Detection of exceptional genomic words: a comparison between species

Ana Tavares, João Rodrigues, Carlos Bastos, Armando Pinho, Paulo Ferreira, Paula Brito and Vera Afreixo

April 15, 2018

Detection of exceptional genomic words: a comparison between species

Ana Tavares, *University of Aveiro*, ahtavares@ua.pt
João Rodrigues, *University of Aveiro*, jmr@ua.pt
Carlos Bastos, *University of Aveiro*, cbastos@ua.pt
Armando Pinho, *University of Aveiro*, ap@ua.pt
Paulo Ferreira, *University of Aveiro*, pjf@ua.pt
Paula Brito, *University of Porto*, mpbrito@fep.up.pt
Vera Afreixo, *University of Aveiro*, vera@ua.pt

Abstract. In this study we explore the potentialities of the inter-word distances to detect exceptional genomic words (oligonucleotides) in several species, using whole-genome analysis. We confront the empirical results obtained from the complete genomes with the corresponding results obtained from the random background. We develop a procedure, based on some statistical properties of the global distance distributions in DNA sequences, to discriminate words with exceptional inter-word distance distribution and to identify distances with exceptional frequency of occurrence. We identify the statistically exceptional words in whole-genomes, i.e., words with unexpected inter-word distance distributions, and we suggest species signatures based on exceptional word profiles.

Keywords. inter-oligonucleotide distances, DNA sequence, exceptional genomic word, stochastic model, goodness of fit.

1 Introduction

Several authors tried to identify exceptional words using different statistical criteria. A standard approach to detect exceptional words relies on their frequency. For example, based on genomic word frequencies and on comparisons between those frequencies and the random background (e.g. [10, 15, 16]).

The distance between two successive occurrences of a pattern in strings has been thoroughly studied and theoretical results have been deduced, in particular the generating functions of the waiting times to return to a specific pattern (e.g., [14, 18]). The probability mass function of the waiting times to return for the first time to a specific genomic word, or inter-word distance distribution, can be obtained by the Markov chain embedding technique, first developed by Fu (see, for example, [6]).

There are some interesting and counter-intuitive relations between frequency and distance distributions. Thus, the two perspectives are worth of separate investigation.

The inter-nucleotide distance (i.e., the distance between successive occurrences of the same nucleotide) has been previously explored to compare the complete genomes of several organisms; this comparison was based on genome distance distributions explored by [2]. The inter-nucleotide distance was also explored in the context of genome annotation by [11]. In [3], the inter-dinucleotide distance distribution was proposed and a comparison between all dinucleotide distributions in the human genome was performed. Note that in [3] overlapping dinucleotides were excluded from analysis, so that the expected distance distribution under an independent nucleotide model is a geometric distribution. Based on an inter-CpG distance, a CpG-island detection algorithm was proposed by [8], where a geometric distribution was used as a reference for comparison.

In this paper, we describe a procedure to highlight exceptional words that is based on inter-word distance distributions, rather than word frequencies. The subtraction of the random background from the counting result (under an independent nucleotide placement assumption) has been suggested as a way of emphasizing the contribution of selective evolution ([12, 5]). Based on this biologic perspective, we take a nucleotide independent model as the departing point and evaluate the discrepancy between real sequences and random background.

2 Materials and methods

Materials

In this study, we used the complete DNA sequences of 30 species, listed in Table 1, downloaded from the website of the National Center for Biotechnology Information (`http://www.ncbi.nlm.nih.gov/genomes`). For each species, we processed the available assembled chromosomes as separate sequences. In each sequence, we studied every word formed by k consecutive unambiguous nucleotides, with $1 < k \leq 5$. The analysis included words partially overlapping preceding or succeeding words. All ambiguous or unsequenced nucleotides, i.e., all non-ACGT symbols, are considered word delimiters.

Methods

Inter-word distance

Consider the alphabet formed by the four nucleotides $\mathcal{A} = \{A, C, G, T\}$, and let s be a symbolic sequence of length N defined in $\mathcal{A}$. For each nucleotide $x \in \mathcal{A}$, consider a numerical sequence, d^x (or simply d), that represents the inter-nucleotide distances between each occurrence of symbol x and the previous occurrence of the same symbol, i.e., the differences between the positions occupied by successive occurrences of symbol x. As an example, we show the four inter-nucleotide distance sequences for $s = AAACGTCGATCCGTG$:

$$d^A = (1, 1, 6),\ d^C = (3, 4, 1),\ d^G = (3, 5, 2),\ d^T = (4, 4).$$

A genomic word, or oligonucleotide (w), is a sequence of length k defined in $\mathcal{A}$. We can extend the notion of inter-nucleotide distance to the case of oligonucleotides. Assuming that the

Table 1: List of DNA builds used for each species

Species	Biological taxonomy	Abbr.
Homo sapiens (human)	animalia	H.sapiens
Macaca mulatta (Rhesus macaque)	animalia	M.mulatta
Pan troglodytes (chimpanzee)	animalia	P.troglodytes
Mus musculus (mouse)	animalia	M.musculus
Rattus norvegicus (brown rat)	animalia	R.norvegicus
Eqqus caballus (horse)	animalia	E.caballus
Cannis lupus familiaris (dog)	animalia	C.lupus
Bos taurus (cow)	animalia	B.taurus
Monodelphis domesticus (opossum)	animalia	M.domesticus
Ornithorhynchus anatinus (platypus)	animalia	O.anatinus
Danio rerio (zebrafish)	animalia	D.rerio
Apis mellifera (honey bee)	animalia	A.mellifera
Arabidopsis thaliana (thale cress)	plantae	A.thaliana
Vitis vinifera (grape vine)	plantae	V.vinifera
Saccharomyces cerevisiae str	fungi	S.cerevisiae
Schizosaccharomyces pombe	fungi	C.pombe
Escherichia coli	bacteria	E.coli
Helicobacter pylori	bacteria	H.pylori
Streptococcus pneumoniae	bacteria	S.pneumoniae
Streptococcus mutans LJ23	bacteria	S.mutansLJ
Streptococcus mutans GS	bacteria	S.mutansGS
Aeropyrum pernix str.K1	archaea	A.pernix
Nanoarchaeum equitans	archaea	N.equitans
Candidatus korarchaeum	archaea	C.korarchaeum
Caldisphaera lagunensis	archaea	C.lagunensis
Aeropyrum camini	archaea	A.camini
NC001341 virus	virus	vir.001341 virus
NC001447 virus	virus	vir.001447 virus
NC004290 virus	virus	vir.004290 virus
NC011646 virus	virus	vir.011646 virus

sequence is read through a sliding window of length k, we can define the inter-oligonucleotide (inter-w) distance sequence d^w as the differences between the positions of the first symbol of consecutive occurrences of that oligonucleotide. For example, the inter-CG distance sequence for the short DNA segment above is $d^{CG} = (3, 5)$.

Reference distribution under a nucleotide independence model

Let $w = x_1x_2x_3 \ldots x_k \in \mathcal{A}^k$ be a generic oligonucleotide and D be the random variable that represents the inter-oligonucleotide distance, from a sequence whose nucleotides are independently generated.

The reference distribution of inter-w distances can be deduced using a state diagram, which represents the progress made towards identifying w as each symbol is read from the sequence. The state diagram has $k+1$ states. The first k states, $S_0, S_1, \ldots, S_{k-1}$, represent intermediate points in the process and state S_k is the final, absorbing state. In the diagram, being in state S_i means that the last i symbols read from the sequence match a prefix of w. As each new symbol is read, a transition occurs from S_i to a new state S_j, until the final, or absorbing, state S_k is reached, meaning that a new occurrence of w has just been identified in the sequence.

We define the distance to the next occurrence of w, starting from an initial state S_I ($I < k$), as the number of steps (transitions) it takes to walk through the diagram from S_I until the final state S_k is reached. The initial state is given by the longest word overlap of w, different from w.

To illustrate this procedure, we present the state diagram for inter-ACG distances in Figure 1. In this specific case, the probability of transition between two non-absorbing states, S_i to S_j, is given by element m_{ij} $(0 \leq i, j \leq 2)$ of the the transition matrix

$$M_{ACG} = \begin{bmatrix} 1-p_A & p_A & 0 \\ 1-p_A-p_C & p_A & p_C \\ 1-p_A-p_G & p_A & 0 \end{bmatrix}.$$

where p_x denotes the nucleotide probability $(x \in \mathcal{A})$. Distance one between two occurrences of ACG is only possible from state S_2. Thus, the probabilities of distance one, from each non-absorbing state are

$$P(D=1) = \begin{bmatrix} P(D=1|S_0) \\ P(D=1|S_1) \\ P(D=1|S_2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ p_G \end{bmatrix}.$$

For higher distances, $d > 1$, the probabilities can be found by combining the transition probabilities for the first step with the probabilities for distance $d-1$, which leads to the recurrence relation

$$\begin{bmatrix} P(D=d|S_0) \\ P(D=d|S_1) \\ P(D=d|S_2) \end{bmatrix} = M_{ACG} \times \begin{bmatrix} P(D=d-1|S_0) \\ P(D=d-1|S_1) \\ P(D=d-1|S_2) \end{bmatrix}$$

where M_{ACG} is the transition matrix of non-absorbing states. Since ACG has only null word overlap besides itself, we must consider S_0 as the initial state. Therefore, under an independent symbol model, the reference probability distribution of inter-ACG distances is given by

$$f(d) = P(D=d|S_0).$$

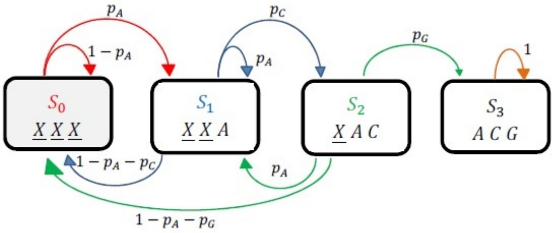


Figure 1: State diagram associated to inter-ACG distances (initial state S_0).

For the generic word w, the reference distance distribution under the independent nucleotide model is given by $f(d) = P(D=d|S_I)$, with

$$\begin{bmatrix} P(D=d|S_0) \\ \vdots \\ P(D=d|S_{k-1}) \end{bmatrix} = M^{d-1} \times \begin{bmatrix} P(D=1|S_0) \\ \vdots \\ P(D=1|S_{k-1}) \end{bmatrix},$$

and

$$P(D=1) = \begin{bmatrix} 0 & \dots & 0 & p_{x_k} \end{bmatrix}^T.$$

where p_{x_k} is the occurrence probability of nucleotide x_k and M is the transition matrix of non-absorbing states.

Our approach to obtain the exact distribution of inter-word distances is a special case of Fu's procedure based on finite Markov chain embedding [6, 7]. To find the transition matrix for a given word requires "a deep understanding of the structure of the specified pattern" [6]. Next, we propose a general expression to compute the transition matrix of non-absorbing states $M = [m_{ij}]$, with $i, j = 0, \ldots, k-1$, based on the concept of word overlap.

Let us denote by $\mathcal{L}(w_1, w_2)$ the length of the longest overlap (a suffix of w_1 that matches with a prefix of w_2) between words w_1 and w_2. Being in S_i means we have just read symbols that match w^i. The next symbol x, appended to w^i, determines the next state. A transition from S_i to S_j with $j > 0$ is only possible if $\mathcal{L}(w^i x_j, w) = j$, so its probability is

$$\text{for } j > 0, \quad m_{ij} = \begin{cases} p_{x_j} & , \quad \mathcal{L}(w^i x_j, w) = j \\ 0 & , \quad \text{otherwise} \end{cases} .$$

And the probability of a transition from S_i to S_0 ($j = 0$) is given by the complementary probability

$$m_{i0} = \begin{cases} 1 - p_{x_{i+1}} - \sum_{s=1}^{i} m_{is} & , \quad i \geq 1 \\ 1 - p_{x_{i+1}} & , \quad i = 0 \end{cases} .$$

The reference distribution under independent nucleotide structure, that we just described, can easily be computed for any whole-genome and for any genomic word, using only four input parameters: the nucleotide frequencies in the sequence.

Measures

To evaluate the goodness of fit between the inter-oligonucleotide distance distribution and the corresponding reference distribution we used the chi-square statistic and the phi coefficient. We also used an effect size measure, Cohen's d, to identify the existence of exceptional distances inside the distribution of a single word.

Due to the sensitivity of these measures to low frequencies that occur for longer distances, we made a cutoff at the 99th percentile of the empirical distribution, $d_{0.99}$. Then, we grouped all distances larger than $d_{0.99}$ in one residual class, $\widetilde{d} = d_{0.99} + 1$.

The empirical distance distribution is given by

$$q_i = \frac{n_i}{N'}, \text{ for } i = 1, \ldots, d_{0.99}$$

and the remaining frequency, $q_{\widetilde{d}}$, where n_i is the number of occurrences of distance i and N' is the total number of inter-w distances. In order to match the size of the reference distribution to the empirical distribution we also made a cutoff in the reference distribution, at $d_{0.99}$.

To extract the exceptional words of each species, we compare the empirical distribution to the corresponding reference distribution under the nucleotide independence (model I). A word is considered exceptional if the empirical inter-word distance and the reference distribution are distinct in a statistically precise way. There are two cases to consider: either the two distributions show a global misfit or there is at least one distance value that deviates significantly from the reference distribution. In the first case, the empirical distribution shows a global misfit to the random background; in the second case, the misfit is more noticeable for specific distances.

To test the goodness of the fit between the empirical and the reference distributions, for each oligonucleotide w, we can use a chi-square statistic, denoted by X^2_w,

$$X^2_w = \sum_{i=1}^{d} \frac{(n_i - f_i \cdot N')^2}{f_i \cdot N'}.$$

To obtain an effect size measure to evaluate the lack of goodness of fit, we use the phi coefficient, denoted by φ_w,

$$\varphi_w = \sqrt{\frac{X^2_w}{N'}}.$$

A perfect fit between the distributions corresponds to $\varphi_w = 0$. We consider a value above 0.10 as a descriptor for small effect size, above 0.30 for medium effect size, above 0.50 for large effect size ([4]), above 0.60 for strong effect size and above 0.80 for a very strong effect size ([13])

For each inter-w distance distribution we are interested in identifying and evaluating the existence of exceptional distances, i.e., distances that occur with a frequency much higher than the expected value. In order to obtain a standard score able to compare how exceptional a distance is over all oligonucleotides of the same length, we use Cohen's d given by

$$CD_i = \frac{q_i - f_i}{\sqrt{f_i(1 - f_i)}}.$$

For reporting and interpreting Cohen's d, we considered a value above 0.20 as a descriptor for small effect size, above 0.50 for medium effect size and above 0.80 for large effect size ([4]). We established those acceptance thresholds as the levels above which the distance is considered exceptional or very exceptional, respectively.

To identify the most exceptional distance inside a distribution, if there is one, we use Cohen's d effect size. After computing Cohen's d for all distances up to the 99th percentile, we identify the distance d for which the maximum Cohen's d is attained and consider it the candidate to the most exceptional distance of the distribution, i.e., $C_d = \max\{CD_i : i = 1, \ldots, d_{0.99}\}$.

The expected values for distances less than or equal to k (the word length) can be null for certain words. For example, the distances between the word AAA in the text $AAAAAAA\cdots$ can never be 2 or 3. Such zero distances were not considered in the computation of the mentioned measures.

3 Results and discussion

Exceptional distance distributions in human genome

We are interested in exceptional distributions, i.e., empirical distributions that either show a significant global misfit to the reference distribution or that exhibit frequencies much higher than expected for specific distances. For all words, we observe the existence of statistical significant differences between empirical and reference distributions (*p-value* < 0.001).

In order to evaluate the lack of fit phenomenon over all words of the same length, we computed the phi coefficient, φ_w, and sorted the word distance distributions according to the value of φ_w. We observe that CG-rich words (i.e., words comprising one or more CG) and words with long word overlap lead to the poorest goodness of fit, in relation to the reference model (see Table 2).

This means that these word distributions have a global misfit or a few distances with exceptional misfit to the reference distribution, in a whole-genome analysis. Let us note that the top-two dinucleotides correspond to well known local motifs (recurrent CG pairs in CpG islands and the TATA binding boxes on transcription start sites). Other high-scoring words may be related to biological motifs.

Conversely, we observe that words with no overlap and without CGs attained the lowest divergences.

Table 2: Phi coefficient between empirical and reference distributions, in the *Homo Sapiens* genome. The maximum and minimum φ_w, the words distributions which present the ten largest and the ten smallest values of φ_w, organized by word length (k).

k	1	2	3	4	5
max(φ_w)	0.191	1.72e+05	3.11e+05	3.84e+12	8.84e+19
min(φ_w)	0.136	0.209	0.116	0.101	0.127
highest φ_w	C	CG	CGA	CGCG	ACGCG
2nd highest	G	TA	TCG	CGAC	CGCGT
3rd highest	-	CC	CGC	GTCG	CGTCG
4th highest	-	GG	GCG	ATCG	CGACG
5th highest	-	GC	ACG	TACG	CGCGA
6th highest	-	AT	CGT	CGTA	TCGCG
7th highest	-	AC	CCG	TCGA	CGGCG
8th highest	-	GT	CGG	TTCG	CGCCG
9th highest	-	-	ATA	CGAA	CGATA
10th highest	-	-	TAT	TCGT	TATCG
⋮					
10th lowest	-	-	TGT	ACTT	CTCTA
9th lowest	-	-	ACA	AAGT	TAGAG
8th lowest	-	AA	CAA	GACA	TCAGT
7th lowest	-	TT	TTG	TGTC	TGACT
6th lowest	-	AG	ACT	ATCT	AGTCA
5th lowest	-	CT	AGT	AGAT	ACTGA
4th lowest	-	TC	TCA	ATGC	AAGCT
3rd lowest	-	GA	TGA	GCAT	AGCTT
2nd lowest	T	CA	ATG	GCTT	AGAGT
lowest φ_w	A	TG	CAT	AAGC	ACTCT

It is known that the human genome has low CG content ([9]). For inter-oligonucleotide distances, the information about CG content ($k = 2$) or CG-rich word ($k > 2$) contents in the sequence is not included in model I. Under this assumption, CG-rich words reach higher phi coefficients and, as a consequence, these words will be identified as exceptional words.

Using Cohen's d, we explored the existence of exceptional distances inside a single distribution, i.e., specific distances with an occurrence probability much higher than expected. Consider, for example, the unexpected spike at distance 24 in the inter-TGCA distance distribution, $C_{24} = 0.616$ (Figure 2).

Note that a high Cohen's d could result from a generalized misfit between the empirical and the reference distribution, rather than from a genuine exceptionality of that distance. Thus, we suggest a practical decision based on the goodness of fit between empirical and reference distance distributions: for one empirical distance distribution that presents moderate to strong discrepancy ($0.2 < \varphi_w < 0.8$) we use 0.5 as the cut point on Cohen's d to identify exceptional distances. For the human genome, only eleven inter-word distributions have been identified as comprising exceptional distances. We do not observe the presence of exceptional distances in

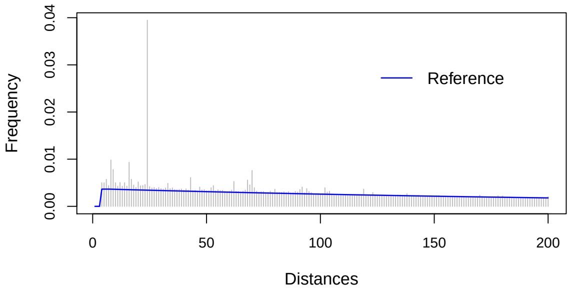


Figure 2: Empirical distance distribution *vs* reference distribution: w =TGCA, $\varphi_w = 0.694$, $C_{24} = 0.616$.

distance distributions for word lengths less than 4 (see Table 3). Figure 3 shows two inter-word distance distributions that comprise an exceptional distance, by our criteria. This procedure detects exceptional words based on their atypical distance distribution along the sequence and not on their frequency of occurrence.

Table 3: Number of distance distributions with moderate or strong lack of fit ($0.2 < \varphi_w < 0.8$) that present an exceptional distance, organized by strength of effect size and word length.

Strength of Cohen's d maximum	word length			
	2	3	4	5
medium effect size ($0.5 \leq C_d < 0.8$)	0	0	1	10
large effect size ($C_d \geq 0.8$)	0	0	0	0

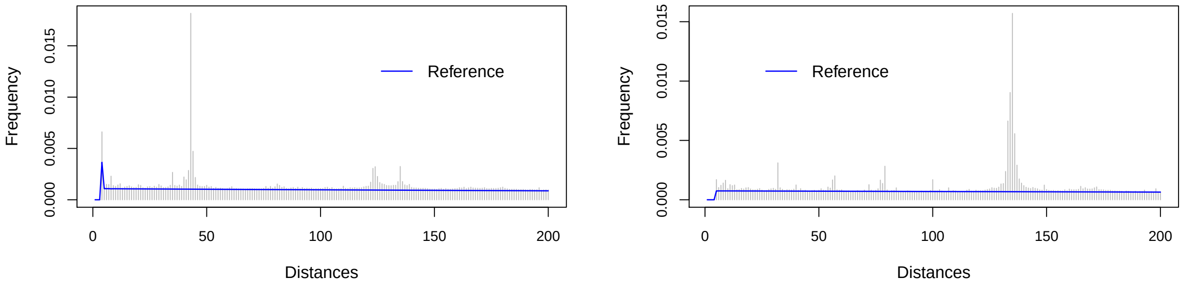


Figure 3: Empirical distance distribution *vs* reference distribution: w =TCACT, $\varphi_w = 0.633$, $C_{43} = 0.533$ (left); w =ATCCC, $\varphi_w = 0.791$, $C_{135} = 0.577$ (right).

This procedure may lead to the identification of new motifs. For example, a word with a perfectly ordinary overall frequency of occurrence may exhibit an abnormal "preference" for occurring at a distance d from the previous occurrence and a slightly decreased preference for occurring at other distances.

Analysis of multiple organisms

Taking into account the empirical distance behaviour and the random background (model I), we introduce exceptionality word criteria and define dichotomic vectors, that may be used as a genomic signature of species.

Consider the following exceptionality word criteria:

- Misfit criterion: the word shows a very strong dissimilarity effect between distributions, $\varphi_w > 0.8$, highlighting the contribution of selective evolution [12];
- Peak criterion: the word has a small or medium dissimilarity effect between distributions and presents a peak with medium or large effect size, $0.2 < \varphi_w < 0.8 \wedge C_d > 0.5$.

Consider, for each specie, a dichotomic vector that marks as nonzero the words identified as exceptional accordingly to one of the criteria. These vectors allows to build dendrograms, which could then be interpreted as phylogenetic trees.

We performed a hierarchical analysis of the 30 species listed in Table 1, considering each one of the exceptionality criteria. The dendrograms were build using the average linkage method. The similarity matrix was computed using the Euclidean distance. In the case of the *misfit criterion*, the dendrogram displays a first branching between eukaryotes and non-eukaryotes (Figure 4a). Inside the eukaryote cluster, we observe that some related species are grouped in the same branch. For instance, primates (*H.sapiens*, *P.troglodytes* and *M.mulatta*), the rodentia (*M.musculus* and *R.norvegicus*) and the fungi (*S.cerevisiae* and *C.pombe*). In the second branch it is observed that, in general, bacteria and archaeotas are closer to each other and separated from the virus. We also notice that the bacteria *S.mutansLJ*, *S.mutansSG* and *S.pneumoniae* are in the same cluster. We emphasize that only the animal organisms reveal distance distributions that verify the *peak criterion*. Restricting the analysis to animal organisms, we obtain a dendrogram which reveals the group of primates and the group of rodentia (Figure 4b).

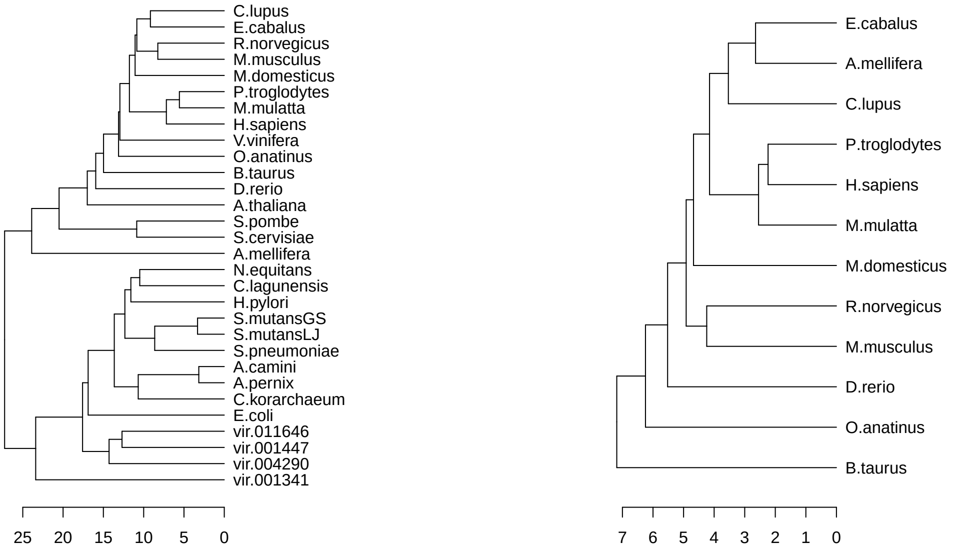
Thus, the binary vector of exceptional words defined by the *misfit criterion* may be used as a genomic signature in all the studied species, while the *peak criterion* can only be used as genomic signature in animal species.

We also constructed dendrograms for the 10 mammal species, using both criteria separately. The obtained dendrograms present some similarities (the split distance between dendrograms is 0.43). We observe that primates are clustered together, as well as the rodentia (Figure 5). These dendrograms support several evolutionary relationships between species. For example, the split distance between our dendrograms and those presented in [17], based in alignment and non-alignment algorithms, is around 50%, which is lower than in random scenarios (see [1]).

4 Conclusions and future research

In this work we studied the inter-word distances in the complete genomes of up to 30 species, for word length k varying between 1 and 5.

We intended to detect exceptional words by comparing the empirical distribution of the inter-word distances with the theoretical one under independent nucleotide model, taking the word overlap structure into account. We evaluated the discrepancy between real sequences and the random background, as a way of emphasizing the contribution of selective evolution. The comparison of the empirical distance frequencies with those that would be observed if the



(a) *misfit criterion* (b) *peak criterion*

Figure 4: Dendrogram of the 30 organisms, with binary vector of exceptional words defined by all words of length 2 to 5 by *misfit criterion* (left); and of the animals by *peak criterion* (rigth).

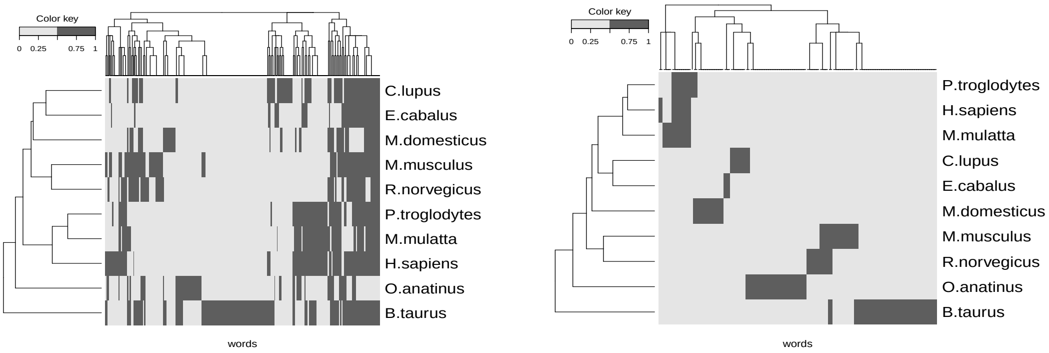


Figure 5: Heat map of mammal species *vs* exceptional words. Binary vectors of exceptional words defined by *misfit criterion* (left) and by *peak criterion* (rigth), considering only vectors with variation.

random background model were valid, allowed us to highlight distinct distance distributions for classes of genomic words.

We introduced a statistical procedure to automatically identify genomic words whose distance distributions show a significant discrepancy from the random background. Our procedure allows to detect some words with a very high lack of fit. These were, in general, words with CG-rich content (as expected). Moreover, we found words with a moderate to strong lack of fit and an unexpected strong spike. Only less than 1 percent of the words of length 4 and 5 show this kind of exceptional distance distribution.

We believe that this procedure, which detects statistically exceptional distributions, may lead to the identification of new motifs. For example, a word with a perfectly ordinary overall frequency of occurrence may exhibit an abnormal "preference" for occurring at a distance d from the previous occurrence and a slightly decreased preference for occurring at other distances.

We also found that the differences mimic, to a certain extent, the evolutionary relationships between the species, which were used to construct dendrograms and perform evolutionary comparisons. In the mammalian organisms, we found matching word dissimilarity values.

In future we intend to extend our procedure to longer words, and evaluate if the method allow to point out known patterns with biological significance. Furthermore, since whole genome are highly heterogeneous, we also expect to perform analysis for detection of regions with exceptional inter-nucleotide distances.

5 Acknowledgements

This work was supported by Portuguese funds through the iBiMED - Institute for Biomedicine, IEETA - Institute of Electronics and Informatics 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 UID/CEC/00127/2013.

Bibliography

[1] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, and P. J. Ferreira. Genome analysis with distance to the nearest dissimilar nucleotide. *Journal of theoretical biology*, 275(1):52–58, 2011.

[2] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, and P. J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, Dec. 2009.

[3] C. A. C. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. a. M. O. S. Rodrigues, and P. J. S. G. Ferreira. Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions. *Journal of Integrative Bioinformatics*, 8(3):172, 2011.

[4] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum, 1988.

[5] S. Ding, Q. Dai, H. Liu, and T. Wang. A simple feature representation vector for phylogenetic analysis of dna sequences. *Journal of Theoretical Biology*, 265(4):618–623, Aug. 2010.

[6] J. C. Fu. Distribution theory of runs and patterns associated with a sequence of multi-state trials. *Statistica Sinica*, 6(4):957–974, 1996.

[7] J. C. Fu and W. W. Lou. *Distribution theory of runs and patterns and its applications: a finite Markov chain imbedding approach.* World Scientific, 2003.

[8] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver. CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, 7(1):446, 2006.

[9] J. Karro, M. Peifer, R. Hardison, M. Kollmann, and H. von Grünberg. Exponential decay of GC content detected by strand-symmetric substitution rates influences the evolution of isochore structure. *Molecular biology and evolution*, 25(2):362–374, 2008.

[10] M. Lothaire. *Applied combinatorics on words*, volume 105. Cambridge University Press, 2005.

[11] A. S. S. Nair and T. Mahalakshmi. Visualization of genomic data using inter-nucleotide distance signals. In *Proceedings of IEEE Genomic Signal Processing*, 2005.

[12] J. Qi, B. Wang, and B.-I. Hao. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. *Journal of molecular evolution*, 58:1–11, 2004.

[13] L. Rea and R. Parker. *Designing and conducting survey research: a comprehensive guide*. Public Administration Series. Jossey-Bass Publishers, 1992.

[14] S. Robin. A compound Poisson model for word occurrences in DNA sequences. *Applied Statistics*, 51, Part 4:437–451, Aug. 2002.

[15] S. Robin, F. Rodolphe, and S. Schbath. *DNA, Words and Models: Statistics of Exceptional Words*. Cambridge University Press, 2005.

[16] S. Robin, S. Schbath, and V. Vandewalle. Statistical tests to compare motif count exceptionalities. *BMC Bioinformatics*, 8(1):84, 2007.

[17] G. E. Sims, S.-R. Jun, G. A. Wu, and S.-H. Kim. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences*, 106(40):17077–17082, 2009.

[18] T. V. Stefanov. The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: an algorithmic approach. *Journal of Applied Probability*, 40:881–892, 2003.