Very-large-scale compositional organization of the human genome

P. Bernaola-Galván[1], P. Carpena[1], A.V. Coronado[1] G. Barturen[2], 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.*

Human DNA is known to have a very complex compositional structure, since there are genomic elements with clear compositional features of many different scales as CpG islands, genes, repeat elements (SINEs, LINEs), segmental duplications, etc. The largest compositional organization well documented and systematically analyzed are the isochores, firstly identified by Bernardi and coworkers by analytical ultracentrifugation of bulk DNA [1, 2] and which are present in the genomes of warm-blooded vertebrates. At DNA sequence level, isochores are large segments with a typical size around 10^5 bp and relatively homogeneous G+C composition harboring the rest of genomic elements. In this context, the human genome is currently viewed as a mosaic of isochores, which define the large scale compositional organization of the genome.

However, we found that human isochores seem to be organized at larger scales into compositional segments of about two orders of magnitude larger than isochores [3]. This high-level organization of isochores can be indirectly unveiled by means of compositional autocorrelation analysis. This method shows that the G+C content of isochores is not independent and is actually correlated up to very large distances, indicating the existence of clusters of isochores of similar composition (Fig. 1a). The use DNA walks also shows the existence of such huge DNA segments ($\sim 15-20$ Mb) with definite G+C composition and typical sizes in agreement with the sizes of the isochore clusters obtained via autocorrelation analysis (Fig. 1b).

We propose the use of a segmentation algorithm to systematically detect these compositional superstructures on the basis of rigorous statistical criteria. The segmentation is a standard method widely used in DNA analysis [5] that finds the change-points dividing a non stationary sequence into homogeneous segments. Nevertheless, none of these approaches is able to obtain such gigantic compositional structures. Instead, they detect much smaller segments — around 10^5 bp on average in the best case [6]. The reason for this over segmentation resides in the fact that most of these segmentation techniques take as the reference for homogeneity a random i.i.d. sequence, i.e.: A sequence should remain undivided only if its heterogeneities are similar to those found in a random i.i.d. sequence. However, it is known that DNA sequences are far from behaving like random sequences and that they present long-range correlations of complex nature in their G+C composition, showing heterogeneities at all scales and that these correlations can be modeled as Fractional Gaussian noise with $\beta \simeq 0.6$ [8]. For such long-range fractal correlated sequences, most of the segmentation techniques detect spurious change-points which are simply due to the heterogeneities induced by the correlations and not to real nonstationarities. To avoid this over segmentation, we present a segmentation algorithm which takes as the reference for homogeneity, instead of a random i.i.d. sequence, a correlated random sequence modeled by a fractional noise with the same degree of correlations as the sequence to be segmented [7].

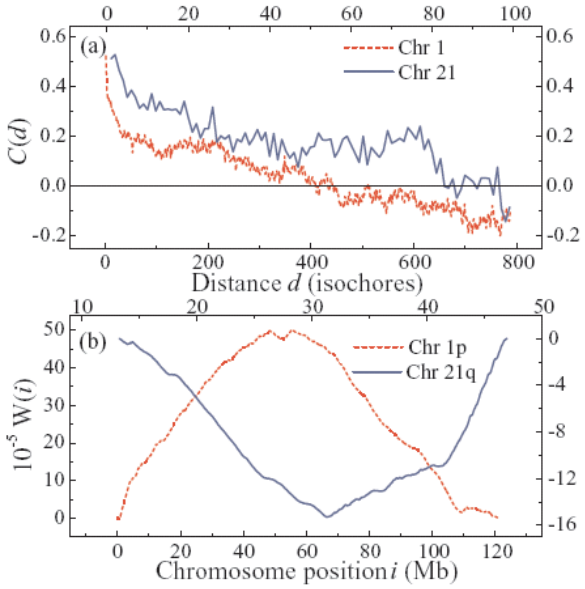

FIG. 1: a) Autocorrelation function $C(d)$ *vs.* the distance d (in isochores) obtained from the series of the G+C content of the isochores [4] in human chromosomes 1 (bottom+left axes) and 21 (top+right axes). The slowly decreasing behavior indicates a strong correlation of isochore G+C content up to large distances, showing that isochores are organized in large clusters of similar composition. The typical cluster sizes are 400 and 80-85 isochores in chromosomes 1 and 21, respectively. As the isochore average size is 1.0×10^5 bp (Chr 1), and 1.5×10^5 bp (Chr 21) we find $\langle s \rangle$ values of about 40 and 12 Mb, respectively. b) DNA walks obtained for the p-arm of human chromosome 1 (bottom+left axes) and the q-arm of human chromosome 21 (top+right axes). In both cases large regions of almost constant slope (i.e. with well-defined G+C content) can be seen. The segmentation algorithm we introduce below splits these two sequences into three regions with definite G+C content, with average sizes of 40 Mb (Chr 1p) and 11.5 Mb (Chr 21q), in close agreement with their $\langle s \rangle$ values.

As an example of the application of our segmentation algorithm we use a human sequence, the q-arm of the chromosome 21 with a length of 33.7 Mb. We obtain only

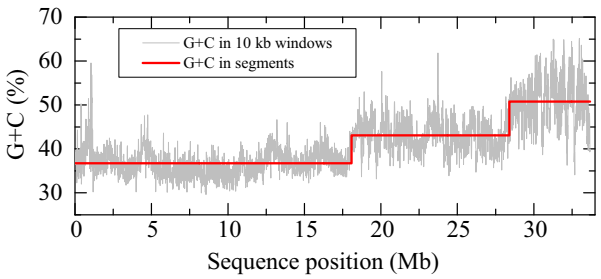

FIG. 2: Segmentation of the sequence of the q-arm of human chromosome 21. The DNA sequence has been modeled as a Gaussian noise with $\beta = 0.556$ which is the scaling exponent obtained for this sequence by means of DFA analysis ($\alpha = 0.788$). The G+C composition along the sequence is also shown by averaging the composition in 10 kb non-overlapping windows. This coarse graining of the data has been done only for representation purposes and does not affect to the segmentation procedure.

three segments which clearly correspond to the three regions of different G+C composition described above (Fig. 1b). Similar results have been obtained when segmenting all human chromosome sequences, showing the existence of previously unknown huge compositional structures in human DNA [3].

Finally, we show evidence of the biological relevance of these superstructures by analyzing the Gene Ontology terms [9]: we show that gene pairs embedded in each superstructure have a higher probability to share a large number of GO terms.

Acknowledgments

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

[1] G. Bernardi, Annu. Rev. Genet. **29** 445 (1995).
[2] G. Bernardi et al., Science **228**, 953 (1985).
[3] P. Carpena et al., Phys. Rev. E **83**, 031908 (2011).
[4] T. Schmidt & D. Frishman, Genome Biol. **9**, R104 (2008).
[5] W. Li et al., Computers & Chemistry **26**, 491 (2002).
[6] J.L. Oliver et al., Nucleic Acids Res. **32**, W287 (2004).
[7] P. Bernaola-Galván et al., European Physical Journal B (in press).
[8] P. Carpena, et al., Phys. Rev. E **75**, 032903 (2007).
[9] The gene ontology consortium, Nat. Genet. **25**, 25 (2000). See also http://www.geneontology.org/.