Prediction of human isochores using a hidden Markov model

Christelle Melodelima[1], Laurent Guéguen[1], Didier Piau[2] and Christian Gautier[1]

[1]UMR CNRS 5558 Biométrie et Biologie Evolutive, Université Claude Bernard, Lyon, France and
[2]UMR CNRS 5208, Université Claude-Bernard, Lyon, France

ABSTRACT

Mammalian genomes are organised into a mosaic of regions (in general longer than 300kb), having different fairly homogeneous *G+C* content. If the *G+C* content remains the basic characterising definition of isochores, the latter have also been associated with many other biological properties. For instance, genes are more compact and their density is highest in *G+C* rich isochores. Various approaches to locate isochores in the human genome were developed but such methods used only the base composition of the DNA sequences. The present paper proposes a new method, based on a hidden Markov model, which takes into account several biological properties associated with the isochore structure of a genome. By using this method, isochore structures were clearly defined and appear to be a basic organisation of the human genome. Since many important biological functions depend on the isochore structure, our model may provide numerous insights for understanding the human genome.

Contact: melo@biomserv.univ-lyon1.fr

Keywords: hidden Markov model, isochore, human genome.

1 INTRODUCTION

Isochores were originally identified as a result of a gradient density analysis of fragmented genomes (Macaya 1976): mammalian genomes are thus a mosaic of regions (DNA segments longer than 300 kb on average) having different homogeneous *G+C* content. Higher, Lower and Medium-density genomic segments are respectively called *H, L* and *M* isochores. The isochore concept has been considered a "fundamental level of genome organisation" (Eyre-Walker and Hurst 2001) and has increased our appreciation of the complexity and compositional variability of eukaryotic genomes (Nekrutenko and Li 2000). 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* isochores than in *L* ones (Mouchiroud 1991). Genes in *H* isochores are more compact with a smaller proportion of intronic sequences and code for shorter proteins than do genes in *L* isochores (Zouback 1996). 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 (Aota 1986, D'Onofrio 1991, Clay 1996). Moreover, the insertion process of repeated elements depends on the isochore regions. SINE (short-interspersed nuclear element) sequences, and particularly *Alu* sequences, are preferentially found in *H* isochores, while LINE (long-interspersed nuclear element) sequences are preferentially found in *L* isochores (Jabbari 1998).

The recent availability of the draft human genome sequence allowed for a direct test of the isochore model and it was hoped that isochores could be identified at the sequence level. Since then, the existence of isochores in the human genome has been the object of an active debate. Different approaches have been developed for isochore prediction. Sliding windows of arbitrary length and step over long, heterogeneous and correlated sequences may lead to misleading results (Li 2001). A *G+C*-plot thus routinely accompanies the publication of every new genome sequence, the long range patterns appearing on the plots being usually identified only by eye. This happens, for instance, with the isochores tentatively identified on the human chromosomes 21 (Hattori et al. 2000) and 22 (Dunham et al. 1999).

Other methods based on sliding windows use a random (uncorrelated) model to test sequence homogeneity (Nekrutenko and Li 2000). Häring and Kypr (2001) denied the existence of isochores in the human chromosomes 21 and 22 and Lander (2001) concluded that isochores do not appear to deserve the prefix "iso". The methodological problem with these works is precisely the random model adopted in which nucleotides are free to change. This leads to the conclusion that only highly repetitive DNA sequences are homogeneous. However, when the heterogeneity within isochore families was quantified (Cunny et al. 1981), it was shown that the homogeneity of isochores is only relative, hence their definition as "fairly homogeneous" regions (Bernardi 2001).

An alternative tool to analyse genome heterogeneity is compositional segmentation. Windowless methods have been developed to calculate the *G+C* content, and some have been used to identify isochores in various genomes. These methods have also been called "DNA segmentation methods" (Bernaola-Galvan et al. 2001). Among them, the method of entropic segmentation (Li et al. 2002, Oliver et al. 2004) and the Z-curve method (cumulative *G+C* profile) which leads to a unique representation of DNA sequences (Zhang et al. 2003). All these windowless methods conclude that the concept of homogeneity of *G+C* content is *relative* and that the isochore structure indeed exists in the human genome. However, these different methods use only the base composition of the DNA sequence to predict isochores.

Compositionally homogeneous segments of genomic DNA often correspond to meaningful biological units. Hidden Markov models with a small number of states are a natural model for the description of the compositional properties of chromosome-size DNA sequences. The first application of HMMs to the analysis of genetic data was done by Churchill (1989) and aimed at analysing the compositional heterogeneity of natural DNA sequences. More recently, Peshkin (1999) has shown that HMMs can be used for

further structural analysis or for direct biological interpretation. The objective of the present paper is therefore to propose a method, based on a hidden Markov model, which allows to detect and to analyse the isochore structure along the human genome in reasonable time. To improve the isochore prediction, we introduce 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, length of exons and introns according to the isochore class, *etc*).

2 MATERIEL

Gene sequences were extracted from Hovergen (Homologous Vertebrate Genes Database, March 2003 release 43) (Duret et al. 1994), and concern only the human genome. To ensure the data concerning the intron/exon organisation was correct, we restricted our analysis to genes of which the RNA transcripts have been sequenced. To avoid distortion of the statistical analysis, redundancy was discarded. This procedure yielded a set of 5034 multi-exon genes and 817 single-exon (that is, intronless) genes. Three classes were defined based upon the *G+C* frequencies at the third codon position ($G+C_3$). The limits were set so that all three classes contained approximately the same number of genes. This yielded the classes *H*=[100%, 72%], *M*=]56%,72%[and *L*=[0%,56%] which are roughly the same as used in other papers (Duret et al. 1995, Zouback et al 1996). Sets of sequences partitioned into the *H, L* and *M* classes (training set) are used to build three HMMs adapted to the organisation structure of each of the three isochore classes *H, L, M*. To test the model, the data concerning all human chromosomes are retrieved from ENSEMBL.

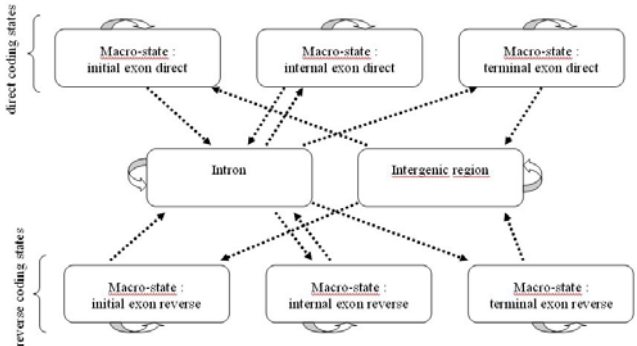
3 METHODS

To detect isochores and analyse their structure along the human genome, we propose a new method, based on a hidden Markov model. To characterize the three isochore regions (*H, L* and *M*), three HMM models were adjusted to each of the regions and compared. In an HMM model, the duration of stay in each state follows a geometric law. If the empirical length distributions of intergenic and intronic regions are geometric, this is not the case for exons (Burge et al. 1997, Berget 1995, Hawkins 1988) as shown by the bell-shaped histograms obtained. Thus, hidden Markov models cannot represent precisely the length distribution of exons. To model the empirically obtained bell-shaped length distributions of the exons, we use sums of a variable number of geometric laws with equal or different parameters (Melodelima et al. 2004). Thus, each region (intergenic, intronic and exonic) is represented by a macro-state in the HMM (Figure 1). Exons are made of a succession of codons and each of the three positions in a codon (0, I, II) has characteristic statistical properties. This implies the need to separate exons in three states (Borodovsky et *al*. 1993 and Burge et al. 1998). HMMs take into account the dependency between a base and its *n* preceding neighbours. In this case, the order of the model is *n*. For our study, *n* has been chosen equal to 5 as in the studies of Borodovsky et *al*. 1993 and Burge & al. 1998. The emission probabilities of the HMM are therefore estimated by the frequencies of 6-letter words in the different regions (intron, initial exon, internal exon and terminal exon) that compose the training set. Moreover this model takes into account the direct and reverse strands of the DNA sequences. Exon states are separated into two categories that represent the direct coding state and the reverse coding state. The three models are trained and adjusted separately on the three sets *H, L* and *M*.

Our HMM method is used to identify isochores in the human genome. We divided the DNA of each human chromosome into overlapping 100 kb segments. Two successive segments overlap by half of their length. For each segment and for each model (*H, L* and *M*), the probability P[*Mod* | *S*] has been computed with *Mod* being the model used and *S* the segment that is tested. For each segment, the three HMMs were well discriminated. In all cases, the probability P[*Mod* | *S*] of the best HMM has appeared to be higher than 0.9. Our method allows to identify isochores larger than 50kb. So each window is clearly associated with *H, L* or *M* following the model that maximise P[*Mod* | *S*]. To be coherent with proceeding definition we consider than an isochore is a region made of window associated with the same class and of length greater than 300 kb. The distribution of the three models has been represented on a graphic along the human chromosomes. To check the coherence of the isochore prediction, the graphic is given with two other plots: a plot of the distribution of the gene density and a plot of the *G+C* content along the chromosomes.

The *G+C* rich regions are well known to present a great variability, thus the $G+C_3$ content is not always sufficient to discriminate the isochore class of a gene. For instance, inside an isochore *H*, some genes have a low $G+C_3$. However our model shows clear homogeneous isochore classes. To explain this result, each macro-sates (exon, intron, gene) detected by the model have been separately analysed. For each isochore, the prediction of the HMMs associated to macros-states *H* and *L* for each region was thus compared to the $G+C_3$ of the gene to its *G+C* content, or to the length of the region.

Figure 1: Basic representation of the different macro-states which characterise the HMM *H*. Curving arrows represent the duration state in a macro-state. Dashed arrows show the transition between macro-states.

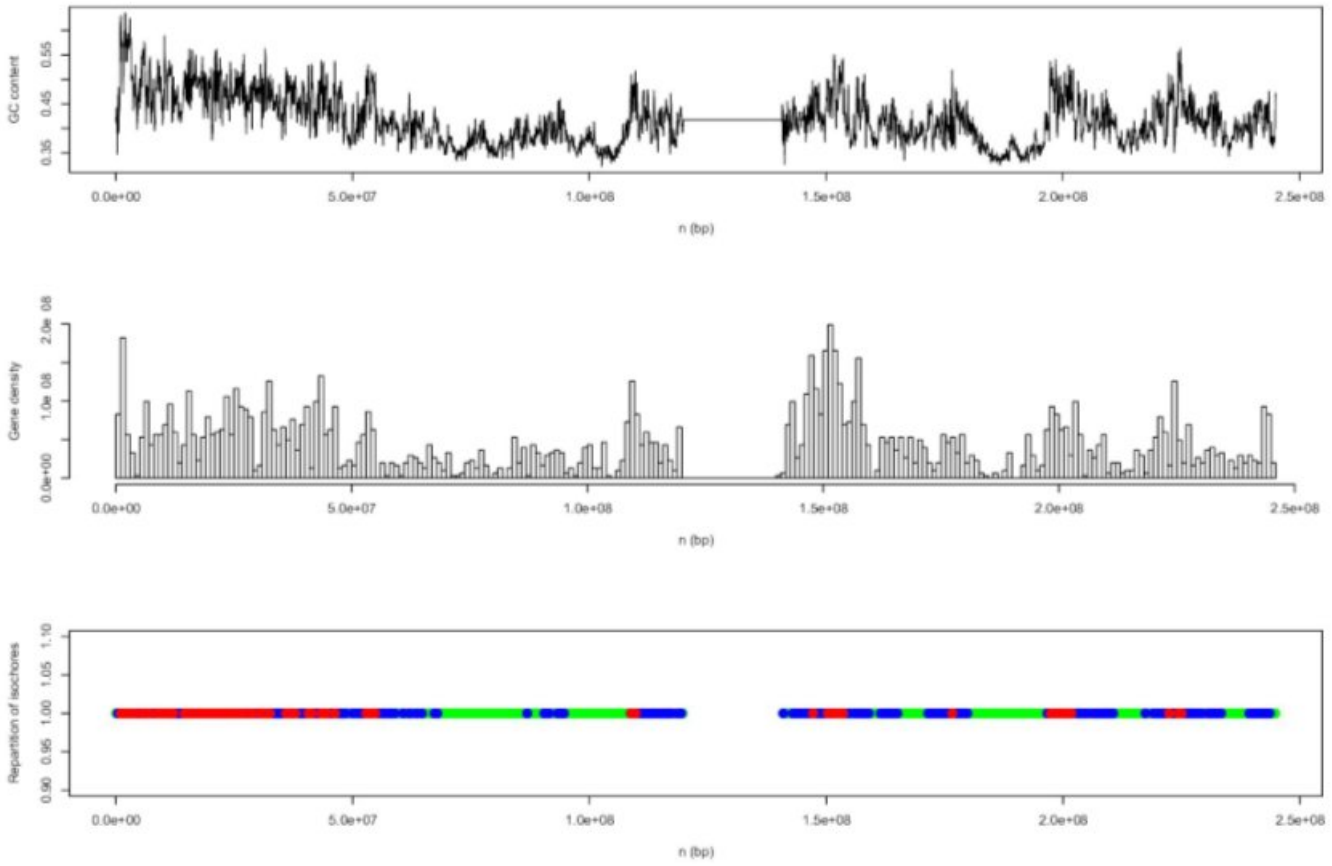


4 RESULTS

4.1 Isochores chromosome map

The distribution of the *G+C* content along the human chromosomes fits fairly well with the isochore organisation for all chromosomes. For instance, the detected *H, L* and *M* isochores appear colored respectively in red, green and blue in Figure 2 (chromosome 1 and 6). The maps display the mosaic organisation of the human genome (Bernardi et al. 1985, Bernardi 2001, Pavlicek et al. 2001) composed by many regions of fairly homogeneous *G+C* content (Li et al. 2003). The average *G+C* content for isochores on all human chromosomes is respectively of 0.515 ± 0.035, 0.45 ± 0.012 and 0.395 ± 0.017 for the *H, M* and *L* isochores. A Wilcoxon's test shows that the *G+C* content of the *H* isochores is significantly different from the one for the *L* isochores (p-value=0.000129). The same test shows that the *G+C* content in the *M* isochores is significantly different from the one in the *L* isochores (p-value=0.000516) and *H* isochores (p-value=0.03175). Such generalised mosaic structure along all the human chromosomes contradicts the suggestion of Eyre-Walker and Hurst (2001) that the isochore structure accounts for only some parts of the genome and confirms the results obtained by Oliver et al (2002)

Figure 2a: Repartition of isochores along human chromosomes. To check the coherence of the isochore prediction, the graphic is given with two other plots: a plot of the distribution of the gene density and a plot of the *G+C* content along the chromosome 1.



4.2 Isochore size variation with the G+C content

The different types of isochores (*H*, *L* and *M*) show a variation in size, depending on the *G+C* content. *G+C* poor isochores (*L*) are significantly larger than *G+C* rich isochores (*H*) (the p-value of the Wilcoxon test is $9.29.10^{-10}$). The average length for the *L* isochores is 7.71 Mb, whereas the average length for the *H* isochores is 2.93 Mb. This relationship was previously observed for the isochores detected by DNA centrifugation (Bettecken et al. 1992, Pilia et al. 1993, De Sario et al. 1996).

4.3 Variation of gene density in human isochores

In the isochores detected by DNA centrifugation, several authors (Bernardi et al. 1985, Mouchiroud et al. 1991, Zoubak et al. 1996, Bernardi 2000) observed that gene density increases from a very low average in *L* isochores to a 20-fold higher average in *H* isochores. Our results agree with this observation. For all chromosomes, the isochore structure fits nicely with the gene density distribution along each chromosome. The gene density in the *H* windows (15 genes per Mb) is higher than the one in the *L* windows (3.67 genes per Mb) leading to a significant Wilcoxon test (p-value=$4.776.10^{-5}$). The same difference is observed when we compare the characteristics of the *M* windows with those of the *H* and *L* windows.

4.4 Analysis of the structure of the isochores *H* and *L*

G+C rich regions are well known for presenting a great variability and are difficult to recognize by a simple *G+C*-plot. Indeed, genes with respectively G+C3 > 0.72, $\in$ [0.56,0.72] and <0.56 have relative frequencies of 48%, 35% and 17% in the *H* isochores we determine. Therefore, a detailed study of the *H* isochores has been conducted to understand why our method found homogeneous *H* isochores. The behaviour of the HMMs *H* and *L* on the genes of the *H* isochores which have a $G+C_3$ superior to 0.72 or inferior to 0.56 has thus been studied. Our method classified 82% of the genes with $G+C_3$ superior to 0.72 in the *H* class. Thus, the HMM "gene" *H* describes correctly the genes with a high $G+C_3$ content. However, 60% of the genes with $G+C_3$ inferior to 0.56 were classified by our method in the *H* class. Our method shows two types of genes with $G+C_3$ inferior to 0.56: genes which are recognized by the HMM "gene" *L* (40%) and genes which are recognized by the HMM "gene" *H* (60%). This fact indicates that something different from the $G+C_3$ content could contribute to characterise these gene. Table 1 shows the influence on the genes predictions of several regions, using the prediction of the HMM adapted to each region.

Figure 2b: Repartition of isochores along human chromosomes. To check the coherence of the isochore prediction, the graphic is given with two other plots: a plot of the distribution of the gene density and a plot of the *G+C* content along the chromosome 6.

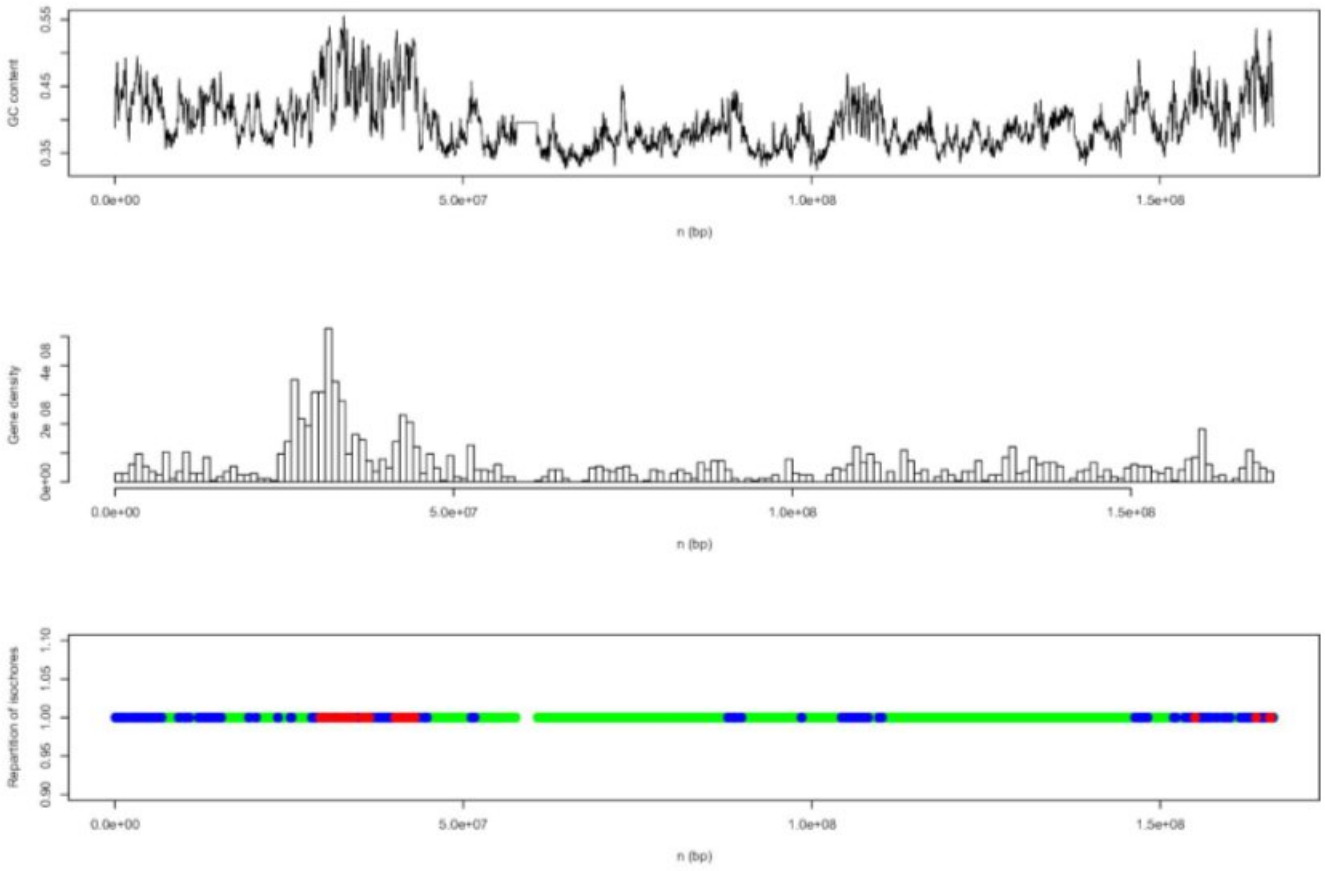


Table 1: Analysis of the prediction of the macro-states on the different regions.

Regions of genes	Sequences (with $G+C_3 <$ 0.72) classified as *H* isochores by an HMM model representing each region	Sequences (with $G+C_3 <$ 0.56) classified as *H* isochores by an HMM model representing each region
Gene	82%	60%
CDS	96%	26%
Introns	93%	29%
5'UTR	86%	60%
3'UTR	86%	61%
intergenic	75%	57%

Legend: In order to compare the *H* and *L* models, the probability of each sequence of a given type (Genes, Introns, …) is computed under the two models (macro-states which characterise the sequence) and the sequence votes for the model which has the greater probability. For instance, the gene sequences ($G+C > 0.72$) vote either for the "Gene *H* model" or for the "Gene *L* model". The conclusion is that on these gene sequences, these "Gene *H* model" obtains 82% better predictions than the "Gene *L* model".

The classification of the genes with $G+C_3$ inferior to 0.56 does not depend on the characteristics of the CDS and of the introns but only on the 5'UTR and 3'UTR regions. To confirm this hypothesis, the correlation between the prediction of the genes and the 5'UTR and 3'UTR regions by their respective HMMs were analysed (see Table 2).

Table 2: Comparison of the genes $G+C_3$ inferior to 0.56 with and 5'UTR predictions in isochore *H*.

Isochore prediction of the HMM "genes"	Isochore prediction of the HMM "5'UTR"	% of genes in this configuration
H	H	50%
H	L	10%
L	H	10%
L	L	30%

In 80% of the cases (Table 2: sum of the lines 1 and 4), the decision of the HMM "gene" and the HMM "5'UTR" was similar (the same results were obtained when the genes and the 3'UTR were compared). The UTR regions predicted in the *H* isochore have a *G+C* content (0.510 ± 0.0041) higher than the UTR region predicted in the *L* isochore (0.429 ± 0.0008). Therefore, when the HMM model finds an isochore in the human genome, two facts permit to classify the genes with

$G+C_3$ <0.56 in the *H* isochore. First, the *G+C* content in the UTR regions influence the predictions of the HMM models (60% of the cases). Second, there is a smoothing effect (40% of the cases), *ie*, the window around the gene influences the choice of the model (particularly the influence of the intergenics region).
A similar analysis was performed for the isochore *L*. The G+C poor regions are more homogeneous than the *G+C* rich regions. Thus, the distribution of genes in the *L* isochores is 6%, 19% and 75% for genes in which $G+C_3$ is >0.72, $\in$ [0.56, 0.72] and <0.56 respectively. Table 5 shows that the classification of the genes with $G+C_3$ > 0.72 do not depend on the characteristics of the CDS and of the introns but only on the 5'UTR and 3'UTR regions as it was the case with the H isochores (Table 3). This hypothesis is confirmed by the correlation between the prediction of the genes and the 5'UTR and 3'UTR regions by the models in 78% of the cases (see Table 4).

Table 3: Analysis of the prediction of the macro-states on the different regions in the *L* isochores.

Regions of genes	Sequences (with $G+C_3$ < 0.56) classified as *L* isochores by an HMM model representing each region	Sequences (with $G+C_3$ > 0.72) classified as *L* isochores by an HMM model representing each region
Gene	92%	93%
CDS	72%	33%
Introns	93%	24%
5'UTR	83%	84%
3'UTR	88%	83%
intergenic	91%	91%

Legend: see Table 1

Table 4: Comparison of the genes $G+C_3$ > 0.72 with and 5'UTR predictions in isochore *L*.

Isochore prediction of the HMM "genes"	Isochore prediction of the HMM "5'UTR"	% of genes in this configuration
H	H	0%
H	L	7%
L	H	16%
L	L	77%

4.5 Statistical correlations between the isochore class and the length and G+C content of the different regions which compose the genes.

The study of the length and *G+C* distribution of the different regions which characterise the genes in the different types of isochores confirm some known characteristics: in the *L* isochores, the introns, UTR and CDS regions are longer and their *G+C* content is lower than in the *H* isochores. If we study these regions following the prediction of the model and the isochore classes, we can see that the length and *G+C* content influence the choice of the model. Thus, in the *H* isochores, introns and UTR regions which are linked to a lower $G+C_3$ content of the gene and predicted in class *H* by the model are significantly shorter and their *G+C* content is significantly higher than introns and UTR regions which are associated with a lower $G+C_3$ content of the gene and predicted in class *L*. The same is observed when we study the *L* isochores (data not shown).

DISCUSSION

The use of Markov models to do data exploration has been underestimated in genome analysis probably because these models are used for prediction purposes mainly. Our study shows that simple hidden Markov models could be used to model the human genome organisation and to find new biological structures. The statistical characteristics of the coding regions of vertebrates vary dramatically between the different isochores classes (Thierry & al. 1976). Hidden Markov models were adapted to each isochore class. Only the protein genes and intergenic regions were modelled to limit the number of parameters, states and the CPU cost.
This method has clearly demonstrated that an isochore structure really exists in the human genome (see Figure 2). The distribution of gene density along a chromosome is in good agreement with the isochore structure identified here: the higher gene density regions are located in the *G+C*-isochores with highest *G+C* content. Moreover, the relationship between isochore class and gene structure is clearly shown by our HMM approach. The main, and new, result is that the $G+C_3$ variability inside *H* isochors is for most genes not reflected in the UTR's regions and that leads to a isochore pattern as detected by our method much clearer than will all preceding ones. This emphasizes the fact that $G+C_3$ is not the only statistical property involved in isochore patterning. Biological mechanisms involved in these patterns differences between coding region and UTR may be either neutral or adaptative. Indeed such differences could result either from specific mutations that may accumulate quicker in UTR due to less functional constraints (neutral point of view), or, alternatively from selective constraints on UTR, probably associate with gene expressions.
When the draft human genome sequence was made available, Lander et al. (2001) tried to look for isochores, but they failed to find any. The existence of isochores in the human chromosomes 21 and 22 is questionable based on a sequence analysis (ref). The reason for the debate is due to the lack of a sequence-based definition of isochores. The concept of an isochore is related to the concept of homogeneous domains over large scales (of hundred of kilobases) in genomes, in which the variations of *G+C* content may be considered to be small. The use of hidden Markov models allows to be free of this isochore definition. Indeed, the isochore structure is predicted by HMMs that are not based only on *G+C* content. This method improves the prediction of isochores in comparison with classical methods. Thus, three hidden Markov models were adjusted to each isochore class to take into account other biological properties associated with the isochores classes *H*, *L* and *M* (such as the different length of the exons or introns and the gene density that vary according to the *G+C* content…). These properties were neglected by classical methods. Procedures that compute the *G+C* content by sliding an overlapping or non-overlapping window along a genome cannot determine precisely the boundaries of the isochores. Thus, the *G+C* variations are sometimes larger than statistical fluctuations and isochores are difficult to determine.
The clarification of the isochore structure is a key to understanding the organisation and biological function of the human genome. By using our method, the structure of a region

may be easily analysed. Indeed, HMMs give more information than classical isochore prediction models. Thus, some biological properties can be extracted and associated with an isochore region, such as position and length of genes, exons, introns or density of genes. This last point is neglected by the other methods.

Last year, many genomes have been sequenced. This huge amount of data makes it impossible to analyse patterns to provide biological interpretation “by hand”, so mathematical and computational methods have to be used. Our approach, using HMMs, seems to be very promising for analysing the organisation of genomes.

CONCLUSION

We have developed a computational method to predict isochores in the whole human genome using an HMM. This method allows to predict isochores of 300 kb and clearly points to a mosaic structure of the human genome. The isochores identified were separated into three classes according to their *G+C* content: heavy, light and medium isochores.

ACKNOWLEDGEMENTS

The calculus have been made at the IN2P3 computer centre using a large computer farm (more than 1000 cpu). The authors thank M.F. Sagot, L. Duret and D. Mouchiroud for helpful comment on the manuscript.

REFERENCES

Aota S, Ikemura T. (1986) Diversity in G+C content at the third position of codons in vertebrate genes and its cause. *Nucleic Acids Res.* 14:6345-6355.

Berget SM. (1995) Exon recognition in vertebrate splicing. *J Biol Chem.* 1995 Feb 10;270(6):2411-4. Review.

Bernardi G. (1995) The human genome: organization and evolutionary history. *Annu Rev Genet.* 29:445-76. Review.

Bernardi G. (2001) Misunderstandings about isochores. Part 1. *Gene.* 276(1-2):3-13. Review.

Bernardi G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene.* 241(1):3-17. Review.

Bettecken T, Aissani B, Muller CR, Bernardi G. (1992) Compositional mapping of the human dystrophin-encoding gene. *Gene.* 122(2):329-35.

Bernaola-Galvan P, Carpena P, Roman-Roldon R, Oliver JL. (2001) 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*, Montreal, Canada, ACM Press, New York, pp 217-218.

Borodovsky M, McIninch J. (1993) Recognition of genes in DNA sequence with ambiguities. *Biosystems.* 30(1-3):161-71.

Burge C, Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* 268(1):78-94.

Burge C, Karlin S.(1998) Finding the genes in genomic DNA. *Curr Opin Struct Biol*. 346-54. Review.

Churchill GA. (1989) Stochastic Models for heterogeneous DNA Sequences. *Bull. Mathematical Biology*. 51: 79-94.

Clay O, Caccio S, Zoubak S, Mouchiroud D, Bernardi G. (1996) Human coding and non coding DNA: compositional correlations. *Mol Phyl Evol*. 1:2-12.

Cuny G, Soriano P, Macaya G, Bernardi G. (1981) The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Eur J Biochem.* 115(2):227-33.

De Sario A, Geigl EM, Palmieri G, D'Urso M, Bernardi G. (1996) A compositional map of human chromosome band Xq28. *Proc Natl Acad Sci* U S A. 93(3):1298-302.

D’Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G. (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol*. 32:504-510.

Dunham I, Shimizu N, Roe BA, et al. The DNA sequence of human chromosome 22. *Nature*. 1999 Dec 2;402(6761):489-95.

Duret L, Mouchiroud D, Gouy M. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res.* 22(12):2360-5.

Duret L, Mouchiroud D, Gautier C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol.* 40(3):308-17.

Eyre-Walker A, Hurst LD. (2001) The evolution of isochores. Nat Rev Genet. 2(7):549-55. Review.

Haring D, Kypr J. (2001) Mosaic structure of the DNA molecules of the human chromosomes 21 and 22. *Mol Biol Rep*. 28(1):9-17.

Hattori M, et al. (2000) The DNA sequence of human chromosome 21. Nature. 405(6784):311-9. Erratum in: *Nature* 2000 Sep 7;407(6800):110.

Hawkins JD. (1988) A survey on intron and exon lengths. Nucleic Acids Res. 16(21):9893-908. Review.

Jabbari K, Bernardi G. (1998) CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene*. 224(1-2):123-7.

Lander ES, Linton LM, Birren B, Nusbaum C, et al. (2001) Initial sequencing and analysis of the human genome. Nature. 409(6822):860-921. Erratum in: Nature 2001 Aug 2;412(6846):565. *Nature* 2001 Jun 7;411(6838):720.

Li W .(2001) Delineating relative homogeneous G+C domains in DNA sequences. *Gene*. 276(1-2):57-72.

Li W, Bernaola-Galvan P, Haghighi F, Grosse I. (2002) Applications of recursive segmentation to the analysis of DNA sequences. *Comput Chem.* 26(5):491-510.

Li W, Bernaola-Galvan P, Carpena P, Oliver JL. (2003) Isochores merit the prefix 'iso'. *Comput Biol Chem*. 27(1):5-10.

Macaya G, Thiery JP, Bernardi G. (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. *J Mol Biol*. 108(1):237-54.

Melodelima C., Guéguen L., Piau D., Gautier C. (2004) Modelling the length distribution of exons by sums of geometric laws. Analysis of the structure of genes and *G+C* influence , *JOBIM*, Montréal.

Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G. (1991) The distribution of genes in the human genome. *Gene*. 100:181-7.

Nekrutenko A, Li WH. (2000) Assessment of compositional heterogeneity within and between eukaryotic genomes. *Genome Res*. 10(12):1986-95.

Oliver JL, Carpena P, Roman-Roldan R, Mata-Balaguer T, Mejias-Romero A, Hackenberg M, Bernaola-Galvan P. (2002) Isochore chromosome maps of the human genome. *Gene*. 300(1-2):117-27.

Oliver JL, Carpena P, Hackenberg M, Bernaola-Galvan P. (2004) IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res*. 32(Web Server issue):W287-92.

Pavlicek A, Jabbari K, Paces J, Paces V, Hejnar JV, Bernardi G.(2001) Similar integration but different stability of Alus and LINEs in the human genome. *Gene*. 276(1-2):39-45.

Peshkin L, Gelfand MS. (1999) Segmentation of yeast DNA using hidden Markov models. *Bioinformatics*.15(12):980-6.

Pilia G, Little RD, Aissani B, Bernardi G, Schlessinger D. (1993) Isochores and CpG islands in YAC contigs in human Xq26.1-qter. *Genomics*. 17(2):456-62.

Thiery JP, Macaya G, Bernardi G. (1976) An analysis of eukaryotic genomes by density gradient centrifugation. J Mol Biol. 108(1):219-35.

Zhang CT, Zhang R. (2003) An isochore map of the human genome based on the Z curve method. *Gene*. 317(1-2):127-35.

Zoubak S, Clay O, Bernardi G. (1996) The gene distribution of the human genome. *Gene*.174(1):95-102