On the complex distribution of relevant words in the human genome

P. Carpena[1], P. Bernaola-Galván[1] , A.V. Coronado[1] C. Carretero-Campos[1], M. Hackenberg[2] and J.L. Oliver[2]
[1]*Dpto. de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga, 29071, Málaga, Spain.*
[2]*Dpto. de Genética, Inst. de Biotecnología, Universidad de Granada, 18071 Granada, Spain.*

In the last years our group has developed some tools able to detect relevant words (keywords) in written texts[1, 2]. The subjacent idea is that a word relevant to a text is not uniformly distributed along the text, but is more concentrated in certain regions of the text and more rarefied in others, giving raise to clusters in the spatial distribution of the word. The larger the clustering of the word in the text, the large the relevance of the word to the text considered. Thus, by measuring properly with a single numerical quantity the degree of clustering for all the different words appearing in the text, a ranking of word relevance can be obtained. We have also shown that when blank spaces and punctuation marks are removed from the text thus obtaining a continuous text (similar to DNA), our method still works properly and identifyies correctly relevant words to the text considered [2].

Very recently we have extended these ideas to human DNA. In particular, we have shown that n-mers with high clustering appear with higher probability than expected (are *enriched*) in known functional regions, and with lower probability than expected in known non-functional regions [3]. Thus, we have stablished a link between the clustering of a 'word' (a n-mer) and its posible biological function.

However, little is known about the real spatial distribution of n-mers, relevant or not. Even for n-mers with high clustering, the numerical quantity measuring the clustering does not take into account if the spatial distribution is highly complex or not, i.e., if there is a complex organization of the appearances of the word at different scales, or if the clustering is due to a more trivial spatial organization, as for example the cases of strings of repeated nucleotides ('AAAA...') or the TATA-box.

In this talk we present results on the spatial distribution of n-mers obtained using different tools. First, we analyze the distribution of distances d between consecutive appearances of a word. Our results indicate that, in general, clustered (relevant) n-mers present a probability density $p(d)$ spanning over larger distances than for non-relenvant n-mers (as expected). But, in addition, if the logarithms of consecutive distances ($\log_{10}(d)$) are studied, its probability density $p(\log_{10}(d))$ shows deeper differences between relevant and non-relevant n-mers. While non-relevant n-mers (as TAAGCC in Fig. 1) present a clear monomodal profile in $p(\log_{10}(d))$ with the peak centered at the mean value suggesting an almost homogeneous distribution, for relevant n-mers (as GGCGGC in Fig. 1) the distribution $p(\log_{10}(d))$ is multimodal, reflecting a complex structure in the spatial organization of such n-mers. However, the study of $p(d)$ or of $p(\log_{10}(d))$ only gives information of the organization and interactions of nearest-neighbor appearances of the n-mer in the DNA sequence, but not of of the structure of the total set of appearances of the word.

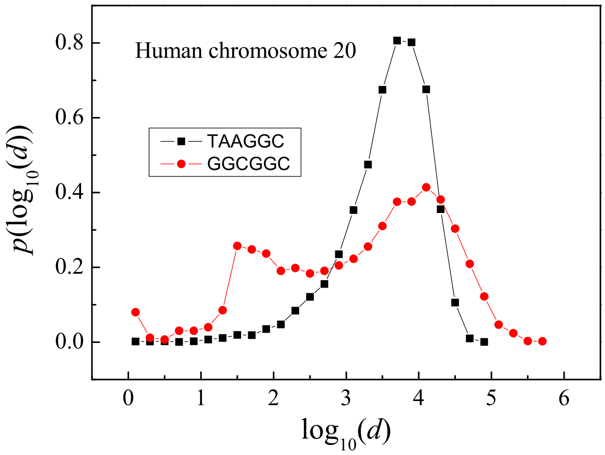

FIG. 1: Probability densities of the logarithms of the distances between consecutive appearances of the relevant word GGCGGC and the non-relevant word TAAGGC in the human chromosome 20.

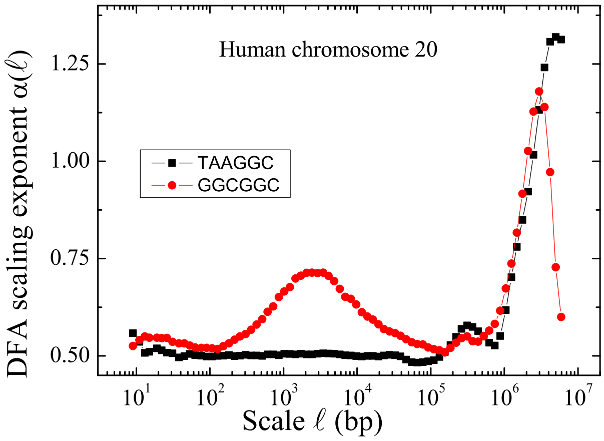

FIG. 2: Local DFA scaling exponent α as a function of the scale of observation ℓ for the relevant word GGCGGC and the non-relevant word TAAGGC in the human chromosome 20.

Therefore, second, and to gain insight into the complexity of the global spatial organization of relevant n-mers, for any n-mer (with $n = 6$) we consider a surrogated sequence from the original DNA sequence in which we put a '1' in the position of the DNA sequence where the studied n-mer appears and '0' otherwise. Once this numerical series is available for any n-mer, we study the long-range correlations properties of the series using detrended fluctuation analysis (DFA) [4], a technique which calculates the fluctuations $F(\ell)$ of the series around its local trend at scale ℓ. Scaling appears if $F(\ell) \propto \ell^{\alpha}$, and correlations are quantified by the exponent α: if $\alpha = 0.5$, the series is uncorrelated, while for $\alpha > 0.5$, there are positive correlations with strength increasing with α. By performing the logarithmic derivative $d\log(F(\ell))/d\ell$, we can calculate the scaling exponent at scale ℓ, $\alpha(\ell)$ and the existence of characteristic scales [5]. For non-relevant n-mers (as TAAGCC in Fig. 2) $\alpha \simeq 0.5$ up to scales of about 10^6 bp, indicating a random uncorrelated organization of the word in this range of scales. Interestingly, for relevant n-mers (as GGCGGC in Fig. 2), $\alpha(\ell)$ presents a clear peak at scales $10^3 - 10^4$ bp, indicating strong compositional complex fluctuations for the word at these scales, probably associated to the existence of functional elements of these sizes linked to the relevant word. In addition, in both relevant and non-relevant cases there exists a large peak of $\alpha(\ell)$ at very large scales ($\ell > 10^6$ bp), very likely related to the gigantic superstructure organization of the human genome [6].

Acknowledgments

We thank the Spanish Junta de Andalucía (Grants P07-FQM3163, and FQM362) for financial support.

[1] M. Ortuño et al., Europhys. Lett. **57**, 759 (2002).

[2] P. Carpena et al., Phys. Rev. E **79** 035102R (2009).

[3] M. Hackenberg et al., J. Theor. Biol. **297**, 127 (2012).

[4] C.-K. Peng et al., Phys. Rev. E **49**, 1685 (1994).

[5] P. Carpena et al., Phys. Rev. E. **75**, 032903 (2007).

[6] P. Carpena et al., Phys. Rev. E **83**, 031908 (2011).