Modelling the length distribution of exons by sums of geometric laws. Analysis of the structure of genes and *G+C* influence.

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

[1]UMR CNRS Biométrie et Biologie Evolutive, Université Claude Bernard, Lyon I, 43, Bd du 11 Novembre 1918, 69622 VILLEURBANNE cedex – France
melo@biomserv.univ-lyon1.fr
[2]LaPCS (Laboratoire de probabilités, combinatoires et statistique), Université Claude Bernard Lyon I, Domaine de Gerland, 50 avenue Tony-Garnier, 69366 Lyon Cedex 07

Abstract. Mathematical and computational methods are essential for gene identification and a more realistic modelling is necessary to better understand genome organization and gene expression. Hidden Markov models are one of the methods widely used for such identification. These models are quite efficient for gene localization but they imply that the lengths of all regions are geometrically distributed. However, in the human genome, the length distribution of the exons does not follow a geometric law. To address this problem, we propose to represent the length distribution of the exons by sums of geometric distributions with equal or different parameters. The model that we obtain has relatively few parameters, and fits very well exon lengths. Moreover, we propose a data processing method, based on a discrimination technique between Hidden Markov Models, which allows to study the structure of coding genes in detail. Our model describes known differences in gene organization between isochore classes and reveals some specific characteristics of intronless genes and a break in the homogeneity of the first coding exons. The use of hidden Markov models with complex states seems therefore to be a promising new approach for the modelling of the organization of a large genome.

Keywords: HMM, geometric laws, exons, length distribution, structure of genes, GC composition.

Introduction

The sequencing of the complete human genome lead to the knowledge of a sequence of three billion pairs of nucleotides (International Human Genome Sequencing Consortium, 2001). The sheer amount of data that this represents makes impossible the experimental search of genes and the analysis of the sequences without the use of automatic data processing methods. For twenty years, mathematical and computational models have been widely developed, for instance, to identify genes in newly sequenced regions (Stormo 2000). The identification of the genes in eukaryotic genomes is more complex than in prokaryotic genomes. The difficulties of the prediction are probably due to the alternation of introns and exons that represent, respectively, the noncoding and coding regions of the gene. The coding regions represent only 3% of the human genome. Moreover, the *G+C* frequency influences the structure of the regions (Duret & al. 1995, Chen & al. 2001). For instance, the density of genes in the *G+C* rich regions is higher than in the *G+C* poor regions. The introns are also smaller and their density is lower in the *G+C* rich regions than in the *G+C* poor regions. Finally, genes from the *G+C* poor regions code for longer proteins than those from *G+C* rich regions (Duret & al. 1995).

Different Markovian approaches have been developed for gene identification: some algorithms use hidden Markov models (HMMs) (GeneMark.hmm of Borodovsky & al. 1998, HMMgene of Krogh 1997**,** VEIL of Henderson & al. 1997), interpolated Markov models (GlimmerM of Salzberg & al 1999), or semi-Markov models (Genscan of Burge & al. 1997). In these models, each state represents a region and the duration of stay in the state represents the length of the region. To build a HMM model, it is necessary to assume that the duration of stay in each state follows a geometric law. The empirical length distributions of the intergenic and of the intronic regions do indeed follow a geometric law. However, histograms representing length distributions of different exons are bell-shaped (Burge & al. 1997, Berget 1995, Hawkins 1988). Thus, hidden Markov models can not represent precisely the length distribution of the exons. One way to overcome this problem is to use semi-Markov models. In this case, the duration of stay in a state depends on the empirical length distribution of the region.

To improve the prediction of the exon and intron lengths and to identify genes with non canonical features, it is important to consider their biological properties. For instance, the introns (1000 to 3000 bp on average) are longer than the exons (100 to 200 bp on average) and their lengths vary with their position in the gene. If hidden Markov models and semi-Markov models are used, exons are most accurately predicted when their length ranges between 75 to 200 bp. Exons smaller than 50 bp or longer than 300 bp are more difficult to predict correctly (Burset & al. 1996, Rogic & al. 2001). Moreover, initial coding exons and terminal exons are more difficult to identify than internal exons. Two hypotheses can explain this difficulty to predict the initial and terminal exons. First, the initial and terminal exons are longer than the internal exons. Second, their structures

are different from the one of internal exons because they contain signals, like the signal peptide in the initial exon. Intronless genes are also complex to identify because they are long (1022 bp on average) and they contain both start and stop codons.

Our study starts by describing a solution for more precisely representing the length distribution of the exons, still within the framework of hidden Markov models. 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. Although this approach has already been suggested (Durbin & al. 1998 page 69), it has not been used to model the entire genome.

In a second part, we propose to use HMMs for data analysis and knowledge discovery, thus adopting an approach not often used. More precisely, a comparison of the estimation ability of different HMM models is used to reveal some properties of human genes. We study mainly exon length and we try to discriminate between first, internal and terminal exons. The interpretation of the discrimination that we obtain leads to the detection of some pseudogenes and shows that first exons follow specific organization rules. In order to improve the quality of the description of the genes by a HMM, our study also considers the influence of the *G+C* frequency on the structure of genes.

Material

The data used in this study is extracted from Hovergen (Homologous Vertebrate Genes Database, March 2003 release 43) (Duret & al. 1994) and concerns only the human genome. The databases contain numerous errors. We chose to restrict the analysis to genes for which RNA transcripts have been sequenced. This ensures that the knowledge of the intron/exon organization is correct. Moreover, both the experimental redundancy and the redundancy due to the presence of large gene families are discarded. These may distort the results of the statistical analyses. We therefore obtain a set of 5034 multi-exon genes and 817 single exon (that is, intronless) genes. We only consider introns situated between coding exons. This set, divided in two equal random parts (training and test sets), is our first data set (set A).

A second data set (set B) is extracted from set A, to study the influence of the $G+C_{III}$ on gene structure, where the $G+C_{III}$ denotes the *G+C* content at the third codon position. Due to degeneracy of the genetic code, the $G+C_{III}$ provides the best discrimination criterion between isochore classes. Genes are clustered according to the $G+C_{III}$ frequency of their CDS and we split them into three classes having each the same number of genes. We obtain the following classes, denoted by *H*, *M* and *L*: *H*=[100%, 72%], *M*=]56%,72%[and *L*=[0%,56%], where the numbers in brackets gives the range of the percentages of each class. Such percentages represent the $G+C_{IIi}$ frequency of CDS. Observe than the percentages obtained for the classes calculated according to other criteria that have been used in the literature are roughly the same. For instance, Duret & al. 1995 obtains *H*=[100%, 75%], *M*=]57%,75%[et *L*=[0%,57%].

Sets A and B are used for modelling the length distribution by sums of geometric laws and for the analysis of the structure of genes.

Methods

a) Modelling of the length distribution by sums of geometric laws

The estimation of the length distribution of the exons and introns is realized from a sample $x_1 \ldots x_n$ of data set sequences. Each x_i is considered as the realization of an independent variable of some given laws. We tested the following laws:

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

$$P[X = k] = C_{k-1}^{m-1} \times p^m \times (1-p)^{k-m} \qquad \textit{Eq.I}$$

* The sum of two geometric laws with different parameters $p_1 > p_2$: (see Annex)

$$P[K = k] = p_1 \times p_2 \frac{(1-p_2)^{k-1} - (1-p_1)^{k-1}}{p_1 - p_2} \qquad \textit{Eq.II}$$

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

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

We want to estimate the length distribution of each region. The law which fits best with the empirical distribution is the law with the smallest Kolmogorov-Smirnov distance. For each region, to estimate the parameters of the different laws, we minimize the Kolmogorov-Smirnov distance. We have:

$$D_{KS}=sup_{x\in data}|\,F(x)-G(x)\,| \qquad Eq.IV$$

where D_{KS} is the Kolmogorov-Smirnov distance, F is the theorical density distribution, G is the empirical density distribution and $x \in$Data. However, the classical Newton or gradient algorithm can not be minimized for the Kolmogorov-Smirnov distance because this distance is not differentiable. We therefore discretize the parameter space with a step of 10^{-5}. We compute the Kolmogorov-Smirnov distance for all these parameters. The parameter associated with the smallest Kolmogorov-Smirnov distance is then chosen. The complexity of the algorithm is linear with the number of parameters discretized. For example, to estimate the parameters of the sum of two geometrical laws of different parameters with a step of 10^{-5}, the algorithm runs in less than one minute. This method ignores the number of parameters.

We chose to use the distance of Kolmogorov-Smirnov to estimate the parameters of the different laws rather than maximum likelihood method for different reasons. The maximum likelihood method is defined by:

Definition: let x be a discrete variable with probability: $P[x|\vartheta_1\ldots\vartheta_k]$, where $\vartheta_1\ldots\vartheta_k$ are k unknown constant parameters which need to be estimated, obtained by an experiment which resulted in N independent observations, $x_1,\ldots,x_N$. then the likelihood function is given by:

$$L(x_1,\ldots,x_N|\vartheta_1\ldots\vartheta_k)=\Pi_{i=1\ldots N}P[x_i|\vartheta_1\ldots\vartheta_k] \qquad Eq.V$$

The logarithmic function is:

$$\Lambda=ln\,(L(x_1,\ldots,x_N|\vartheta_1\ldots\vartheta_k))=\Sigma_{i=1\ldots N}P[x_i|\vartheta_1\ldots\vartheta_k] \qquad Eq.VI$$

The maximum likelihood estimators of $\vartheta_1\ldots\vartheta_k$, are obtained by maximizing L or Λ.

The choice between these two methods of estimating the parameters (Kolmogorov-Smirnov distance and maximum likelihood) was empirical. We made many simulations of length distributions and estimated the distribution by both the Kolmogorov-Smirnov and the maximum likelihood methods. For each simulation, the maximum likelihood method fitted the end of the length distribution, thus neglecting many small exons (see Figure 1). Intuitively, the maximum likelihood method is a global estimation, and therefore tends to adapt as well as possible to all the data. In our case, this means that the maximum likelihood method will try to fit well also the length of the long exons which are rarer. This will lead to a less good estimation of the length of small exons which are more numerous. The maximum likelihood method thus fails when the end of the distribution is longer. We therefore preferred to use a "biased" method to better represent the length of the a majority of the exons.

The above methodology is applied to sets A and B.

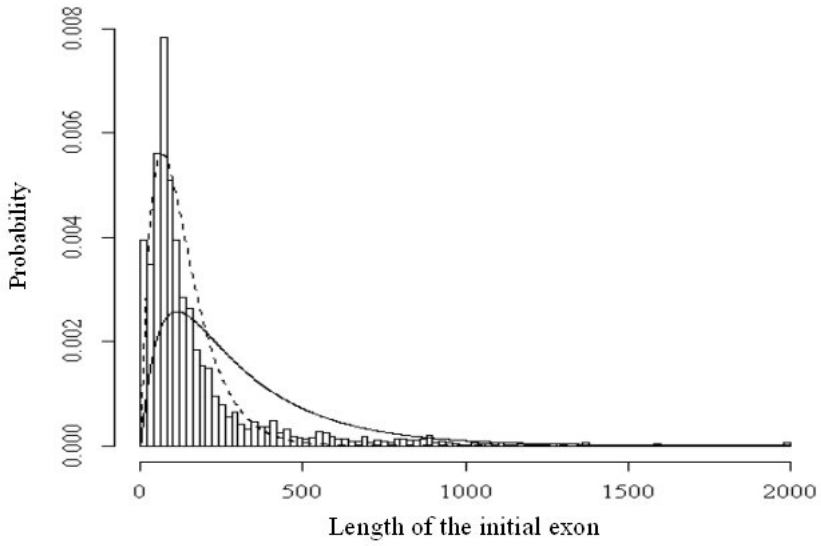

Figure 1: *Empirical distribution of the length of the initial coding exon. The histogram represents the empirical distribution of the length of the initial exons in a multi-exons gene. The dotted line describes the theoretical distribution, obtained by the Kolmogorov-Smirrnov distance. The full line characterizes the binomial distribution, obtained by the maximum likelihood method.*

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. The state duration is characterized by the parameters of the sum of these geometric laws.

For example :

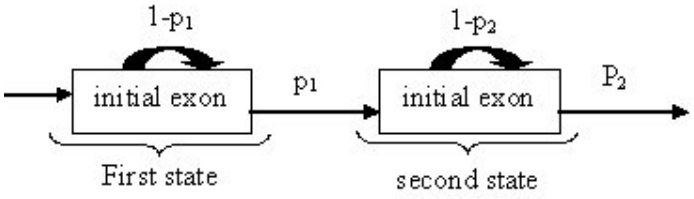

Figure 2: *The initial exon length distribution is modelled by a sum of two geometrical laws of parameters p_1 and p_2, this region is represented by two states, with the duration state: p_1 and p_2. The probabilities of emission of these two states are the same.*

Various studies (Burge & al. 1997, Rogic & al. 2001 and Chen & al. 2002) have shown that the length distribution of the exons depends on their position in the gene. We can differentiate four types of exons: initial coding exons, internal exons, terminal exons (see Figure 3) and single-exon genes.

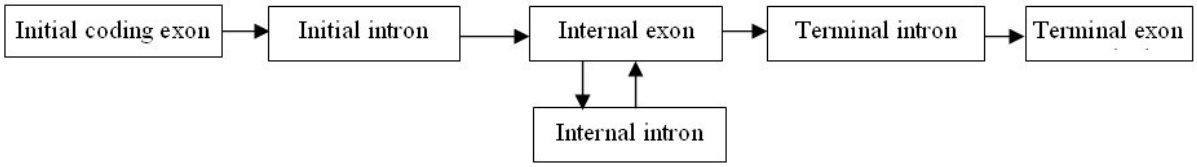

Figure 3: *A representation of our definition of gene.*

b) Analysis of the structure of a gene

The DNA sequence is heterogeneous along the genome but is composed of a succession of homogenous regions such as coding and non-coding regions. HMMs are used to localize these different regions. Each type of region corresponds to a state of the HMM. To take into account the existence in exons of a reading frame, most models represent each type of exons by three separated states of the HMM, depending on the reading frame (Borodovsky & *al*. 1993 and Burge & al. 1998). Like these authors, we used here HMMs of order 5 to take into account the dependance between two codons. When a letter is emitted by the HMM, we take into account the 5 letters that have been emitted before. Thus, the emission probabilities of the HMMs are estimated by the frequencies of the 6-letter words in the different regions (introns, initial exon, internal exon and terminal exon) that compose the training set. There is therefore a HMM model for each region. For example, the HMM below represents the initial exon:

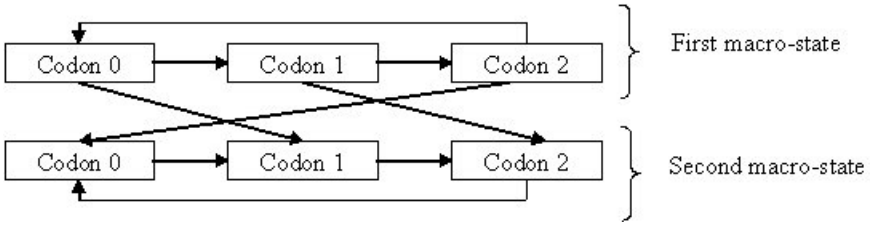

Figure 4: *This figure represents the initial exon HMM that is separated into two macros-states to take into account the length distribution of initial exons (sum of two geometrical laws). Each macro-state is split into three states that represent the position of the nucleotide in the codon, the two states Codon 0 have the same emission probabilities (idem for Codon 1 and Codon 2).*

We want to know if all these models are well adapted to their region. Indeed, we can suppose that the best model for a region is the model that has been trained on this region. To check this hypothesis, all HMMs are then pairwise compared by applying them on each region (introns, initial exon, internal exon and terminal exon) of the test set in order to determine the model which has the best probability of emitting the sequence tested (*Eq. VII*). We have:

$$D=\{ \log P(S/HMM_1) - \log P(S/HMM_2) \} / |S|. \qquad \text{Eq. VII}$$

where D is a discrimination measure, S is the sequence being tested, $|S|$ is its length and HMM_1, HMM_2 are the two models tested. $P(S/HMM_1)$ is computed using the forward algorithm (Rabiner 1989).
The HMM having the best probability, for a majority of the sequences of a same region, is kept to characterize this region.

To distinguish the different HMMs according to the *G+C* frequency, a second study was realized. The analysis of the different types of exons according to their *G+C* frequency was completed by a correspondence analysis on the emission probabilities of the different HMMs.
The procedure above is applied to sets A and B.

Results

a) Modelling the length distribution by sums of geometric laws

When set A is used, the histograms of the length distribution of single exons (that is, of intronless genes) exhibit a bi-modality (see Figure 5). The two distinct modes are probably due to annotation errors in the database. Indeed, we compared intronless genes to the complete human genome with the Blast similarity search program (Altschul & al 1990). The result shows that many small intronless genes are pseudogenes (genes which have lost their function). The distribution obtained after removing these pseudogenes is a bell-shaped distribution like the distribution for the other types of exons.

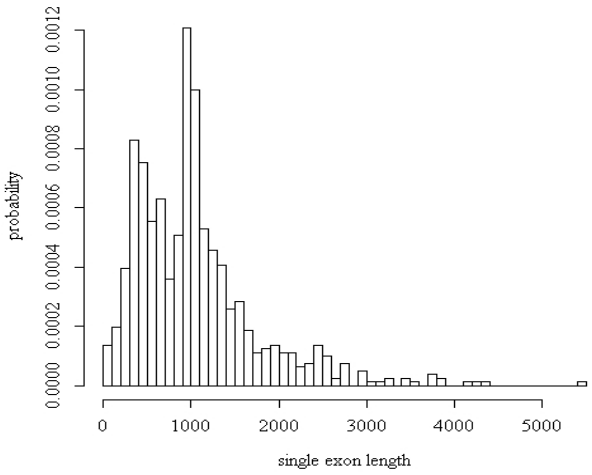

Figure 5: *Empirical distribution of the length of single exons (that is, of intronless genes).*

Exon lengths vary according to the position of the exon in the gene. Initial and terminal exons are longer than internal ones. Introns also present a positional variability in their lengths. The internal and terminal introns have similar length distributions, when compared to the length distribution of the initial introns. The p-value of a Wilcoxon nonparametric test is 2.10^{-16} and $6.367.10^{-12}$ when we compare internal/initial introns and terminal/initial introns respectively (see Table 1). Minimizing the Kolmogorov-Smirnov distance provides a good fit to the empirical distribution of the exons length (see Figure 1). This does not take into account the number of parameters of the geometric laws, but it does not seem that this implies an overparametrization of the chosen models. We thus model the lengths of the initial exons by the sum of 2 geometric laws of parameters $1.7.10^{-2}$ and $1.35.10^{-2}$ (see Figure 1). Terminal exons are described by the sum of 3 geometric laws of parameters ($1.16.\ 10^{-2}$, $5.5.10^{-3}$, 0.1). Internal exons are characterized by the sum of 5 geometric laws of the same parameter $3.8.10^{-2}$ (that is, by a negative binomial distribution). Finally, single exons are modelled by a negative binomial distribution of parameters 3 and $2.84\ 10^{-3}$. The initial intron length distribution is modelled by a geometric law of parameter 9.10^{-4}. Similar results are obtained for the other intron types.

Position in the gene	Mean length (en bp)	Median length (bp)
initial coding exon	177	104
internal exon	141	123
terminal exon	238	148
initial intron	3362	945
internal intron	1615	641
terminal intron	1452	547

Table 1: *Length of the exons and of the introns according to their 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	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	2646	816	3962	872	4942	1529
internal intron	992	345	1446	437	2841	1322
terminal intron	1247	422	1334	579	2691	1136

Table 2: *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.*

Table 2 shows the results obtained using set B and taking into account the influence of the *G+C* frequency at the third codon position in the gene. It is well known that exon and intron lengths depend upon *G+C* content. In heavy isochores (*G+C* rich), introns are fewer and shorter than in light (A+T rich) isochores (Chen 2002). Initial and terminal exons are longer in regions with a high *G+C* content. In all cases, the Wilcoxon test is significant. As an example, the frequencies of initial exons having a length greater than 300bp are respectively 22.4%, 13.5% and 9% in the H, M and L isochore classes. However, internal exon length does not vary with isochore class (the Student test is not significant). The length distribution of the exons exhibits a bell-shaped pattern in all three *G+C* classes. We modelled the lengths of the initial exons in classes H and L by a sum of 2 geometric laws of parameters $5.5.10^{-3}$ and $8.7.10^{-2}$ (see Table 3), and $9.23.10^{-2}$ and $7.10.\ 10^{-3}$ respectively. Initial exons of class M are described by a sum of 3 geometric laws of parameters 0.252 , $7.52.\ 10^{-2}$ and $7.52.\ 10^{-3}$.

Laws	Parameters p	K-S distance
Bin (2,p)	0.0117	0.1084
Bin (3,p)	0.0185	0.16
Bin(4,p)	0.02634	0.1826
Σ 2	0.0055-0.087	0.0447
Σ 3	0.122-0.0622-0.00622	0.0614

Table 3: *Results of the estimation of the parameters of the different laws obtained for the initial exons of class H minimizing the Kolmogorov-Smirnov distance.*

Notation: bin(n,p) represents the binomial negative law of parameters n,p. Σ 2 represents a sum of 2 geometric laws of different parameters. Σ 3 represents a sum of 3 geometric laws of different parameters. K-S is the abbreviation forf Kolmogorov-Smirnov.

b) Analysis of the structure of genes

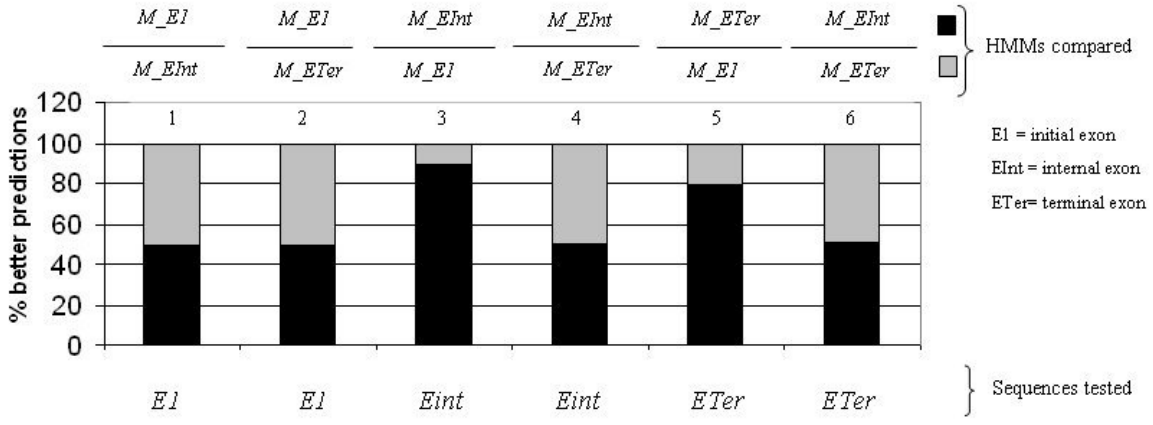

Figure 6: *Models learned on different sequences (initial, internal and terminal exons) were pairwise compared on the same sequences to determine the best predictions.*
For instance, in histogram 1, the likelihood of each first exon is computed under models learned on M_E1 and M_Eint. The black bar represents the percentage of first exons having a higher likelihood for the first exon model and the grey bar for the second exon model. Histograms 1-2: The models M_E1, M_Eint and M_ET have

same predictive power on initial exons. Histograms 3-5: The models M_Eint and M_ET predict well, respectively, internal and terminal exons compared to the model M_E1 (82% and 75%). Histograms 4-6: The models M_Eint and M_ET have the same predictive power on initial and terminal exons.

When set A is used, the HMM discrimination reveals two main characteristics: 1) models for internal exons (denoted by *M_Eint*) and terminal exons (denoted by *M_ETer*) have approximately the same prediction behaviour (see Figure 6, histograms 4 and 6); 2) *M_Eint* and *M_ETer* are clearly different from the model for initial exons denoted by *M_E1*. More precisely, the likelihood of *M_E1* is weaker on internal exons (resp. terminal exons) than the likelihood of *M_Eint* (resp. *M_ETer*) (see Figure 6, histograms 3 and 5). However *M_E1* is not able to recognize first exons (see Figure 6, histograms 1 and 2).

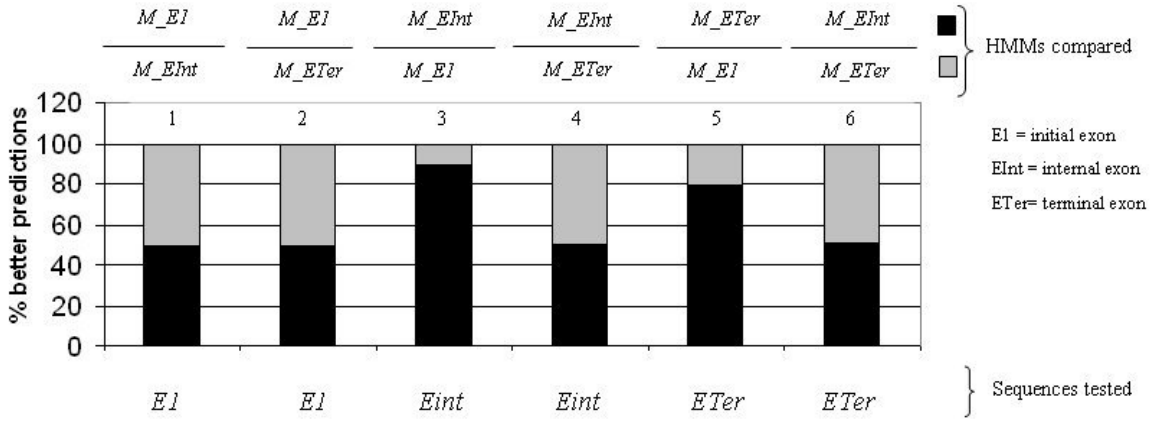

Figure 7: *The models learned on different sequences (initial, internal and terminal exons) were pairwise compared on the initial exons to determine the best predictions.*
The model M_IE$_{80}$ predicts better initial exons than all different models tested.

This clearly suggests that first exons are incorrectly modelled. The specific statistical characteristics of initial exons could result from the existence of signals overlapping the beginning of genes. To explore this hypothesis, we split the initial exon HMM model into two HMMs. The first is trained on the *n* first nucleotides of the initial exon for a given value of *n*, and the second one on the remaining of the initial exon. This new initial exon model is called *M_E1$_n$*. The pairwise comparisons between the *M_E1$_n$* models obtained (see Figure 7) show that the *M_E1$_{80}$* model allows for a better discrimination. The initial exons are better predicted by the *M_E1$_{80}$* model than by the simple initial exon model in 70% of the cases (see Figure 7, histogram 1). Moreover, among all the *M_E1$_n$* models, the *M_E1$_{80}$* model leads to the highest likelihood (see Figure 7, histograms 2 to 6). These facts suggest that the break of homogeneity in the initial exon happens around the 80th base. Finally, this separation allows an improved discrimination between internal and initial exon models on the initial exons (49% to 61% in favour of the *M_E1$_{80}$* model: Figure 6, histogram 1 and Figure 5, histogram 7) and on the internal exons (89% to 92% in favour of *M_Eint*, not represented in the Figure). The same results are observed in 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 containing a signal peptide are better recognized by the first HMM of the *M_EI$_{80}$* model than by the second HMM of the *M_EI$_{80}$* model in 75% of the cases. Moreover, we have compared these results with those obtained with the SignalP program (Nielsen 1998). The initial exons predicted as having a signal peptide by the SignalP program are better recognized by the *M_EI$_{80}$* model than by the internal exon model in 70% of the cases. When the SignalP program does not predict a signal peptide, the *M_EI$_{80}$* model and the internal exon model give the same results.

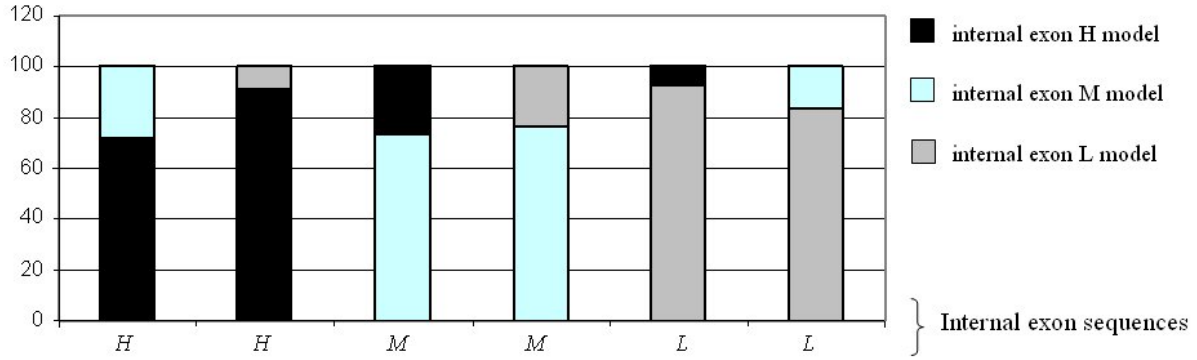

Figure 8: *The models learned on different sequences (internal exons of classes H, M and L) were pairwise compared on the same sequences to determine the best predictions.*

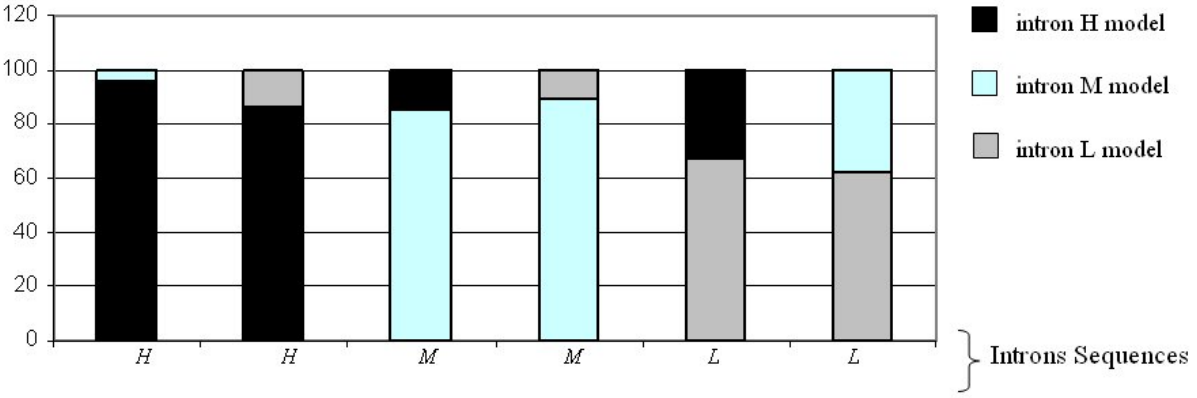

Figure 9: *The models learned on different sequences (introns of classes H, M and L) were pairwise compared on the same sequences to determine the best predictors.*

Using set B, we trained a HMM for exons in each isochore class (*H, M, L*). Internal exons having a high *G+C* frequency are better predicted by the internal exon model *H* than by the internal exon model *M* (71.8%) or the internal exon model *L* (91.4%) (see Figure 8, histograms 1 and 2). Likewise, the internal exons of class *M* are better predicted by the internal exon model *M*, and the internal exons of class *L* are better predicted by the internal exon model *L* (see Figure 8, histograms 3 to 6). Moreover, the initial and terminal exons of classes *H, M* and *L* are better predicted by their respective models (*H, M* and *L*). The results concerning introns are different. The introns of class *H* and *M* are better predicted by, respectively, HMMs *H* and *M* (see Figure 9, histograms 1 to 4). However, for the introns of class *L*, there is a lack of discrimination between the intron models *H, M* and *L* (see Figure 9, histograms 5, 6). It is therefore important to consider different HMMs according to the *G+C* frequency of the region studied.

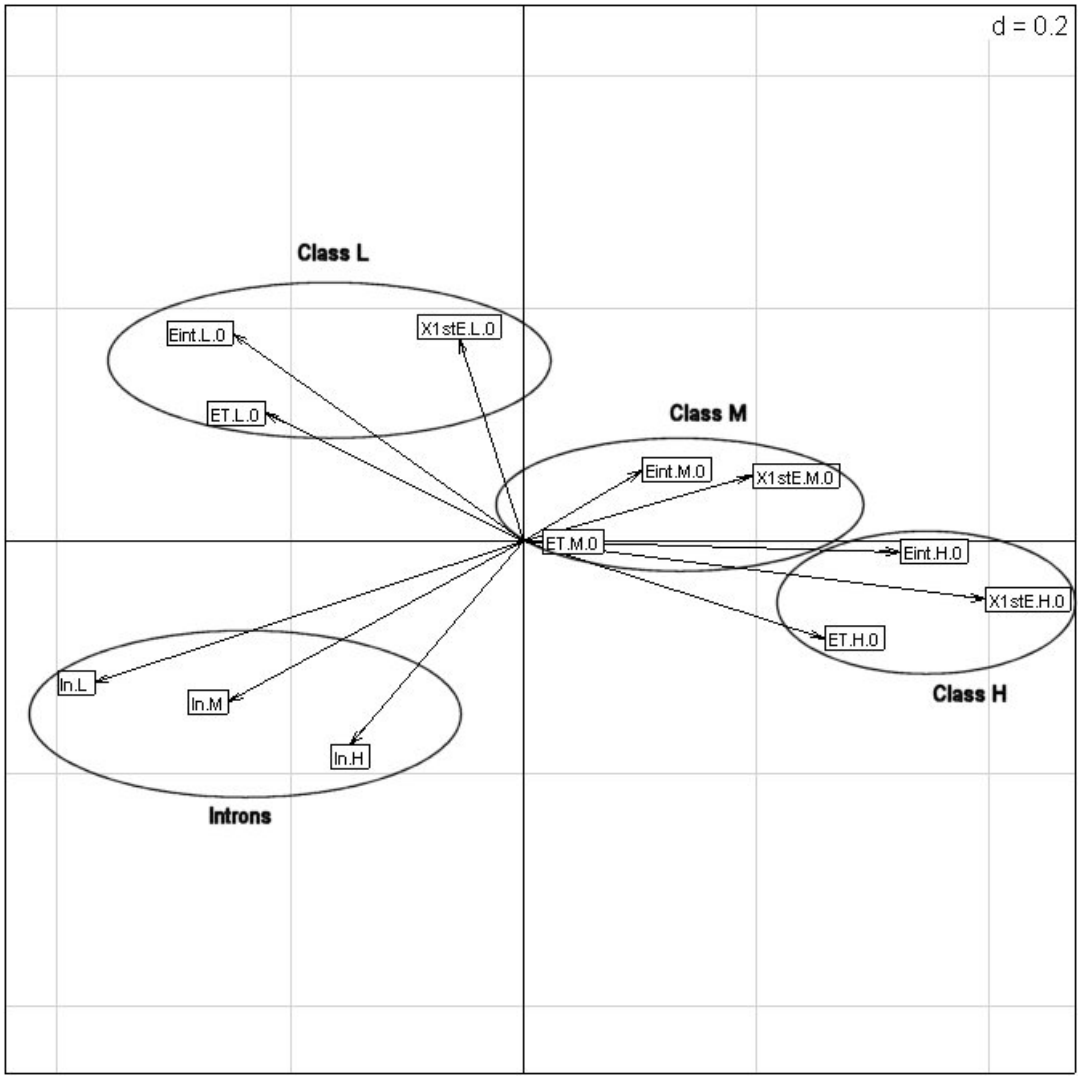

Figure 10: *Correspondence analysis of the emission probabilities of the different state models in reading frame 0.*
EInt.H.0=internal exon model of class H and reading frame 0
EInt.M.0=internal exon model of class M and reading frame 0
EInt.L.0=internal exon model of class L and reading frame 0
ETer.H.0=terminal exon model of class H and reading frame 0
ETer.M.0=terminal exon model of classM.H and reading frame 0
ETer.L.0=terminal exon model of class L and reading frame 0
E1.H.0=initial exon model of class H and reading frame 0
E1.H.0=initial exon model of class M and reading frame 0
E1.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

A correspondence analysis of the frequency of the 6-letter words in the different types of sequences gives also the same results. Figure 10 shows that the frequency of words with 6 letters in exons and introns are clearly separated into four groups when the classes *H*, *L* and *M* were compared for reading frame 0. The same results are obtained for the reading frames 1 and 2. We thus see three different groups that represent the classes *H*, *L* and *M* of the exons, and a fourth group which represents the introns without distinction of the *G+C* classes considered. Figure 11 indicates that the difference among the exons according to their reading frame is very important. We can thus see three different groups. The first group represents the exons (initial, internal and final) in reading frame 0. However, the emission probabilities of the internal and terminal exons with poor *G+C* content in reading frame 0 are different from the other probabilities in reading frame 0. Indeed, the emission probabilities of the internal and terminal exons with poor *G+C* content are closer to those of the group "reading frame 2" than to those of the group "reading frame 0". These two states are therefore not characteristic of the

reading frame 0. The poor *G+C* frequency can explain this difference because the introns are not characteristic and the variability of the third position of the codon is more important than the variability of the first and second position of the codons. The second group shows the exons in reading frame 1. Finally, the last group represents the exons in reading frame 2 and the introns. The emission probabilities of the different models are very different depending on their reading frames and depending on to their *G+C* frequencies. Moreover, the emission probabilities of the exons in reading frame 2 and of the introns are similar.

Figure 11: Correspondence *analysis of the emission probabilities of the different state models.*

Discussion

The length of the exons and introns varies along the genome. The length distribution depends on the position in the gene and on the *G+C* composition. It is important to consider these properties to model correctly the structure of genes. These properties are often neglected by the different existing hidden Markov models. This work shows that it is possible to represent more precisely the length distribution of the exons using Hidden Markov models. We showed that the empirical exon length distribution is well-fitted by sums of geometric laws. In this way, we improved the description of the genes by HMMs. The complexity of the semi-Markov model algorithm is higher than the complexity of hidden Markov model algorithm. Thus, they require a difficult optimization to be able to perform as efficiently as a hidden Markov model algorithm. Thus, this study proposes an alternative method to semi-Markov models for the modelling of the genome. This work also shows that the minimization of the Kolmogorov-Smirnov distance allows to obtain a better fit of the model to the empirical length distribution of exons than the maximum likelihood method.

The prediction obtained with hidden Markov models and semi-Markov models by the recent algorithms are good, but many problems subsist: in particular it is difficult to predict small exons (< 75bp), initial and terminal exons and genes with many exons (Rogic 2001). Our study shows that the estimation of the length

distribution of exons by the maximum likelihood method neglects small exons. The estimation of the length distribution by the Kolmogorov-Smirnov distance may therefore improve the prediction of small exons by hidden Markov models and semi-Markov models. The bad predictions of the intronless genes obtained by these two models are probably due to the presence of wrong annotations in the database, for instance, to undetected pseudogenes. Our study shows that the majority of small intronless genes are pseudogenes. To improve the predictions, we tried to take into account as many biological properties as possible. The correlation between the lengths of the exons and their positions in the gene is known and is used to improve the predictions. A generalization to introns could improve the prediction of genes with many exons. Our study shows that the *G+C* frequency has a great influence on the length of exons and introns. The bad predictions of the initial and terminal exons are improved if the *G+C* frequency is taken into account. Indeed, initial and terminal exons are longer in the *G+C* rich class. Initial exons are more numerous in the *H* class. Introns are longer if the genes are *G+C* poor.

We also show that HMMs can be used to uncover new biological properties of genomes. The existence of a break in the homogeneity of the initial exons is revealed by the better result obtained with the M_EI_{80} model. Such a break can be due to the presence of a peptide signal at the beginning of the first exons. The length of a peptide signal (45 to 90 bases) corresponds to the average duration of stay in the start state of our M_IE_{80} model (average of 70 bases). This point is confirmed by the better results obtained by the M_IE_{80} start state when discriminating signal peptide sequences. Moreover, our method may be used to check the validity of some database annotations. Indeed, using it, we noticed that many pseudogenes are annotated as intronless genes. It would therefore be interesting to consider these two results (hypothetical peptide signal and wrong annotations) in the various currently available methods for predicting gene, to enhance the quality of the predictions they give. Finally, this paper shows the importance of the *G+C* frequency along the genome for the HMM modelling. HMMs that adapt to the *G+C* content of a region lead to an improvement in the prediction of exons and introns.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W, Lipman, D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403-410.

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

[3] Borodovsy, M., McIninch, J. (1993). Recognition of genes in DNA sequences with ambiguities. *Biosystems*, **30(1-3)**, 161-171.

[4] Borodovsky, M., Lukashin, A.V (1998) GeneMark.hmm, New solutions for gene finding. *Nucleic Acids Research*, **26(4)**, 1107-1115.

[5] Burge, C., Karlin, S.(1997) Prediction of complete gene structure in human genomic DNA. *Journal of Molecular Biology*, **268**, 78-94.

[6] Burge, C., Karlin, S. (1998) Finding the genes in genomic DNA. *Curr.Opin.Struc.Biol*. **8**, 346-354.

[7] Burset, M., Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353-367.

[8] Chen, C., Gentles, A.J,.Jurka, J., Karlin, S. (2002) Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22. *PNAS*, **99**, 2930-3935.

[9] Durbin, R., Eddy, S. , Krogh, A., Mitchison, J. (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. *Cambridge University Press*.

[10] Duret, L., Mouchiroud, D., Gouy, M. (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Research*, **22(12)**, 2360-2365.

[11] Duret, L., Mouchiroud, D., Gautier, C. (1995) Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *Journal of Molecular Evolution*, **40**, 308-317.

[12] Hawkins, J.D. (1988) A survey on intron and exon lengths. *Nucleic Acids Research*, **16**, 9893-9908.

[13] Henderson, J., Salzberg, S., Fasman, K.H. (1997) Finding genes in DNA with a hidden Markov model. *Journal of Computational Biology*, **4**, 127-141.

[14] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. (2001) *Nature*, **409**, 860-919.

[15] Krogh, A. (1997) 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*, 179-186.

[16] Rabiner, L.R (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *In Proceedings of the IEEE*, **77-2**, 257-285.

[17] Nielsen, H., Krogh, A. (1998) Prediction of signal peptides and anchors by a hidden Markov model. *In Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology* (ISMB 6), AAAI Press, Menlo Park, California, 122-130.

[18] Rogic, S., Mackworth, A.K., Ouellette, F.B. (2001) Evaluation of Gene-Finding Programs on Mammalian Sequences. *Genome Research*, **11**, 817-832.

[19] Salzberg, S.L, Pertea, M., Delcher, A., Gardner, M.J, Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24-31.

[20] Stormo, G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Research*. **10**, 394-397.

Annex:

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

$$P[X = k] = p_1 \times p_2 \frac{(1-p_2)^{k-1} - (1-p_1)^{k-1}}{p_1 - p_2} \qquad \textit{Eq.II}$$

We suppose that the variable X and Y follow respectively a geometrical law of parameter p_1 and p_2.

$$P[S = X + Y = s] = \sum_{i \geq 1} P[Y = i, X + Y = s]$$

$$P[S = X + Y = s] = \sum_{i \geq 1} P[Y = i, X = s - i]$$

$$s - i \geq 1$$

$$P[S = X + Y = s] = \sum_{i=1}^{s-1} P[Y = i] \times P[X = s - i]$$

$$P[S = X + Y = s] = \sum_{i=1}^{s-1} p_1 (1-p_1)^{i-1} p_2 (1-p_2)^{s-i-1}$$

$$P[S = X + Y = s] = p_1 p_2 (1-p_2)^{s} \sum_{i=1}^{s-1} (1-p_1)^{i-1} (1-p_2)^{-i-1}$$

$$P[S = X + Y = s] = p_1 p_2 (1-p_2)^{s-2} \sum_{j=1}^{s-2} \left(\frac{1-p_1}{1-p_2}\right)^{j}$$

$$P[S = X + Y = s] = p_1 p_2 (1-p_2)^{s-2} \sum_{i=1}^{s-1} \left(\frac{1-p_1}{1-p_2}\right)^{i-1}$$

$$P[S = X + Y = s] = p_1 p_2 (1-p_2)^{s-2} \frac{1 - \left(\frac{1-p_1}{1-p_2}\right)^{s-1}}{\left(\frac{1-p_1}{1-p_2}\right)} \qquad p_1 > p_2$$

$$P[S = X + Y = s] = p_1 \times p_2 \frac{(1-p_2)^{s-1} - (1-p_1)^{s-1}}{p_1 - p_2}$$

By same method, we obtain the sum of three geometric laws with different parameters $p_1 < p_2 < p_3$:

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