Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions

Carlos A. C. Bastos[1,2*], Vera Afreixo[1,3], Armando J. Pinho[1,2], Sara P. Garcia[1], João M. O. S. Rodrigues[1,2], Paulo J. S. G. Ferreira[1,2]

[1]Signal Processing Lab, IEETA, University of Aveiro, 3810-193 Aveiro, Portugal

[2]Department of Electronics, Telecommunications and Informatics, University of Aveiro, 3810-193 Aveiro, Portugal

[3]Department of Mathematics, University of Aveiro, 3810-193 Aveiro, Portugal

Summary

We study the inter-dinucleotide distance distributions in the human genome, both in the whole-genome and protein-coding regions. The inter-dinucleotide distance is defined as the distance to the next occurrence of the same dinucleotide. We consider the 16 sequences of inter-dinucleotide distances and two reading frames. Our results show a period-3 oscillation in the protein-coding inter-dinucleotide distance distributions that is absent from the whole-genome distributions. We also compare the distance distribution of each dinucleotide to a reference distribution, that of a random sequence generated with the same dinucleotide abundances, revealing the CG dinucleotide as the one with the highest cumulative relative error for the first 60 distances. Moreover, the distance distribution of each dinucleotide is compared to the distance distribution of all other dinucleotides using the Kullback-Leibler divergence. We find that the distance distribution of a dinucleotide and that of its reversed complement are very similar, hence, the divergence between them is very small. This is an interesting finding that may give evidence of a stronger parity rule than Chargaff's second parity rule.

1 Introduction

Finding and understanding correlation structures in genomic sequences has been the goal of many studies, and several methodologies have been employed, including heterogeneities in compositional biases, segmentation techniques, entropy measures, correlation functions, Fourier analysis, wavelet analysis or the analysis of self-similarities (e.g. [20, 12, 15, 4, 3, 2, 9, 7]). We aimed at contributing to this goal by studying the distribution of the distances between similar n-mers. We started with nucleotides, by exploring the inter-nucleotide distance (i.e. the distance to the next occurrence of the same nucleotide) in the genomes of organisms from the three domains of life, and using the distributions for inferring phylogenies [1].

Here, we address the distribution of inter-dinucleotide distances (i.e. the distance to the next occurrence of the same dinucleotide) and focus our analysis on the human genome. Dinucleotides have a prominent role in genome biology, hence, studying their content and distribution is key

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

in any integrative genome analysis strategy. Motivated by biological properties, e.g. the existence of CpG islands in close proximity to regulatory regions of protein-coding genes and the association of CpGs with higher mutational rates, the distributional and compositional patterns of dinucleotides in genomes have long been explored. Examples include understanding inhomogeneities in genomes from overall occurring frequencies and distributional patterns in specific regions and using them for ab initio gene finding methodologies (e.g. [3, 14, 19, 18]) or to uncover genomic signatures (e.g. [8, 10]). Another famous example are Chargaff's parity rules, which describe global frequency rules in the double- and single-stranded DNA molecule and have a curious consequence for the symmetric distribution of a given n-mer and its reversed complement. Recently, extensions of these rules for n-mers of size up to 10 have been observed in many organisms [5, 6, 16]. To the best of our knowledge, the analysis of the non-local distribution of dinucleotides in the short- and long-range scales remains, however, an open question.

We consider the 16 sequences of inter-dinucleotide distances, as well as the global distribution as sequentially read from the genome. Moreover, in order to account for all possible dinucleotides, we consider both reading frames. This methodology allows for the comparison of the 16 individual inter-dinucleotide distance distributions. We use the Kullback-Leibler divergence for comparing the inter-dinucleotide distance distributions of the 16 dinucleotides in both the whole-genome and protein-coding distributions. The Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions, measuring the expected number of extra bits required to code samples from a distribution when using a code based on another distribution [13]. We also present an error analysis of the comparison of these distributions with an independent (i.i.d) random process, in order to extract the random background from the observed inter-dinucleotide distance distributions. From the perspective of molecular evolution, DNA sequences reflect both the results of random mutation and of selective evolution. It has been proposed that one should subtract the random background from the simple counting result, in order to highlight the contribution of selective evolution [17]. Hence, by presenting an analysis of the relative error, we are able to highlight the selective evolution of each dinucleotide.

2 Methods

2.1 Inter-dinucleotide distance sequence

Consider the alphabet $\mathcal{A} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$ and let $s = (s_k)_{k\in\{1,\ldots,N\}}$ be a symbolic sequence defined in $\mathcal{A}$. Consider a numerical sequence, d^x, that represents the inter-dinucleotide distances of dinucleotide x, with $x \in \mathcal{A} \times \mathcal{A}$.

The distance between dinucleotides may be computed by considering each dinucleotide as a new symbol from an alphabet with 16 symbols. There are two possible reading frames, one starting at the first nucleotide of the sequence and the other starting at the second dinucleotide. As an example, consider a genomic sequence

AAACAAACTGACACAAAACACT ...,

whose two possible dinucleotide reading frames (R_1 and R_2) output the following dinucleotide sequences,

$$R_1\text{: } \underbrace{AA}\,\underbrace{AC}\,\underbrace{AA}\,\underbrace{AC}\,\underbrace{TG}\,\underbrace{AC}\,\underbrace{AC}\,\underbrace{AA}\,\underbrace{AA}\,\underbrace{CA}\,\underbrace{CT}\cdots$$
$$R_2\text{: } \underbrace{A}\,\underbrace{AA}\,\underbrace{CA}\,\underbrace{AA}\,\underbrace{CT}\,\underbrace{GA}\,\underbrace{CA}\,\underbrace{CA}\,\underbrace{AA}\,\underbrace{AC}\,\underbrace{AC}\,\underbrace{T}\cdots$$

The distance sequence for each dinucleotide is a vector containing the distances between consecutive occurrences of that dinucleotide. Continuing with the previous example, the beginning of two of the 16 inter-dinucleotide distance sequences for the two reading frames is

$$\begin{array}{cc} d_{R_1}^{AA} = (2, 5, 1, \cdots) & d_{R_2}^{AA} = (2, 5, \cdots) \\ d_{R_1}^{AC} = (2, 2, 1, \cdots) & d_{R_2}^{AC} = (1, \cdots) \\ \cdots & \end{array}$$

We consider also the global dinucleotide distance sequence, which is an extension of the global inter-nucleotide distance previously proposed [1]. The i-th element of this sequence is the distance between the i-th dinucleotide and its next occurrence. Taking the R_1 reading frame as an example, the global dinucleotide distance sequence is

$$d_{R_1} = (2, 2, 5, 2, \cdots),$$

where the the first element in the vector is the distance between the first dinucleotide in R_1 (AA) and its next occurrence (in the third position). Similarly, the second number refers to the second dinucleotide (AC) and so forth. The global inter-nucleotide distance for the R_2 reading frame is

$$d_{R_2} = (2, 4, 5, \cdots).$$

The global inter-dinucleotide distance sequence may also be seen as a merging of the individual inter-dinucleotide distance sequences. Consequently, the observed global inter-dinucleotide distance distribution results from the sum of the individual inter-dinucleotide distance distributions.

2.2 Comparison with an independent random process

Consider p^{AA}, p^{AC}, p^{AG}, p^{AT}, $\cdots p^{\mathrm{TT}}$ the occurrence probabilities of dinucleotides AA, AC, AG, AT, ... TT, respectively. If the dinucleotide sequences were generated by an independent and identically distributed (i.i.d.) random process, then each of the inter-dinucleotide distance sequences, d^x, where x now represents a dinucleotide in an alphabet of 16 symbols, would follow a geometric distribution. In fact, the probability distribution of the inter-dinucleotide distances of symbol x in a random sequence is

$$f^x(k) = p^x(1 - p^x)^{k-1}, \qquad k = 1, 2, ...$$

and the corresponding global distance sequence distribution is

$$f(k) = \sum_{x \in \mathcal{A} \times \mathcal{A}} p^x p^x (1 - p^x)^{k-1}.$$

From the perspective of molecular evolution, DNA sequences reflect both the results of random mutation and selective evolution. It has been proposed that one should subtract the random background from the simple counting result, in order to highlight the contribution of selective evolution [17]. Therefore, we present an analysis of the relative error, in order to highlight the selective evolution of each dinucleotide. The relative error is defined as

$$r(k) = \frac{f(k) - f_{\mathrm{o}}(k)}{f(k)} \ , \tag{1}$$

where $f_{\mathrm{o}}(k)$ is the observed relative frequency of the distance k, and $f(k)$ is the relative frequency of the reference distribution.

To summarize the relative differences between the observed and the reference distribution, we compute the cumulative absolute relative error up to the i^{th} distance,

$$S_r(i) = \sum_{k=1}^{i} |r(k)| \ . \tag{2}$$

2.3 Genomic data

We used the complete genome of *Homo sapiens* build 36.3, obtained from the National Center for Biotechnology Information (NCBI) website (`ftp://ftp.ncbi.nih.gov/genomes/`).

All chromosomes of the human genome were processed separately and the resulting distance counts were added to compute the global distance distribution. For the protein-coding distribution, data was retrieved from the RNA folder in build 37.2. All symbols in the genomic sequences that did not correspond to one of the four standard nucleotides (A, C, G and T) were removed before further processing.

3 Results

The distance distributions of each dinucleotide in the two reading frames were compared by computing the Kullback-Leibler divergence between the two distributions. The maximum divergence value between the corresponding distributions in the two reading frames is $< 8 \times 10^{-4}$ in the whole-genome distribution, and $< 2 \times 10^{-4}$ in the protein-coding distribution. Since, in both cases, the divergences between the two reading frames distributions are small, the distributions of each frame were merged into a single distribution by adding the corresponding distance counts.

In order to compare the distance distribution of the 16 dinucleotides, we computed the Kullback-Leibler divergence between the 16 dinucleotides in the human genome. The comparison results are shown in Table 1 for the whole-genome distribution, and in Table 2 for the protein-coding distribution. For the whole-genome distribution, there are very small ($< 7 \times 10^{-5}$) values for the divergence between the distance distributions of each dinucleotide and that of its reversed complement (AA-TT, AC-GT, AG-CT, CA-TG, CC-GG, GA-TC). For the protein-coding distribution, all divergence values for the reverse complement pairs are $> 2 \times 10^{-3}$ and $< 16 \times 10^{-3}$.

Table 1: Kullbak-Leibler divergence between the 16 dinucleotides in the whole-genome distance distribution. The highlighted values (in grey) correspond to reversed complement dinucleotide pairs.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA		0.275	0.106	0.061	0.112	0.211	1.190	0.105	0.147	0.610	0.211	0.273	0.090	0.147	0.111	0.000
AC	0.316		0.086	0.143	0.110	0.041	0.649	0.086	0.025	0.135	0.041	0.000	0.074	0.025	0.111	0.318
AG	0.105	0.076		0.015	0.006	0.052	0.875	0.000	0.025	0.220	0.052	0.075	0.013	0.025	0.006	0.106
AT	0.058	0.125	0.014		0.013	0.078	0.950	0.014	0.048	0.288	0.078	0.124	0.016	0.048	0.013	0.059
CA	0.106	0.090	0.005	0.013		0.069	0.905	0.005	0.036	0.228	0.069	0.089	0.020	0.036	0.000	0.107
CC	0.240	0.044	0.074	0.099	0.104		0.657	0.074	0.032	0.104	0.000	0.044	0.040	0.032	0.105	0.242
CG	2.966	2.144	2.894	2.755	3.291	1.682		2.891	2.312	1.565	1.692	2.158	2.043	2.313	3.311	2.969
CT	0.105	0.076	0.000	0.015	0.006	0.052	0.875		0.025	0.220	0.052	0.076	0.013	0.025	0.006	0.106
GA	0.167	0.025	0.028	0.054	0.043	0.030	0.753	0.028		0.194	0.030	0.024	0.019	0.000	0.043	0.169
GC	0.521	0.093	0.221	0.281	0.258	0.079	0.571	0.222	0.137		0.079	0.094	0.180	0.137	0.260	0.525
GG	0.239	0.044	0.074	0.099	0.104	0.000	0.657	0.074	0.031	0.104		0.044	0.040	0.032	0.105	0.242
GT	0.314	0.000	0.085	0.142	0.109	0.041	0.651	0.085	0.024	0.135	0.041		0.074	0.024	0.110	0.316
TA	0.097	0.072	0.016	0.019	0.028	0.038	0.830	0.016	0.019	0.228	0.038	0.071		0.019	0.028	0.098
TC	0.167	0.025	0.028	0.054	0.043	0.030	0.754	0.028	0.000	0.194	0.030	0.024	0.019		0.043	0.169
TG	0.105	0.091	0.005	0.013	0.000	0.069	0.907	0.005	0.036	0.229	0.069	0.090	0.020	0.036		0.107
TT	0.000	0.277	0.107	0.062	0.113	0.212	1.192	0.106	0.149	0.613	0.212	0.275	0.091	0.149	0.112	

Table 2: Kullbak-Leibler divergence between the 16 dinucleotides in the protein-coding distance distribution. The highlighted values in grey correspond to reversed complement dinucleotide pairs and the values highlighted in bold correspond to dinucleotide pairs with divergence below 10^{-2}.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA		0.185	0.058	0.103	0.089	0.047	0.223	0.074	0.064	0.187	0.070	0.194	0.142	0.114	0.071	**0.007**
AC	0.189		0.098	0.033	0.096	0.112	0.034	0.077	0.052	0.071	0.077	**0.004**	0.015	0.011	0.117	0.132
AG	0.052	0.084		0.023	**0.003**	0.018	0.132	**0.003**	**0.009**	0.045	0.011	0.110	0.070	0.039	**0.002**	0.031
AT	0.092	0.033	0.027		0.026	0.029	0.061	0.018	0.014	0.021	0.012	0.051	0.019	0.012	0.039	0.053
CA	0.075	0.082	**0.003**	0.021		0.024	0.130	**0.003**	0.014	0.029	0.013	0.111	0.072	0.040	**0.003**	0.049
CC	0.040	0.111	0.019	0.029	0.026		0.130	0.023	0.027	0.046	**0.005**	0.137	0.073	0.060	0.029	0.019
CG	0.257	0.042	0.209	0.076	0.209	0.159		0.185	0.131	0.105	0.129	0.040	0.020	0.069	0.248	0.187
CT	0.066	0.066	**0.003**	0.014	**0.003**	0.022	0.117		0.009	0.036	0.011	0.090	0.056	0.028	**0.006**	0.039
GA	0.061	0.048	**0.010**	0.014	0.015	0.026	0.092	0.009		0.048	0.014	0.067	0.043	0.015	0.016	0.035
GC	0.133	0.068	0.042	0.019	0.031	0.040	0.086	0.034	0.038		0.021	0.100	0.052	0.044	0.049	0.089
GG	0.059	0.076	0.013	0.012	0.014	**0.005**	0.100	0.012	0.014	0.024		0.102	0.051	0.036	0.021	0.031
GT	0.208	**0.004**	0.130	0.052	0.132	0.138	0.035	0.106	0.073	0.105	0.103		0.018	0.021	0.155	0.148
TA	0.151	0.015	0.098	0.021	0.100	0.079	0.019	0.079	0.053	0.058	0.056	0.019		0.018	0.124	0.096
TC	0.115	0.010	0.044	0.012	0.046	0.059	0.051	0.031	0.016	0.050	0.035	0.020	0.015		0.058	0.072
TG	0.063	0.095	**0.002**	0.030	**0.003**	0.026	0.147	**0.006**	0.014	0.048	0.018	0.125	0.085	0.048		0.043
TT	**0.007**	0.130	0.034	0.058	0.057	0.022	0.162	0.042	0.036	0.123	0.036	0.141	0.091	0.072	0.048	

Other divergence values are of the same order of magnitude, namely, the pairs AG-CA, AG-GA, AG-TG, CA-CT, CT-TG have divergence values $< 10^{-2}$. As expected, considering the size of the data set, the differences between distance distributions are statistically significant.

Figure 1 shows the observed whole-genome distance distribution for the first 100 distances of each dinucleotide (the distributions of the corresponding reversed complements are not shown). Seven of the plots in Figure 1 show an approximately exponential decay starting at the first distance, and three of them (CC, CG and GC) show an increase at the first distances. Because of the typical exponential decay in all distance distributions, we chose to analyse the relative

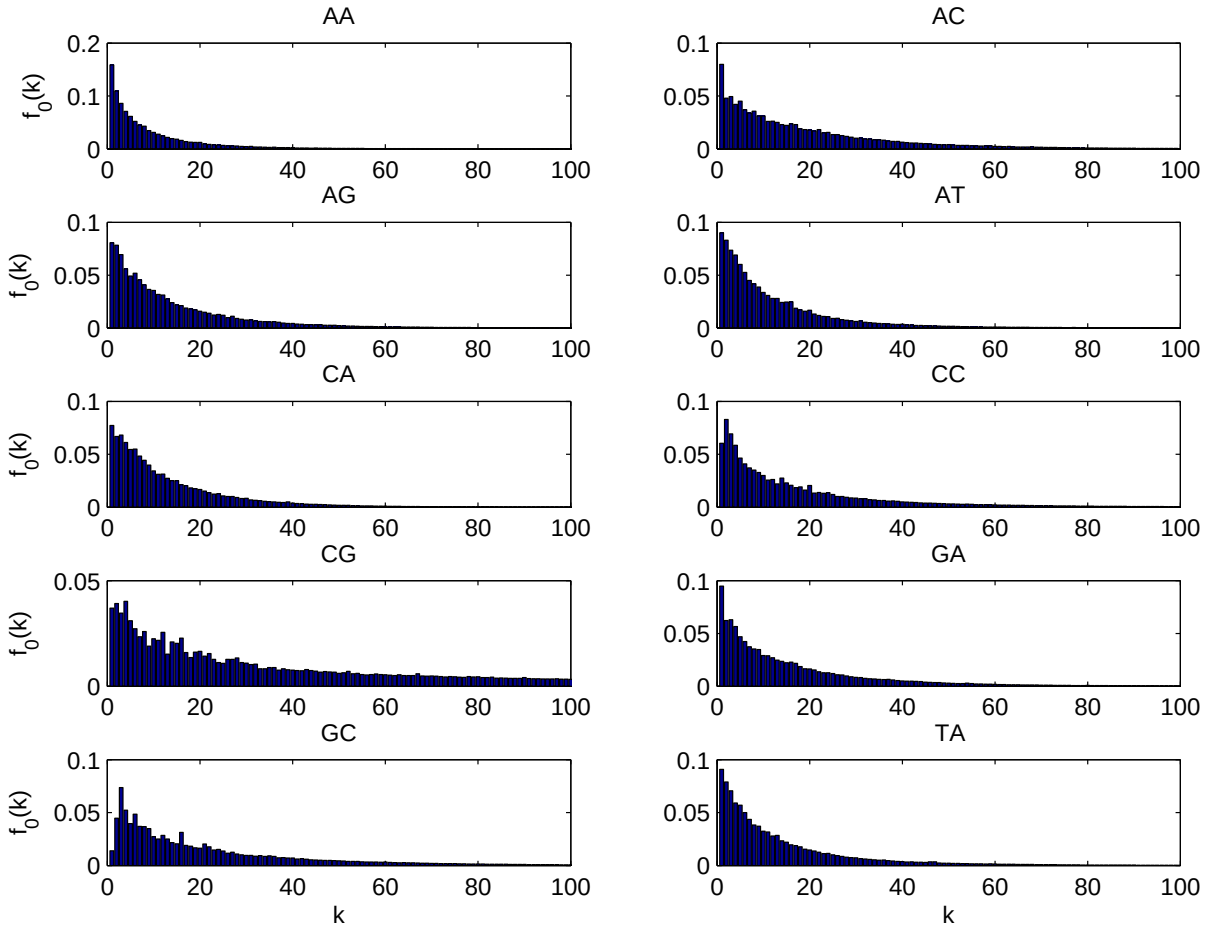

Figure 1: Distribution of the first 100 distances of 10 dinucleotides in the whole human genome (the distributions of the non displayed dinucleotides are very similar to those of the corresponding reversed complement).

error profiles instead.

Figure 2 displays the relative error profiles for each dinucleotide, for both the whole-genome (top panel) and protein-coding (bottom panel) distributions. The main difference between these profiles (whole-genome and protein-coding) is the periodic oscillatory behavior in the protein-coding profiles. In the whole-genome profiles (top panel), the most noteworthy feature is the singular behaviour of the AA and TT dinucleotides, with a sharp increase of the absolute value of the error for distances larger than 50. As for the protein-coding profiles (bottom panel), this increase in the absolute relative error for larger distances is more evident for AA, followed by TT, CC and GG (all dinucleotides with self-overlapping).

Figure 3 shows the comparison of absolute cumulative errors for the first 100 distances of each dinucleotide, in both the whole-genome (top panel) and protein-coding (bottom panel) distributions. We observe a distinct global behavior of dinucleotide CG in both scenarios (whole-genome and protein-coding). For distances smaller than 60, it has the highest cumulative error and seems to have a logarithmic-like behavior, while some of the other dinucleotides show an exponential-like behavior.

In the whole-genome profiles (top panel), the cumulative errors of dinucleotides AA and TT reveal a large exponential increase that surpasses the cumulative relative error of CG near the

60th distance. Between distances 90 and 100 the relative error of dinucleotides AT and TA also surpasses the relative error of CG. As for the protein-coding profiles (bottom panel), the same exponential increase is observed for dinucleotides AA, TT, CC and GG, but at decreasing rates. The behavior of dinucleotide TA is slightly different here, with a notorious increase for distances larger than 50.

The value of the cumulative absolute error of the whole-genome distribution is higher than the cumulative absolute error of the protein-coding distribution.

Figure 4 displays the global relative error, i.e. the relative error of the global distribution, for both the whole-genome (left panel) and protein-coding (right panel) distributions. These profiles summarize the 16 profiles of individual dinucleotides. Again, the most obvious difference, is the periodic oscillations of period 3 in the protein-coding distribution.

We used the Discrete Fourier Transform (DFT) to characterize the periodicity observed in the global relative error of the protein-coding distribution (bottom panel in Figure 2 and right panel in Figure 4). Figure 5 shows the spectrum of the relative error for the whole-genome (left panel) and protein-coding (right panel) global distributions. From the protein-coding panel in Figure 5 (right), a local peak at $k = N/3$ can be seen, which corresponds to a period of three samples. It is known that the symbolic autocorrelation spectrum of protein coding DNA regions typically has a peak at $k = N/3$ frequency [3, 19, 18]. It has been shown in [2] that the symbolic autocorrelation spectrum and the indicator sequences spectrum are equivalent concepts. Note that this peak is much more pronounced for the inter-dinucleotide distance distributions than for the corresponding inter-nucleotide distance distributions [1].

4 Conclusion

In this work, we have studied the distribution of inter-dinucleotide distances in the human genome, which contains information about dinucleotide repetition structures in the genome. We compared the distributions of the whole-genome and the protein-coding regions of the human genome, in order to uncover possible differences. Moreover, we considered both reading frames, in order to account for all possible dinucleotides. We have found that the distribution of inter-dinucleotide distances does not depend significantly on the reading frame, but it has different properties for distinct regions of the genome. For example, the protein-coding distributions have a pronounced oscillatory behavior not present in the whole-genome counterpart.

Three dinucleotides – AA, TT and CG – have the overall highest absolute cumulative relative error in the first 100 inter-dinucleotide distances (Figure 3), though the behavior of CG is considerably different (for example, its cumulative error does not have an exponential increase until distance ~ 400). It is well know that CG is under-represented in the human genome (e.g. [8, 10]) and that CG clustering is species-specific [11]. This clustering behavior is naturally reflected in the inter-dinucleotide distance distribution. In the protein-coding distributions, some differences are observed, but the dinucleotide CG still presents a particular profile, and dinucleotides AA, TT are now joined by dinucleotides CC and GG as those with the overall highest absolute cumulative relative error.

The protein-coding distance distributions show a synchronized period-3 oscillatory behavior for all dinucleotides. We are convinced that this periodicity may be at least partially explained

by the triplet nature of the protein-coding genetic code and of codon usage. We also believe that this periodicity may be used to improve gene location algorithms that partly rely on the indicator sequences spectrum from DFT analysis, as we found a pronounced peak in the amplitude spectrum.

One of the most interesting findings that resulted from this work was the observation that the whole-genome distance distribution of an arbitrary dinucleotide is almost identical to that of its reversed complement. Although the distance distributions of the protein coding sequences also show similarities between dinucleotides and its reverse complements, this behavior is less marked. It is well known that the frequency of occurrence of a given n-mer is very similar to that of its reversed complement, even considering only one of the strands. This is known as Chargaff's second parity rule and it has been observed for values of n up to 10 in most organisms [5, 6, 16]. It is not obvious that Chargaff's second parity rule and its extensions fully explain this similarity, as Chargaff's rule solely pertains occurring frequencies, and our inter-dinucleotide distance distributions accounts for the spacing amongst dinucleotide occurrences. In fact, we are convinced that this observation may be related to a new parity rule, stronger than Chargaff's, that not only relates the number of occurrences of the n-mers with their corresponding reversed complements, but also relates the distances at which they occur. However, this is still a conjecture needing further study.

Acknowledgements

Sara P. Garcia acknowledges funding from the European Social Fund and the Portuguese Ministry of Science, Technology and Higher Education. Partially funded by FCT.

References

[1] Vera Afreixo, Carlos A. C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J. S. G. Ferreira. Genome analysis with inter-nucleotide distances. *Bioinformatics*, 25(23):3064–3070, 2009.

[2] Vera M. A. Afreixo, Paulo J. S. G. Ferreira, and Dorabella M. S. Santos. Fourier analysis of symbolic data: A brief review. *Digital Signal Processing*, 14(6):523–530, 2004.

[3] Vera M. A. Afreixo, Paulo J. S. G. Ferreira, and Dorabella M. S. Santos. Spectrum and symbol distribution of nucleotide sequences. *Phys. Rev. E*, 70(3):031910, 2004.

[4] Mahmood Akhtar and Julien Epps. Signal processing in sequence analysis: Advances in eukaryotic gene prediction. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):310–321, 2008.

[5] Guenter Albrecht-Buehler. Asymptotically increasing compliance of genomes with Chargaff's second parity rules through inversions and inverted transpositions. *Proc. Nat. Acad. Sci. USA*, 103(47):17828–33, 2006.

[6] Guenter Albrecht-Buehler. Inversions and inverted transpositions as the basis for an almost universal "format" of genome sequences. *Genomics*, 90:297305, 2007.