DNA Sequences at a Glance

Armando J. Pinho*, Sara P. Garcia, Diogo Pratas, Paulo J. S. G. Ferreira

Signal Processing Lab, IEETA/DETI, University of Aveiro, Aveiro, Portugal

Abstract

Data summarization and triage is one of the current top challenges in visual analytics. The goal is to let users visually inspect large data sets and examine or request data with particular characteristics. The need for summarization and visual analytics is also felt when dealing with digital representations of DNA sequences. Genomic data sets are growing rapidly, making their analysis increasingly more difficult, and raising the need for new, scalable tools. For example, being able to look at very large DNA sequences while immediately identifying potentially interesting regions would provide the biologist with a flexible exploratory and analytical tool. In this paper we present a new concept, the "information profile", which provides a quantitative measure of the local complexity of a DNA sequence, independently of the direction of processing. The computation of the information profiles is computationally tractable: we show that it can be done in time proportional to the length of the sequence. We also describe a tool to compute the information profiles of a given DNA sequence, and use the genome of the fission yeast *Schizosaccharomyces pombe* strain 972 h^{-} and five human chromosomes 22 for illustration. We show that information profiles are useful for detecting large-scale genomic regularities by visual inspection. Several discovery strategies are possible, including the standalone analysis of single sequences, the comparative analysis of sequences from individuals from the same species, and the comparative analysis of sequences from different organisms. The comparison scale can be varied, allowing the users to zoom-in on specific details, or obtain a broad overview of a long segment. Software applications have been made available for non-commercial use at http://bioinformatics.ua.pt/software/dna-at-glance.

Citation: Pinho AJ, Garcia SP, Pratas D, Ferreira PJSG (2013) DNA Sequences at a Glance. PLoS ONE 8(11): e79922. doi:10.1371/journal.pone.0079922

Editor: Cynthia Gibas, University of North Carolina at Charlotte, United States of America

Received November 27, 2012; **Accepted** September 30, 2013; **Published** November 21, 2013

Funding: Supported by the European Fund for Regional Development (FEDER) through the Operational Program Competitiveness Factors (COMPETE) and by the Portuguese Foundation for Science and Technology (FCT) in the context of projects FCOMP-01-0124-FEDER-010099 (FCT reference PTDC/EIA-EIA/103099/2008) and FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011). SPG acknowledges funding from the European Social Fund and the Portuguese Ministry of Education and Science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: ap@ua.pt

Introduction

This paper is about looking at DNA sequences or, more precisely, at graphical representations of DNA sequences. In other words, it is about the summarization of DNA data bearing in mind graphical representations, a problem related to some of the current challenges in large-scale computing [1]. The idea is old, as the sayings "a picture is worth a thousand words" and the century-old advertisement title "one look is worth a thousand words" show. In fact, the association of graphical information to DNA sequences has been pursued for long. Sequence logos [2] and the chaos game representation (CGR) [3] are two well-known examples. Most often, the underlying motivation is to look for and to display information related to the degree of randomness of the sequences, hoping to find meaningful structure. The degree of randomness is intimately related with the complexity, predictability, compressibility, repeatability and, ultimately, with the information theoretic notion of entropy of a sequence. Other methods, use the graphical paradigm for presenting several parameters that can be obtained from a DNA sequence. For example, the Genome Atlas of Jensen *et al.* [4] allows the visualization of information related to repeats, nucleotide composition, and structural parameters, in microbial genomes (the genome of *E. coli* is analyzed in [5] using this approach).

Some methods provide visual information of global properties of the DNA sequences. For example, CGR uses the distribution of points in an image to express the frequency of the oligonucleotides that compose the sequence [6]. From these CGR images, other global representations can be derived, such as genomic signatures [7,8] or entropic profiles [9].

Originally [9], entropic profiles were estimated using global histograms of the oligonucleotide frequencies, calculated using CGR images. Later, they have been generalized by Vinga *et al.* [10] in order to calculate and visualize local entropic information. Other approaches for estimating the randomness along the sequence have also been proposed. For example, Crochemore *et al.* [11] used the number of different oligonucleotides that are found in a window of predefined size for estimating the entropy. Troyanskaya *et al.* [12] proposed the linguistic complexity, also calculated on a sliding window, as a measure of the local complexity of the DNA sequence.

Both the global and the local estimates of the randomness of a sequence provide useful information and both have shortcomings. The global estimates do not show how the characteristics change along the sequence and the local estimates fail to take into consideration the global properties of the sequence. This last drawback was addressed by Clift *et al.* [13] using the concept of sequence landscape. Using directed acyclic word graphs, they were able to construct plots displaying the number of times that oligonucleotides from the target sequence occur in a given source sequence. If the target and source sequences coincide, then the

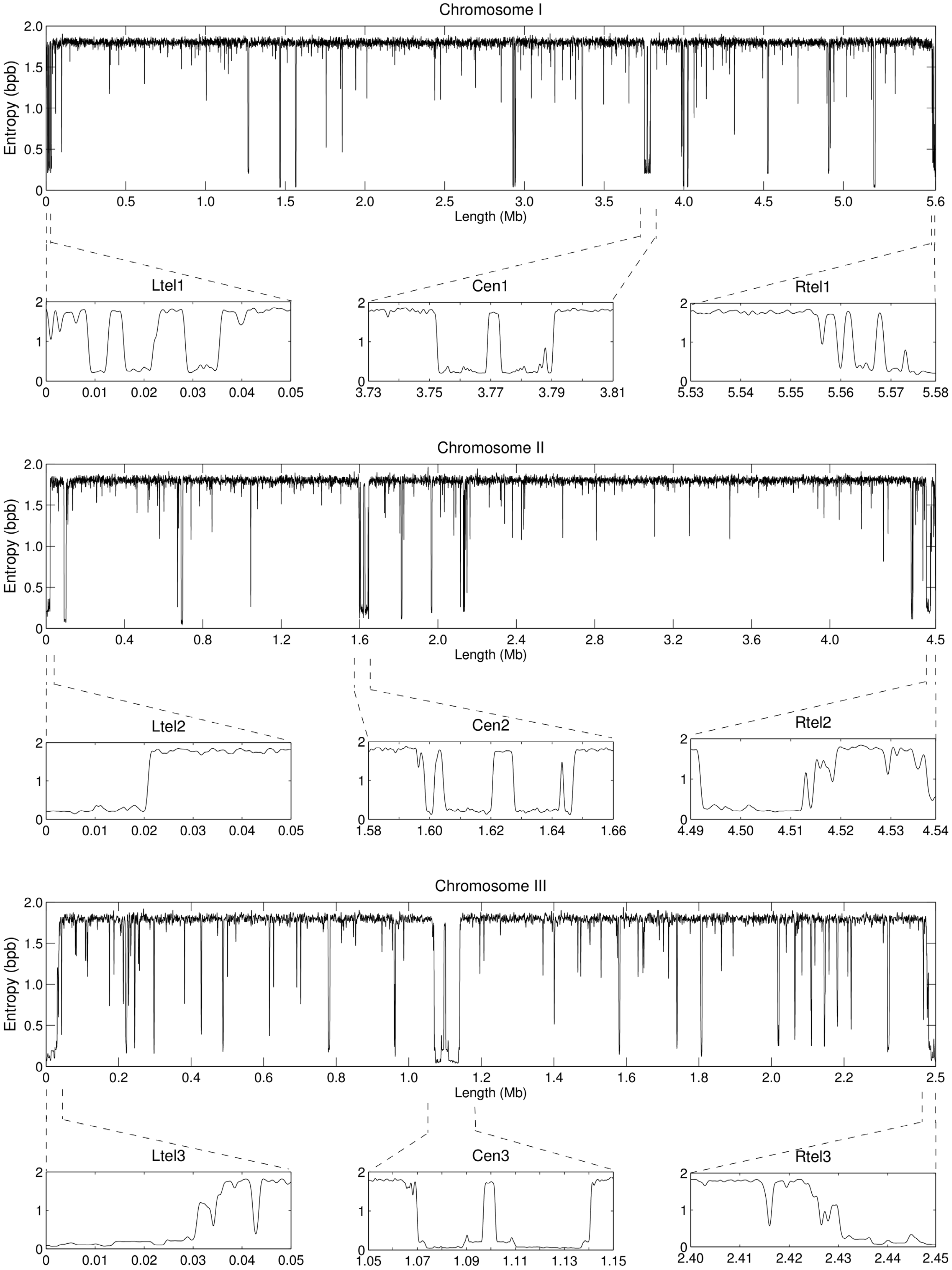
Chromosome I
Entropy (bpb)
2.0
1.5
1.0
0.5
0
0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0 5.6
Length (Mb)
Ltel1
0 0.01 0.02 0.03 0.04 0.05
Cen1
3.73 3.75 3.77 3.79 3.81
Rtel1
5.53 5.54 5.55 5.56 5.57 5.58
Chromosome II
0 0.4 0.8 1.2 1.6 2.0 2.4 2.8 3.2 3.6 4.0 4.5
Length (Mb)
Ltel2
0 0.01 0.02 0.03 0.04 0.05
Cen2
1.58 1.60 1.62 1.64 1.66
Rtel2
4.49 4.50 4.51 4.52 4.53 4.54
Chromosome III
0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.5
Length (Mb)
Ltel3
0 0.01 0.02 0.03 0.04 0.05
Cen3
1.05 1.07 1.09 1.11 1.13 1.15
Rtel3
2.40 2.41 2.42 2.43 2.44 2.45
2
1
0

Figure 1. Information profiles of the chromosomes of *S. pombe* highlighting centromeric and telomeric regions. The profiles are the result of eight competitive finite-context models with context depths 2, 4, 6, 8, 10, 12, 14 and 16. They represent the minimum of the combined direct and reversed profiles, low-pass filtered with a Blackman smoothing window of 1,001 bp. Probabilities were estimated with $\alpha = 1/20$ for contexts 14 and 16. For clarity, the full chromosome profiles were sampled every 20 bp. Zoomed in profiles Ltel and Rtel display telomeric and subtelomeric regions of each chromosome, and zoomed in Cen profiles display the respective centromeric regions.
doi:10.1371/journal.pone.0079922.g001

landscape provides information about self-similarities (repeats) of the target sequence.

The sequence landscapes of Clift *et al.* [13] seem to have been the first attempt of displaying local information while taking into account the global structure of the sequence. This idea was also pursed by Allison *et al.* [14], using a model that considers a sequence as a mixture of regions with little structure and regions that are approximate repeats. Based on this statistical model, they have produced information sequences, which quantify the amount of surprise of having a given base at a given position, knowing the remaining left (or right) part of the sequence. When plotted, these information sequences provide a quick overview of certain properties of the original symbolic sequence, allowing for example to easily identify zones of rich repetitive content [15–17].

The interest of complexity measures for DNA sequence analysis has been explored by several researchers, such as in [18–20]. The key measure is known as Kolmogorov complexity, and was independently introduced by Solomonoff [21,22], Kolmogorov [23] and Chaitin [24], and further developed in [25,26]. The Kolmogorov complexity of a string is the shortest program for a universal computer which outputs the string and stops. This measure is non-computable and is usually approximated by other computable measures, such as, Lempel-Ziv complexity measures [18,27], linguistic complexity measures [28], or compression-based complexity measures [17,29,30].

The information sequences of Allison *et al.* [14] are intimately related to data compression. The importance of data compression for pattern discovery in the context of DNA sequences was already recognized by Grumbach *et al.* [31] and, since then, it has been reinforced by others (e.g. [15,32]). In fact, the existence of regularities in a sequence renders it algorithmically compressible. The algorithmic information content of a sequence is the size, in bits, of the shortest accurate description of the sequence.

Compression-based complexity measures have an intuitive definition (regular sequences are compressible, as opposed to random ones) and their efficiency is easily quantifiable by the number of bits generated by the encoder. DNA is non-stationary, with regions of low information content (i.e., low entropy or low complexity) alternating with regions of average entropy close to two bits per base. This property is modeled by most DNA compression algorithms with a low-order finite-context model for the high-entropy regions and a Lempel-Ziv dictionary-based or copy-based approach for the repetitive, low-entropy regions. XM [16] has been one of the most successful approaches for compressing DNA sequence data. Also, because it provides a probabilistic distribution for each DNA base being encoded, it offers a natural way of obtaining the above mentioned information sequences [17].

In this paper we "look at DNA sequences" by using information profiles derived from a probabilistic model. The model consists of a combination of several finite-context models, each of a different depth. Such models have been shown to adequately capture the statistical properties of DNA sequences [30,33,34] but are direction-dependent, that is, the results depend on which direction the DNA sequence is processed. In this work we remove this directional dependency by combining the amount of information that a certain DNA base carries in each processing direction.

The information profiles are found using an algorithm based on finite-context models that needs time proportional to the length of the sequence. We present a proof-of-concept study of the potential of information profiles in genome analysis, namely, for detecting genomic structural and functional regularities. We uncover genomic regularities on a large-scale, such as, centromeric and telomeric regions of a chromosome, or transposable elements. In this context, we use the genome of the fission yeast *Schizosaccharomyces pombe* strain 972 h^- as case-study. We also present the potential of information profiles in a comparative genomics approach, using five human chromosomes 22. This example provides evidence that the proposed method scales well when applied to larger genomic sequences.

Materials and Methods

Genomic Data

We use chromosomes I (accession number NC003424.3), II (accession number NC003423.3) and III (accession number NC003421.2) of the genome of the fission yeast *Schizosaccharomyces pombe* strain 972 h^- [35], retrieved from the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). We also use the human chromosome 22 of the reference genome assembly GRCh37.p9 (accession number NC000022.10, [36]), the alternate Celera assembly (accession number AC000065.1, [37]), the genome of J. Craig Venter (HuRef assembly, accession number AC000154.1, [38]), the genome of a Han Chinese individual (YH assembly, [39]), and the genome of a Korean individual (KOREF 20090224 assembly, [40]). The first three mentioned versions of human chromosome 22 were also retrieved from NCBI, the YH chromosome 22 was retrieved from the Beijing Genomics Institute (BGI) website (ftp://public.genomics.org.cn/BGI/yanhuang/fa/), and the KOREF chromosome 22 was retrieved from ftp://bioftp.org/BiO/Store/Genome/KOREF_KoreanReferenceGenome/KOREF_20090224/fasta/.

Information Profiles based on Finite-context Models

To "look at DNA" at different scales we rely on information profiles that quantitatively measure the local complexity of the DNA sequence. The profiles provide a visual representation of the sequence, and can be interpreted in a simple way. The less regular the behaviour, the higher the numerical values. Thus, visual inspection immediately shows regions of low complexity (for example, repetitions), regions of high complexity, and other patterns of possible interest.

The probabilistic models required to draw the information profiles are, not surprisingly, related to data compression and information-theoretic concepts such as entropy. Compression and modeling are intertwined: the problem of discovering an efficient representation of the information source can be stated as a data modeling problem.

A probabilistic model of a DNA sequence is a mathematical description of the sequence, viewed as an information source. The model provides an estimate of the probability of the next DNA symbol. The entropy of the model sets a lower bound on the compression performance. Conversely, compression performance yields a bound on the entropy. However, a given compression

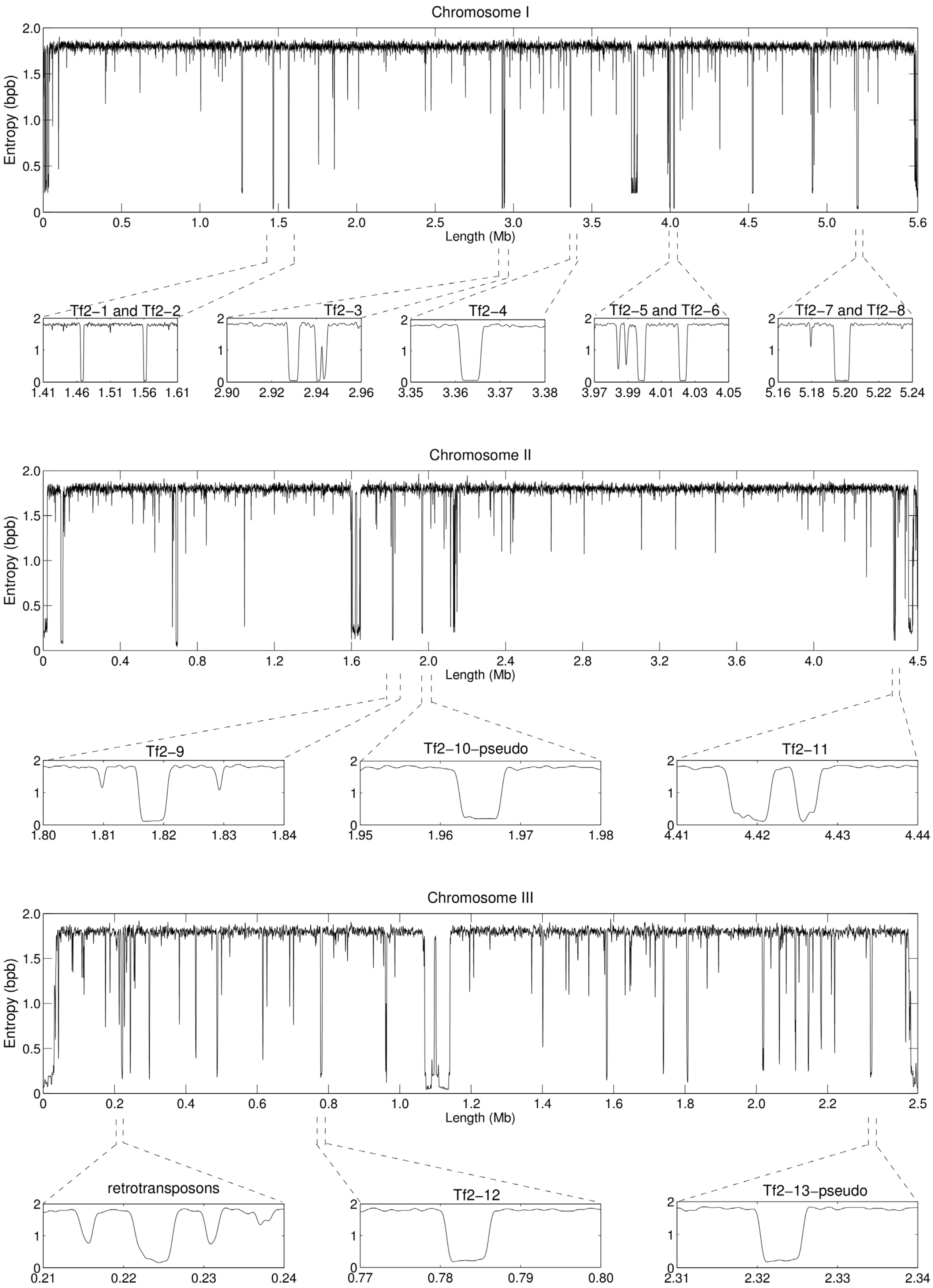
Chromosome I
Entropy (bpb)
2.0
1.5
1.0
0.5
0
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.6
Length (Mb)
Tf2−1 and Tf2−2
1.41
1.46
1.51
1.56
1.61
Tf2−3
2.90
2.92
2.94
2.96
Tf2−4
3.35
3.36
3.37
3.38
Tf2−5 and Tf2−6
3.97
3.99
4.01
4.03
4.05
Tf2−7 and Tf2−8
5.16
5.18
5.20
5.22
5.24
2
1
0
Chromosome II
Entropy (bpb)
2.0
1.5
1.0
0.5
0
0
0.4
0.8
1.2
1.6
2.0
2.4
2.8
3.2
3.6
4.0
4.5
Length (Mb)
Tf2−9
1.80
1.81
1.82
1.83
1.84
Tf2−10−pseudo
1.95
1.96
1.97
1.98
Tf2−11
4.41
4.42
4.43
4.44
Chromosome III
Entropy (bpb)
2.0
1.5
1.0
0.5
0
0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.5
Length (Mb)
retrotransposons
0.21
0.22
0.23
0.24
Tf2−12
0.77
0.78
0.79
0.80
Tf2−13−pseudo
2.31
2.32
2.33
2.34

Figure 2. Information profiles of the chromosomes of *S. pombe* highlighting Tf2-type retrotransposons. Parameters are the same as in Fig. 1. Zoomed in profiles display the 13 full length Tf2 elements and an additional display of other Tf2-type retrotransposons.
doi:10.1371/journal.pone.0079922.g002

method may have limited potential or interest in connection with information profiles.

Of the myriad of coding methods proposed for compressing genomic sequences (e.g. [16,30,31,33,34,41–49]), most are based on search procedures for finding exact or approximate repeats, in the sequence itself or in its reversed complement. Although this may lead to interesting compression rates, it generally requires a significant computational effort. Recently, it has been shown that appropriate combinations of finite-context models are able to give competitive [30] or even superior [34] compression results, at a smaller computational cost.

Finite-context models are probabilistic models based on the assumption that the information source is Markovian, i.e., that the probability of the next outcome depends only on some finite number of (recent) past outcomes referred to as the context. The proposed approach is based on a mixture of finite-context models. We assign probability estimates to each symbol in $\mathcal{A}=\{\mathrm{A,C,G,T}\}$, regarding the next outcome, according to a conditioning context computed over a finite and fixed number $k>0$ of past outcomes $x_{n-k+1..n}=x_{n-k+1}\ldots x_n$ (order-k finite-context model with $|\mathcal{A}|^k$ states).

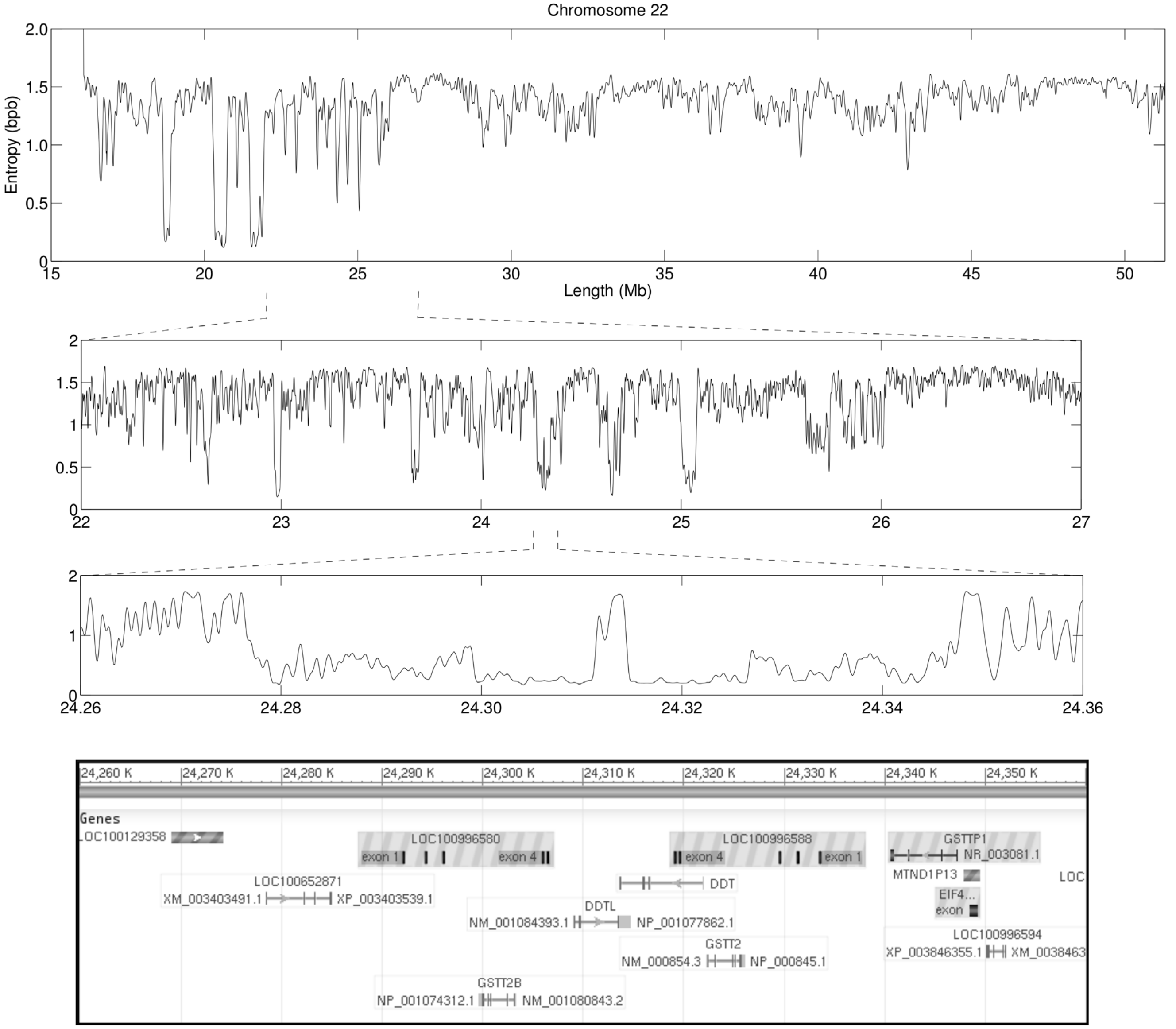

Figure 3. Information profile of chromosome 22 of the GRCh37 human reference genome assembly. Parameters are similar to those of Fig. 1, except for the smoothing window, which has a value of 100,001 bp in the upper plot, 10,001 bp in the middle plot, and 1,001 bp in the lower plot. Zoomed-in profiles reveal regularities at increasingly larger resolution, including several genes from duplicated gene families in the lower plot. The lower panel was downloaded from the NCBI website and it identifies annotated genes in the zoomed-in region.
doi:10.1371/journal.pone.0079922.g003

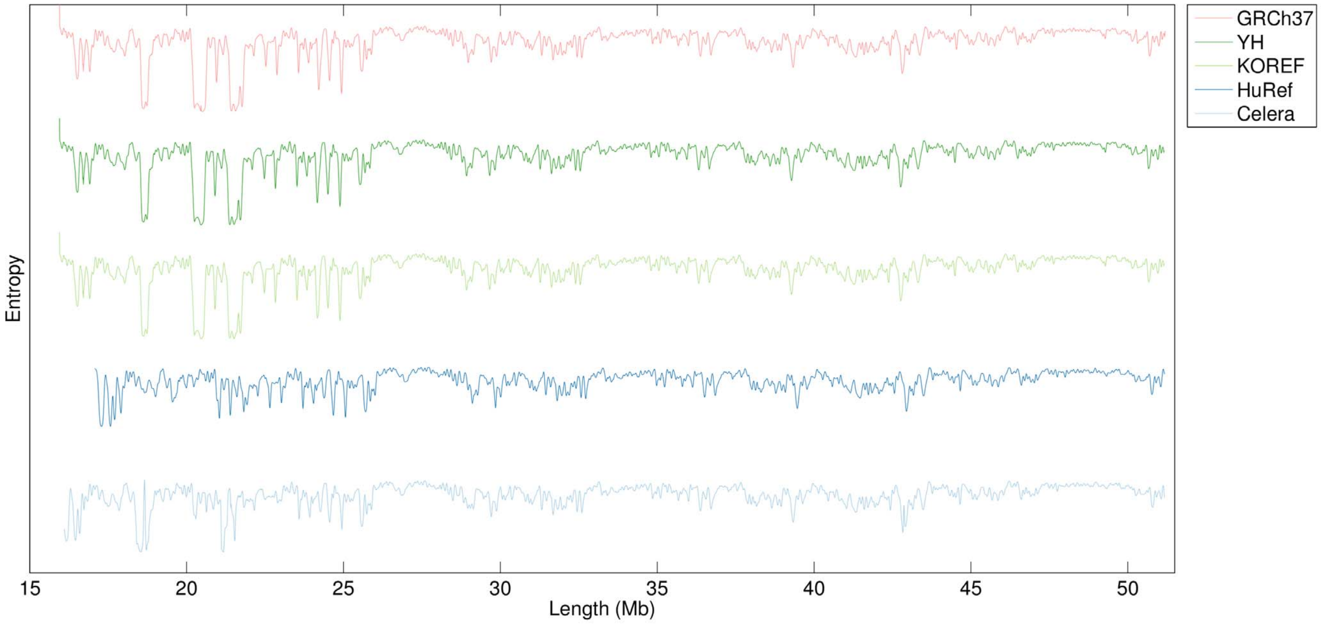

Figure 4. Information profiles of chromosome 22 in five human genome assemblies. Parameters are similar to those of Fig. 1, except for the smoothing window, which has a value of 100,001 bp. GRCh37 is the reference human genome assembly, YH is the genome assembly of a Han Chinese individual, KOREF is the genome assembly of a Korean individual, HuRef is the genome assembly of J. Craig Venter, and Celera is an alternate human genome assembly. These plots highlight, on the one hand, the high similarity between the GRCh37 assembly and the YH and KOREF assemblies. On the other hand, it is also easy to find out regions where both the HuRef and Celera assemblies diverge considerably from the GRCh37 reference assembly (e.g., in the range 20–23 Mb).
doi:10.1371/journal.pone.0079922.g004

The probability estimates $P(x_{n+1}|x_{n-k+1..n})$ are calculated using symbol counts that are accumulated while the sequence is processed, making them dependent not only on the past k symbols, but also on n. We use the estimator

$$P(s|x_{n-k+1..n}) = \frac{C(s|x_{n-k+1..n}) + \alpha}{C(x_{n-k+1..n}) + |\mathcal{A}|\alpha}, \tag{1}$$

where $C(s|x_{n-k+1..n})$ represents the number of times that, in the past, symbol s was found having $x_{n-k+1..n}$ as the conditioning context and where

$$C(x_{n-k+1..n}) = \sum_{a \in \mathcal{A}} C(a|x_{n-k+1..n}) \tag{2}$$

is the total number of events that has occurred so far in association with context $x_{n-k+1..n}$. Parameter α allows balancing between the maximum likelihood estimator and an uniform distribution (when the total number of events, n, is large, it behaves as a maximum likelihood estimator). For $\alpha = 1$, (1) reduces to the well-known Laplace estimator.

The per symbol information content average provided by the finite-context model of order-k, after having processed n symbols, is given by

$$H_{k,n} = -\frac{1}{n} \sum_{i=0}^{n-1} \log_2 P(x_{i+1}|x_{i-k+1..i}) \tag{3}$$

bits per symbol. When using several models simultaneously, the $H_{k,n}$ can be viewed as measures of the performance of those models until that instant. Therefore, the probability estimate can be given by a weighted average of the probabilities provided by each model, according to

$$P(x_{n+1}) = \sum_k P(x_{n+1}|x_{n-k+1..n})\, w_{k,n}, \tag{4}$$

where $w_{k,n}$ denotes the weight assigned to model k and

$$\sum_k w_{k,n} = 1. \tag{5}$$

Our modeling approach is based on a mixture of probability estimates. In order to compute the probability estimate for a certain symbol, it is necessary to combine the probability estimates given by (1) using (4). The weight assigned to model k can be computed according to

$$w_{k,n} = P(k|x_{1..n}), \tag{6}$$

i.e., by considering the probability that model k has generated the sequence until that point. In that case, we would get

$$w_{k,n} = P(k|x_{1..n}) \propto P(x_{1..n}|k)P(k), \tag{7}$$

where $P(x_{1..n}|k)$ denotes the likelihood of sequence $x_{1..n}$ being generated by model k and $P(k)$ denotes the prior probability of model k. Assuming

$$P(k) = \frac{1}{K}, \tag{8}$$

where K denotes the number of models, we also obtain

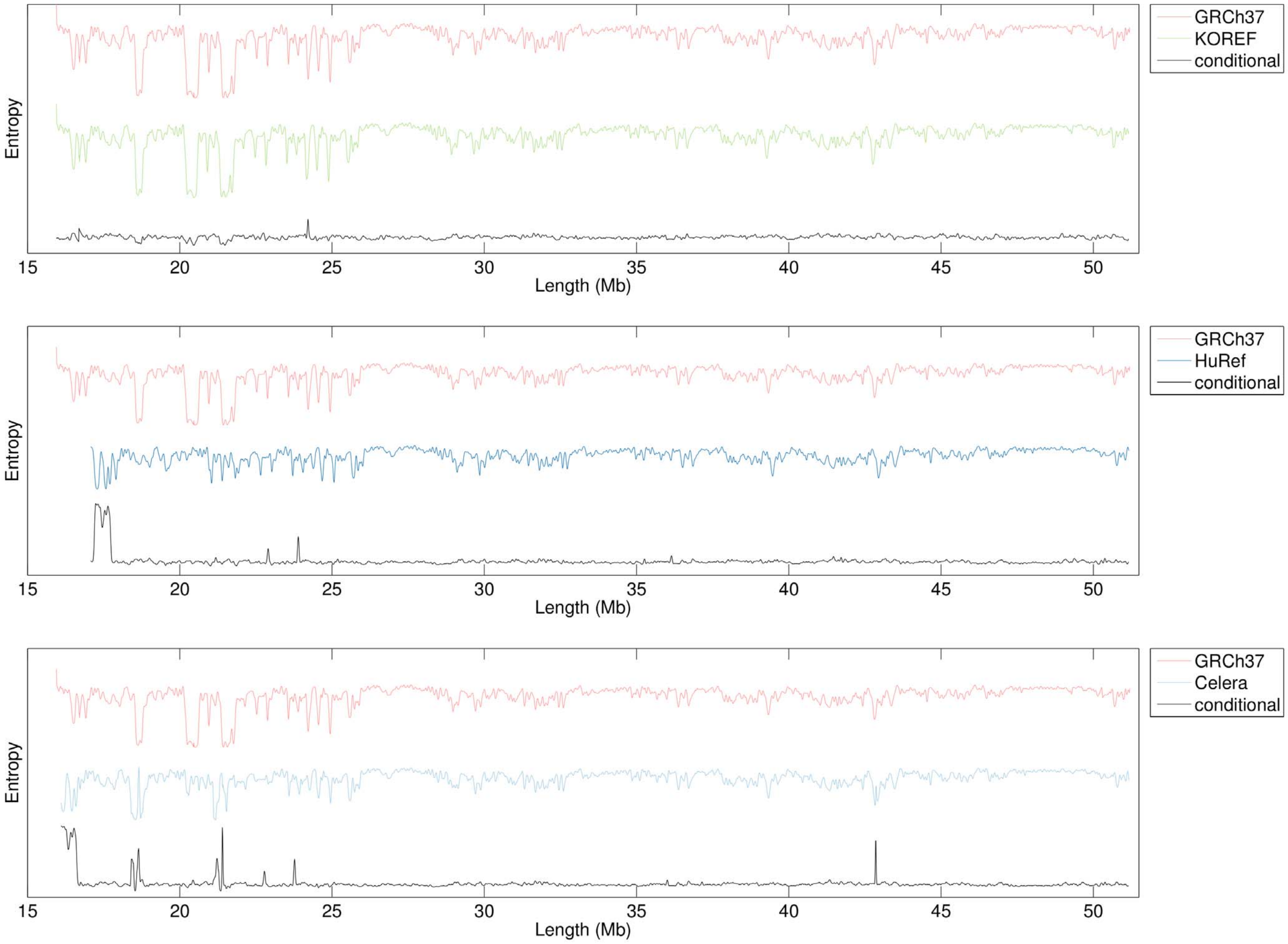

Figure 5. Information and conditional profiles of three pairwise comparisons of human chromosomes 22. Parameters are similar to those of Fig. 1, except for the smoothing window, which has a value of 100,001 bp. GRCh37 is the reference human genome assembly, KOREF is the genome assembly of a Korean individual, HuRef is the genome assembly of J. Craig Venter, and Celera is an alternate human genome assembly. The conditional profiles were obtained using the statistics of the finite-context models trained over the GRCh37 human genome assembly. Peaks in these profiles highlight regions of sequence divergence in the KOREF, HuRef and Celera chromosomes, with respect to the GRCh37 one. As an example of the additional information conveyed by these conditional profiles, we highlight the peak in the Celera assembly around base 43 Mb. Whereas slightly perceivable in the non-conditional profiles, the divergence of the two assemblies (GRCh37 and Celera) at this particular location is much more evident in the conditional profile.

doi:10.1371/journal.pone.0079922.g005

$$w_{k,n} \propto P(x_{1..n}|k). \tag{9}$$

Calculating the logarithm we get

$$\log_2 P(x_{1..n}|k) = \log_2 \prod_{i=1}^{n} P(x_i|k,x_{1..i-1}) = \tag{10a}$$

$$= \sum_{i=1}^{n} \log_2 P(x_i|k,x_{1..i-1}), \tag{10b}$$

which is related to the number of bits that would be required by model k for representing the sequence $x_{1..n}$. It is, therefore, the accumulated measure of the performance of model k until instant n. DNA sequences are known to be non-stationary. Due to this, the performance of a model may vary considerably from region to region of the sequence. In order to extract the best possible performance from each model, we adopted a progressive forgetting mechanism. The idea is to allow each model to progressively forget the distant past and, consequently, to give more importance to recent outcomes. Therefore, we rewrite (11) as

$$\sum_{i=1}^{n} \log_2 P(x_i|k,x_{1..i-1}) = \tag{11a}$$

$$= \gamma \sum_{i=1}^{n-1} \log_2 P(x_i|k,x_{1..i-1}) + \log_2 P(x_n|k,x_{1..n-1}), \tag{11b}$$

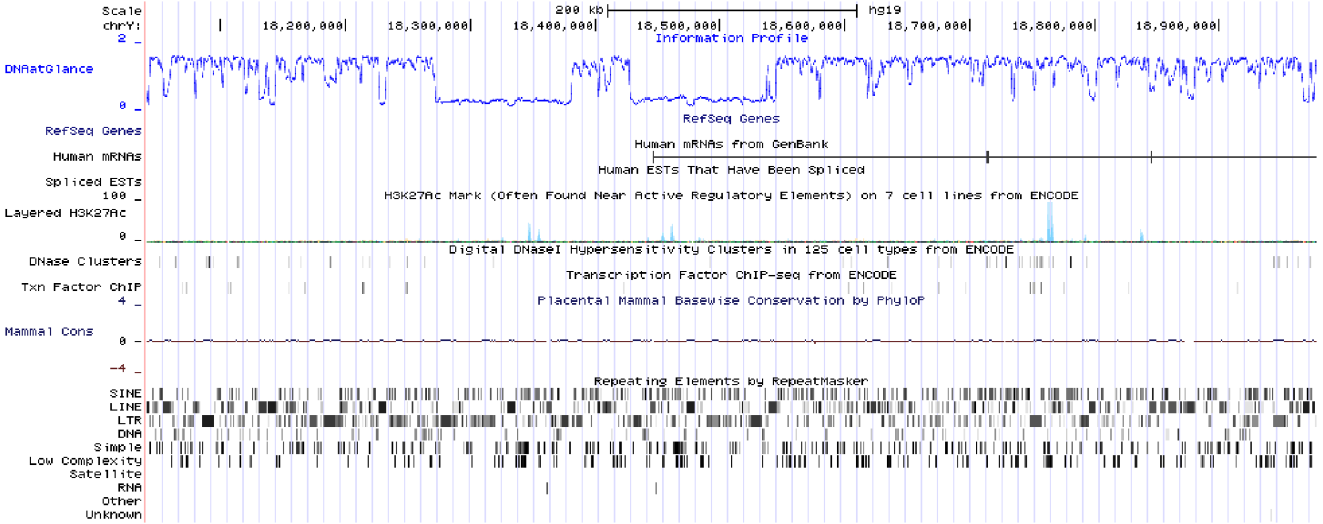
Scale
chrY:
200 kb
hg19
18,200,000
18,300,000
18,400,000
18,500,000
18,600,000
18,700,000
18,800,000
18,900,000
Information Profile
2
DNAatGlance
0
RefSeq Genes
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
100
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
0
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
4
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
0
-4
Repeating Elements by RepeatMasker
SINE
LINE
LTR
DNA
Simple
Low Complexity
Satellite
RNA
Other
Unknown

Figure 6. Information profile of part of human chromosome Y integrated in the UCSC Genome Browser. The custom track named "DNAatGlace" was uploaded to the browser in WIG format.
doi:10.1371/journal.pone.0079922.g006

where $\gamma \in [0,1)$ dictates the forgetting factor to be used. Defining

$$p_{k,n} = \prod_{i=1}^{n} P(x_i|k, x_{1..i-1}) \tag{12}$$

and removing the logarithms, we can rewrite (0) as

$$p_{k,n} = p_{k,n-1}^{\gamma} P(x_n|k, x_{1..n-1}) \tag{13}$$

and, finally, set the weights to

$$w_{k,n} = \frac{p_{k,n}}{\sum_k p_{k,n}}. \tag{14}$$

This probabilistic model yields an estimate of the probability of each symbol in the DNA sequence, and as such it allows us to quantify the degree of randomness or surprise along one direction of the sequence.

Results and Discussion

Chromosomes are processed both in the downstream, or direct ($5' \rightarrow 3'$), and upstream, or reversed ($3' \rightarrow 5'$), directions. This dual processing aims at eliminating the directionality bias introduced when only one of the two possible directions is taken into consideration. Therefore, the information content of each DNA base is calculated by running the statistical model in one direction, then in the other direction, and finally by taking the smallest value obtained.

Fission Yeast

Figure 1 displays the information profiles of the three chromosomes in the genome of *S. pombe*, obtained independently for each chromosome. The profiles are the result of the combination of eight finite-context models with context depths of 2, 4, 6, 8, 10, 12, 14 and 16. They represent the minimum of the combined direct and reversed profiles, low-pass filtered with a Blackman window of 1,001 bp. Probabilities were estimated with $\alpha = 1/20$ in Eq. 1 for the larger contexts of $k = 14$ and $k = 16$. For clarity, the full chromosome profiles shown result from sampling every 20 bp.

Low-information regions in Fig. 1 are associated with the presence of repetitive sequences. For example, chromosome III has more and often more prominent low-information regions than chromosomes I and II, which is in compliance with some properties of this chromosome concerning repetitive structures, such as, the presence of tandem rDNA repeats [50] or the density of transposable element remnants in this chromosome being twice that of chromosomes I and II [35]. Annotated with Ltel and Rtel are regions of low-information content pertaining the telomeric (where available; see http://www.sanger.ac.uk/Projects/S_pombe/telomeres.shtml) and subtelomeric regions of each chromosome. Annotated with Cen are regions of low-information content pertaining the centromeric regions of each chromosome.

Telomeres in *S. pombe* consist of ~300 bp long tandemly repeated $5'$-GGTTACA$_{0-6}$C$_{0-1}$G$_{0-6}$-$3'$sequences, with the GGTTAC repetitive unit being the most commonly found [51,52]. Chromosomes I and II share some subtelomeric sequences, while the telomeric repeats at both ends of chromosome III are immediately flanked by tandem arrays of rRNA genes [50,53]. The highly repetitive content of these regions, including degenerate and tandem telomeric repeats [50], and duplicated and highly-similar subtelomeric regions [35,50], is captured in the low-information Ltel and Rtel regions of the profiles in Fig. 1.

Mammalian centromeres contain a large number of tandemly arranged repetitive sequences. Wood *et al.* [35] reported an estimated length of 35 kb for the centromere of chromosome I, 65 kb for the centromere of chromosome II, and 110 kb for the centromere of chromosome III, in inverse proportion to the lengths of the respective chromosomes, namely, 5.7 Mbp, 4.6 Mbp, and 3.5 Mbp. However, updated centromere positions are cen1: 3,753,687–3,789,421 bp, cen2: 1,602,264–1,644,747 bp, and cen3: 1,070,904–1,137,003 bp (http://www.sanger.ac.uk/Projects/S_pombe/centromere.shtml), which correspond to a decrease of ~30% in the length of annotated centromeric regions cen2 and cen3 with respect to previous values [35]. These updated lengths of the centromeric regions and inverse proportionality to the chromosome size are recovered in the information profiles of Fig. 1.

Cen1 consists of a non-conserved central core (cnt1) of 4.1 kb flanked by two 5.6-kb imperfect inverted imr1 repeats (imr1L, imr1R) that display sequence identity with each other, and two pairs of 4.4-kb dg and 4.8-kb dh repeats (dg1, dh1) separated by cen253, a repeat of ~0.3 kb. The maps of the other two centromeres have the same basic structure, with central cnt regions flanked by imr repeats and by variable numbers of dg and dh repeats separated by cen253. Moreover, there are many tRNA genes in the centromeric regions, with clusters flanking cen2 and cen3 and also within the imr regions of all three centromeres [35,52]. This centromeric mirror-like repetitive structures are captured in the Cen regions of the profiles in Fig. 1, with the core cnt regions evident by higher-information central peaks, the imr and dg/dh repeats accounting for regions of low-information content, and t-RNA genes contributing to other peaks of higher-information e.g. at the frontiers of cen2.

Figure 2 displays again the information profiles of the three chromosomes in the genome of *S. pombe*, obtained and sampled similarly as in Fig. 1. Highlighted are again low-information regions associated with the presence of repetitive sequences, now focusing on transposable elements.

Two related families of long terminal repeat (LTR)-retrotransposons, named transposon of fission yeast 1 (Tf1) and 2 (Tf2), have been identified in *S. pombe* [54]. Retrotransposons are mobile DNA elements ubiquitous in eukaryotic genomes, which remain active in most mammalian genomes. They mobilize via an RNA intermediate that is then reverse transcribed and reintegrated into the genome by a copy-and-paste mechanism, thereby duplicating the element. Strain 972 h^{-} of *S. pombe* contains 13 full length Tf2 elements of length ~4.9 kb and no Tf1 elements. It also contains many single LTRs derived from Tf1 and Tf2 elements [52]. Annotated in Fig. 2 are all 13 full length Tf2 elements, plus an additional display of other Tf2-type retrotransposons. Plot Tf23 in chromosome I displays that retrotransposon in the first low-information region, followed by a large retrotransposon in the second low-information region, which is in accordance with the annotations in [52]. The wider low-information region in plot Tf27 and Tf28 includes both elements. Plot Tf211 in chromosome II displays that retrotransposon in the second low-information region, preceded by a large retrotransposon. The plot annotated with retrotransposons in chromosome III showcases some of the LTRs derived from Tf1 and Tf2 elements, where a large repeat is associated to the wider low-information region and two smaller elements are identified in the two additional low-information regions.

This accurate matching of the low-information regions in Figs. 1 and 2 to annotated repetitive genomic structures, such as the centromeric and telomeric regions of a chromosome or its transposable elements, proves information profiles may be useful in *de novo* discovery of large-scale genomic regularities. Clearly, it is not possible to infer the genomic sequence *per se* from the information profiles, or the location of genomic regularities within base pair resolution. However, it is possible to discover the presence of regularities on a genome-wide scale, which may be useful for an exploratory genome analysis or for genome comparisons.

Human Chromosome 22

To illustrate the potential of information profiles in the analysis of larger and more complex genomes, we use the human chromosome 22 as case-study. This ~51 Mbp chromosome is the second smallest human autosome and it was the first to be fully sequenced [55].

Figure 3 displays the information profile of chromosome 22 of the GRCh37 reference human genome assembly. As before, the profiles are the result of the combination of eight finite-context models with context depths of 2, 4, 6, 8, 10, 12, 14 and 16. They represent the minimum of the combined direct and reversed profiles, low-pass filtered. Probabilities were estimated with $\alpha = 1/20$ in Eq. 1 for the larger contexts of $k = 14$ and $k = 16$. As the first 15 Mbp remain unsequenced (containing solely Ns), the upper plot in Fig. 3, which displays the information profile of the whole chromosome, ignores this region. In order to display the global information of such a large chromosome, the profile should be heavily low-pass filtered. Here, we used a smoothing window size of 100,001 bp.

The first striking feature of this chromosome-wide profile is that the average information content is considerable lower than that of the chromosomes of *S. pombe* (Figs. 1 and 2). This is a direct consequence of the fact that ~42% of the human chromosome 22 comprises interspersed and tandem repeats [55].

The middle plot in Fig. 3 shows a 5 Mbp zoomed-in region of the chromosome (22–27 Mbp), filtered with a smoothing window of 10,001 bp. Clearly, additional detail and regularities are observable at increasingly larger resolution. Highlighted in the lower plot is a region of low entropy, as consequence of being densely occupied by genes from duplicated gene families, filtered with a smoothing window of 1,001 bp. For completeness, the final plot in Fig. 3 shows an image of gene annotations taken from the NCBI nucleotide browser corresponding to the displayed region. One example of such gene families are the glutathione S-transferases, with several genes and pseudogenes annotated to this region. Gene LOC100652871, a glutathione S-transferase theta-4-like, is located in region 24,278,480–24,284,985 bp. Another glutathione S-transferase theta-4-like gene, LOC 100996594, is located in region 24,350,125–24,352,036 bp of the reversed complement. Gene LOC100996580, a glutathione S-transferase theta-1-like, is located in region 24,292,105–24,306,644 bp. Another glutathione S-transferase theta-1-like gene, LOC100996588, is located in region 24,319,105–24,333,620 bp of the reversed complement. Gene GSTT2 (ID 2953), a glutathione S-transferase theta 2, is located in region 24,322,314–24,326,106 bp. Gene/pseudogene GSTT2B (ID 653689), a glutathione S-transferase theta 2B is located in region 24,299,601–24,303,368 bp of the reversed complement. Finally, the pseudogene GSTTP1 (ID 25774), a glutathione S-transferase theta pseudogene 1, is located in region 24,340,595–24,347,258 bp of the reversed complement.

We selected this particular region of chromosome 22 for showcasing because it reinforces the reason why this new tool may be valuable. During our own exploration of this analysis method, we were browsing through the information profile of the human chromosome 22, observing regions of roughly 1 Mbp in length, when that striking and almost symmetric profile of about 100,000 bp caught our attention. Many more similarly interesting regions can be observed along the chromosome. Hence, the tool here proposed provides a handy procedure for quickly detecting potentially interesting genomic regions.

To illustrate the potential of information profiles in the context of comparative genomics, we use again the human chromosome 22 as case-study.

Figure 4 displays the information profiles for five human chromosomes 22. As before, the profiles are the result of the combination of eight finite-context models with context depths of 2, 4, 6, 8, 10, 12, 14 and 16. They represent the minimum of the combined direct and reversed profiles, and low-pass filtered with a smoothing window of 100,001 bp. Here, similarities and differences between the sequences are clearly visible. Both YH and KOREF human genome assemblies were obtained from resequencing experiments that used the GRCh37 reference human genome assembly for mapping their short reads. This is the main reason why both profiles are so similar to that of GRCh37. On the other hand, the HuRef and Celera assemblies are two *de novo* assemblies [37,38], hence their profiles are considerably more dissimilar to that of GRCh37.

Figure 5 repeats the information profiles for four human chromosomes 22, as well as, the conditional profiles of three pairwise comparisons between those chromosomes. Here, the discovery strategy was different, as the conditional profiles were obtained using the statistics of the finite-context models trained over the GRCh37 chromosome, with the same parameters as described above. As such, the baseline-like regions highlight sequence similarity between both chromosomes, whereas peaks highlight regions of clear sequence divergence in the KOREF, HuRef and Celera chromosomes, with respect to the GRCh37 one. The main observation stemming from these conditional profiles pertains the large-scale structural variation between these human chromosomes. For the *de novo* assemblies (HuRef and Celera), most of the observed variation occurs in the beginning of the profiles. The peak a little over the 24 Mb mark is common to all three conditional profiles, hinting at the possibility of this being a highly-variable region in the human population.

Conclusion

We introduced an algorithm to detect genomic regularities within a *blind* discovery strategy. The algorithm uses information profiles built using an efficient DNA sequence compression method. The results described support our claim that information profiles provide a valuable discovery tool for the genome-wide individual or comparative analysis of genomes, through the detection of biologically-relevant genomic regularities. We used the genome of the fission yeast *Schizosaccharomyces pombe* strain 972 h^{-} for illustration. This model-organism was chosen because of its genome size, which renders visualization of the information profiles easier. Nevertheless, to give evidence that the tool is also applicable to larger genomes, we included information profiles at several scales of the human chromosome 22. Using five human chromosomes 22, we also showcased the potential of this methodology for comparative genomics analyses.

Our algorithm relies on the efficient probabilistic modeling of the genomic sequence based on finite-context (Markov) models. The approach is sufficiently flexible and powerful to enable addressing various biological questions and quickly obtaining the corresponding information profiles for a first-hand assessment. Indeed, the creation of information profiles does not require unusual computational facilities. Building an information profile requires a computation time that varies only linearly with the size of the sequence. For example, the information profile of human chromosome 22 was created in a laptop computer in less than five minutes. Moreover, the amount of computer memory required does not depend on the size of the sequence, but only on the depth of the finite context models used for modeling the sequence.

To facilitate the exploration of the information profiles here introduced, we made available two software applications: one is highly flexible but command-line based; the other has a graphical user interface and was designed to be very easy to use. Both applications are freely available for non-commercial use and can be downloaded from http://bioinformatics.ua.pt/software/dna-at-glance.

Due to its exploratory nature, these software applications currently offer a number of options that allow for many combinations of the parameters. Nevertheless, for ease of use, they can also be ran with default parameters. A detailed explanation of these parameters and some examples of their use is included in the packages.

In Figure 6 we show an example of integration of the information profiles in the UCSC Genome Browser, in this case displaying a segment of human chromosome Y. The data were uploaded as a custom track in WIG format (that can be produced by the supporting software applications). As can be seen in this example, there are two relatively large regions of low information content that are not easily guessed by inspection of the output provided by the RepeatMasker tool, giving evidence of the complementary nature of the approach described in this paper.

The ability "to look at a DNA sequence" and immediately being able to visually identify regions of potential interest is, in our opinion, a valuable tool for the biologist. The work that we present in this paper is an important step in that direction.

Author Contributions

Conceived and designed the experiments: AJP SPG DP PJSGF. Performed the experiments: SPG DP. Analyzed the data: AJP SPG. Wrote the paper: AJP SPG DP PJSGF.

References

1. Wong PC, Shen HW, Jonhson CR, Chen C, Ross RB (2012) The top 10 challenges in extreme-scale visual analytics. IEEE Computer Graphics and Applications 32: 63–67.
2. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. Nucleic Acids Research 18: 6097–6100.
3. Jeffrey HJ (1990) Chaos game representation of gene structure. Nucleic Acids Research 18: 2163–2170.
4. Jensen LJ, Friis C, Ussery DW (1999) Three views of microbial genomes. Research in Microbiology 150: 773–777.
5. Pedersen AG, Jensen LJ, Brunak S, Staerfeldt HH, Ussery DW (2000) A DNA structural atlas for Escherichia coli. Journal of Molecular Biology 299: 907–930.
6. Goldman N (1993) Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences. Nucleic Acids Research 21: 2487–2491.
7. Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B (1999) Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. Molecular Biology and Evolution 16: 1391–1399.
8. Fertil B, Massin M, Lespinats S, Devic C, Dumee P, et al. (2005) GENSTYLE: exploration and analysis of DNA sequences with genomic signature. Nucleic Acids Research 33: W512–W515.
9. Oliver JL, Bernaola-Galván P, Guerrero-García J, Román-Roldán R (1993) Entropic profiles of DNA sequences through chaos-game-derived images. Journal of Theoretical Biology 160: 457–470.
10. Vinga S, Almeida JS (2007) Local Renyi entropic profiles of DNA sequences. BMC Bioinformatics 8.
11. Crochemore M, Vérin R (1999) Zones of low entropy in genomic sequences. Computers & Chemistry : 275–282.
12. Troyanskaya OG, Arbell O, Koren Y, Landau GM, Bolshoy A (2002) Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity. Bioinformatics 18: 679–688.
13. Clift B, Haussler D, McConnell R, Schneider TD, Stormo GD (1986) Sequence landscapes. Nucleic Acids Research 14: 141–158.
14. Allison L, Stern L, Edgoose T, Dix TI (2000) Sequence complexity for biological sequence analysis. Computers & Chemistry 24: 43–55.
15. Stern L, Allison L, Coppel RL, Dix TI (2001) Discovering patterns in Plasmodium falciparum genomic DNA. Molecular & Biochemical Parasitology 118: 174–186.
16. Cao MD, Dix TI, Allison L, Mears C (2007) A simple statistical algorithm for biological sequence compression. In: Proc. of the Data Compression Conf., DCC-2007. Snowbird, Utah, pp. 43–52.
17. Dix TI, Powell DR, Allison L, Bernal J, Jaeger S, et al. (2007) Comparative analysis of long DNA sequences by per element information content using different contexts. BMC Bioinformatics 8: S10.
18. Gusev VD, Nemytikova LA, Chuzhanova NA (1999) On the complexity measures of genetic sequences. Bioinformatics 15: 994–999.
19. Nan F, Adjeroh D (2004) On the complexity measures for biological sequences. In: Proc. of the IEEE Computational Systems Bioinformatics Conference, CSB-2004. Stanford, CA.
20. Pirhaji L, Kargar M, Sheari A, Poormohammadi H, Sadeghi M, et al. (2008) The performances of the chi-square test and complexity measures for signal

recognition in biological sequences. Journal of Theoretical Biology 251: 380–387.

21. Solomonoff RJ (1964) A formal theory of inductive inference. Part I. Information and Control 7: 1–22.
22. Solomonoff RJ (1964) A formal theory of inductive inference. Part II. Information and Control 7: 224–254.
23. Kolmogorov AN (1965) Three approaches to the quantitative definition of information. Problems of Information Transmission 1: 1–7.
24. Chaitin GJ (1966) On the length of programs for computing finite binary sequences. Journal of the ACM 13: 547–569.
25. Wallace CS, Boulton DM (1968) An information measure for classification. The Computer Journal 11: 185–194.
26. Rissanen J (1978) Modeling by shortest data description. Automatica 14: 465–471.
27. Lempel A, Ziv J (1976) On the complexity of finite sequences. IEEE Trans on Information Theory 22: 75–81.
28. Gordon G (2003) Multi-dimensional linguistic complexity. Journal of Biomolecular Structure & Dynamics 20: 747–750.
29. Chen X, Kwong S, Li M (1999) A compression algorithm for DNA sequences and its applications in genome comparison. In: Asai K, Miyano S, Takagi T, editors, Genome Informatics 1999: Proc. of the 10th Workshop. Tokyo, Japan, pp. 51–61.
30. Pinho AJ, Ferreira PJSG, Neves AJR, Bastos CAC (2011) On the representability of complete genomes by multiple competing finite-context (Markov) models. PLoS ONE 6: e21588.
31. Grumbach S, Tahi F (1993) Compression of DNA sequences. In: Proc. of the Data Compression Conf., DCC-93. Snowbird, Utah, pp. 340–350.
32. Rivals E, Delgrange O, Delahaye JP, Dauchet M, Delorme MO, et al. (1997) Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. Computer Applications in the Biosciences 13: 131–136.
33. Pinho AJ, Neves AJR, Bastos CAC, Ferreira PJSG (2009) DNA coding using finite-context mod- els and arithmetic coding. In: Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP-2009. Taipei, Taiwan, pp. 1693–1696.
34. Pinho AJ, Pratas D, Ferreira PJSG (2011) Bacteria DNA sequence compression using a mixture of finite-context models. In: Proc. of the IEEE Workshop on Statistical Signal Processing. Nice, France.
35. Wood V, Gwilliam R, Rajandream MA, Lyne M, Stewart A, et al. (2002) The genome sequence of Schizosaccharomyces pombe. Nature 415: 871–80.
36. The International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
37. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304–1351.
38. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biology 5: 2113–2144.
39. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–66.
40. Ahn SM, Kim TH, Lee S, Kim D, Ghang H, et al. (2009) The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. Genome Research 19: 1622–1629.
41. Rivals E, Delahaye JP, Dauchet M, Delgrange O (1996) A guaranteed compression scheme for repetitive DNA sequences. In: Proc. of the Data Compression Conf., DCC-96. Snowbird, Utah, p. 453.
42. Chen X, Kwong S, Li M (2001) A compression algorithm for DNA sequences. IEEE Engineer431 ing in Medicine and Biology Magazine 20: 61–66.
43. Matsumoto T, Sadakane K, Imai H (2000) Biological sequence compression algorithms. In: Dunker AK, Konagaya A, Miyano S, Takagi T, editors. Genome Informatics 2000: Proc. of the 11th Workshop. Tokyo, Japan, pp. 43–52.
44. Manzini G, Rastero M (2004) A simple and fast DNA compressor. Software–Practice and Experience 34: 1397–1411.
45. Korodi G, Tabus I (2005) An efficient normalized maximum likelihood algorithm for DNA sequence compression. ACM Trans on Information Systems 23: 3–34.
46. Behzadi B, Le Fessant F (2005) DNA compression challenge revisited. In: Combinatorial Pattern Matching: Proc. of CPM-2005. Jeju Island, Korea: Springer-Verlag, volume 3537 of *LNCS*, pp. 190–200.
47. Pinho AJ, Neves AJR, Afreixo V, Bastos CAC, Ferreira PJSG (2006) A three-state model for DNA protein-coding regions. IEEE Trans on Biomedical Engineering 53: 2148–2155.
48. Korodi G, Tabus I (2007) Normalized maximum likelihood model of order-1 for the compression of DNA sequences. In: Proc. of the Data Compression Conf., DCC-2007. Snowbird, Utah, pp. 33–42.
49. Pinho AJ, Neves AJR, Ferreira PJSG (2008) Inverted-repeats-aware finite-context models for DNA coding. In: Proc. of the 16th European Signal Processing Conf., EUSIPCO-2008. Lausanne, Switzerland.
50. Wood V (2006) *Schizosaccharomyces pombe* comparative genomics; from sequence to systems. In: Sunnerhagen P, Piskur J, editors. Comparative Genomics, Springer-Verlag, volume 15 of *Topics in Current Genetics*. pp. 233–285.
51. Sugawara N (1989) DNA sequences at the telomeres of the fission yeast *S. pombe*. Ph.D. thesis, Harvard University, Cambridge, MA.
52. Rhind N, Chen Z, Yassour M, Thompson DA, Haas BJ, et al. (2011) Comparative functional genomics of the fission yeasts. Science 332: 930–936.
53. Hiraoka Y, Chikashige Y (2004) Telomere organization and nuclear movements. In: Egel R, editor, The molecular biology of *Schizosaccharomyces pombe*, Springer-Verlag, Genetics, Genomics and Beyond, chapter 12. pp. 191–205.
54. Levin HL, Weaver DC, Boeke JD (1990) Two related families of retro-transposons from *Schizosac-charomyces pombe*. Molecular and Cellular Biology 10: 6791–6798.
55. Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, et al. (1999) The DNA sequence of human chromosome 22. Nature 402: 489–95.