High-level organization of isochores into gigantic superstructures in the human genome

P. Carpena,[1,2,*] J. L. Oliver,[3] M. Hackenberg,[3] A. V. Coronado,[1,2] G. Barturen,[3] and P. Bernaola-Galván[1]

[1]*Departamento de Física Aplicada II, Universidad de Málaga, ES-29071, Málaga, Spain*

[2]*Division of Sleep Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA*

[3]*Departamento de Genética, Instituto de Biotecnología, Universidad de Granada, ES-18071 Granada, Spain*

(Received 28 May 2010; revised manuscript received 10 January 2011; published 15 March 2011)

Human DNA shows a complex structure with compositional features at many scales; the isochores—long DNA segments ($\sim 10^5$ bp) of relatively homogeneous guanine-cytosine (G + C) content—are the largest well-documented and well-analyzed compositional structures. However, we report here on the existence of a high-level compositional organization of isochores in the human genome. By using a segmentation algorithm incorporating the long-range correlations existing in human DNA, we find that every chromosome is composed of a few huge segments ($\sim 10^7$ bp) of relatively homogeneous G + C content, which become the largest compositional organization of the genome. Finally, we show evidence of the biological relevance of these superstructures, pointing to a large-scale functional organization of the human genome.

DOI: 10.1103/PhysRevE.83.031908 PACS number(s): 87.14.gk, 87.10.Vg, 87.18.Wd

I. INTRODUCTION

Human DNA is known to have a very complex compositional structure, since there are genomic elements with clear compositional features of many different scales such as CpG islands, genes, repeat elements (SINEs, LINEs), segmental duplications, etc. The largest compositional organization well documented and systematically analyzed are the isochores, first identified by Bernardi and coworkers by analytical ultracentrifugation of bulk DNA [1,2], which are present in the genomes of warm-blooded vertebrates. At the DNA sequence level, isochores are large segments with a typical size of around 10^5 base pairs (bp) and relatively homogeneous guanine-cytosine (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 report here that human isochores seem to be organized at larger scales into compositional segments of about two orders of magnitude larger than isochores. We proceed in three steps. First, we use two independent methods to show the organization of isochores into larger compositional structures. On the one hand, we use compositional autocorrelation analysis to show 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. On the other hand, we use DNA walks to show the existence of huge DNA segments (∼15–20 Mb) with definite G + C composition and typical sizes in agreement with the sizes of the isochore clusters obtained via autocorrelation analysis. We term these segments as superstructures from now on. Second, we introduce a segmentation algorithm which incorporates the long-range correlations existing in human DNA and is capable of detecting systematically these compositional superstructures on the basis of rigorous statistical criteria. And third, by analyzing the gene ontology (GO) terms [3], we show that gene pairs embedded in each superstructure have a higher probability of sharing a large number of GO terms, thus pointing to the biological relevance of these structures.

II. EVIDENCE OF ISOCHORE ORGANIZATION

Several computationally detected isochore data sets are available in the bibliography [4,5]. However, an isochore data set obtained from a consensus between them has been proposed [6] recently, and this latter data set is the one we will use here.

On average, isochores are long DNA segments with an average size of about 10^5 bp [6] and with a G + C composition rather homogeneous internally, and clearly different from adjacent isochores. We can define the numerical series $\{x_i, i = 1,2,\ldots,n\}$, where x_i is the G + C fraction of the ith isochore of the sequence containing n isochores. If the isochores are the highest level of compositional organization in the genome and not organized into larger structures, the values of the time series $\{x_i\}$ should be uncorrelated (i.e., with independent isochore G + C values). We can show the correlation properties of the series $\{x_i\}$ by calculating the autocorrelation function $C(d)$:

$$C(d) = \frac{1}{\sigma^2}\left[\frac{\sum_{i=1}^{n-d} x_i x_{i+d}}{n-d} - \frac{\sum_{i=1}^{n-d} x_i}{n-d}\frac{\sum_{i=1}^{n-d} x_{i+d}}{n-d}\right], \quad (1)$$

where d is the distance (measured in isochores) and σ^2 is the variance of the set $\{x_i\}$. In contrast to the expectation, we obtain in all the human chromosomes that $C(d)$ is a slowly decreasing function significantly greater than 0 up to very large distances measured in isochores, and which can become negative for even larger distances [Fig. 1(a)]. This slowly decreasing behavior with even sign alternating 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 size of these clusters $\langle s \rangle$ can be estimated as the d value for which $C(d) \simeq 0$ [7]. We obtain that $\langle s \rangle$ depends on the chromosome [Fig. 1(a)] and shows a chromosome average size of about 130 isochores (∼ 15 Mb), indicating the typical scale for the isochore organization.

*pcarpena@ctima.uma.es

A second and independent method able to show the existence of gigantic segments with definite G + C content is the use of "DNA walks" [8], an intuitive way to visualize the compositional structure of DNA sequences. To generate the walk, the DNA sequence $S(i)$, with $i = 1,2,\ldots\ell$, (ℓ being the sequence length), is first mapped into a numerical series $N(i)$ using the SW rule: $N(i) = 1$ (-1) when $S(i) =$ C or G (A or T). We use this particular rule because we are interested in the G + C organization of the sequence. The walk of the DNA sequence at position j, $W(j)$, is then

$$W(j) = \sum_{i=1}^{j}(N(i) - \langle N\rangle), \quad \langle N\rangle = \frac{\sum_{i=1}^{\ell} N(i)}{\ell}. \qquad (2)$$

Regions with positive (negative) slope in the walk indicate that the G + C content in that part of the sequence is above (below) the global mean G + C value, and a *constant slope* indicates a region of approximately constant G + C content. We obtain that the DNA walks for human chromosome arms [9] are composed of few regions [see Fig 1(b) for two examples] of approximately constant slope, indicating that the chromosomes are actually formed by a small number of gigantic compositional superstructures of definite G + C content, which are on average two orders of magnitude larger than the isochores (see Table I). This scale is also in agreement with the typical isochore clusters predicted by autocorrelation analysis (caption of Fig. 1), indicating that both methods are identifying the same structural organization. In the next section, we present a segmentation algorithm aimed at detecting systematically these superstructures.

III. THE SEGMENTATION ALGORITHM. CHANGE OF NULL HYPOTHESIS

Our second aim is to design an algorithm that is able to find superstructures automatically based on rigorous statistical criteria. This could be achieved through entropic segmentation [10], a convenient algorithm widely used in the literature [11–15] as it incorporates the Jensen-Shannon divergence (D_{JS}) measure to split the DNA sequence into pieces with different composition or "segments":

$$D_{\mathrm{JS}}(S_1,S_2) = H_t - \frac{n}{\ell}H_1 - \frac{\ell - n}{\ell}H_2. \qquad (3)$$

In this formula, S_1 and S_2 are DNA sequences of lengths n and $\ell - n$ to be compared, H_1 and H_2 are the corresponding Shannon entropies of both sequences using the SW mapping rule defined above, and H_t is the Shannon entropy of the concatenation of sequences S_1 and S_2. In brief, the algorithm works as follows: (i) Move a pointer along the DNA sequence to be segmented. At any position i of the pointer ($i = 1,2,\ldots\ell$) compare the subsequences to the left (S_1) and to the right (S_2) of the pointer using D_{JS}. (ii) Find the point in the sequence ($i_{\max}$) where the maximum compositional difference between the left and the right subsequences ($D_{\mathrm{JS,\,max}}$) is measured. (iii) Decide whether the value $D_{\mathrm{JS,\,max}}$ is statistically significant (i.e., with a p value lower than a certain threshold p_0 fixed at the beginning of the segmentation). If so, then split the sequence into two segments at position $i_{\max}$ and otherwise leave the sequence undivided. (iv) If the sequence is split, then repeat iteratively the algorithm in the resulting subsegments. The segmentation ends when none of the resulting segments can be split, and at this point we say that the sequence is segmented with a p_0 significance level (usually, p_0 is taken as 0.05 or 0.01).

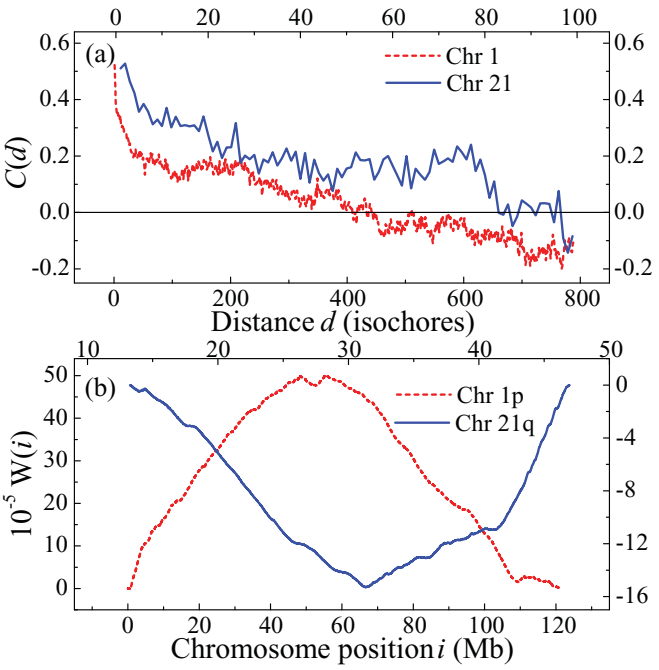

FIG. 1. (Color online) (a) Autocorrelation function $C(d)$ vs distance d (in isochores) obtained from the series of the G + C content of the isochores [6] in human chromosomes 1 (bottom+left axes) and 21 (top+right axes). The slowly decreasing behavior indicates a strong correlation of isochore G + C 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.

TABLE I. Average length and average absolute difference in G + C content between adjacent segments obtained for SSs, isochores from two methods [4,6], and chromosome bands [26].

p value	Average length (bp)	$\lvert\Delta(\mathrm{G}+\mathrm{C})\rvert$ (%)
Superstructures:		
0.05	15.6×10^6	7.10
0.01	21.6×10^6	7.10
Isochores:		
0.05 (IsoFinder)	1.15×10^5	4.96
0.05 (Consensus)	1.25×10^5	4.24
Chromosome bands:		
Expt.	3.55×10^6	2.55

The critical point in the segmentation algorithm is step (iii), since this is the point where it is decided whether the segmentation continues or not, and therefore defines the stopping criterium. To achieve this, we use the *hypothesis testing* (HT) strategy, where a DNA sequence model is chosen and the null distribution is derived from it, i.e., the $D_{\mathrm{JS,\,max}}$ values expected by chance alone. Each $D_{\mathrm{JS,\,max}}$ value found in the segmentation of the real DNA sequence is then tested against the null hypothesis that the observed $D_{\mathrm{JS,\,max}}$ can be reached also in our null distribution. We reject the null hypothesis if the p value is lower than the predefined significance p_0, and in this case the cut is accepted [10,12,13].

Usually, the null hypothesis chosen when segmenting a real DNA sequence of length ℓ is a random independent identically distributed (i.i.d.) uncorrelated sequence of the same length and the same G + C composition [10,12]. Therefore, the p value of a given $D_{\mathrm{JS,\,max}} = x$ found in the real DNA sequence is calculated as the probability of obtaining that or a greater value when trying to segment this random i.i.d. uncorrelated sequence, i.e.,

$$P_{\ell,\mathrm{rand}}(x) = \mathrm{Prob}\{D_{\mathrm{JS,\,max}} \geqslant x\}. \tag{4}$$

Unfortunately, even for this simple null hypothesis, the distribution $P_{\ell,\mathrm{rand}}(x)$ does not have a closed analytical form and should be numerically calculated by means of Monte Carlo simulations [14].

Although the HT strategy described above is the most widely used in segmentation problems [10,12,13,16], it is not unique. An alternative way to address this problem is the *model selection* (MS) strategy, where segmentation is viewed as the selection between two models describing the target sequence: with the cut (model 1) and without it (model 2) [17]. In this framework one has to compute the log-likelihood (L) for the two models, L_1 and L_2, to obtain its increment ($\Delta L = L_2 - L_1$), and find the position of the cut in the sequence where ΔL is maximum (ΔL_{max}). Since a new cut always increases L (ΔL is always positive [14]), one needs to set a threshold (ΔL_0) above which the cut is accepted, i.e., when $\Delta L_{\mathrm{max}} > \Delta L_0$.

Despite the fact that both strategies (HT and MS) look different, they are quite similar. Indeed, it can be shown [14] that $\Delta L = \ell \times D_{\mathrm{JS}}$ with ℓ being the sequence length, and therefore $\Delta L_{\mathrm{max}} = \ell \times D_{\mathrm{JS,\,max}}$. Thus, in both cases the problem is reduced to check if a particular $D_{\mathrm{JS,\,max}}$ value is large enough. While in the HT strategy a cut with a given $D_{\mathrm{JS,\,max}}$ value is accepted on a probabilistic basis (the statistical significance), in the MS one the cut is accepted whenever $\Delta L_{\mathrm{max}} = \ell \times D_{\mathrm{JS,\,max}} > \Delta L_0 \equiv \ell \times D_{\mathrm{JS},0}$. Thus, a MS segmentation with a threshold ΔL_0 is equivalent to a HT segmentation in which a cut with a certain $D_{\mathrm{JS,\,max}}$ value is accepted when $D_{\mathrm{JS,\,max}} > D_{\mathrm{JS},0} = \Delta L_0/\ell$.

A. The over-segmentation problem

However, when applying the segmentation algorithm to human DNA sequences, one finds thousands of small segments [14] (a few hundreds bp on average), and therefore well below both superstructure and isochore scales. For example, for the q-arm of human chromosome 21 (the smallest chromosome) used in Fig. 1(b), we find that the sequence is divided into 44854 segments with an average length of 751 bp. The key reason for such "over segmentation" relies on the fact that the segmentation algorithm, under the above definition of the statistical significance in Eq. (4), splits the DNA sequence wherever it does not behave like a random uncorrelated sequence. Nevertheless, real DNA sequences are far from behaving like random sequences at all: in contrast, they present long-range correlations of complex nature [18] reaching scales up to tens of megabase-pairs. Thus, a more convenient DNA model to be used as a statistical reference (null hypothesis) when segmenting real DNA sequences should possess two main properties: (i) It should be homogeneous in order to detect heterogeneities in real DNA sequences. (ii) It should incorporate long-range correlations, which are present in real DNA. The standard null hypothesis (4) of a random i.i.d. sequence incorporates property (i), but lacks property (ii).

B. Changing the null hypothesis

A model for long-range correlated (LRC) and homogeneous DNA sequences can be derived from fractional Gaussian noises (fGn) which are well-known stochastic models of stationary LRC sequences [19]. Homogeneous (stationary) LRC DNA sequences can be generated as follows: (i) Use the Fourier filtering method [20], which has proven to be useful in many contexts [21], to create a fGn with a power spectrum $S(f)$ of the type $S(f) = 1/f^{2\alpha-1}$ with $\alpha < 1$. Correlations are quantified by the input exponent α. For a pure random i.i.d. sequence, $\alpha = 0.5$, and the larger α, the stronger the correlations present in the sequence. (ii) Map the fGn into a DNA binary sequence $N(i)$ according to the SW rule: when fGn(i) is larger than a certain threshold t_0, $N(i) = 1$ (i.e., G or C) and $N(i) = -1$ otherwise (i.e., A or T). The G + C content of $N(i)$ is then controlled by t_0. In our case, we fix a t_0 value to create LRC DNA sequences with a G + C fraction identical to the one in the human genome (0.41). As the fGn follows a Gaussian distribution $N(0,1)$, t_0 is obtained from

$$\int_{t_0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.41, \tag{5}$$

giving $t_0 \simeq 0.227$.

A comparison of the behavior of D_{JS} in a random i.i.d. and a LRC DNA sequence created with this method is shown in Fig. 2. Despite that both sequences are homogeneous and have a constant mean G + C content, the D_{JS} values for the LRC sequence are in general much larger than for the random one, and this is true in particular for both $D_{\mathrm{JS,\,max}}$ values. If the random sequence is used as the null hypothesis (4) for the segmentation algorithm, the LRC sequence could be split practically at any point in spite of its homogeneity, since the major part of its D_{JS} profile is above the $D_{\mathrm{JS,\,max}}$ value obtained in the random sequence [Fig. 2(c)], thus leading to the over segmentation observed in real LRC DNA sequences.

For this reason, we propose here to change the null hypothesis (4). Instead of modeling a homogeneous DNA sequence as an uncorrelated random i.i.d. sequence, we model it using a homogeneous LRC DNA sequence derived from a fGn (as described above) with the same long-range correlations as the target DNA sequence. Under this alternative null hypothesis we propose, the compositional differences found

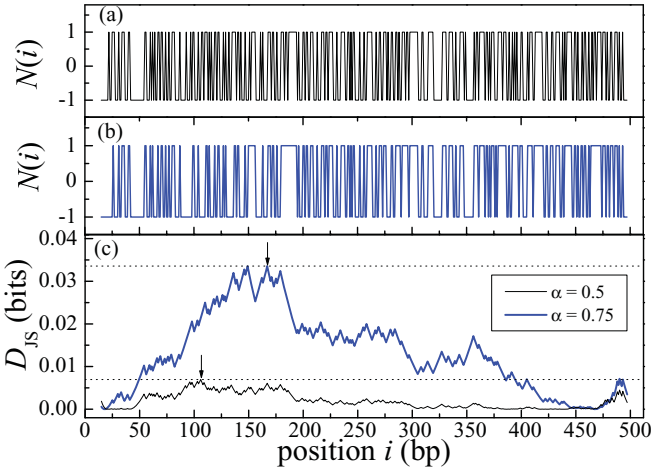

FIG. 2. (Color online) (a) A random ($\alpha = 0.5$) numeric DNA sequence $N(i)$ of 512 bp and a G + C content of 0.5. (b) A numeric LRC ($\alpha = 0.75$) DNA sequence $N(i)$ of 512 bp with a G + C content of 0.5. (c) The profile of the $D_{\rm JS}$ values at any position in the sequences of parts (a) and (b). In both cases, the position where $D_{\rm JS}$ reaches its maximum is marked with an arrow, and the $D_{\rm JS,\,max}$ value itself with a dashed line.

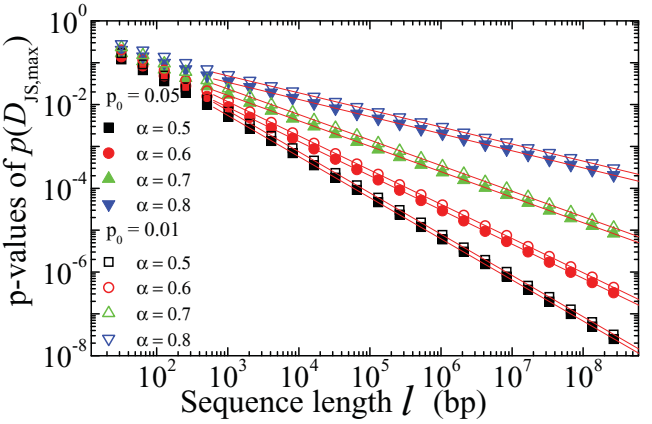

FIG. 3. (Color online) Behavior of percentiles 95 and 99 (p values 0.05 and 0.01) of the distribution $p_{\ell,\alpha}(D_{\rm JS,\,max})$ vs the sequence length ℓ for synthetic LRC DNA sequences of different α values. Solid lines show power-law fittings.

in a real DNA sequence will be compared to the expectation in a homogeneous *and* LRC DNA sequence with the same correlations and the same length.

This null hypothesis introduces two changes into the segmentation algorithm: (i) The degree of long-range correlations present in the target DNA sequence must be quantified as the first step in the algorithm. To achieve this, we choose detrended fluctuation analysis (DFA) [22], a widely used method to measure long-range correlations with convenient properties [23]. DFA calculates the fluctuations $F(l)$ of the sequence around the local trend at different scales l. Long-range correlations are identified when $F(l) \sim l^{\alpha}$, and then correlations are quantified by a single parameter, the scaling exponent α which by construction is the same exponent we used to create LRC DNA sequences. For a pure random sequence, $\alpha = 0.5$, and the larger α, the stronger the long-range correlations present in the sequence. In practice, we calculate $F(l)$ for any human chromosome arm of length ℓ by varying l from small values (typically, $l_{\rm min} = 8$ up to an upper scale of $l_{\rm max} = \ell/10$) [23]. The exponent α is determined by a linear fitting of a log-log plot of $F(l)$ vs l. (ii) We need to replace the statistical significance (4) with the one expected in homogeneous LRC DNA sequences. To achieve this, we use the Monte Carlo method for any pair (α,ℓ) to obtain numerically the probability distribution $P_{\ell,\alpha}(x)$, i.e., the probability of obtaining $D_{\rm JS,\,max} \geqslant x$ when trying to segment a homogeneous LRC DNA sequence of length ℓ and a degree of correlations α:

$$P_{\ell,\alpha}(x) = {\rm Prob}\{D_{\rm JS,\,max} \geqslant x\}. \qquad (6)$$

This statistical significance replaces Eq. (4). Numerically, it is more convenient to obtain the ℓ and α dependence of certain percentiles of $P_{\ell,\alpha}(x)$, and we have considered percentiles 95 and 99 (i.e., p values 0.05 and 0.01) as shown in Fig. 3, which are the customary values in hypothesis testing strategies. In brief, our calculation is as follows: for a given sequence length ℓ, and a given value of correlations (α) we generate $2^{37}/\ell$ LRC sequences using the method described above, from which the percentiles 95 and 99 of the distribution (i.e., p values 0.05 and 0.01) are calculated. Next, we keep α constant and repeat the calculation for different values of ℓ. We have generated sequence lengths in powers of 2 in the range $[2^6,2^{28}]$ to study numerically the ℓ dependence of both percentiles. We repeat the same procedure for different values of the correlations by varying α in the range [0.5,0.9], which is the appropriate range for the correlations observed in human DNA sequences [24]. Both percentiles behave similarly: for a given value of the correlations (α), they behave as power laws of ℓ (the fittings in Fig. 3) with negative exponents increasing with α (decreasing with α in absolute value). For any fixed ℓ, the percentiles increase with α, indicating that the larger the correlations in a homogeneous sequence of fixed length, the larger the compositional differences expected by chance in that sequence. Once the values of both percentiles are available in our grid in the (α,ℓ) space, when segmenting a real DNA sequence of length ℓ_1 with correlations α_1 in which a particular $D_{\rm JS,\,max}$ value is found, we check its statistical significance by interpolation in the grid: we estimate the percentile 99 (95) corresponding to the point (α_1,ℓ_1) and verify whether the $D_{\rm JS,\,max}$ value is larger than percentile 99 (95) or not. If the answer is positive, then the cut is accepted with a p value of 0.01 (0.05).

IV. RESULTS

With our algorithm, each human chromosome is divided into a few huge segments or superstructures (SSs) from now on. Detailed results for all the chromosomes can be seen in our web page [25], including the DNA sequences we have used. A basic statistics of SSs and a comparison to isochores and chromosome bands are shown in Table I. Chromosomal bands are the only genomic entity of a similar scale [26] to SSs. However, our results in Table I show that they are smaller than SSs and that they do not define the compositional structure so clearly as SSs. SSs are not only two orders of magnitude larger than isochores on average,

but also the G + C differences between neighboring SSs are larger than for neighboring isochores (see Table I). This fact is somewhat counterintuitive, since one expects that compositional fluctuations increase when observed at smaller scales, and it points to the statistical relevance of SSs. Besides, SS borders are robust when examined at different scales: when the G + C content of a DNA sequence is plotted in a great range of increasing scales of observation, the SS borders are detectable in the whole range while other compositional structures of smaller size (as the isochores) disappear for large enough scales, where only SSs are observable. As an example, in Fig. 4(a) we plot the results of the G + C composition of chromosome 21q obtained with a Gaussian wavelet with a characteristic scale varying in the range $(10^4, 1.25 \times 10^6)$ (bp). Also, SSs are in agreement with the intuitive segmentation 'by eye' suggested by the patchy structure of DNA walks [Fig. 1(b)]. As an example we show in Fig. 4(b) the walk of chromosome 21q and the SSs obtained with our algorithm. Note that not only do the SS borders match perfectly the major changes of the slope in the walk, but also the G + C content of each SS corresponds to a well-defined slope in the walk, coinciding with the large-scale compositional regions shown with the wavelet in Fig. 4(a). In this context, isochores are given by the fluctuations of the walk slope around the dominant mean slope giving the SS G + content [Fig. 4(c)].

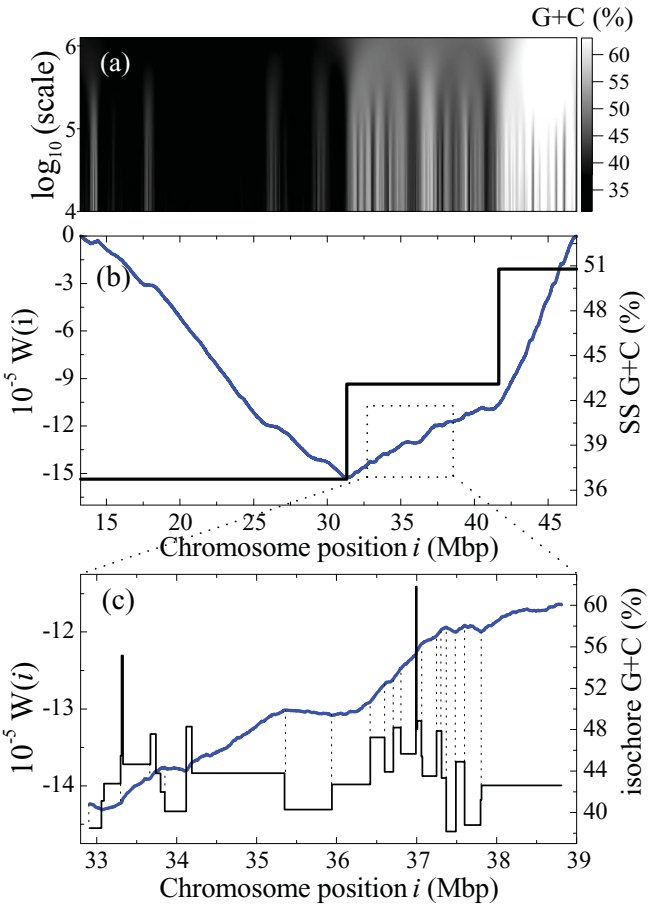

FIG. 4. (Color online) (a) A Gaussian wavelet plot of the G + C content of Chr 21q. The wavelet scale varies in the range $(10^4, 1.25 \times 10^6)$ (bp) (left axis). The chromosome coordinates are given by the bottom axis of part (b). (b) Walk of Chr 21q (left+bottom axes), and the superstructures (right+bottom axes) obtained with our algorithm with $p_0 = 0.01$. (c) A zoom of a small region of the walk of part (b) marked with a box (left+bottom axes) and the isochores [6] contained in that region with their G + C contents shown in the right axis.

SSs define the largest compositional organization of the human genome, but another important question is their possible biological relevance, specially as compared to isochores, since the latter are recognized to have functional relevance. To address this question, we perform a semantic analysis on the functional terms of gene ontology (GO) [3]. The GO initiative is aimed at standardizing the representation of gene and gene-product attributes across species and databases. It provides a controlled vocabulary of terms for describing gene-product characteristics and gene-product annotation data. This vocabulary is organized in three main categories: biological processes, cellular components, and molecular functions. Each of these categories has different hierarchical levels (10 levels for the category of cellular components and 14 levels for each of the other two categories), ordered by their functional specificity (from more general to more specific). The terms associated to a given gene product are thus not mutually exclusive.

We have studied the *semantic similarity* (m), which is defined as the number of GO terms shared by two genes [27–29]. In particular we used RefSeq genes [30] and we obtained the association between RefSeq genes and GO terms from the annotation-modules database [31]. If SSs (or isochores) were biologically relevant, they would be more likely related to certain functions. So, gene pairs co-located within the same SS or isochore should share common functionalities, and therefore larger m values, than gene pairs chosen at random. In this way, for the SSs and three different isochore sets [4–6] of the human genome, we have calculated $p(m)$, the probability that two genes co-located within the same SS or isochore share m GO terms (i.e., have an m semantic similarity). We have also calculated $p(m)$ for a randomized gene-pair data set, which will be useful for comparison. In all cases, we have considered GO terms of all levels (1–14) in the calculation, although similar results were obtained when individual GO levels 3 or 4 were considered. Our results are shown in Fig. 5, where we plot the cumulative probability $P(m) = \sum_{i=1}^{m} p(i)$ for all cases (SSs, isochores, and random gene pairs). As expected, we see that $P(m)$ increases very fast for randomly selected pairs of genes, indicating the low chance for these genes to share a large number of GO terms. In the other cases (SSs and isochores) we see qualitatively that $P(m)$ increases more slowly than in the random case, indicating that gene pairs belonging to these structures have a higher probability of sharing a large number of GO terms, in agreement with the biological relevance of these structures.

To compare quantitatively the biological relevance of the different compositional organizations, we have calculated in all cases the average semantic similarity $\langle m \rangle$ defined as

$$\langle m \rangle = \sum_{m=1}^{\infty} m p(m). \tag{7}$$

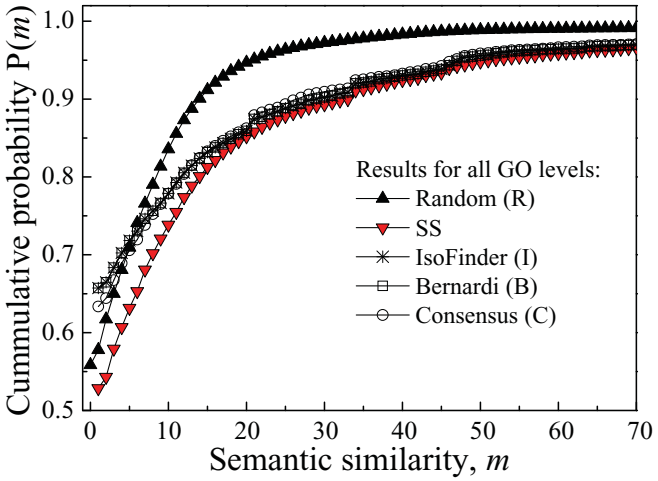

FIG. 5. (Color online) The cumulative probability $P(m)$ obtained for genes pairs co-located in the same segment of different segmentations: Superstructures (SS) and three different sets of isochores (C [6], I [4], B [5]), as well as random (R) gene-pairs used for comparison.

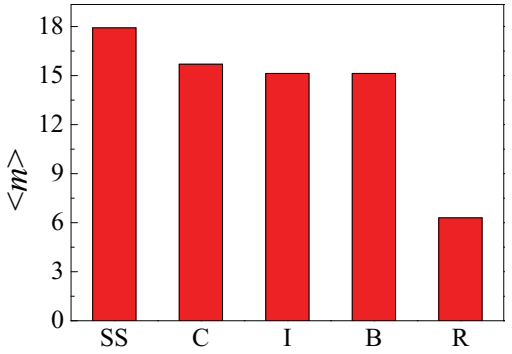

FIG. 6. (Color online) The average semantic similarity $\langle m \rangle$ for pairs of genes co-located in the same segment of different segmentations: superstructures (SS) and three different sets of isochores (C [6], I [4], B [5]), as well as random (R) gene pairs used for comparison.

The values of $\langle m \rangle$ for the five cases considered (SSs, isochores, and random gene pairs) are shown in Fig. 6. While random gene pairs share on average around six GO terms, genes co-located in the same SS or isochore share about three times more GO terms. This significant increase of the $\langle m \rangle$ values points to the biological relevance of both SS and isochores. In particular, all the isochore sets provide very similar results with a typical value of about $\langle m \rangle \simeq 15$. Strikingly, the average semantic similarity $\langle m \rangle$ is larger for gene pairs co-located in the same SS (about $\langle m \rangle \simeq 18$) than for gene pairs belonging to the same isochore. Although at first sight the difference between both $\langle m \rangle$ values (15 for isochores and 18 for SSs) does not seem to be impressive, this result is actually very surprising. Note that genes hosted by the same isochore are practically neighbors (on average, each isochore contains few genes) and therefore one might expect those genes to be functionally similar, while genes hosted by the same SS can be very distant, since on average every SS hosts around 200 genes. The fact that genes belonging to the same SS share on average a similar or even larger number of GO terms (i.e., have similar or more functionalities in common) than genes belonging to the same isochore, *which are on average 100 times closer*, points to a very large scale functional organization of the genome. In this sense, SSs are the high-level organization of isochores. Further research is needed to establish the possible relation of SSs with other chromosome domains (as co-expression or functional domains) or with the chromosome architecture.

ACKNOWLEDGMENTS

We thank the Spanish Government (Grant BIO2008-01353 to J.L.O., P.C., and P.B., mobilities PR2009-0285 to P.C. and JC2009-00067 to A.V.C., and Juan de la Cierva Grant to M.H.), the Spanish Junta de Andalucía (Grants P07-FQM3163, P06-FQM1858), and the Basque Country (Programa de formación de investigadores grant to G.B.).

[1] G. Bernardi, Annu. Rev. Genet. **29**, 445 (1995).
[2] G. Bernardi *et al.*, Science **228**, 953 (1985).
[3] The Gene Ontology Consortium, Nat. Genet. **25**, 25 (2000). See also [http://www.geneontology.org/].
[4] J. L. Oliver *et al.*, Nucleic Acids Res. **32**, W287 (2004).
[5] M. Costantini *et al.*, Genome Research **16**, 536 (2006).
[6] T. Schmidt and D. Frishman, Genome Biol. **9**, R104 (2008).
[7] P. Bernaola-Galván, P. Carpena, R. Román-Roldán, and J. L. Oliver, Gene **300**, 105 (2002).
[8] C. K. Peng *et al.*, Nature (London) **356**, 168 (1992).
[9] Chromosome sequences (genome assembly hg18) downloaded from [http://genome.ucsc.edu/]. We generate chromosome arms by excluding the centromere and filling the remaining gaps with random sequence matching the composition of the flanks around each gap.
[10] P. Bernaola-Galván, R. Román-Roldán, and J. L. Oliver, Phys. Rev. E **53**, 5181 (1996).
[11] W. Li, Phys. Rev. Lett. **86**, 5815 (2001).
[12] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Roman-Roldan, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).
[13] V. Thakur, R. K. Azad, and R. Ramaswamy, Phys. Rev. E **75**, 011915 (2007).
[14] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Roman-Roldan, J. Oliver, and H.E. Stanley, Phys. Rev. E **65**, 041905 (2002).
[15] P. Carpena and P. Bernaola-Galván, Phys. Rev. B **60**, 201 (1999).
[16] Note that the use of D_{JS} is just a particular choice. HT can be carried out with any other measure of compositional difference; see, for example, P. Bernaola-Galván, P. C. Ivanov, L. A. NunesAmaral, and H.E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001).

[17] W. Li, Gene **276**, 57 (2001).
[18] P. Carpena, P. Bernaola-Galvan, A. V. Coronado, M. Hackenberg, and J. L. Oliver, Phys. Rev. E **75**, 032903 (2007).
[19] B. B. Mandelbrot and J. W. van Ness, SIAM Rev. **10**, 422 (1968).
[20] C.-K. Peng, S. Havlin, M. Schwartz, and H.E. Stanley, Phys. Rev. A **44**, 2239 (1991); H. A. Makse, S. Havlin, M. Schwartz, and H.E. Stanley, Phys. Rev. E **53**, 5445 (1996).
[21] P. Carpena, P. Bernaola-Galván, and P. Ch. Ivanov, Phys. Rev. Lett. **93**, 176804 (2004); A. Esmailpour, M. Esmaeilzadeh, E. Faizabadi, P. Carpena, and M. R. Tabar, Phys. Rev. B **74**, 024206 (2006).
[22] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley, and A.L. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[23] K. Hu, P. C. Ivanov, Z. Chen, P. Carpena, and H. E. Stanley, Phys. Rev. E **64**, 011114 (2001); A. V. Coronado and P. Carpena, J. Biol. Phys. **31**, 121 (2005).
[24] $\alpha \in [0.77,0.83]$ for human chromosomes (see [25]).
[25] [http://bioinfo2.ugr.es/SSinfoSup/].
[26] T. S. Furey and D. Haussler, Hum. Mol. Genet. **12**, 1037 (2003).
[27] P. W. Lord *et al*., Bioinformatics **19**, 1275 (2003).
[28] H. K. Lee *et al*., Genome Res. **14**, 1085 (2004).
[29] H. Li *et al*., BMC Genomics **7**, 103 (2006).
[30] K. D. Pruit *et al*., Nucleic Acids Res. **37**, D32 (2008).
[31] M. Hackenberg and R. Matthiesen, Bioinformatics **24**, 1386 (2008).