A Markovian Approach for the Analysis of the Gene Structure

Christelle Melo de Lima[1], Laurent Guéguen[1], Christian Gautier[1] and Didier Piau[2]

[1]UMR 5558 CNRS Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1
43 boulevard du 11-Novembre-1918
69622 Villeurbanne Cedex 69622 – France.
melo@biomserv.univ-lyon1.fr
[2] Institut Camille Jordan UMR 5208, Université Claude Bernard Lyon 1
Domaine de Gerland, 50 avenue Tony-Garnier
69366 Lyon Cedex 07 – France.

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. Numerous methodological difficulties are encountered when using HMMs to segregate genes from transposons or retroviruses, or to determine the isochore classes of genes. The aim of this paper is to analyse these methodological difficulties, and to suggest new tools for the exploration of genome data. We show that HMMs can be used to analyse complex genes structures with bell-shaped distributed lengths, modelling them by macro-states. Our data processing method, based on discrimination between macro-states, allows to reveal several specific characteristics of intronless genes, and a break in the homogeneity of the initial coding exons. This potential use of markovian models to help in data exploration seems to have been underestimated until now, and one aim of our paper is to promote this use of Markov modelling.

Keywords: HMM, macro-state, gene structure, $G + C$ content

1 Introduction

The sequencing of the complete human genome led to the knowledge of a sequence of three billion pairs of nucleotides [19]. 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 3% of the human genome is known to code for proteins. Another difficulty is that the statistical characteristics of the coding region vary dramatically from one species to the other, and even from one region in a given genome to the other. For example, vertebrate isochores ([29], [3]) exhibit such a variability in relation to their $G + C$ frequencies. Thus it is necessary to use different models for different regions if one seeks to detect patterns in genomes.

A classical way of modelling genomes uses hidden Markov Models (HMMs) ([22], [18], [23]). 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], [2], [17]), hence they cannot be represented by geometrical distributions. Semi-Markov models are one option to overcome this

problem [6]. These models 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], [27], [15]). 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 [27]. Thus the estimation problem is more difficult for these variable duration HMMs than for standard HMMs [27].In other words, semi-Markov models are efficient tools to detect protein genes, but they are much more complex than HMMs. 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.

The modelling of a gene may be used to annotated complete genomes, as Genscan [6] in Ensembl, but also to explore data in order to detect exceptional patterns and to help in their biological interpretation. Thus, the use of Markov models for the purpose of data exploration has been underestimated in genome analysis. This objective requires simple parameters and a relative small amount of computer resources, to be able to perform numerous analyses of the data. For this purpose, we show how to use macro-states HMMs models for complete genome analysis.

2 Materials

Gene sequences were extracted from Hovergen (Homologous Vertebrate Genes Database) [11] for the human genome. To ensure that 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. To simplify the model, UTRs (including their introns) were not separated from intergenic regions. As a consequence, in the present paper, the word "intron" means an intron which is located between two coding exons.

The statistical characteristics of the coding and noncoding regions of vertebrates differ dramatically between the different isochore classes [4]. The isochore has been classified as a "fundamental level of genome organisation" [13] and this concept has increased our appreciation of the complexity and variability of the composition of eukaryotic genomes [25]. 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 [24]). 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 [12]. 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 ([10], [9]). 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 [20]. Thus, we took into account the isochore organisation of the human genome. Three classes were defined 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%, 72%], M=]56%,72%[and L=[0%,56%], which were used to build a training set. These classes were roughly the same as those used by other authors ([24], [30]). These sets were used to model the distributions of the lengths of the exons and the introns, and to analyse the structure of genes.

3 Methods

3.1 Estimation of the parameters

Estimation of emission probabilities: The DNA sequence is heterogeneous along the genome, but it consists of a succession of homogenous regions, such as coding and non-coding regions. HMMs are used to distinguish between these different types of regions.

Exons consist of a succession of codons, and each of the three possible positions in a codon (1, 2, 3) has characteristic statistical properties. This implies the need to divide exons into three states ([7], [5]). 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 was taken to be equal to 5, as in the studies of Borodovsky and Burge ([7], [5]). 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. Even if introns have not codon structure, the use of 6-letter words allow to improve the discrimination between coding and no-coding region. Therefore there is an HMM for each region.

Thus, the emission probabilities of the model were estimated by using the maximum likelihood method in order to highlight why some sequences are not correctly predicted although it is the case for other sequences of the same region. In other words, we relied on the error of predictions of the HMMs, rather than analyse somewhat blindly the genomes to do an exploration of the human genome.

Estimation of the structure of the macro-states: An alternative to the semi-Markov models is suggested to model the bell-shaped empirical length distributions of the exons. We propose to use sums of a variable number of geometric laws with equal or different parameters. 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 to estimate the parameters and compute the most likely state sequences still apply [15]. The length distribution of the exons and introns was estimated from the training set (data set sequences are named $x_1 \dots x_n$). Each x_i was considered to be the realization of an independent variable of a given law. We tested the following laws:

– The sum of m geometric laws of same parameter p (*i.e.* a binomial negative law):

$$P[X = k] = C_{k-1}^{m} p^m (1-p)^{k-m}$$

– The sum of two geometric laws with different parameters $p_1 > p_2$:

$$P[X = k] = p_1 \times p_2 \times \frac{(1-p_2)^{k-1} - (1-p_1)^{k-1}}{p_1 - p_2}$$

– The sum of three geometric laws with different parameters $p_1 < p_2 < p_3$:

$$P[X=k] = \frac{p_1 \times p_2 \times p_3}{p_2 - p_3} \times \left(\frac{(1-p_1)^{k-1} - (1-p_3)^{k-1}}{p_3 - p_1} - \frac{(1-p_2)^{k-1} - (1-p_3)^{k-1}}{p_3 - p_2} \right)$$

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. However, the classical Newton or gradient algorithm cannot minimise for the Kolmogorov-Smirnov distance, because this distance cannot be differentiable. We therefore discretised the parameter space with a step of 10^{-5}, and fixed the minimum value. Parameter estimations were not based on the maximum likelihood, which would have matched the end of the exon length distribution thus neglecting many small exons (Figure 1 a). Indeed, that it is for a geometrical law or a convolution of geometrical laws, the parameter p is estimated by the reverse of the mean ($E[X] = 1/p$) by the method of the maximum likelihood. 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 human gene. Again, in order to provide simple but efficient models, equal transitions between states of a macro-state were favoured when it was possible.

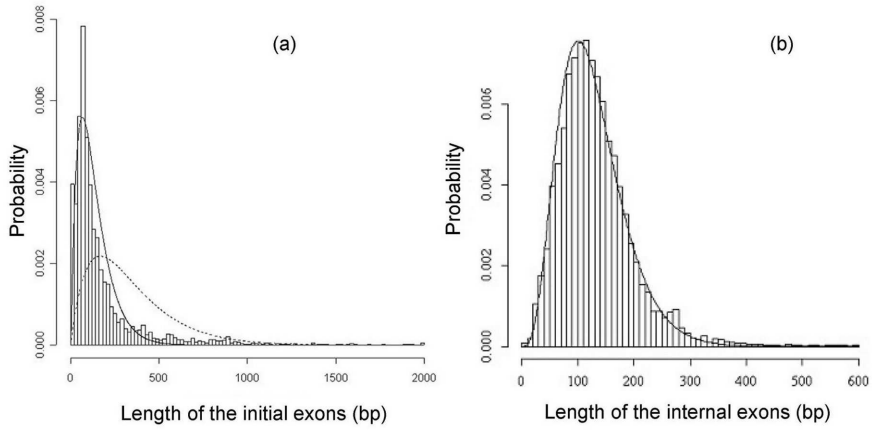

Figure 1. (a) The histogram represents the empirical distribution of the length of the initial exons in a multi-exons gene. The dotted line describes the theoretical distribution, obtained from the Kolmogorov-Smirnov distance. The continuous line characterises the binomial distribution, obtained by the method maximum likelihood. (b) The histogram represents the empirical distribution of the length of the internal exons. The dotted line describes the theoretical distribution, obtained from the Kolmogorov-Smirnov distance.

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], [28], [8]) have shown that the length distribution of the exons depend on their position in the gene. We took all exon types into account: initial coding exons, internal exons, terminal exons and single-exon genes.

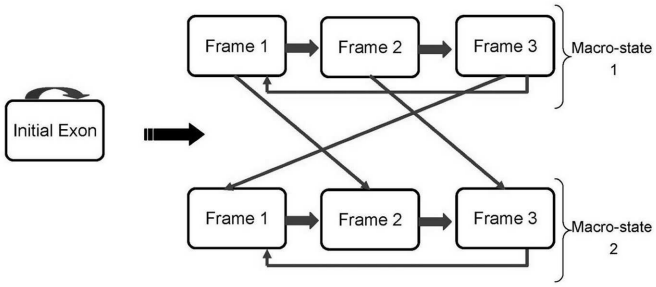

Figure 2. Figure representing initial exon HMM

3.2 Models selection

Algorithm of Models selection In order to measure the adequacy of a model with a genomic region, the theory of HMMs proposes two solutions: the probability of the observed sequence conditioned by the optimal trajectory in the hidden states (Viterbi) or the probability of the sequence x under the model M, $P[x|M]$.
The first method neglects the fact that many trajectories are biologically equivalent. The second method sums the probabilities corresponding to internal structures of a sequence, which were different. Thus, a model that predicts a bad internal structure can be associated to a high value of the probability. For example, these two techniques of selection of models in the context of HMMs were compared:
We consider a HMM of type $M1M0$ with 2 states, called A and B, and 2 observations, called 0 and 1. We assume that the transition probabilities from A to B and from B to A are both $t = 9.53643.10^{-7}$, and that A emits 0, respectively B emits 1, with probability p. We note $M_{0.9}$ the HMM with the probability $p = 0.9$ and $M_{0.6}$ the HMM with $p = 0.6$. We choose the sequence $x = 0^n1^n$ for given value of n=10, the aim is to choose $M_{0.9}$ and $M_{0.6}$.
If the maximisation of the probability of the sequence was used, it is needed to compute $P(x|M)$:

$$P(x|M_{0.9}) = 6.97.10^{-11} < P(x|M_{0.6}) = 1.27.10^{-6}.$$

In this case the model $M_{0.6}$ is better than the model $M_{0.9}$.
If the probability of the observed sequence conditioned by the optimal trajectory in the HMM was used, it is needed compute the probability given by the Viterbi algorithm: $P(x/s_{op}, M)$. For the models $M_{0.9}$ and $M_{0.6}$, the optimal sequence is composed of n states A followed by n states B. In general case ($0.5 < p < 1$), this probability is:

$$P(x|s_{op}, M_{0.9}) = 0.1215 > P(x|s_{op}, M_{0.6}) = 3.65.10^{-5}.$$

Thus, $M_{0.9}$ is better than $M_{0.6}$. This very schematic example shows opposite conclusions for the two methods and amphases the fact none of these approaches has a universal validity. On the other hand, if we consider HMMs that correspond to

a macro-state, the situation is biologically clearer. All trajectories in a macro-state are biologically equivalent. The method of the optimal trajectories is therefore not adapted to this problem, while, the situation is well described by the probability of the sequence under the model. Thus, the probabilities that are summed correspond to the same biological structures.

Analysis of the gene structure using HMMs selection The use of HMMs for classifying sequences raises the question of the evaluation of their discriminating power. The method chosen here is to split the set of sequences of known nature into two sets: one for training and one to compare the different models.
All HMMs (introns, initial exons, internal exons and terminal exons models) are then compared pairwise for all the sequences in a given type of region (intron, initial exon, internal exon and terminal exon sequences) of the test set, in order to identify the model which is the most likely to represent the test sequence. This gives the discrimination measure D, with

$$D = P(S/HMM_1)/P(S/HMM_2),$$

where S is the sequence being tested, and HMM_1, HMM_2 are the two models tested. The computations were realized with the package SARMENT [16]. The best HMM for most of the sequences in a given region is used to characterise this region. Each model is finally characterised by the frequency with which it recognises the sequences. This approach allows to show the types of sequences that were not well recognised by their corresponding model. Finally, the analysis of the different types of exons was completed by a correspondence analysis.

4 Results – Discussion

4.1 Inclusion of explicit distributions of the durations of the states in HMMs

In order to model the bell-shaped empirical length distributions of exons (Figure 1), we have used sums of geometric distributions with equal or different parameters. The length of an exon depends on its position within the gene. Initial and terminal exons tend to be longer than internal ones (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, but these types of introns are both smaller than the initial introns (Table 1). As is well known, the lengths of exons and introns depend on their $G + C$ content [8]. 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 initial exons are longer in $G+C$ rich regions (i.e. displays a significant Wilcoxon non-parametric test). However, the length of the internal and terminal exons does not vary with the class of isochores (*i.e.* displays a non significant Wilcoxon non-parametric test). The length of the exons displays clearly a bell-shaped pattern, for the three $G + C$ classes. Since 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 1), we used it to model their length distribution by a sum of geometric laws and estimated the parameters of these laws (see Method for a comparison with the maximum likelihood approach).

Position in the gene	Length (bp) in class H		Length (bp) in class M		Length (bp) in class L	
	Mean	Median	Mean	Median	Mean	Median
Initial coding exon	223	123	176	102	160	87
Internal exon	144	126	143	125	144	120
Terminal exon	244	165	237	145	218	138
Initial intron	4027	3189	4139	3540	5315	4857
Internal intron	1461	958	1767	1310	2850	2433
Terminal intron	1394	884	1764	1282	2819	2415

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

Laws	Parameters p	K-S distance
$G_2(p)$	0.0117	0.1084
$G_3(p)$	0.0185	0.16
$G_4(p)$	0.02634	0.1826
$G(p_1, p_2)$	0.0055-0.087	0.0447

Table 2. Parameters estimation of different laws obtained for initial exons of class H minimising Kolmogorov-Smirnov distance (K-S)

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 $p_i = 1/D_i$. 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 $G_n(1/p)$. Finally $G_n(D)$ is a geometric distribution with expectation D and parameter $p = 1/D$, which we write $G(D)$.

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 can be modelled by sums of geometric laws. The estimated distributions are $G_2(58.82, 74.07)$ for initial exons (Figure 1 a), $G_3(86.21, 181.81, 10)$ for terminal exons, $G_5(26.32)$ for internal exons (Figure 1 b), $G_3(351.11)$ for intronless genes, and the geometric distribution $G(111.11)$ for initial introns. Other types of introns are also modelled by a geometrical distribution.

The distributions of the lengths of the single exons (that is, the intronless genes) exhibit a clear bi-modality (Figure 3). By using the software Blast [1] to search regions of the human genome similar to our intronless genes, we have found that many small intronless genes are often repeated along the human genome. The comparison of all these repeated intronless genes to a database of pseudogenes [21] revealed that many small intronless genes are actually pseudogenes, *i.e.*, genes that have lost their function. After the elimination of these pseudogenes, the distribution of the lengths of the real intronless genes is bell-shaped, like the distributions for the other types of exons.

4.2 Evaluation of the models

The macro-states used for initial exons (M_E1), internal exons (M_Eint) and terminal exons (M_Eter) were evaluated. In order to compare the models two-by-two, the likelihoods of each sequence of a given type (initial exons, internal exon, etc.) with respect to the two models were compared and the model with the greater likelihood is voted for by this sequence (see Method). For example, the sequences of initial exons vote between the model for the initial exons (M_E1) and the model for the internal

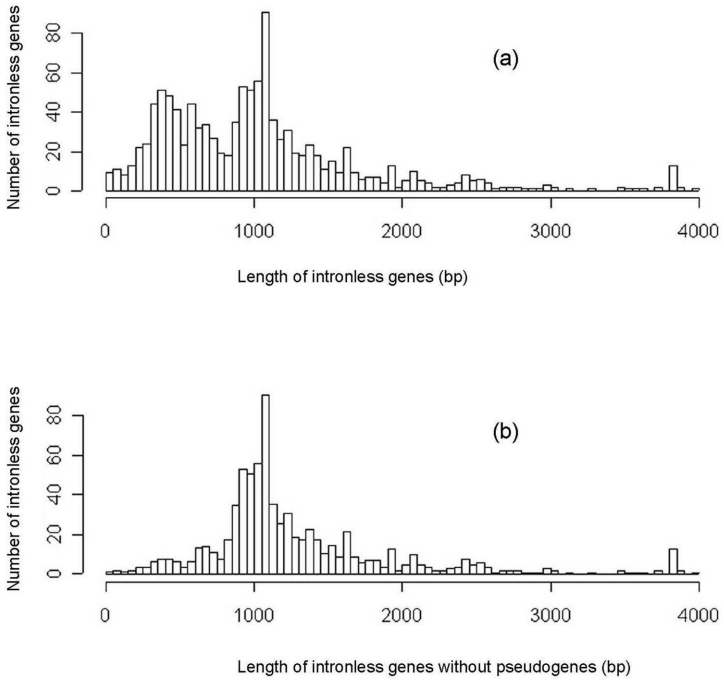

Figure 3. (a) The histogram represents the empirical distribution of the length of the intronless genes. (b) The histogram represents the empirical distribution of the length of the intronless genes without pseudogenes.

exons (M_Eint), assuming roughly equal proportions (Figure 4, Histogram 1). In conclusion, on the sequences of initial exons, the models M_E1 and M_Eint have similar predictive powers. Figure 4 gives results for the isochore class H. We stress the following points.

1. Internal exons and terminal exons share similar statistical properties. This is shown by the similar predictive powers of the models M_Eint and M_Eter (Figure 4, Histograms 4 and 6).
2. The initial exons are clearly discriminated from the other exons. This is shown by the smaller likelihood of the internal exons in M_E1 than in M_Eint (Figure 4, Histograms 3 and 5).
3. The modelling of the initial exons is inadequate. This is shown by the small likelihood of the initial exons in M_E1 (Figure 4, Histograms 1 and 2).

The specific statistical characteristics of the initial exons might result from the existence of signals located at, or covering, the beginning of the genes. To explore this hypothesis, we have split our HMM for the initial exons into two HMMs. The first one models the first n nucleotides of the initial exon, and the second the remaining part of the initial exon. This new initial exon model is called M_E1_n. Pairwise comparisons between the models M_E1_n for various values of n (Figure 5) show that the M_E1_{80} model yields the better discrimination. This suggests that the break of homogeneity in the initial exon happens around the 80th base. Finally, this separation provides a better discrimination between the models of the internal and initial exons on the one hand and the model of the initial exons on the other hand (49% to 61% in favour of the M_E1_{80} model [Figure 4, histogram 1 and Figure 5, histogram 7]) and from the

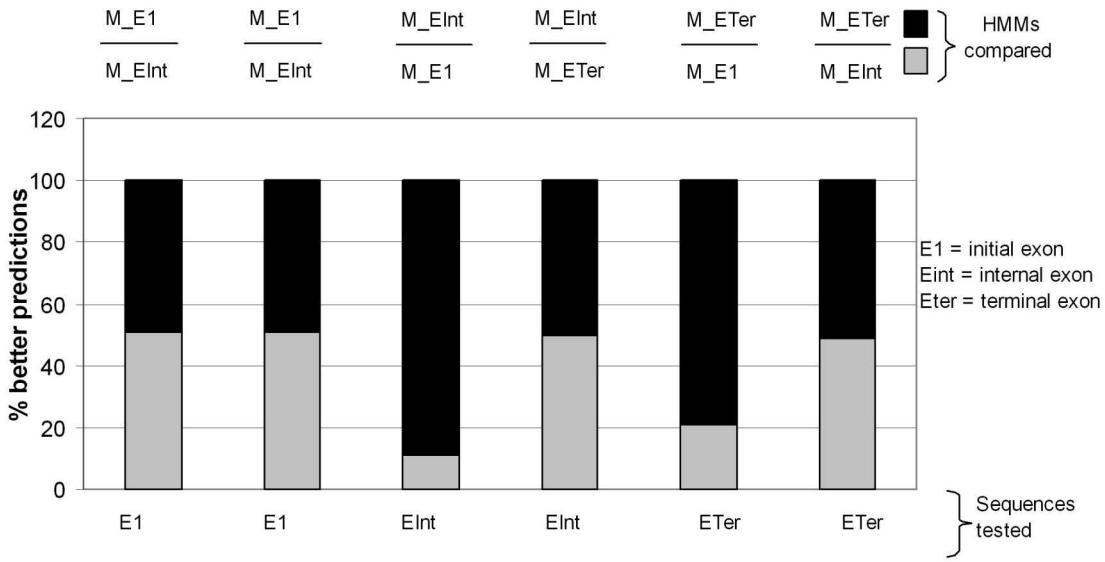

Figure 4. Models learned from different sequences (initial, internal and terminal exons) were compared pairwise using the sequences used to determine the best predictions. For instance, in histogram 1, the likelihood of each first exon was computed using models learnt on $E1$ and $Eint$. The black bar represents the percentage of first exons having a higher likelihood for the first exon model, and the grey bar those with a higher liklihood for the second exon model. Histograms 1-2: The models $E1$, $Eint$ and ET have same predictive power on initial exons. Histograms 3-5: The models $Eint$ and ET provide a good prediction of the internal and terminal exons compared to the $E1$ model (82% and 75%, respectively). Histograms 4-6: The models $Eint$ and ET have the same predictive power for initial and terminal exons.

internal exons (89% to 92% in favor of M_Eint, not shown in the Figure). Similar results were found for the terminal exons.

The break in the homogeneity of the first exon could be explained by the presence of a signal peptide. The first exons which contain a signal peptide are better recognised by the first HMM of the M_E1_{80} model than by the second one in 75% of the cases. These results were also compared with those obtained by SignalP [26]. The initial exons which, according to SignalP, contain a signal peptide, were more accurately recognised by the M_E1_{80} model than by the internal exon model in 70% of the cases. When SignalP does not predict a signal peptide, the M_E1_80 and the internal exon models yield similar results.

The significance of the modelling of isochores is highlighted by the results described in the previous paragraph, which show the effect of the distributions of the lengths of exons and introns. This claim was confirmed by our study of the influence of the isochore class on the words frequencies, in the different types of regions. For every type of exons (*i.e.*, initial, internal and terminal), the model trained with a specific isochore class performed better on this class than the others (Figure 6). The situation as concerns the classes of introns is somewhat different. The introns from classes H and M are better predicted by our HMMs H and M, respectively (Figure 7, Histograms 1 to 4), whereas the three models H, L, and M, are more or less equivalent for the introns of class L (Figure 7, Histograms 5, 6). This analysis clearly reveals some major statistical differences between the three isochore classes, and the importance of taking into account this heterogeneity of the genome in a con-

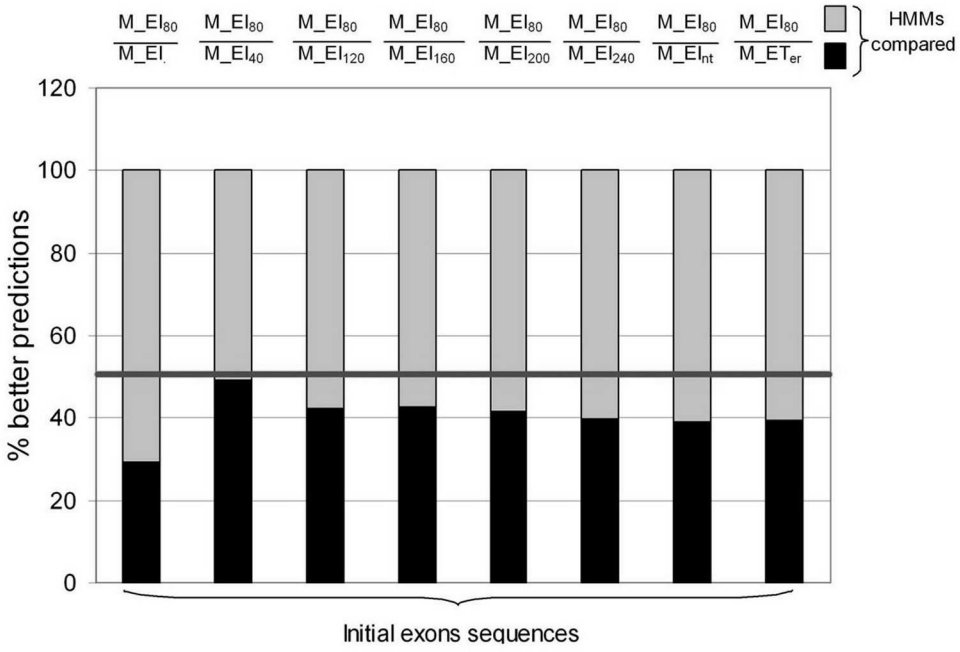

Figure 5. The models learned from different sequences (initial, internal and terminal exons) were compared pairwise to the initial exons to identify the best predictions. The $IE80$ model provides a better prediction of initial exons than any other model tested.

text of prediction of genes. The poor recognition of introns in L isochores by all these models might result from an over-simplistic modelling. We point out that repeated elements, particularly LINEs, were not taken into account. Their higher frequency in the isochores of class L could explain the response of the model.

Many other data exploration tools exist. Multivariate analysis is on among the most popular methods that uses exactly the same data as HMMs. Indeed, if sequences are represented by frequencies of 6-bases words (see method), then a correspondence

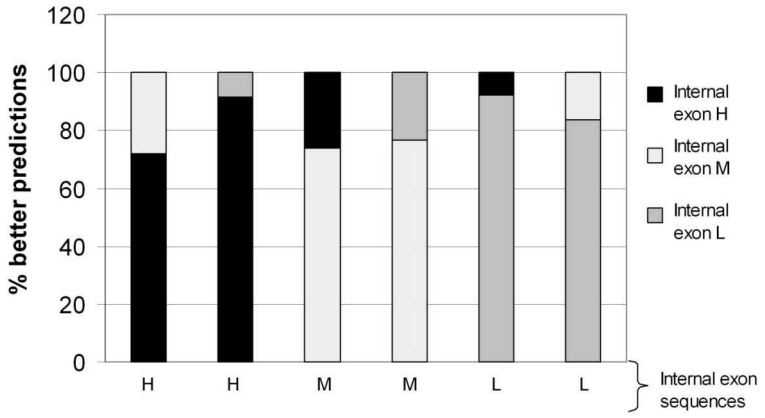

Figure 6. The models learned from different sequences (internal exons of classes H, M and L) were compared pairwise on the same sequences to determine the best predictions.

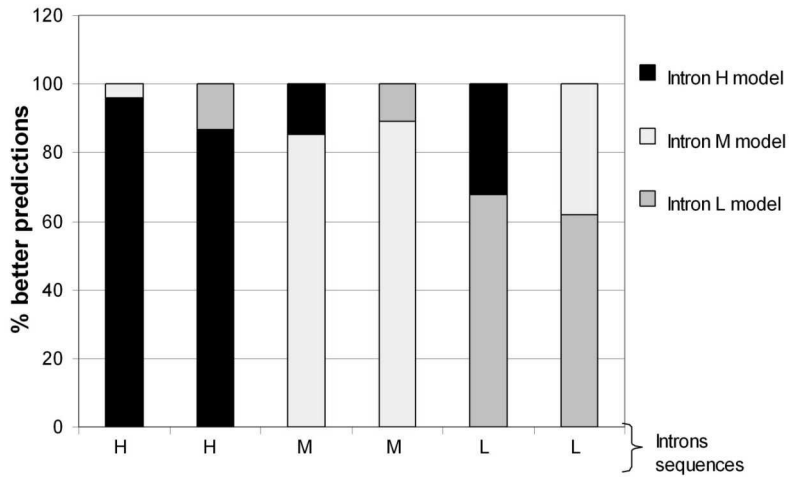

Figure 7. The models learned from different sequences (introns of classes H, M and L) were compared pairwise on the same sequences to determine the best predictions.

analysis will take into account exactly the same data as the one which is used to estimate the parameters of an HMM (see method). Figures 8 and 9 show the general patterns found by correspondence analysis. The frequencies of words of length 6 in the exons and the introns are neatly divided into four groups: H exons, M exons, L exons, and introns (Figure 8). When the reading frames are also taken into account (Figure 9), they are separated on the first factor, showing that the statistical differences between the codon positions represent the main statistical pattern in coding sequences.

5 Conclusion

The use of Markov models for the purpose of data exploration has been underestimated in genome analysis. This study is the first large scale exploration of the use of macro-states. Our approach allows to discriminate most genomic regions and is based on a selection among HMM models using macro-states. Macro-states allow to model distributions of lengths which are not geometric. Our strategy yields a comprehensive description of the human genome that highlights the following features:

1. The particular structure of intronless genes revealed the large number of errors of annotation in the databases for these genes: most small intronless genes are actual pseudogenes.
2. The great statistical differences between the three classes of isochores, and therefore the importance of taking into account this heterogeneity of the genome for the purpose of gene prediction. Initial exons exons are longer in the H class ($G+C$ rich). Introns are longer in the L class ($G+C$ poor).
3. Initial exons exhibit a very specific pattern, due to the fact that half of them contain a peptide signal. An average duration of stay in the first state of M_E180 of 80 bases long was observed, this is consistent with biological knowledge about such signals, which are 45 to 90 bases long. Initial exons without a peptide signal, and the second parts of the initial exons with a peptide signal, are statistically similar to internal exons and terminal exons, respectively.

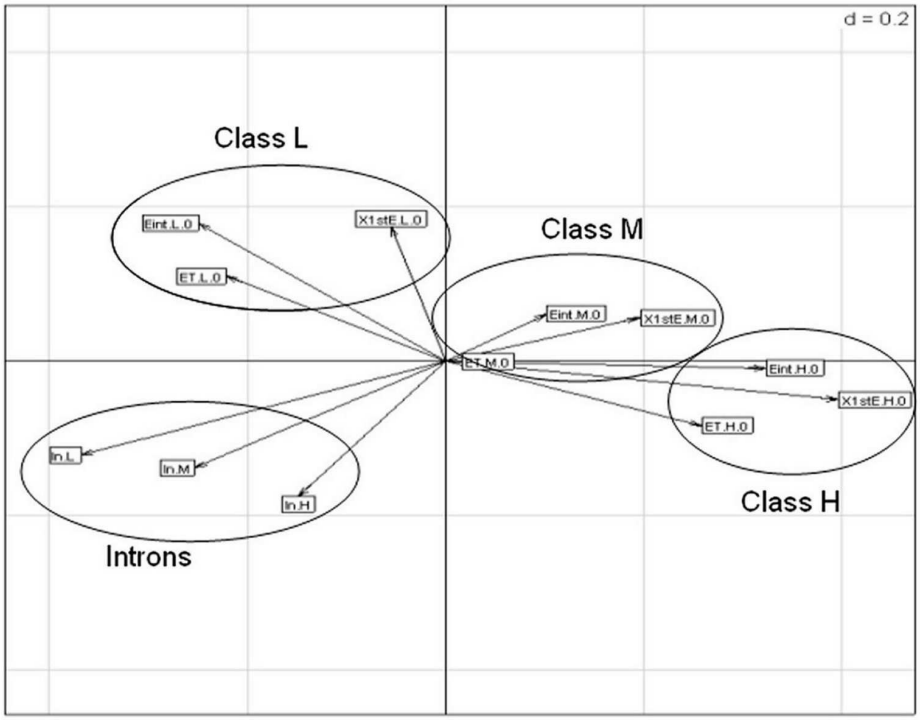

Figure 8. Correspondence analysis of the emission probabilities of the different states models in reading frame 0. The first axis (36.2% of total variability) represents the $G + C$ gradient. $Eint.H.0$=internal exon model of class H and reading frame 0; $Eint.M.0$=internal exon model of class M and reading frame 0; $Eint.L.0$=internal exon model of class L and reading frame 0; $ETt.H.0$=terminal exon model of class H and reading frame 0; $ETt.M.0$=terminal exon model of class M and reading frame 0; $ETt.L.0$=terminal exon model of class L and reading frame 0; $First.E.H.0$=initial exon model of class H and reading frame 0; $first.E.H.0$=initial exon model of class M and reading frame 0; $first.E.H.0$=initial exon model of class L and reading frame 0. $IN.H$=intron model of class H; $IN.M$=intron model of class M; $IN.L$=intron model of class L.

Macro-states HMMs models are based on exactly the same data as. Multivariate analysis but allows to identified the general patterns with a much lower cost in CPU resources. This is very close to the principle of some "old" gene prediction methods (see RECSTA [14]). However, the markovian approach has important advantages: it is not necessary to know the limits of the regions before the analysis, and more importantly, the model is more versatile; hence, new hypotheses can be explicitly introduced, as was done for the signal peptide.

6 Acknowledgements

The computations have been made at the IN2P3 computer centre using a large computer farm (more than 1500 CPUs). We thanks Marie-France Sagot for assistance in preparing and reviewing the manuscript.

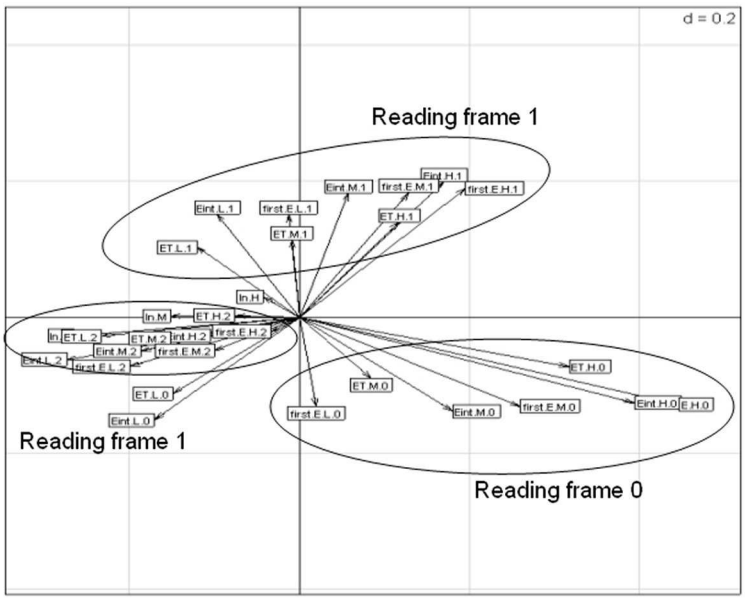

Figure 9. Correspondence analysis of the emission probabilities of the different state models. The first axis (59.5% of the total variability) represents the reading frame gradient.

References

[1] S. Altschul, W. Gish, E. Myers, and D. Lipman: *Basic local alignment search tool.* J. Mol. Biol., 1990, pp. 215–410.

[2] S. Berget: *Exon recognition in vertebrate splicing.* The Journal of Biological Chemistry, 270(6) 1995, pp. 2411–2414.

[3] G. Bernardi: *Isochores and the evolutionary genomics of vertebrates.* Gene, 241(1) 2000, pp. 3–17.

[4] G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier: *The mosaic genome of warm-blooded vertabrates.* Science, 228(4702) 1985, pp. 953–958.

[5] M. Borodovsky and J. McIninch: *Recognition of genes in dna sequences with ambiguities.* Biosystems, 30(1-3) 1993, pp. 161–171.

[6] C. Burge and S. Karlin: *Prediction of complete gene structure in human genomic dna.* Journal of Molecular Biology, 268 1997, pp. 78–94.

[7] C. Burge and S. Karlin: *Finding the genes in genomic dna.* Curr.Opin.Struc.Biol., 8 1998, pp. 346–354.

[8] C. Chen, A. Gentles, J. Jurka, and S. Karlin: *Genes, pseudogenes, and alu sequence organization across human chromosomes 21 and 22.* PNAS, 99 2002, pp. 2930–3935.

[9] O. Clay, S. Caccio, S. Zoubak, D. Mouchiroud, and G. Bernardi: *Human coding and non coding dna: compositional correlations.* Mol. Phyl. Evol., 1 1996, pp. 2–12.

[10] G. D'Onofrio, D. Mouchiroud, B. Aïssani, C. Gautier, and G. Bernardi: *Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins.* J. Mol. Evol., 32 1991, pp. 504–510.

[11] L. Duret, D. Mouchiroud, and M. Gouy: *Hovergen: a database of homologous vertebrate genes.* Nucleic Acids Research, 22(12) 1994, pp. 2360–2365.

[12] L. Duret, D. Mouchiroud, and C. Guatier: *Statistical analysis of vertebrate sequences reveals that long genes are scarce in gc-rich isochores.* Journal of Molecular Evolution, 40 1995, pp. 308–317.

[13] A. Eyre-Walker and L. Hurst: *The evolution of isochores.* Nat. Rev. Genet., 2(7) 2001, pp. 549–55.

[14] G. Fichant and C. Gautier: *Statistical method for predicting protein coding regions in nucleic acid sequences.* CABIOS, 3 1987, pp. 287–295.

[15] Y. Guédon: *Estimating hidden semi-markov chains from discrete sequences.* Journal of Computational and Graphical Statistics, 12(3) 2003, pp. 604–639.

[16] L. Guéguen: *Sarment: Python modules for hmm analysis and partitioning of sequences.* Bioinformatics, 21(16) 2005, pp. 3427–8.

[17] J. Hawkins: *A survey on intron and exon lengths.* Nucleic Acids Research, 16 1988, pp. 9893–9908.

[18] J. Henderson, S. Salzberg, and H. Fasman: *Finding genes in dna with a hidden markov model.* Journal of Computational Biology, 4 1997, pp. 127–141.

[19] International Human Genome Sequencing Consortium: *Initial sequencing and analysis of the human genome.* Nature, 409 2001, pp. 860–919.

[20] K. Jabbari and G. Bernardi: *Cpg doublets, cpg islands and alu repeats in long human dna sequences from different isochore families.* Gene, 224(1-2) 1998, pp. 123–127.

[21] A. Khelifi, L. Duret, and D. Mouchiroud: *Hoppsigen: a database of human and mouse processed pseudogenes.* Nucleic Acids Research, 3D 2005, pp. 59–66.

[22] A. Krogh: *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, pp. 179–186.

[23] V. Lukashin and M. Borodovsky: *Gene-mark.hmm : new solutions for gene finding.* Nucleic Acids Research, 26 1998, pp. 1107–1115.

[24] G. Mouchiroud, G. D'Onofrio, B. Aissani, G. Macaya, C. Gautier, and G. B. G: *The distribution of genes in the human genome.* Gene, 100 1991, pp. 181–7.

[25] A. Nekrutenko and W. Li: *Assessment of compositional heterogeneity within and between eukaryotic genomes.* Genome Res., 10(12) 2000, pp. 1986–95.

[26] H. Nielsen and A. Krogh: *Prediction of signal peptides and anchors by a hidden markov model.* In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, 1998, pp. 122–130.

[27] L. Rabiner: *A tutorial on hidden markov models and selected applications in speech recognition.* Poceeding of the IEEE, 77(2) 1989.

[28] S. Rogic, A. Mackworth, and F. Ouellette: *Evaluation of gene-finding programs on mammalian sequences.* Genome Research, 11 2001, pp. 817–832.

[29] J. Thiery, G. Macaya, and G. Bernardi: *An analysis of eukaryotic genomes by density gradient centrifugation.* J Mol Biol, 108(1) 1976, pp. 219–35.

[30] S. Zoubak, O. Clay, and G. Bernardi: *The gene distribution of the human genome.* Gene, 174(1) 1996, pp. 95–102.