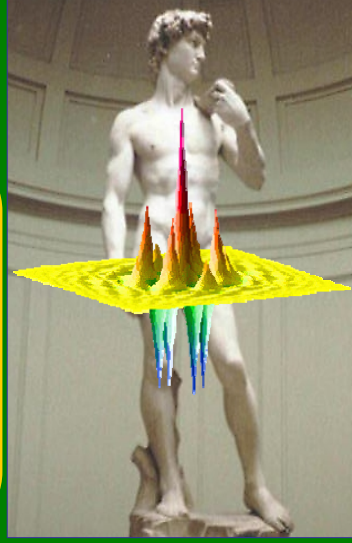
An improved Haar wavelet analysis of the human genome

Eran Elhaik[1], Dan Graur[1], and Kresimir Josic[2]

[1] Department of Biology & Biochemistry, University of Houston, Houston, TX 77204-5001
[2] Department of Mathematics, University of Houston, Houston, TX 77204-3008

UNIVERSITY OF HOUSTON



Introduction

The isochore theory, proposed three decades ago (Macaya et al. 1976. *J. Mol. Biol.* 108:237-254), depicts the mammalian genome as a mosaic of long, fairly homogeneous genomic regions called isochores, each with a characteristic GC content. Low-GC and high-GC segments alternate throughout the genome. Detecting isochores requires the use of segmentation algorithms that detect homogeneous sequences within the genome that are distinguishable in their GC content from adjacent sequences. Many such segmentation algorithms have been proposed in the literature, but none of those methods detected "isochores" satisfactorily. These failures caused scientists to doubt the very existence of isochores (Lander et al. 2001. *Nature* 409:860–921; Haring and Kypr, 2001 *Mol. Biol. Rep.* 28:9-17; Cohen et al. 2005. *Mol. Biol. Evol.* 22:1260-1272).

Here, we propose a new DNA segmentation algorithm based on the one-dimension Haar decomposition function (Goupillaud et al. 1984. *Geoexploration*. 23:85-102). In this method, we define the mean GC content of any two adjacent nucleotides as a "signal" that may assume three values: 0, 0.5, and 1. Our Haar wavelet algorithm analyzes the signal and detects segments with similar GC fluctuations for any arbitrary determined threshold value. We adopt a set of four attributes that are claimed to characterize homogeneous domains and statistically test their veracity against the complete human genome. We compare the segmentation results of our algorithm with the segmentation results of the Janson-Shannon divergence (D_{JS}) algorithm, a binary recursive segmentation procedure (Bernaola-Galvan et al. 1996. *Phys. Rev.* 53E:5181-5189). Our findings indicate that long homogeneous domains exist in the human genome.

Conclusions

We propose an improved Haar wavelet algorithm to study compositional homogeneity in eukaryote genomes. We performed a full genome comparison between the Haar wavelet and the D_{JS} methods using four criteria:

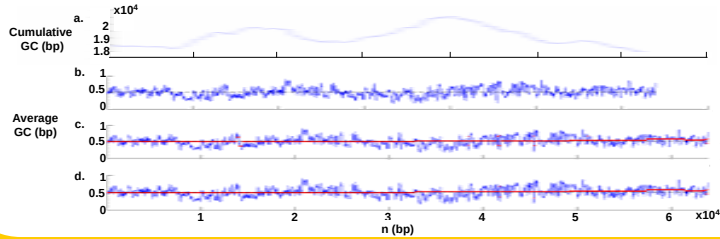
A. Segment homogeneity. Haar wavelet finds relatively fewer homogeneous segments than D_{JS}, and a large number of nonhomogeneous segments (noise). We can, therefore, say that Haar wavelet is relatively noisier than D_{JS}.
B. Minimum length of homogeneous segments. The homogeneous and non homogeneous segments obtained by both methods follow power-law decay distribution. The Haar wavelet method yielded a higher number of long homogeneous segments (> 300 Kb) than the D_{JS} method. The proportion of those homogeneous segments out of the total segments is also higher in the Haar wavelet method. The D_{JS} segmentation results suggest that the D_{JS} is not suitable to detect long homogeneous segments.
C. Homogeneous segment genome coverage. The Haar wavelet homogeneous segments have a higher chromosomal coverage than the D_{JS} homogeneous segments, in most of the Haar thresholds.
D. Segment GC content differentiation level. We showed that the GC content greatly differs between every two adjacent segments in Haar wavelet. The differential level ranges closely to the threshold values (14%–24%). The difference in D_{JS} was 10%, the smallest among all thresholds.

Haar Wavelet Algorithm

Algorithm description

Our algorithm has four steps:

(1) Calculate the signal values. i.e., the mean GC content of any two adjacent nucleotides.
(2) Decompose the signal with a one-dimension Haar function.
(3) Reconstruct the signal using a threshold. Signal values lower than the threshold are assigned a value of 0, and signal values higher than (1 – threshold) are assigned a value of 1.
(4) Segment the reconstructed signal based on the following rule: if the difference between two neighboring subsignals is smaller than the threshold, then the segment is elongated; otherwise it is terminated, and a new segment is started.



An illustration of the Haar wavelet segmentation algorithm used on a hypothetical DNA sequence:

(a)the cumulative GC profile of the sequence.

(b) The average GC content of two adjacent nucleotides.

The results after applying the algorithm with (c) 0.13 threshold and (d) 0.15 threshold. The original signal is plotted in the background. The bold lines in red are the segments.

Methods and Attributes

Comparative Analysis

Here we show a comparison of our segmentation algorithm and the D_{JS} segmentation method using four criteria:

A. Segment homogeneity.

B. Minimum length of homogeneous segments.

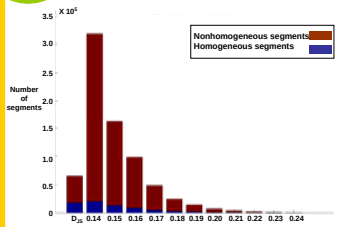
C. Homogeneous segment genome coverage.

D. GC content difference between adjacent segments.

To compare the compositional homogeneity of the segments resulting from the two segmentation methods, we used the nonparametric Ansari-Bradley test (Ansari and Bradley, 1960. *Ann. Math. Stat.* 31:1174-1189). Each chromosome was divided into 2,048-bp-long non overlapping windows, and the GC-content values for each window were calculated for the entire chromosome and for the segment in question. A one tailed test has been applied with $H_0: \sigma^2_{segment} \geq \sigma^2_{chromosome}$, $H_1: \sigma^2_{segment} < \sigma^2_{chromosome}$.

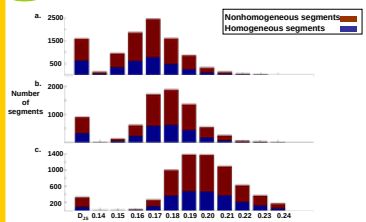
Results

A



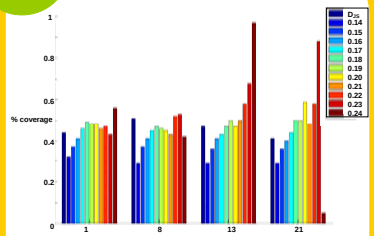
A. Homogeneous and nonhomogeneous segment length distribution. The results shown are for the entire genome. The leftmost bar represents the D_{JS} results. Subsequent bars represent the Haar wavelet with 11 different thresholds. The homogeneous segments are shown in blue.

B



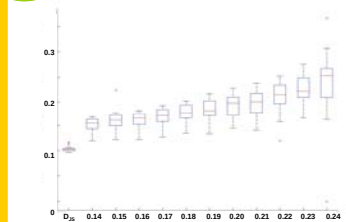
B. Long homogeneous and nonhomogeneous segment-length distribution.
The leftmost bar represents the D_{JS} results. Subsequent bars represent the Haar wavelet with 11 different thresholds. The length distribution is grouped into 3 size categories: (a) 300 Kb - 500 Kb, (b) 500 Kb – 1 Mb, and (c) 1 Mb and above. The homogeneous segments are shown in blue.

C



C. Chromosomal coverage of homogeneous segments.
The results are shown for four human chromosomes (1, 8, 13, and 21). Each chromosome is represented by 12 bars. The leftmost bar represents the D_{JS} results. Subsequent bars represent the Haar wavelet with 11 different thresholds. The Haar wavelet chromosomal coverage of homogeneous segments changes among the different threshold. Thresholds 0.18 and above show significantly higher chromosomal coverage than D_{JS} in most chromosomes. The "critical" value for the Haar threshold varies significantly among the chromosomes, i.e., it is difficult to decide a priori on a threshold.

D



D. GC content differences between adjacent segments.
The box has lines at the lower quartile, median, and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. Outliers are data with values beyond the ends of the whiskers. The median values for the Haar wavelet segmentation found to be around their threshold value and unambiguously show that D_{JS} adjacent segments are less differ in their GC content than any of the Haar wavelet thresholds.