen	1546846	1192026 (6)	0
C Korarch	1590757	492310 (8)	0
N equitan	490885	13223383 (7)	0
NC001341	4491	22318466 (4)	10 (7)
NC001447	11965	22249495 (5)	4 (10)
NC004290	5234	22308126 (5)	4 (9)
NC008724	288046	20639968 (7)	0
NC011646	34952	22060458 (5)	3 (10)

Table 3: Pearson's correlation coefficient between sequence length and VR values for all generated random sequences.

k	2	3	4	5	6	7	8	9	10	11	12
$\|r\|$	0.02	0.03	0.03	0.01	0.02	0.02	0.01	0.03	0.07	0.08	0.09

of organisms under study present higher VR values than in the random sequences profiles. Bacteria, archaea and fungi groups have the most similar exceptional symmetry profiles.

The simulated data represents the control groups for the symmetry phenomenon without exceptional symmetry. Naturally, the random sequences generated under the independence hypothesis show no exceptional symmetry. Using the random sequences, we compute the standard error for the mean of VR values, and we compute confidence interval at the 95% level for the mean (mean-2*SE; mean+2*SE).

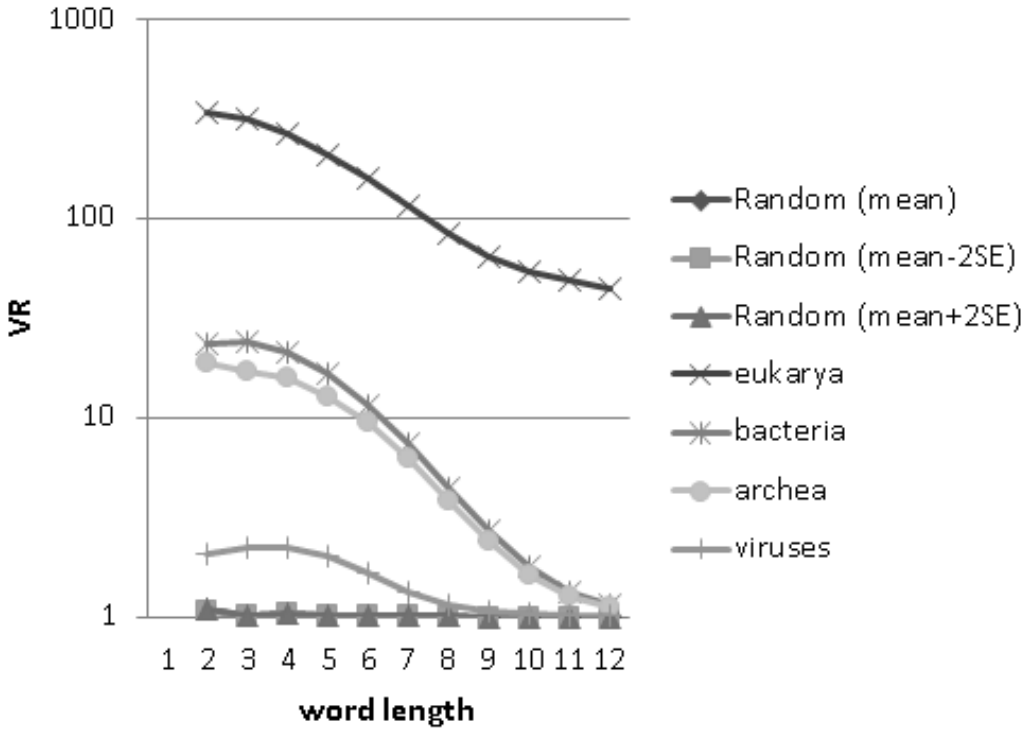

Figure 2: Mean VR values for eukarya, bacteria, archaea, viruses and random sequences (mean and mean $\pm$ 2 $\times$ standard error).

By comparing the VR measures for each cellular organism to the measures of the random data, we observe that all the studied cellular organisms present exceptional single strand DNA word symmetry, for all word lengths.

Some of the viruses studied here do not present significant exceptional symmetry for short word lengths, since VR values are lower than the upper bound (mean +2SE) of the confidence interval obtained in the uniform symmetry context (see Fig. 3). The shapes of the exceptional symmetry profiles for NC001341, NC001447 and NC004290 viruses are similar. In NC008724 and NC011646 exceptional symmetry profiles the effect is stronger than in the other viruses under study. In a previous work [1], a valley was reported in intermediate word lengths for these two viruses, but we now recognise this as an artifact caused by missing words. The corrected df_s and df_u proposed here eliminate this artifact. Furthermore, this correction eliminates the tendency for the increasing of the VR profile for the longer k-mers, which was perceptible in the uncorrected profiles shown in [1] for most species (viruses, archea, bacteria, fungi and protozoa).

The highest variation coefficient inside each group (eukarya, bacteria, archaea and viruses) for all studied word lengths is obtained in eukarya. Figure 4 shows the subgroup profiles for: animalia, plantae, fungi and protists. The highest exceptional symmetry values were obtained for animalia.

Note that *Rattus norvegicus* shows the highest exceptional symmetry values when compared with the other species in this study. We also observed that the maximum exceptional value for cellular organisms was obtained between word lengths 2 and 5, and for viruses it was obtained for the longer word lengths under study.

Although the exceptional symmetry measure (VR) is independent from the sequence length in random sequences, there might be some association in real sequences. In fact, we observed that exceptional symmetry is strongly related to the length of the organism's genome. For each organism we obtained the average of the VR values (over the 12 word lengths) and the genome size. The Pearson correlation coefficient between mean VR values and genome size is 0.81.

In order to explore the potentialities of the exceptional symmetry profile to compare the evolutionary relation between species, we generate a dendrogram. The exceptional symmetry profile of each species is represented as a vector $[\log_{10}(VR(1)), \ldots, \log_{10}(VR(12))]$. We use the Euclidean distance between the symmetry profiles of each pair of species and the UPGMA linkage criteria between groups. Figure 5 shows the resulting dendrogram. This dendrogram presents some accepted evolutionary relationships between living organisms. For instance, the dendrogram presents several subgroups: viruses group; archaea, bacteria and fungi group; vertebrate group; eukarya group (without fungi).

4 Conclusion

We proposed a new measure to assess exceptional symmetry based on the comparison of symmetric word counts. Simulation with random sequences show that this measure is insensitive to genome size and word length.

Leveraging on this measure,we created a kind of organism genomic signature which is an exceptional symmetry profile (the vector with the first 12 measures of exceptional symmetry). Although the exceptional symmetry profile is a vector with only twelve elements, it seems to contain enough information to cluster organisms in major groups in a dendrogram.

Based on exceptional symmetry profile, we observed that all cellular organisms under study present exceptional symmetry: we conjecture that exceptional symmetry is an universal law of cellular organisms. We also found some viruses with a VR profile showing a behavior opposite to exceptional symmetry, $VR < 1$, in short words. The animals group showed the highest exceptional symmetry in this study.

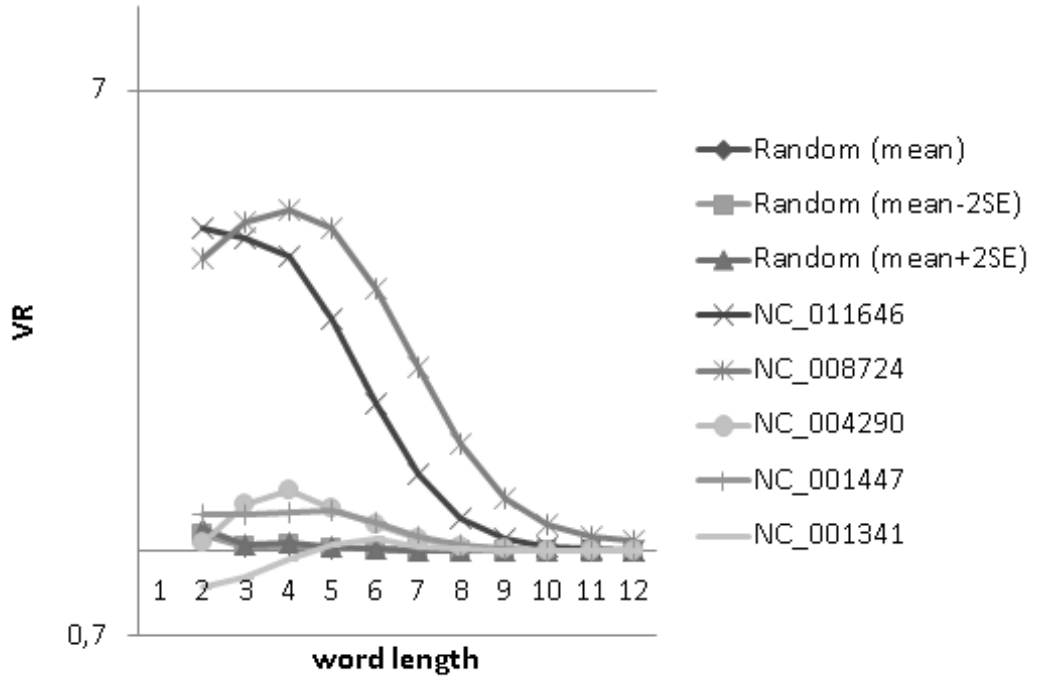

Figure 3: VR **values for viruses; random (mean and mean** $\pm$ **2** $\times$ **standard error).**

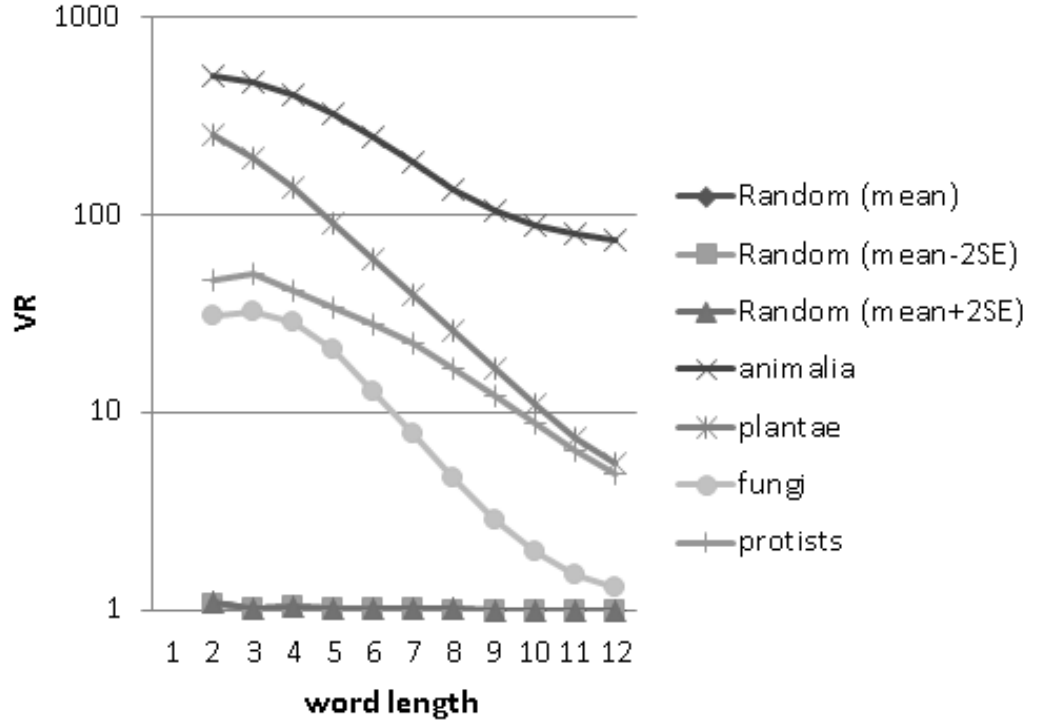

Figure 4: VR **values for animalia, plantae, fungi, protists and random (mean and mean** $\pm$ **2** $\times$ **standard error).**

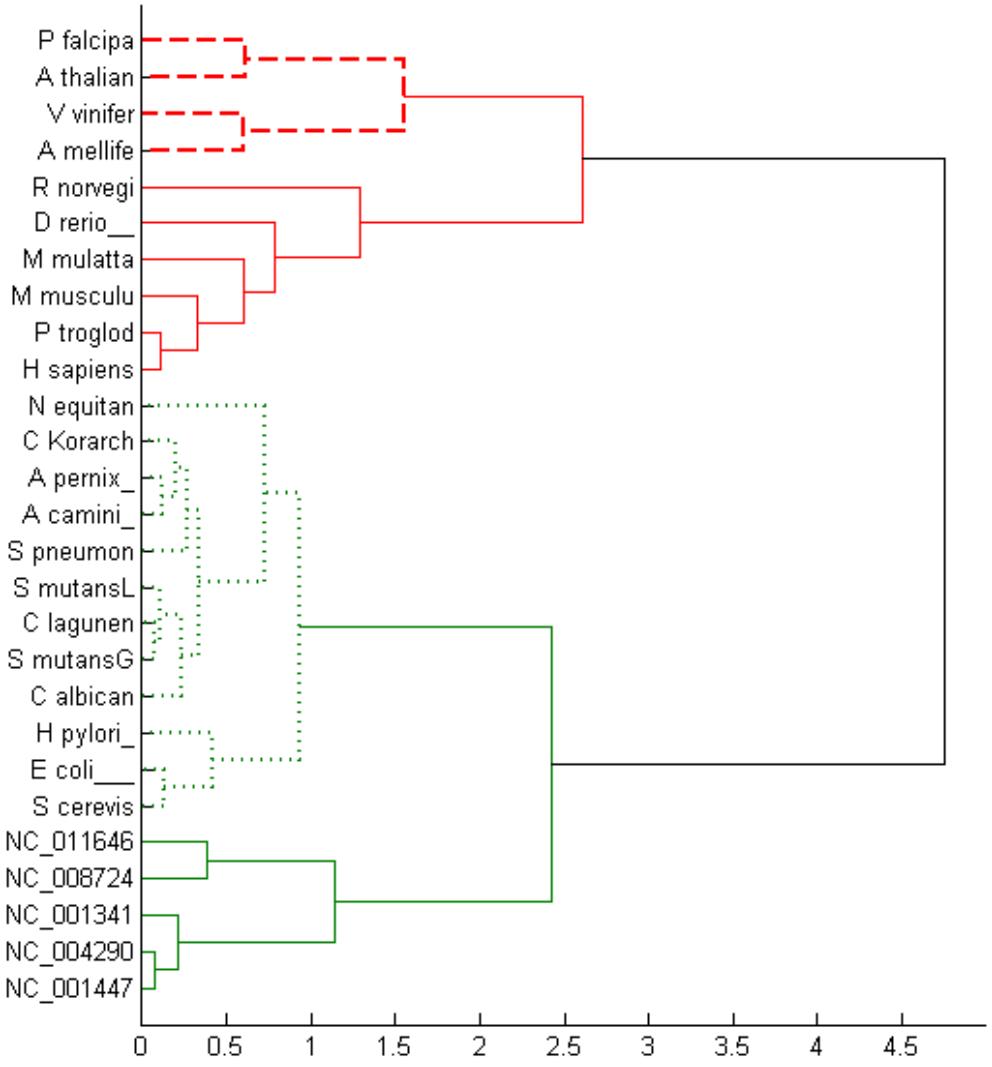

Figure 5: Dendrogram of all species under study. Using exceptional symmetry profiles for each species and Euclidean distance and UPGMA.

Acknowledgements

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

[1] V. Afreixo, J.O.S. Rodrigues, and C.A.C. Bastos. Exceptional single strand DNA word symmetry: universal law? In *8th International Conference on Practical Applications of Computational Biology and Bioinformatics (PACBB 2014). Advances in Intelligent Systems and Computing*, 294:137–143, 2014.

[2] V. Afreixo, J.M.O.S. Rodrigues, C.A.C. Bastos. Analysis of single-strand exceptional word symmetry in the human genome: new measures. *Biostatistics*, doi: 10.1093/biostatistics/kxu041, 2014.

[3] V. Afreixo, C.A.C. Bastos, S.P. Garcia, J.M.O.S. Rodrigues, A.J. Pinho, and P.J.S.G. Ferreira. The breakdown of the word symmetry in the human genome. *Journal of Theoretical Biology*, 335:153–159, 2013.

[4] V. Afreixo, S.P. Garcia, and J.M.O.S. Rodrigues. The breakdown of symmetry in word pairs in 1,092 human genomes. *Jurnal Teknologi*, 66(3):1–8, 2013.

[5] G. Albrecht-Buehler. Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. *Genomics*, 90:297–305, 2007.

[6] P.-F. Baisnée, S. Hampson, and P. Baldi. Why are complementary DNA strands symmetric? *Bioinformatics*, 18(8):1021–1033, 2002.

[7] D.R. Forsdyke. *Evolutionary Bioinformatics*. Springer, Berlin, 2010.

[8] D.R. Forsdyke and S.J. Bell. Purine loading, stem-loops and Chargaffs second parity rule: a discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics*, 3(1):3–8, 2004.

[9] J.D. Karkas, R. Rudner, and E. Chargaff. 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.

[10] S.-G. Kong, W.-L. Fan, H.-D. Chen, Z.-T. Hsu, N. Zhou, B. Zheng, and Hoong-Chien Lee. Inverse symmetry in complete genomes and whole-genome inverse duplication. *PLoS ONE*, 4(11):e7553, 2009.

[11] M. Mascher, I. Schubert, U. Scholz, and S. Friedel. Patterns of nucleotide asymmetries in plant and animal genomes. *Biosystems*, 111(3):181–189, 2013.

[12] K. Okamura, J. Wei, and S.W. Scherer. Evolutionary implications of inversions that have caused intra-strand parity in DNA. *BMC Genomics*, 8:160, 2007.

[13] V.V. Prabhu. Symmetry observations in long nucleotide sequences. *Nucleic Acids Research*, 21(12):2797–2800, 1993.

[14] D. Qi and A. J. Cuticchia. Compositional symmetries in complete genomes. *Bioinformatics*, 17(6):557–559, 2001.

[15] R. Rudner, J.D. Karkas, and E. Chargaff. 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.

[16] R. Rudner, J.D. Karkas, and E. Chargaff. 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.

[17] S.-H. Zhang and Y.-Z. Huang. Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. *Bioinformatics*, 26(4):478–485, 2010.

[18] S.-H. Zhang and Y.-Z. Huang. Strand symmetry: Characteristics and origins. In *4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE), 2010*, pages 1–4, 2010.