Introduction


Here, we propose a new DNA segmentation algorithm based on the one-dimensional Haar wavelet transform (Goupillaud et al. 1984. Geoexploration 22:1228-1237). This method allows us to 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 examines the signal and detects segments with similar GC fluctuations for any arbitrary determined threshold value. We adapt a set of four attributes that are claimed to characterize homogeneous domains and statistically test their reality against the complete human genome. We compare the segmentation results of our algorithm with the segmentation results of the Janson-Shannon divergence (DJS) algorithm, a binary recursive segmentation procedure (Berman-Galvan et al. 1996. Phys. Rev. E 53:1-1-15). Our findings indicate that long homogeneous domains exist in the human genome.

Haar Wavelet Algorithm

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 subsegments is smaller than the threshold, then the segment is elongated; otherwise it is truncated, and a new segment is started.

Algorithm description

A. Segment homogeneity. Haar wavelet finds relatively fewer homogeneous segments than DJS, and a large number of nonhomogeneous segments (noise). We can, therefore, say that Haarwavelet is relatively noisier than DJS.
B. Minimum length of homogeneous segments. The homogeneous and nonhomogeneous 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 DJS method. The proportion of those homogeneous segments out of the total segments is also higher in the Haar wavelet method. The DJS segmentation results suggest that the DJS 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 DJS homogeneous segments, in most of the human chromosomes.
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 DJS adjacent segments is 10%, the smallest among all thresholds.

Comparative Analysis

To compare the compositional homogeneity of the segments resulting from the two segmentation methods, we used the non-parametric Ansari-Bradley test (Ansari and Bradley, 1960. Ann. Math. Stat. 31:1174-1185). 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 threshold values for homogenous segments H1: 1.5, and H2: 2.1. Each chromosome is represented by 12 bars. The leftmost bar represents the DJS results. Subsequent bars represent the Haar wavelet with 11 different thresholds. The resulting box plot 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 homogenous segments found to be around their threshold value and unambiguously show that DJS adjacent segments are less different in their GC content than any of the Haar wavelet thresholds.