A markovian approach for the segmentation of chimpanzee genome.

Christelle Melodelima and Christian Gautier

UMR 5558 CNRS Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 43 boulevard du 11 Novembre 1818, 69622 Villeurbanne Cedex - France and PRABI (Rhône Alpes Bioinformatics Center).
{melo,cgautier}@biomserv.univ-lyon1.fr

Abstract. Hidden Markov models (HMMs) are effective tools to detect series of statistically homogeneous structures, but they are not well suited to analyse complex structures such as DNA sequences. Numerous methodological difficulties are encountered when using HMMs to model non geometric distribution such as exons length, to segregate genes from transposons or retroviruses, or to determine the isochore classes of genes. The aim of this paper is to suggest new tools for the exploration of genome data. We show that HMMs can be used to analyse complex gene structures with bell-shaped length distribution by introducing macrostates. Our HMMs methods take into account many biological properties and were developped to model the isochore organisation of the chimpanzee genome which is considered as a fondamental level of genome organisation. A clear isochore structure in the chimpanzee genome, correlated with the gene density and guanine-cytosine content, has been identified.

Key words: Hidden Markov model, DNA sequence, isochore modelling

1 Introduction

The chimpanzee is an excellent model organism in biomedical research due to the similarities between many of its physiological processes and those of human. The availability of the chimpanzee genome sequence has already largely influenced the research in many fields, and more profound impact is certainly to follow. The sequencing of the complete chimpanzee genome led to the knowledge of a sequence of 4.4 billion pairs of nucleotides. Such amounts of data make it impossible to analyse patterns or to provide a biological interpretation analysis unless one relies on automatic data-processing methods. For twenty years, mathematical and computational models have been widely developed in this setting. Numerous methodological efforts have been devoted to multicellular eukaryotes since a large proportion of their genome has no known function. For example, only 1 to 3% of the chimpanzee genome is known to code for proteins. Another difficulty is that the statistical characteristics of the coding region vary dramatically from one specie to the other, and even from one region in a given genome

to the other. For example, vertebrate isochores [1], [2] exhibit such a variability in relation to their guanine-cytosine ($G + C$) frequencies. Thus it is necessary to use different models for different regions if one seeks to detect patterns in genomes.

One way of modelling genomes uses hidden Markov Models (HMMs) [3], [4], [5]. To each type of genomic region (exons, introns, etc.), one associates a state of the hidden process, and the distribution of the stay in a given state, that is, of the length of a region, is geometric. While this is indeed an acceptable constraint as far as intergenic regions and introns are concerned, the empirical distributions of the lengths of exons are clearly bell-shaped [6], [7], [8], hence they cannot be represented by geometrical distributions. Semi-Markov models are one option to overcome this problem [6]. Although these models are widely used, they are very versatile, since they allow to adjust the distribution of the duration of the stay in a given state directly to the empirical distribution. The trade off is a strong increase in the complexity of most algorithms implied by the estimation and the use of these models. For example, the complexities of the main algorithms (forward-backward and Viterbi) are quadratic in the worst case with respect to the length of the sequence for hidden semi-Markov chains and linear for HMMs [6], [9], [10]. This may limit their range of application as far as the analysis of sequences with long homogeneous regions is concerned. Another difficulty is the multiplication of the number of parameters that are needed to describe the empirical distributions of the durations of the states, and which must be estimated, in addition to usual HMM parameters [9]. Thus the estimation problem is more difficult for these variable duration HMMs than for standard HMMs [9]. In other words, semi-Markov models are efficient tools to detect protein coding genes, but they are much more complex than HMMs.

In this paper, HMMs were used to detect isochores which were originally identified as a result of gradient density analysis of fragmented genomes [11]. Mammalian genomes are a mosaic of regions (DNA segments on average more than 300 kb in length) with differing, homogeneous $G + C$ contents. High, Medium and Low-density genomic segments are known as H, M and L isochores in order of decreasing $G + C$ content respectively. The isochore has been classified as a "fundamental level of genome organisation" [12] and this concept has increased our appreciation of the complexity and variability of the composition of eukaryotic genomes [13]. Existing isochore prediction methods only use the overall base composition of the DNA sequence ([14], [15], [16], [17], [18]). The aim of this paper is to suggests a new approach using HMMs and allowing to take into account many biological properties, such as $G + C$ content, gene density, length of the different regions, the reading frame of exons. We suggest to use HMM for modelling the exon length distribution by sum of geometric laws. To do this a state representing a region is replaced by a juxtaposition of states with the same emission probabilities. This juxtaposition of states is called macro-states. Macro-state HMMs models were used for complete genome analysis. Therefore, a method based on a hidden Markov model, which makes it possible to detect the isochore structure has been developped and tested on the chimpanzee genome.

2 Materials

Gene sequences were extracted from Ensembl for the chimpanzee genome. This procedure yielded a set of 22524 genes. The statistical characteristics of the coding and noncoding regions of vertebrates differ dramatically between the different isochore classes [13]. Many important biological properties have been associated with the isochore structure of genomes. In particular, the density of genes has been shown to be higher in H than in L isochores [20]. Genes in H isochores are more compact, with a smaller proportion of intronic sequences, and they code for shorter proteins than the genes in L isochores [16]. The amino-acid content of proteins is also constrained by the isochore class: amino acids encoded by $G+C$ rich codons (alanine, arginine.) being more frequent in H isochores [21] and [22]. Moreover, the insertion process of repeated elements depends on the isochore regions. SINE (short-interspersed nuclear element) sequences, and particularly Alu sequences, tend to be found in H isochores, whereas LINE (long-interspersed nuclear element) sequences are preferentially found in L isochores [23]. Thus, we took into account the isochore organisation of the chimpanzee genome. Three classes were defined and based on the $G + C$ frequencies at the third codon position ($G + C_3$). The limits were set so that the three classes contained approximately the same number of genes. This yielded classes H=[100%, 70%], M=]58%,70%[and L=[0%,58%], which were used to build a training set. These classes were the same compared with those used by other authors [20], [24] in the human genome. Each class H, L and M, was randomly divided into two equal parts, a training set and a test set. The training sets were used to model the length distributions of the exons and the introns, and to analyse the structure of genes. To test the model, data on all chimpanzee chromosomes were retrieved from ENSEMBL.

3 Method

3.1 Estimation of the HMM parameters

Estimation of emission probabilities

The DNA sequence consists of a succession of different regions, such as gene and intergenic regions. A gene is a succession of coding (exon) and non-coding (intron) region. In this study, HMMs are used to discriminate between these different types of regions. Exons consist of a succession of codons, and each of the three possible positions in a codon (0, 1, 2) has specific statistical properties. Thus, exons were divided into three states [25], [26].
HMMs take into account the dependency between a base and its n preceding neighbours (n defined the order of the model). For our study, n was taken to be equal to 5, as in the studies of Borodovsky [24] and Burge [25]. The emission probabilities of the HMM were therefore estimated from the frequencies of 6-letter words in the different regions (intron, initial exon, internal exons and

terminal exon) that made up the training set.

Estimation of the structure of the macro-states

We suggest to use sums of a variable number of geometric laws with equal or different parameters in order to model the bell-shaped empirical length distributions of the exons. Thus a "biological state" is represented by a HMM and not by a single Markov state. The emission of probabilities of every state in this HMM are the same. A key property of this macro-state approach is that the conditional independence assumptions within the process are preserved with respect to HMMs. Hence, the HMM algorithms used to estimate the parameters and compute the most likely state sequences still apply [10].
The length distribution of the exons and introns was estimated from the training set (data set sequences are named $x_1...x_n$). Each x_i was considered to be the realization of an independent variable of a given law. We have tested the following laws:

1. the sum of m geometric laws of same parameter Θ (i.e. a binomial negative law):
$$P[X = k] = C_{k-1}^{m-1} \times \Theta^m \times (1 - \Theta)^{k-m} \ , \tag{1}$$
2. the sum of two geometric laws with different parameters $\Theta_1 > \Theta_2$:
$$P[X = k] = \Theta_1 \times \Theta_2 \frac{(1 - \Theta_2)^{k-1} - (1 - \Theta_1)^{k-1}}{\Theta_1 - \Theta_2,} \ , \tag{2}$$
3. the sum of three geometric laws with different parameters $\Theta_1 < \Theta_2 < \Theta_3$:
$$P[X = k] = \frac{\Theta_1 \times \Theta_2 \times \Theta_3}{\Theta_2 - \Theta_3} \times \left\{ \frac{(1 - \Theta_1)^{k-1} - (1 - \Theta_3)^{k-1}}{\Theta_3 - \Theta_1} - \frac{(1 - \Theta_2)^{k-1} - (1 - \Theta_3)^{k-1}}{\Theta_3 - \Theta_2} \right\} . \tag{3}$$

We define $G_n(D_1, ..., D_n)$ as the distribution of the sum of n random variables of geometric distributions, each with expectation D_i and parameter $\Theta_i = 1/D_i$. Thus the expectation of $G_n(D_1, ..., D_n)$ is $D_1 + ... + D_n$. When $D_i = D$ for every i, this is called a negative binomial distribution with parameters $(n, 1/D)$, which we denote $G_n(1/\Theta)$. Finally $G_n(D)$ is a geometric distribution with expectation D and parameter $\Theta = 1/D$, which we write $G(D)$.
To estimate the parameters of the different laws, we minimised the Kolmogorov Smirnov distance for each law. The law which fits best with the empirical distribution is the law with the smallest Kolmogorov-Smirnov distance.

$$D_{KS} = sup_x |F(x) - G(x)| \ , \tag{4}$$

where D_{KS} is the Kolmogrorov-Smirnov distance, F is the theorical density distribution, G is the empirical density distribution. However, the classical Newton

or gradient algorithm cannot minimise the Kolmogorov-Smirnov distance, since this distance cannot be differentiable. For this reason, we have discretised the parameter space with a step of 10^{-5}. Parameters estimations were not based on the maximum likelihood, which would have matched the end of the exon length distribution while neglecting many small exons (Figure 1a). The definition of the maximum likelihood method is as follow: let x be a discrete variable with probability $P[x|\Theta_1...\Theta_k]$ (where $\Theta_1...\Theta_k$ are k unknown constant parameters which need to be estimated) obtained by an experiment which result in N independant observations, $x_1, ..., x_N$. Then the likelihood fonction is given by:

$$L(x_1, ..., x_N|\Theta_1...\Theta_k) = \prod_{i=1...N} P[x_i|\Theta_1...\Theta_k] \ . \tag{5}$$

The logarithm function is:

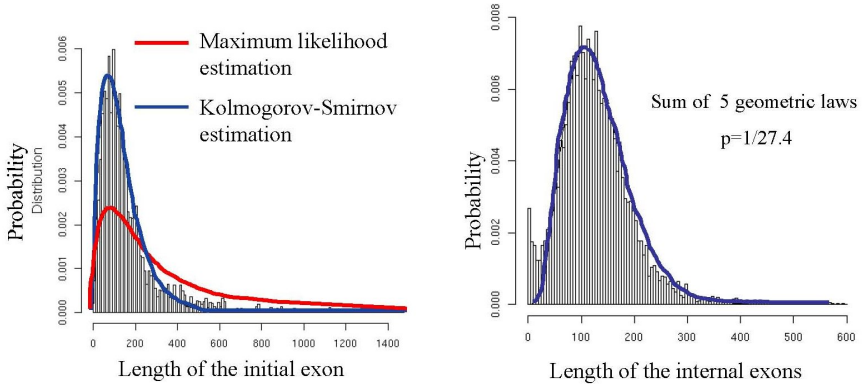
$$\wedge = ln(L(x_1, ..., x_N|\Theta_1...\Theta_k)) = \sum_{i=1...N} P[x_i|\Theta_1...\Theta_k] \ . \tag{6}$$

The maximum likelihood estimators $\Theta_1...\Theta_k$ are obtained by maximizing L or $\wedge$. Indeed, for a geometrical law or a convolution of geometrical laws, the parameter Θ is estimated by the reverse of the mean ($E[X] = 1/\Theta$) using the maximum likelihood method. The extreme values thus tend to stretch the distribution towards the large ones. We therefore have preferred to use the Kolmogorov-Smirnov distance in order to obtain a better modelling of the chimpanzee gene. Moreover, in order to provide simple but efficient models, equal transitions between states of a macro-state were used when it was possible.
Thus, a region is represented by a hidden state of the HMM. If the length distribution of a region is fitted by a sum of geometric laws, the state representing the region is replaced by a juxtaposition of states with the same emission probabilities, thus leading to macros-states (Figure 2). The state duration is characterised by the parameters of the sum of these geometric laws. Various studies [6], [27] have shown that the length distribution of the exons depend on their position in the gene. All exon types were taken into account: initial coding exons, internal exons, terminal exons and single-exon genes.

3.2 Modelling of isochores organisation

To characterize the three isochore regions (H, L and M) along the chimpanzee genome, three HMM models (H, L and M) were adjusted using the training sets, and then compared on all chimpanzee chromosomes. We divided the DNA of each chimpanzee chromosome into window of 100-kb. Two successive window overlapped by half their length. For each window and for each model (H, L and M), the probability $P[m|S]$ was computed as follows:

$$P(m \mid S) = \frac{P(S \mid m)P(m)}{\sum_{m' \in \{H,M,L\}} P(S \mid m')P(m')} \ , \tag{7}$$



(a) (b)

Fig. 1. (a) The histogram shows the empirical distribution of the length of the initial exons in a multi-exons gene. The blue curve shows the theoretical distribution obtained from the Kolmogorov-Smirnov distance. The red curve characterises the binomial distribution, obtained by the maximum likelihood method. (b) The histogram shows the empirical distribution of the length of the internal exons. The blue curve shows the theoretical distribution obtained from the Kolmogorov-Smirnov distance.

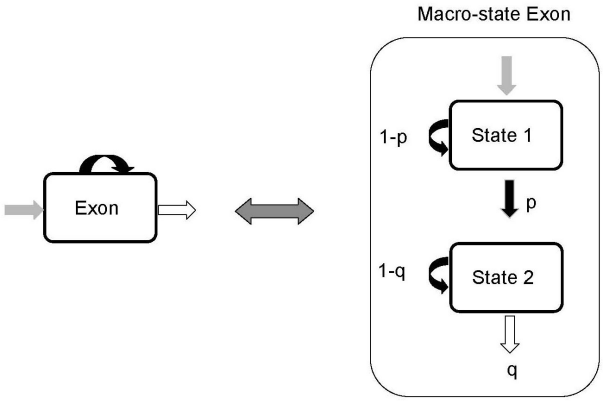


Fig. 2. The macro-state initial exon is composed of two smaller macro-states modelling the distribution of the length $G_2(1/p, 1/q)$ of initial exons in H isochore. Black arrows show the transition inside the macro-state. Grey and white arrows show respectively the different input and output of the macro-state.

m is H, L or M, and S the window that is being tested, $P(S|m)$ is computed by the forward algorithm, and $P(m)$ is estimated by the frequency of genes according to their $G + C_3$ content (due to our definition of $G + C_3$ limits we get $P(H) \approx P(M) \approx P(L) \approx 1/3$, and so our Bayesian approach is numerically very close to a maximum likelihood approach). The computations were realized with the package SARMENT [28]. To be consistent with the preceding definition, we assumed an isochore to be a region consisting of at least 5 consecutive windows associated by the method with the same isochore class. This ensured that all estimated isochore lengths were greater than 300 kb, but meant that some windows can be unassociated to an isochore class.
Several tests were performed in order to check the coherence of isochore prediction: (i) the distribution of isochores was plotted with the distribution of the gene density, and the GC content along the chromosome, (ii) the segmentation allows to define the isochore class of each window along the chimpanzee genome. The isochore repartition of these windows has been compared with a random repartition of these windows. One thousand simulations have been realized. (iii) Furthermore, the ratio of coding regions has been compared between the isochores H and L predicted by our method.

4 Results

4.1 Estimations of the HMMs parameters

Sums of geometric laws with equal or different parameters were used in order to model the bell-shaped empirical length distributions of exons (Figure 2). The length of an exon depends on its position within the gene. Initial and terminal exons tend to be longer than internal exons (Table 1). The length of introns displays also a noticeable positional variability. The distributions of the lengths of internal and terminal introns are relatively similar. However, internal and terminal introns are both smaller compared with initial introns (Table 1). The lengths of introns depend on their $G+C$ content. Table 1 shows that the $G+C$ frequency at the third codon position is negatively correlated with the length of the introns, i.e., high frequencies correspond to short introns, and vice versa. The length of the exons displays clearly a bell-shaped pattern (Figure 1b), for the three $G+C$ classes. The minimisation of the Kolmogorov-Smirnov distance yields a good fit with the empirical distribution of the length of the exons (Figure 1 and Table 2). Therefore, the Kolmogorov-Smirnov distance was chosen to model their length distribution by sum of geometric laws and to estimate the parameters of these laws (see Method for a comparison with the maximum likelihood approach).
We show here only the results for the modelling of the distributions of the lengths in the H class. However, the distributions of the lengths in the classes M and L were modelled by sums of geometric laws. The estimated distributions are $G_2(52.6, 106.4)$ for initial exons (Figure 1), $G_2(58.8, 108.7)$ for terminal exons, $G_5(27.4)$ for internal exons, $G_3(415.2)$ for intronless genes. The geometric dis-

Table 1. Length of the exons and of the introns according to their position in the gene and according to the $G+C$ frequency at third codon position in the gene.

Position in the gene	Length (bp) in class H		Length (bp) in class M		Length (bp) in class L	
	mean	median	mean	median	mean	meadian
initial coding exon	184	123	167	114	162	112
internal exon	140	101	138	108	139	107
terminal exon	193	138	201	130	202	127
initial intron	2746	2318	3864	3274	4474	4227
internal intron	1192	1045	1446	1437	2841	2322
terminal intron	1247	1012	1534	1479	2691	2136

Table 2. Parameters estimation of different laws obtained for initial exons of class H minimising the Kolmogorov-Smirnov distance.

Lois	Paramètres p	Distance K-S
$G_2(\Theta)$	0.0126	0.05761
$G_3(\Theta)$	0.0197	0.08782
$G_4(\Theta)$	0.0226	0.11243
$G(\Theta_1, \Theta_2)$	0.019 - 0.0094	0.05023

tribution for initial introns was $G(1923.1)$. Other types of introns were also modelled by a geometrical distribution.

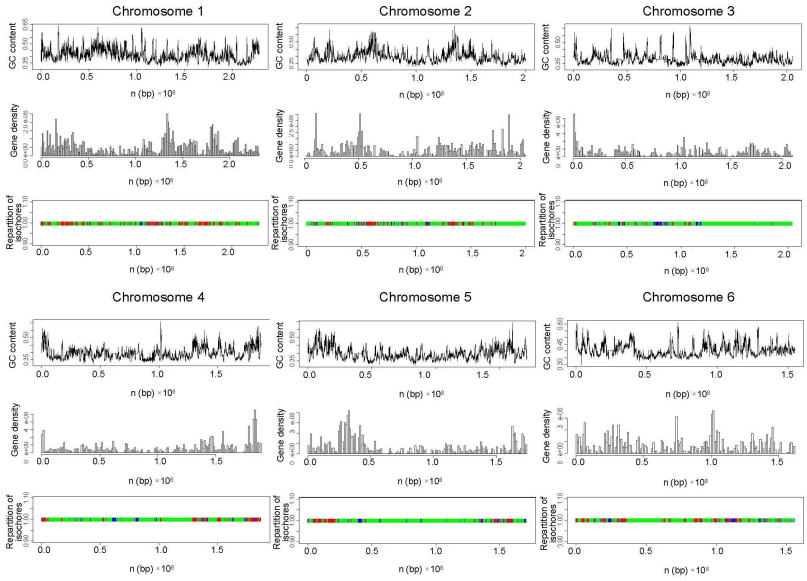
4.2 Modelling of isochore organisation

The quality of discrimination between isochore classes for each windows was measured by $max_{H,L,M}(P(m/S))$. For each of the 59075 windows in the chimpanzee genome the maximum value was greater than 0.75, leading to a very clear association between each window and a unique isochore class. A second important criterion was the isochore length, since the method imposes a minimum length of 300 kb, resulting in some unaffected windows. In the chimpanzee genome, unallocated windows represent only 4.75% of the total number of windows. These windows were not considered to constitute an isochore. Along the chimpanzee genome, the distributions of these unallocated windows was random. Figure 3 shows the chimpanzee genome segmentation obtained by the method described in this paper. Figure 3 is available online at `http://melodelima.chez-alice.fr/chimpanzee_isochores/chimpanzee_isochore.html`.
All the tests performed to verufy our predictions are isochores, were satisfactory. Along the chimpanzee genomes the isochore repartition of windows obtained has been compared with 1000 random repartitions of the same windows. A significant difference between our predictions and random repartitions was observed (respectively the p-value of the χ^2 test were equal to 5.10^{-8}). Furthermore, the percentage of coding region in each isochore class was coherent with the observation obtained along mammals genomes ([20]). The coding regions represent

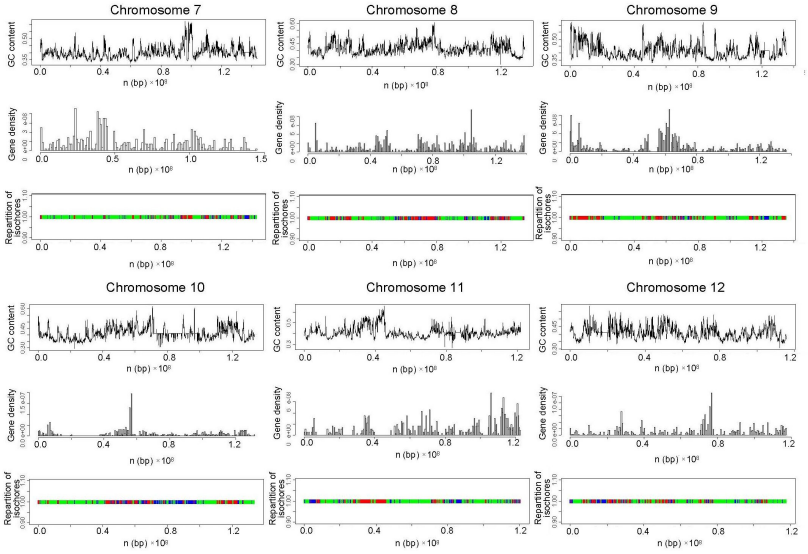
4.2% of the isochores H and only 1.1% of the isochores L. The p-value of the Wilcoxon test was significant (p= 1.10^{-4}).
The segmentation of the chimpanzee genome correlates well with the $G + C$ content. The mean $G + C$ contents were 0.48 ($\sigma = 0.03$), 0.42 ($\sigma = 0.02$) and 0.38 ($\sigma = 0.02$) respectively for the H, M and L isochore classes as defined by our HMM method. The Kruskal-Wallis non-parametric test was significant (p-value $< 10^{-5}$). The length of the isochores detected by their $G + C$ content is known to depend on the isochore class, with $L > M > H$ [29]. This is what we found here. Although the minimum length of isochores that can be predicted by our method was 300 kb, we were also able to predict much longer isochores. The average length for L isochores was 7.2 Mb, whereas the average length for the H and M isochores was 2 Mb and 4.4 Mb respectively. These lengths were significantly different (Kruskal-Wallis p-value $< 10^{-12}$). Figure 3 shows the relationships between isochore class and gene density. For all chromosomes, the isochore structure is correlated with the gene density distribution along the chromosome. The gene density in the H isochores (8.2 genes per Mb) was higher than the gene density in the L isochores (4.3 genes per Mb), leading to a significant Wilcoxon test (p-value = 2.10^{-5}). The same difference was observed when we compared the characteristics of the M isochores (5.6 genes per Mb) with those of the H (p-value = 4.10^{-3}) and L isochores (p-value = 4.10^{-2}).

5 Discussion

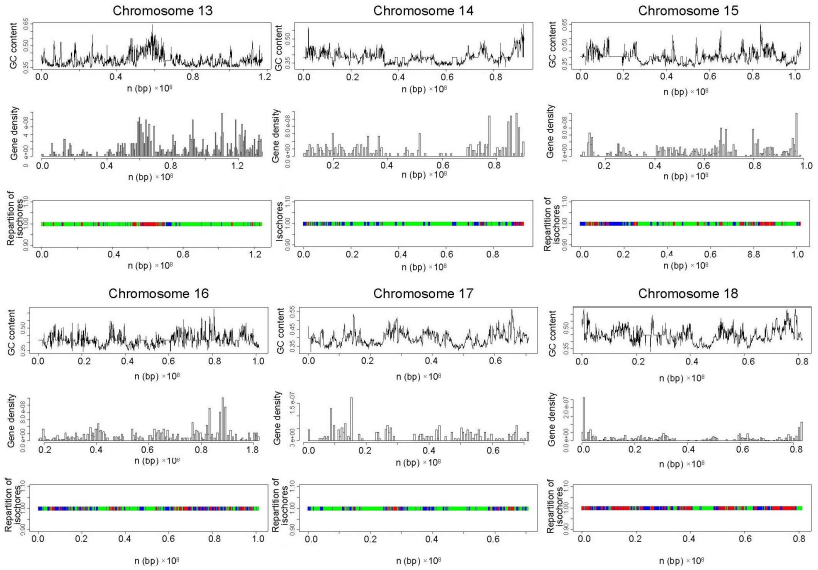
Last year, a large number of genomes were sequenced. This huge amount of data makes it impossible to analyse patterns in order to provide a biological interpretation "by hand". Therefore, mathematical and computational methods have to be used. Our approach, using HMMs, is a very promising method for analysing the organisation of genomes. Our study shows that hidden Markov models could be used to analyse genome organisation. This study was conducted on the chimpanzee genome but our method can be adapted to other eukaryote genomes. To model the bell-shapped length distribution of the exons, we have used sums of a variable number of geometric laws with equal or different parameters. Each region is represented by a macro-state in the HMM. A key property of this macro-state approach is that the conditional independence assumptions within the process are preserved with respect to HMMs. Moreover, we have preferred to use the Kolmogorov-Smirnov distance in order to obtain a better modelling of the chimpanzee genes.
The chimpanzee genomes consists of many nested structures (chromosomes, isochores, genes, exons/introns, reading frame). For the analyses of the isochore organisation of genomes, we have proposed a new method based on HMM, taking into account genes as a local structure. The different approaches already developed for isochore prediction ([14], [15], [16], [17], [18]) use only the overall base composition of the DNA sequence to predict isochores. However, the statistical characteristics of the $G + C$ content differ in the coding and non-coding regions of vertebrate genes. To improve the isochore prediction capacity, we have


Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
GC content
Gene density
Repartition of isochores
n (bp) ×10⁸

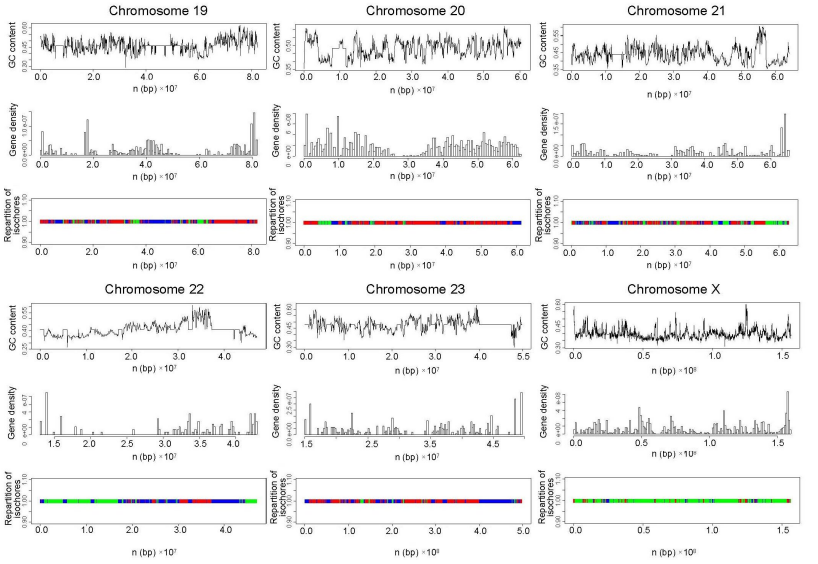
(a)


Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
GC content
Gene density
Repartition of isochores
n (bp) ×10⁸

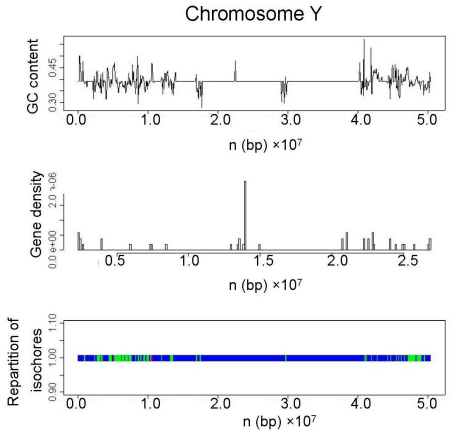
(b)


Chromosome 13
Chromosome 14
Chromosome 15
Chromosome 16
Chromosome 17
Chromosome 18
GC content
Gene density
Repartition of isochores
n (bp) ×10⁸
n (bp) ×10⁷

(c)


Chromosome 19
Chromosome 20
Chromosome 21
Chromosome 22
Chromosome 23
Chromosome X
GC content
Gene density
Repartition of isochores
n (bp) ×10⁷
n (bp) ×10⁸
n (bp) ×10⁹

(d)



(e)

Fig. 3. Distribution of isochores along chimpanzee chromosomes obtained by our method. The detected H, L and M isochores appear respectively in red, green and blue. To check the coherence of isochore prediction, each figure is shown with the distribution of the gene density, and the $G + C$ content along the chromosome. (a) Chromosomes 1 to 6, (b) chromosomes 7 to 12, (c) chromosomes 13 to 18, (d) chromosomes 19 to X, (e) chromosome Y.

introduced the idea of using an HMM that takes into account not only the $G+C$ content of the DNA sequence, but also several biological properties associated with the isochore structure of the genome (such as gene density, differences in the $G+C$ content of different regions of the gene, lengths of exons and introns). Therefore, three HMMs were adjusted to each isochore class in order to take into account biological properties associated with H, L and M isochores. In our case, this supplementary information allowed us to determine the precise boundary of the isochores and the structure of a region may be easily analyzed. The segmentation in this paper are linked to an isochore structure of the chimpanzee genome. There was a significant difference between the isochore repartition in our prediction windows and a random repartition of these windows. Furthermore, there was more coding region in isochore H compared with isochores L in the two fishes. Thus, our method has clearly confirmed the existence of an isochore structure in the chimpanzee genome.

In conclusion, The statistical characteristics of the coding and noncoding regions of vertebrates differ dramatically between the different isochore classes [2]. The clarification of the isochore structure is a key to understand the organisation and biological function of the chimpanzee genome and we show here that hidden Markov models were appropriate for each isochore class. One advantage of the model presented in this paper is that the number of basic states for each isochore class (without taking the frame and coding strand into account) in our model is only 7: first, internal and terminal introns and exons and "intergenic regions". Intergenic region were used to model all non-coding regions of the genome and

the introns inserted between two coding exons. This small number of states has made it possible to conduct a complete chimpanzee genome analysis. This method could be easily adapted to other genomes and could be used to study the evolution of isochores among the vertebrate genomes. The comparative genomic analysis have a key role to push our knowledge further in the comprehension of the structure and function of human genes, to study evolutionary changes among organisms and help to identify the genes that are conserved among species.

acknowledgements We thanks Marie-France Sagot for assistance in preparing and reviewing the manuscript.

References

1. Thiery, J.P., Macaya, G., Bernardi, G.: An analysis of eukaryotic genomes by density gradient centrifugation. J. Mol. Biol., Vol. 108(1). (1976) 219-35
2. Bernardi, G.: Isochores and the evolutionary genomics of vertebrates. review. Gene, Vol. 241(1). (2000) 3–17
3. Krogh, A.: Two methods for improving performance of an HMM and their application for gene-finding. In Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology (1997) 179–186
4. Henderson, J., Salzberg, S., Fasman, K.H.: Finding genes in DNA with a hidden Markov model. Journal of Computational Biology, Vol. 4. (1997) 127–141
5. Lukashin, V.A., Borodovsky, M.: Gene-Mark.hmm: new solutions for gene finding. Nucleic Acids Research, Vol. 26. (1998) 1107–1115
6. Burge, C., Karlin, S.: Prediction of complete gene structure in human genomic DNA. Journal of Molecular Biology, Vol. 268.(1997) 78–94
7. Berget, S.M.: Exon recognition in vertebrate splicing. The Journal of Biological Chemistry, Vol. 270(6). (1995) 2411–414
8. Hawkins, J.D.: A survey on intron and exon lengths. Nucleic Acids Research **16**, (1998) 9893–9908
9. Rabiner, L.: A tutorial on hidden Markov models and selected applications in speech recognition. Poceeding of the IEEE, Vol. 77(2). (1989) 257–286
10. Guédon Y.: Estimating hidden semi-Markov chains from discrete sequences. Journal of Computational and Graphical Statistics, Vol. 12(3). (2003) 604–639
11. Macaya, G., Thiery, J.P., Bernardi, G.: An approach to the organization of eukaryotic genomes at a macromolecular level. J. Mol. Biol., Vol. 108(1). 237–54 (1976)
12. Eyre-Walker, A., Hurst, L.D.: The evolution of isochores. Nat. Rev. Genet., Vol. 2(7). (2001) 549–555 Review
13. Nekrutenko, A., Li, W.H.: Assessment of compositional heterogeneity within and between eukaryotic genomes. Genome Res., Vol. 10(12). (2000) 1986–1995
14. Bernaola-Galvan, P., Carpena, P., Roman-Roldon, R., Oliver, J.L.: Mapping isochores by entropic segmentation of long genome sequences. In: Sankoff D, Lengauer T, RECOMB Proceedings of the Fifth Annual International Conference on Computational Biology. (2001) 217–218
15. Li, W., Bernaola-Galvan, P., Carpena, P., Oliver, J.L.: Isochores merit the prefix 'iso'. Comput. Biol. Chem., Vol. 27(1). (2003) 5–10
16. Oliver, J.L., Carpena, P., Roman-Roldan, R., Mata-Balaguer, T., Mejias-Romero, A., Hackenberg, M., Bernaola-Galvan, P.: Isochore chromosome maps of the human genome. Gene, Vol. 300(1-2). (2002) 117–27

17. Zhang, C.T., Zhang, R.: An isochore map of the human genome based on the Z curve method. Gene, Vol. 317(1-2). (2003) 127–35
18. Costantini, M., Clay, O., Auletta, F., Bernardi, G.: An isochore map of human chromosomes. Genome Research, Vo .16. (2006) 536–541
19. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F.: The mosaic genome of warm-blooded vertabrates. Science, Vol. 228(4702). (1985) 953–958
20. Mouchiroud, D., D'Onofrio, G., Aissani, B., Macaya, G., Gautier, C., Bernardi, G.: The distribution of genes in the human genome. Gene, Vol. 100. (1991)181–187
21. D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, B.: Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J. Mol. Evol., Vol. 32. (1991) 504–510
22. Clay, O., Caccio, S., Zoubak, S., Mouchiroud, D., Bernardi, G.: Human coding and non coding DNA: compositional correlations. Mol. Phyl. Evol. , Vol. 1. (1996) 2–12
23. Jabbari, K., Bernardi, G.: CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene, Vol. 224(1-2). (1998) 123–127
24. Zoubak, S., Clay, O., Bernardi, G.: The gene distribution of the human genome. Gene, Vol. 174(1). (1996) 95–102
25. Burge, C., Karlin, S.: Finding the genes in genomic DNA. Curr.Opin.Struc.Biol., Vol. 8. (1998) 346–354
26. Borodovsky, M., McIninch, J.: Recognition of genes in DNA sequences with ambiguities. Biosystems, Vol. 30(1-3). (1993) 161–171
27. Rogic, S., Mackworth, A.K., Ouellette, F.B.: Evaluation of Gene-Finding Programs on Mammalian Sequences. Genome Research, Vol. 11. (2001) 817–832
28. Guéguen, L.: Sarment: Python modules for HMM analysis and partitioning of sequences. Bioinformatics, Vol. 21(16). (2005) 3427–34278
29. De Sario, A., Geigl, E.M., Palmieri, G., D'Urso, M., Bernardi, G.: A compositional map of human chromosome band Xq28. Proc Natl Acad Sci U S A., Vol. 93(3). (1996) 1298–302