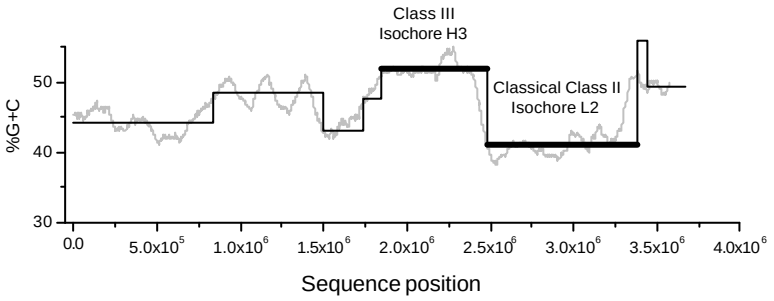
Mapping isochores by entropic segmentation of long genome sequences

Pedro Bernaola-Galván[1] Pedro Carpena[1] Ramón Román-Roldán[2] José L. Oliver[3]

Keywords: isochores, DNA sequences, entropic segmentation, Jensen-Shannon divergence, gene density, gene finding, comparative genomics

Analytical DNA ultracentrifugation revealed that many eukaryotic genomes are mosaics of *isochores*: long DNA segments (>300 kb) relatively homogeneous in GC content (above a size of 3 kb) when compared to the large heterogeneity in the entire genome [1,2]. Important genome features are dependent on such isochore structure –i.e. genes are found predominantly in the GC-richest isochore classes. However, not reliable method is available to reveal isochores at the shorter sequence scale. Sometimes, they are identified 'by eye' on a moving-window graphical plot of GC-content [3,4]. A recently published 'decomposition algorithm' [5] is strongly dependent on the reference sequence taken as 'homogeneous'. Here we show that a step-by-step version of the proven entropic segmentation method [6,7,8] is able to identify the boundaries between long genome regions displaying the typical features of isochores. The main improvement is that the sequence is now segmented in an ordered way: that maximizing the overall compositional complexity of the sequence, which is equivalent to maximize the statistical significance at each scale [8].

We first use the human major histocompatibility complex (MHC) as a 'benchmark' test for the step-by-step segmentation algorithm. This genomic region harbors the only two isochores experimentally characterized at the sequence level so far [9,10]. Previous predictions were based on a moving-window plot of G+C content, and therefore were only approximate. Our step-by-step segmentation algorithm allows now for a more precise location of MHC isochore boundaries:



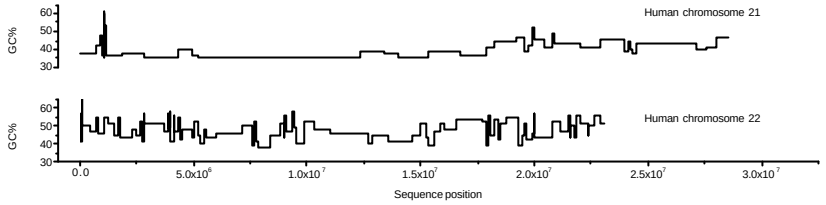
Isochore chromosome map (straight line) of the human MHC region. Bold lines indicate the two experimentally determined isochores. Errors in boundary determination were all below 500 bp. The rough line corresponds to the %G+C in a moving window across the MHC sequence.

[1] Dpto. de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga, E-29071, Spain. E-mail: rick@uma.es, pcarpena@ctima.uma.es

[2] Dpto. de Física Aplicada, Fac. de Ciencias, Universidad de Granada, E-18071, Spain. E-mail: rroman@ugr.es

[3] Dpto. de Genética, Inst. de Biotecnología, Fac. de Ciencias, Universidad de Granada, E-18071, Spain. E-mail: oliver@ugr.es

We then look for what the same algorithm, with no changes to the parameters/stringencies, yields for other, uncharted, regions. As an example, we show the isochore chromosome maps for the larger contigs of chromosomes 21 (NT_002836) and 22 (NT_001454):



Isochore maps of the larger contigs of human chromosomes 21 and 22 obtained by the step-by-step segmentation algorithm.

The long homogeneous regions we found in a corpus of human genome contigs show the features (G+C range, relative proportion of isochore classes, size distribution, and relation with gene density) of the isochores identified through DNA centrifugation. The isochore chromosome maps of completely sequenced genomes we obtained can be most helpful in gene-finding projects or comparative genomics.

References

[1] Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. & Rodier, F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* 228: 953–958.

[2] Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* 241: 3-17.

[3] Dunham, I. et al. 1999. The DNA sequence of human chromosome 22. *Nature* 402: 489-495 .

[4] Hattori, M. et al. 2000. The DNA sequence of human chromosome 21. *Nature* 405: 311-319.

[5] Nekrutenko, A. & W.H. Li. 2000. Assesment of compositional heterogeneity within and between eukaryotic genomes. *Genome Research* 10: 1986-1995.

[6] Bernaola-Galván, P., Román-Roldán, R. and Oliver, J.L. 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys. Rev. E* 53: 5181-5189.

[7] Oliver, J.L., Román-Roldán, R., Pérez, J. and Bernaola-Galván, P. 1999. SEGMENT: Identifying compositional domains in DNA sequences. *Bioinformatics* 15: 974-979.

[8] Román-Roldán, R., Bernaola-Galván, P. and Oliver, J.L. 1998. Sequence Compositional Complexity of DNA through an Entropic Segmentation Method. *Phys. Rev. Lett.* 80: 1344-1347.

[9] Stephens, R., Horton, R., Humphray, S., Rowen, L., Trowsdale, J. and Beck, S. 1999. Gene Organisation, Sequence Variation and Isochore Structure at the Centromeric Boundary of the Human MHC. *J. Mol. Biol.* 291: 789-799.

[10] The MHC sequencing consortium. 1999. Complete sequence and gene map of a human major histocompatibility complex. *Nature* 401: 921-923.